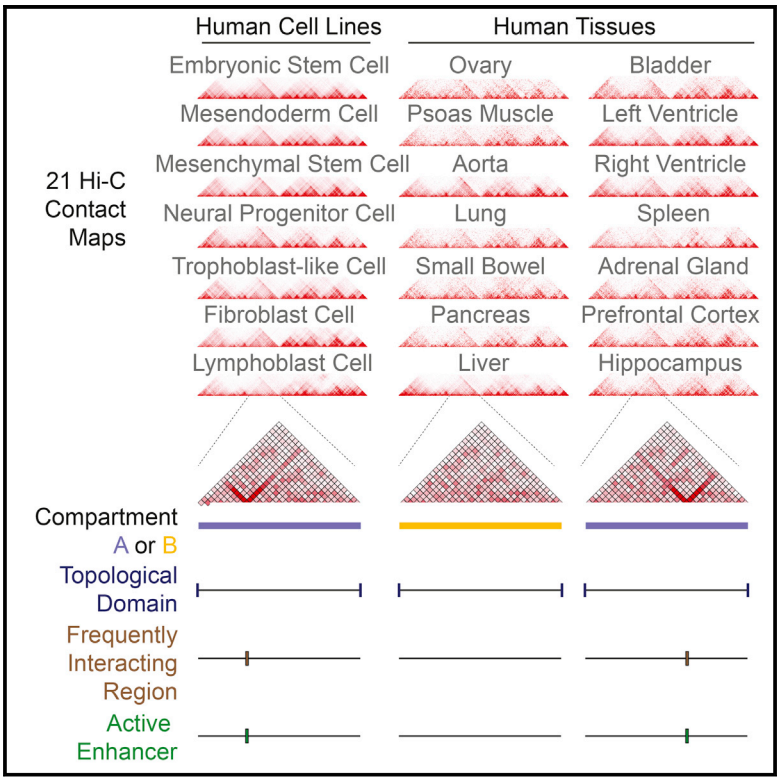
Resource

Cell Reports

A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome

Graphical Abstract



Highlights

- Integrative analysis of chromatin architecture in a broad set of human tissues
- FIREs are an architectural feature of chromatin organization
- FIREs are enriched for super-enhancers and show tissue-specific chromatin interactions
- FIRE formation is partially dependent on CTCF and the Cohesin complex

Authors

Anthony D. Schmitt, Ming Hu, Inkyung Jung, ..., Yiing Lin, Cathy L. Barr, Bing Ren

Correspondence

hum@ccf.org (M.H.), biren@ucsd.edu (B.R.)

In Brief

Schmitt et al. analyze Hi-C maps in 21 human cell lines and primary tissues and uncover a class of genome organizational features termed FIREs. FIREs are local interaction hotspots, highly tissue-specific, and correspond to active enhancers. We discuss the implications of our findings for the study of gene regulation and disease. Explore the Cell Press IHEC web portal at http://www.cell.com/consortium/IHEC.

Accession Numbers

GSE87112

Schmitt et al., 2016, Cell Reports *17*, 2042–2059
November 15, 2016
http://dx.doi.org/10.1016/j.celrep.2016.10.061

OPEN ACCESS
CellPress

Cell Reports
Resource

A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome

Anthony D. Schmitt,[1,2,12,13] Ming Hu,[3,12,14,*] Inkyung Jung,[1,15] Zheng Xu,[4,10,11] Yunjiang Qiu,[1,5] Catherine L. Tan,[1,13] Yun Li,[4] Shin Lin,[6] Yiing Lin,[7] Cathy L. Barr,[8] and Bing Ren[1,9,16,*]

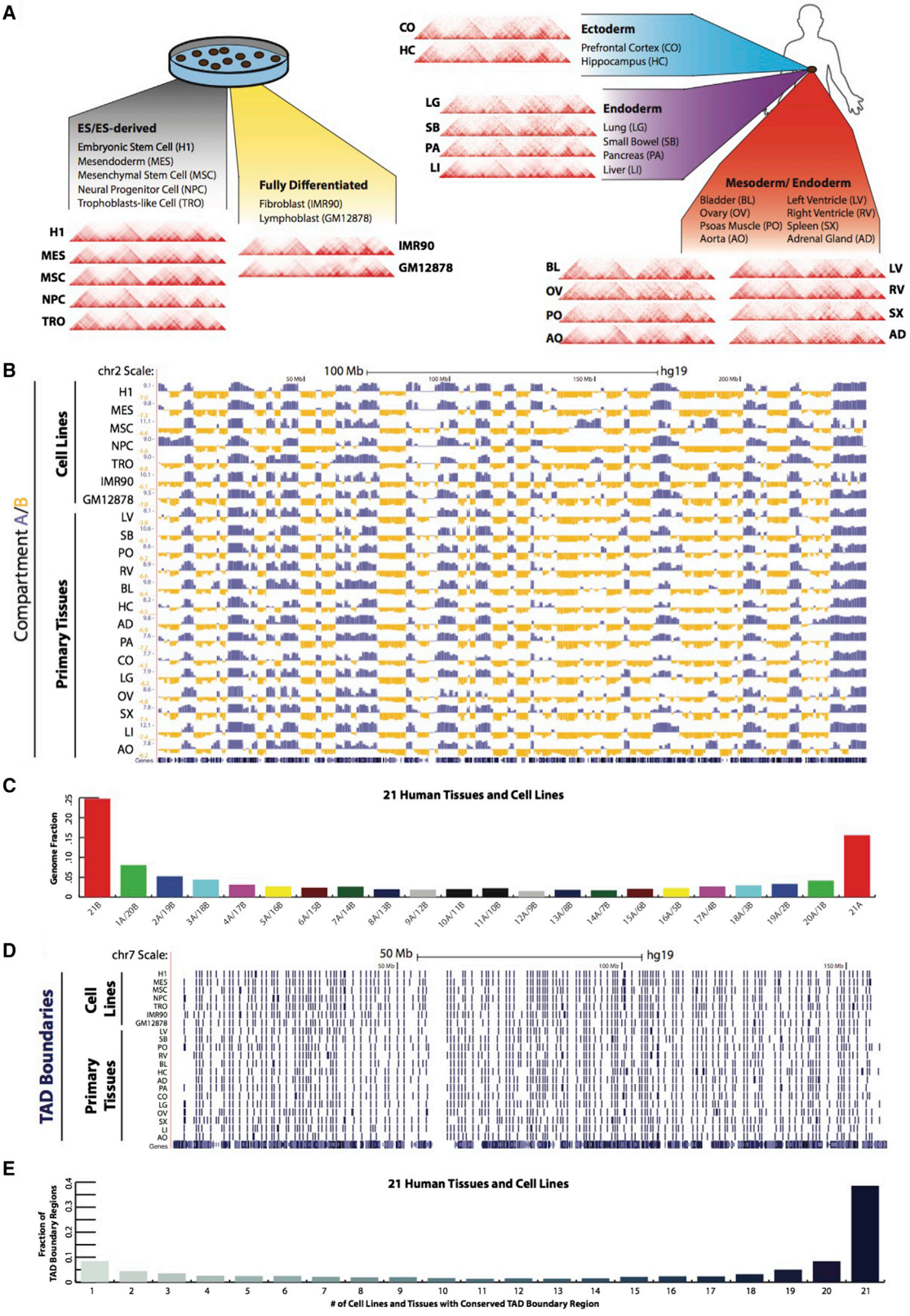
[1]Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA
[2]UCSD Biomedical Sciences Graduate Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[3]Division of Biostatistics, Department of Population Health, New York University School of Medicine, 650 First Avenue, New York, NY 10016, USA
[4]Departments of Genetics, Biostatistics, and Computer Science, University of North Carolina, Chapel Hill, NC 27599, USA
[5]USCD Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[6]Division of Cardiology, Department of Medicine, University of Washington, 850 Republican Street, Seattle, WA 98108, USA
[7]Department of Surgery, Washington University School of Medicine, 660 S Euclid Ave., Campus Box 8109, St. Louis, MO 63110, USA
[8]Krembil Research Institute University Health Network, The Hospital for Sick Children, The University of Toronto, Krembil Discovery Tower, 60 Leonard Ave. 8KD-412, Toronto, ON M5T 2S8, Canada
[9]Department of Cellular and Molecular Medicine, Moores Cancer Center and Institute of Genome Medicine, UCSD School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA
[10]Quantitative Life Sciences Initiative, University of Nebraska, Lincoln, NE 68583, USA
[11]Department of Statistics, University of Nebraska, Lincoln, NE 68583, USA
[12]Co-first author
[13]Present address: Arima Genomics Inc., 6404 Nancy Ridge Dr., San Diego, CA, 92121, USA
[14]Present address: Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, OH 44195, USA
[15]Present address: Department of Biological Sciences, KAIST, Daejeon 34141, South Korea
[16]Lead Contact
*Correspondence: hum@ccf.org (M.H.), biren@ucsd.edu (B.R.)
http://dx.doi.org/10.1016/j.celrep.2016.10.061

SUMMARY

The three-dimensional configuration of DNA is integral to all nuclear processes in eukaryotes, yet our knowledge of the chromosome architecture is still limited. Genome-wide chromosome conformation capture studies have uncovered features of chromatin organization in cultured cells, but genome architecture in human tissues has yet to be explored. Here, we report the most comprehensive survey to date of chromatin organization in human tissues. Through integrative analysis of chromatin contact maps in 21 primary human tissues and cell types, we find topologically associating domains highly conserved in different tissues. We also discover genomic regions that exhibit unusually high levels of local chromatin interactions. These frequently interacting regions (FIREs) are enriched for super-enhancers and are near tissue-specifically expressed genes. They display strong tissue-specificity in local chromatin interactions. Additionally, FIRE formation is partially dependent on CTCF and the Cohesin complex. We further show that FIREs can help annotate the function of non-coding sequence variants.

INTRODUCTION

Chromosome conformation capture (3C)-based techniques have begun to reveal molecular details of nuclear organization in eukaryotic cells (Dekker et al., 2002; Dixon et al., 2012, 2015; Dostie et al., 2006; Fraser et al., 2015; Jin et al., 2013; Lieberman-Aiden et al., 2009; Rao et al., 2014; Seitan et al., 2013; Simonis et al., 2006; Sofueva et al., 2013; Vietri Rudan et al., 2015; Zuin et al., 2014). It is now clear that each chromosome occupies a separate space in the interphase nucleus, known as a "chromosome territory," which is partitioned into distinct neighborhoods or compartments (Lieberman-Aiden et al., 2009; Meaburn and Misteli, 2007). Within each compartment, topologically associating domains (TADs) constrain chromatin interactions (Dixon et al., 2012, 2016; Nora et al., 2012; Sexton et al., 2012). Within each TAD, chromatin interactions between distal *cis*-regulatory elements occur in a cell-type-dependent manner to allow modulation of promoter activity by enhancers (Dryden et al., 2014; Montavon and Duboule, 2013; Phillips-Cremins et al., 2013; Simonis et al., 2006; Tang et al., 2015). Previous 3D genome analyses have been largely limited to cultured cells and a small collection of primary cell types. By contrast, our knowledge of chromatin organization in human tissues is still scarce. Variation in chromatin interaction patterns among diverse tissue types remains poorly defined, and its functional relationship with gene regulation remains to be characterized. This is a critical shortcoming because diseases pertaining to



(legend on next page)

specific organ systems are often not easy to recapitulate in vitro. Therefore, systematic characterization of chromosome architecture across a broad set of well-annotated primary tissues could be of great value for further study of genome function.

Recent studies of chromatin modification landscapes across a large number of human tissues and cell types have greatly improved our understanding of genome function and regulation (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium et al., 2015). The research has revealed that over 12% of the genome possesses cell-type-specific chromatin signatures consistent with them acting as *cis*-regulatory sequences. However, to better understand how these DNA sequences contribute to tissue- and cell-type-specific gene expression patterns, it is necessary to characterize the chromatin architecture in each tissue. Here, we report integrative analysis of chromatin organization maps of 14 human tissues and 7 human cell lines for which complete epigenome datasets have been generated by the Epigenome Roadmap Consortium, ENCODE, or the National Institute of Child Health and Human Development (NICHD) (ENCODE Project Consortium, 2012; Roadmap Epigenomics Consortium et al., 2015). We developed a computational method to discover the spatially active chromatin segments termed frequently interacting regions (FIREs). We find FIREs are enriched for active enhancer regions, harboring super-enhancers as well as disease-associated variants in the corresponding disease-relevant tissue type. In addition, FIREs are substantially conserved between human and mouse genomes of the same cell type, and their formation depends in part on the Cohesin complex and CTCF. Finally, most FIREs exhibit promiscuous interactions in the local chromatin neighborhood. These observations improve our understanding of the role of dynamic chromatin organization in the regulation of tissue-specific gene expression programs in human cells.

RESULTS

Compendium of Chromatin Organization Maps across 21 Human Cell and Tissue Types

We conducted Hi-C analysis on 14 primary human tissues collected from four donors (Figure 1A), for which epigenome datasets had been produced as part of the NIH Epigenome Roadmap project (Roadmap Epigenomics Consortium et al., 2015). We combined the resulting datasets with those previously generated by us for seven cultured cell types using a common experimental protocol that was reported separately (Dixon et al., 2012, 2015; Jin et al., 2013; Selvaraj et al., 2013). The combined datasets were processed using a common data processing pipeline, after merging data from biological replicates deemed as reproducible (Figures S1A–S1E). Collectively, we analyzed >8.6 billion unique contacts, out of which >2.5 billion were long-range (>15 kb) intra-chromosomal contacts, with 809M unique contacts and 254M long-range *cis* contacts per cell line and 214M unique contacts and 53M long-range *cis* contacts per tissue type (Table S1). We first analyzed compartment A/B patterns in each tissue/cell type (Figure 1B; Table S2). As previously reported for cultured human cells (Dixon et al., 2015), we observed substantial compartment A/B switching across primary tissues (Figures 1B and 1C), finding that 59.6% of the genome is dynamically compartmentalized in different tissues and cell types. These data also underscore the significant degree of compartment conservation across the genome, revealing that as much as 40.4% of the genome is invariant, which is a statistically significant degree of invariant genome compartmentalization (chi-square test p value < 2.2e−16) (Figure S1F).

TADs have been reported to be stable across different cell types and experimental conditions and conserved in related species (Dixon et al., 2012, 2015; Rao et al., 2014; Zuin et al., 2014). To investigate the degree of TAD boundary conservation in primary human tissues, we applied the insulation score method (Crane et al., 2015), which is robust in sequencing depth (Figures S1G–S1I) to identify TAD boundaries at 40-kb bin resolution (Table S3). We identified a total of 3,010 distinct TAD boundaries in 21 samples (14 tissues and 7 cell lines). Upon careful inspection of a broad panel of genetic loci (Figures 1A and 1D) as well as systematic comparison across samples (Figures 1D and 1E), we find that TAD boundaries are indeed highly conserved across different cell lines and tissues. These results are highly significant, considering that, by chance, only 1.7% of TAD boundaries are expected to share for all (chi-square test p value < 2.2e−16).

Identification of Frequently Interacting Regions in the Human Genome

As a means to investigate conserved and tissue-specific chromatin interactions, we first used Fit-Hi-C (Ay et al., 2014) to

Figure 1. Global Features of 3D Genome Organization in 7 Cell Lines and 14 Adult Tissues

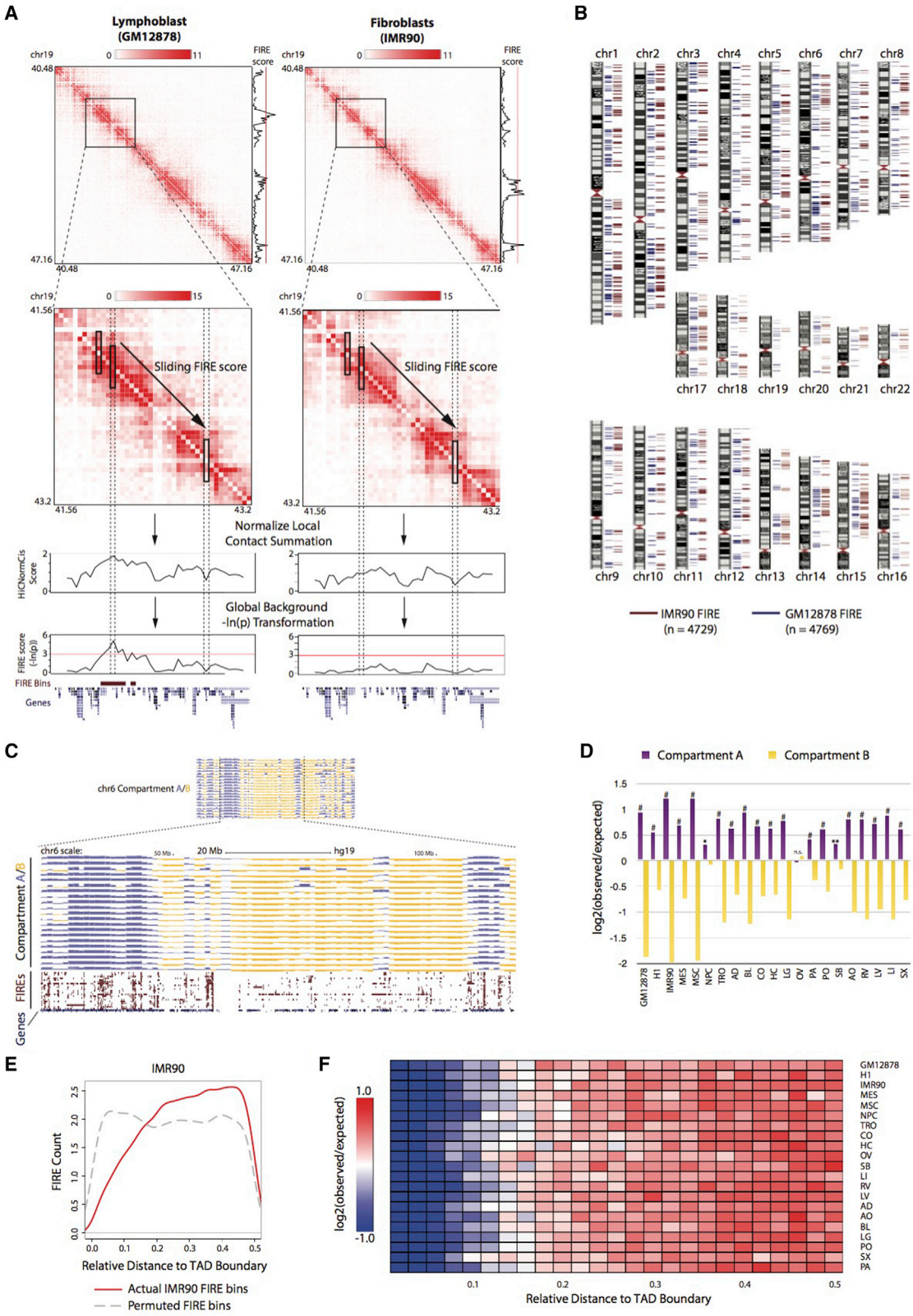
(A) Illustration of the primary 21 Hi-C datasets analyzed, depicting the cell (left panel) or tissue (right panel) origin of the samples as well as the germ layer origin for tissues (right panel). Hi-C interaction patterns across an 11.68-Mb region (chr12:82,840,000–94,520,000) are shown for all 7 cell lines and 14 tissues at 40-kb bin resolution.

(B) Genome browser snapshot showing compartment A/B patterns (PC1 value) across chromosome 2 in 21 samples, with 7 cell lines at the top and 14 primary adult tissues on the bottom. Compartment A/B patterns are at 1-Mb bin resolution. Positive PC1 in blue corresponds to compartment A, and negative PC1 in yellow corresponds to compartment B.

(C) Bar plots showing the degree of conservation of A/B compartment labels of 21 human cell lines and adult tissues. The y axis is the fraction of the genome conserved by the 22 possible combinations of compartment A/B designations. The label below each bar represents the composition of the compartment designations. For example, "16A/5B" represents the genomic region where 16 samples exhibit a compartment A label and the other five samples exhibit a compartment B label.

(D) Genome browser snapshot showing topological domain boundaries across chromosome 7 in 21 samples, with 7 cell lines at the top and 14 primary adult tissues on the bottom. Boundaries are identified at 40-kb bin resolution.

(E) Bar plots showing the degree of topological domain boundary conservation across 21 human cell lines and tissues. For each putative boundary region, we tallied how many samples have a boundary within that region (see Supplemental Experimental Procedures). Shown here is a total fraction of TAD boundary regions, whereby the y axis is the fraction of TAD boundaries conserved at least a certain number of samples, as categorized along the x axis.



(legend on next page)

identify significant chromatin interactions at various significance thresholds (Table S4). However, Fit-Hi-C, like other peak-calling methods (Jin et al., 2013; Rao et al., 2014; Xu et al., 2015, 2016), is sensitive to sequencing depth, and therefore we found considerable variation in total chromatin contacts between samples, precluding any statistically rigorous comparative peak-calling analysis across tissues. However, upon closer examination of the chromatin contacts near the contact matrix diagonal (± 200 kb from the matrix diagonal), we noticed that some regions exhibit unusually high levels of local contact frequency in a tissue-type-dependent manner (Figure 2A). We therefore developed a computational approach to normalize and compare local interaction frequencies across all 21 tissues and cell types. Specifically, we developed a Poisson-regression-based normalization approach (termed as "HiCNormCis") to normalize the total raw local (15–200 kb) *cis* contacts for each 40-kb bin genome-wide (Figure S2A; Supplemental Experimental Procedures). This method removes bias from three sources known to affect Hi-C data, including effective restriction fragment lengths, GC content, and sequence mappability (Hu et al., 2012; Yaffe and Tanay, 2011). Compared to other normalization approaches, such as HiCNorm (Hu et al., 2012), vanilla coverage (Lieberman-Aiden et al., 2009), and iterative correction and eigenvector decomposition (ICE) (Imakaev et al., 2012), HiCNormCis achieved the best performance for bias removal (Figure S2B). Lastly, we used a Gaussian distribution to approximate the normalized total local *cis* contacts (Figure S2C), and converted HiCNormCis output values to −ln(p value), which we define as the final "FIRE score." FIREs (also termed "FIRE bins") are therefore defined as bins with a one-sided p value less than 0.05, corresponding to −ln(p value) greater than 3 (Figure 2A). We found that our FIRE scores were highly reproducible (Figures S2D and S2E), and robust to sequencing depth (Figures S2A and S2F), choice of restriction enzymes in Hi-C library preparation (Figures S2G and S2H), as well as choice of experimental protocols, such as dilution Hi-C or in situ Hi-C (Figure S2I).

We first identified FIREs in GM12878 and IMR90 cells (Figures 2A and 2B). Global analysis of FIREs revealed a dispersed distribution along the genome (Figure 2B). We next determined FIREs in the remainder of tissues and cell lines (Tables S5 and S6) after removing local genomic feature biases (Figure S2J). We then explored how FIREs are positioned in relation to A or B compartments as well as in relation to TAD boundaries (not chromatin "loops"). Careful inspection of FIRE positioning and genome-wide enrichment analyses indicated that FIREs are enriched in compartment A and depleted in compartment B (Figures 2C and 2D; Table S7). We also examined the FIRE distribution within TADs, and found that FIREs are depleted near TAD boundaries and enriched within TADs and toward the TAD center (Figures 2E and 2F).

FIREs, Chromatin Loops, and Insulated Neighborhoods

We further analyzed FIREs at 5-kb resolution using previously published in situ Hi-C data in IMR90 and GM12878 (Rao et al., 2014), and compared FIRE positioning relative to the smaller (~185 kb) chromatin "loops." As expected, FIREs are significantly enriched for chromatin loop anchors (chi-square test p value < 2.2e−16); however, ~90% of FIREs are within loops, and these FIREs demonstrate unique properties to be discussed in the following sections. Our data indicate that FIREs are hotspots of local chromatin interactions that are distinct from compartments, TADs, and chromatin loops (Rao et al., 2014), which are generally anchored by convergent CTCF binding. By contrast, most FIREs are located within TADs and chromatin loops, indicating they represent specific loci "within the loop" at higher resolution. Similarly, FIREs are likely distinct from insulated neighborhoods due to the high positional overlap between the CTCF-mediated "chromatin loops" and "insulated neighborhoods" (Ji et al., 2016). Our analysis of FIREs and insulated neighborhoods at 40-kb resolution in H1 cells indicates that insulated neighborhoods are also enriched for FIREs (chi-square test p value = 5.32e−15), but >70% of insulated neighborhoods do not contain a FIRE (Figure S3D) (also discussed more below).

FIREs Are Tissue-Specific and Located Near Cell Identity Genes

To characterize the tissue-specificity of FIREs, we combined all 21 datasets (7 cell lines and 14 tissues), and performed a

Figure 2. Identification and Positional Enrichment of Frequently Interacting Regions

(A) Illustrative examples showing the FIRE score methodology. Hi-C contact maps from a 6.68-Mb region (chr19:40,480,000–47,160,000) are shown for GM12878 and IMR90 cells at 40-kb bin resolution (top). To the right of the contact maps are line plots showing the fully processed FIRE score for each 40-kb bin. A red line is drawn at the significance cutoff. The second row of contact maps illustrates FIRE scores in a sub-matrix (chr19:41,560,000–43,200,000) of the above contact maps (black box). Line plots directly below show the intermediate stage in the FIRE score calculation, which is the output from HiCNormCis (see Supplemental Experimental Procedures). Genome-wide HiCNormCis normalized counts are then *Z* score transformed and converted to a −ln(p value) scale to obtain the final FIRE score (bottom line plots). Dashed columns highlight two 40-kb bins, one showing a FIRE peak in GM12878 cells, but not in IMR90 cells, and the other showing a low FIRE score in both cell types.

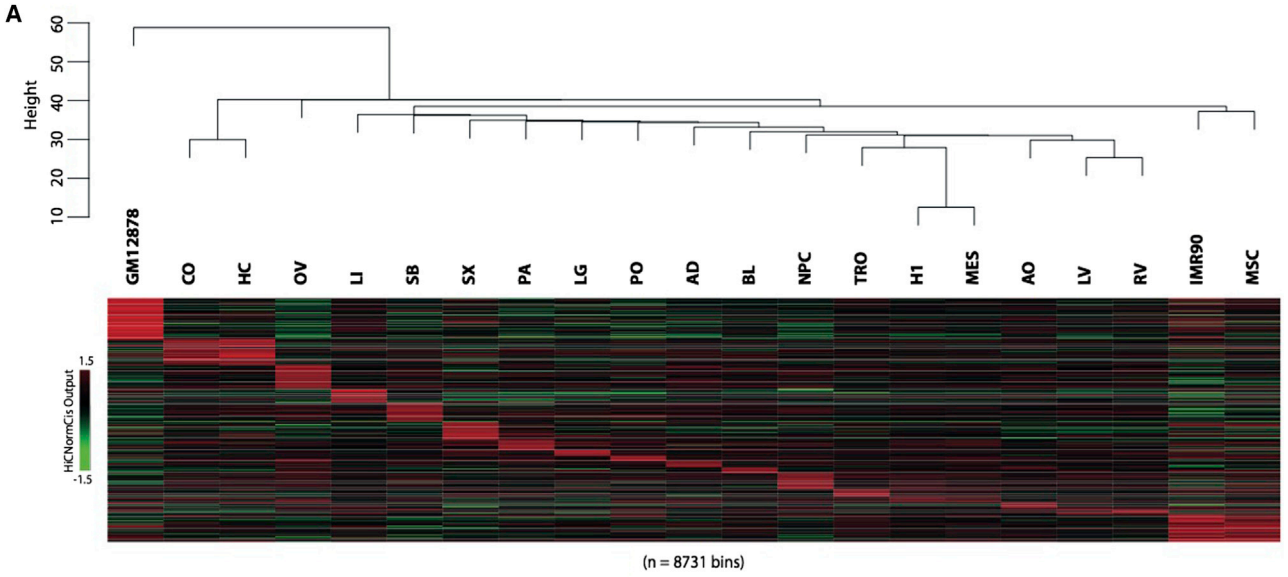
(B) Chromosome ideograms showing the genome-wide positional distribution of FIRE bins in GM12878 (blue, n = 4,769) and IMR90 (maroon, n = 4,729). Genome-wide visualization captures both conserved and specific FIRE bins. Only autosomes are depicted.

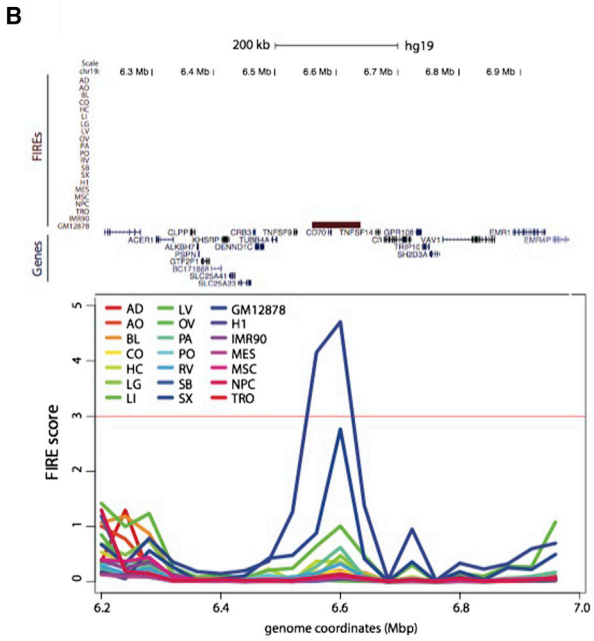
(C) Genome browser snapshot of compartment A/B patterns in 21 samples across chromosome 6 (top), and a genome browser snapshot of a 90-Mb subset of chromosome 6 (chr6:25,000,000–115,000,000) showing compartment A/B patterns for 21 samples (top set, blue/yellow) and FIRE calls (bottom set, maroon).

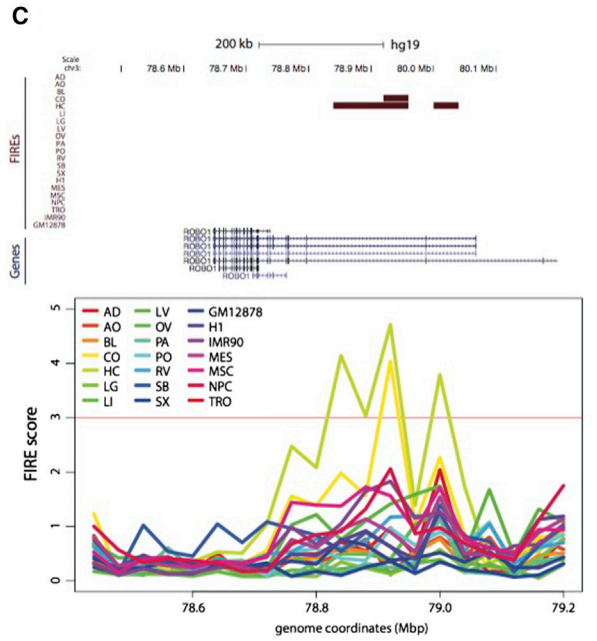
(D) Bar plots showing an enrichment analysis of FIRE positioning within either compartment A or B, illustrating FIREs are enriched in compartment A and depleted in compartment B compared to random permutation of the FIRE bin location within each sample (*p < 5.0e−7; **p < 7.0e−13; #p < 2.2e−16; chi-square test). Statistical tests correspond to the significance of FIRE enrichment in compartment A.

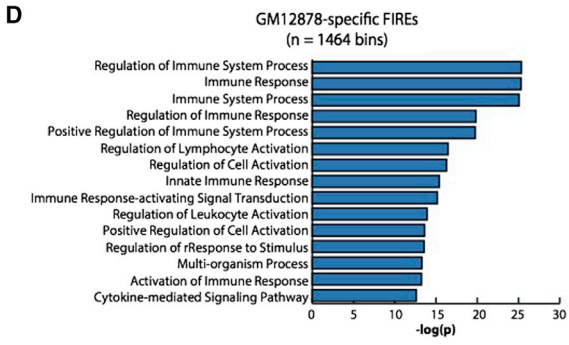
(E) Line plot showing an example of IMR90 FIRE bin positioning relative to TADs (see Supplemental Experimental Procedures). The red line depicts the observed counts (y axis) of actual IMR90 FIRE bins, whereas the gray dashed line shows the counts of permuted FIRE bin locations. The x axis ranges from 0 to 0.5, where 0 represents TAD boundaries and 0.5 represents TAD center points.

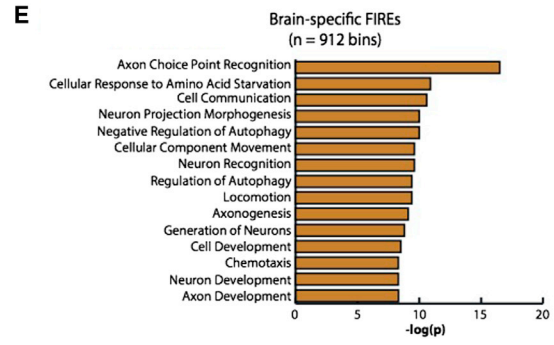
(F) Heat map showing the TAD position enrichment analysis across all 21 samples. Shown are the log2(observed/expected) values for each distance increment, as computed in (E).











(legend on next page)

comparative analysis (Figure 3A; Table S6). Approximately 38.8% (8,142/20,974 bins) of FIREs were identified in only one tissue or cell type, and approximately 57.7% (12,094/20,974 bins) of FIREs were identified in two or fewer, revealing the highly tissue-specific nature of FIREs (Figure S2K). Further, a hierarchical clustering analysis of genome-wide FIRE scores revealed similarities among certain cell types, such as H1 and MES, as well as MSC and IMR90 (Dixon et al., 2015) (Figure 3A). As expected, tissues from the same organ (brain: cortex and hippocampus; heart: left ventricle and right ventricle) clustered together (Figure 3A). Tissue-specific FIREs tend to be positioned in close proximity to genes related to the cellular identity (Figures 3B and 3C). For example, within a GM12878-specific FIRE is the promoter for *CD70*, a gene well known for its role in immune cell activation and maturation (Arens et al., 2004) (Figure 3B). Moreover, ~110 kb from a FIRE region present only in brain tissues is an alternative *ROBO1* promoter, a gene involved in axon guidance during development (Leyva-Díaz et al., 2014) (Figure 3C). To extend these observations to all tissue-specific FIREs and to interpret the functional roles and disease relatedness of these FIREs, we performed GREAT analysis (McLean et al., 2010) (Tables S8 and S9). The results showed that genes in close proximity to tissue-specific FIREs are related to the functionality of that tissue/cell type (Figures 3D and 3E; Tables S8 and S9). Moreover, using only our 5-kb resolution FIRE calls in GM12878 and IMR90, we also found abundant sample-specific FIREs (~57% of FIREs are sample specific), and confirmed that sample-specific FIREs are positioned near cell identity genes (Tables S8 and S9) at a higher resolution. Collectively, these results suggest that FIREs are closely associated with cell identity and tissue function.

FIREs Are Enriched for Active Enhancers and Super-Enhancers

Because FIREs tend to be positioned near genes related to cell identity and tissue function, we posited that FIREs may be enriched for active enhancers. To test this hypothesis, we analyzed previously generated ChIP-seq data for six histone modifications (H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K27me3, and H3K9me3) for these tissues and cell types (Roadmap Epigenomics Consortium et al., 2015). We observed that FIREs display a high density of active chromatin features (e.g., H3K27ac and H3K4me1), and overlap with super-enhancers found in the same tissues (Hnisz et al., 2013) (Figure 4A). We then characterized the histone modification signatures across 1-Mb regions centered at FIREs. FIREs are ubiquitously enriched for two active enhancer marks, H3K4me1 and H3K27ac, and depleted for the repressive chromatin mark H3K27me3 (Figure 4B), whereas enrichment of other marks did not show clear patterns (Figure S3A). FIREs also overlap with typical enhancers and super-enhancers (Hnisz et al., 2013) annotated in the cell lines and tissues where such data are available (Figures 4C and 4D). For example, 35.0% of typical enhancers and 77.8% of super-enhancers annotated in GM12878 cells overlap FIREs (Fisher's exact test p value < 2.2e−16) (Figures 4C and 4D). Importantly, we also found significant enrichment for FIREs at typical enhancers and super-enhancers (chi-square test p value < 2.2e−16) when analyzing FIREs at 5-kb bin resolution (Table S6) using previously published high-resolution Hi-C data in GM12878 and IMR90 (Rao et al., 2014) (Figure S3B). Also, with respect to previously annotated chromatin loops (Rao et al., 2014), we find that the aforementioned 90% of FIREs that do not overlap loop anchors are also significantly enriched for typical and super-enhancers (chi-square test p value < 2.2e−16). For example, we observed GM12878-specific FIREs corresponding to a GM12878-specific super-enhancer, whereas the same locus in IMR90 lacks any enhancer or FIRE, despite sharing a conserved chromatin loop (Figure S3C). These FIRE analyses at 5-kb resolution corroborate our findings at 40-kb resolution, and indicate that FIREs represent distinct structural entities with differing biochemical properties compared to chromatin loops. As anticipated, we also find a significant overlap between FIREs and super-enhancer domains in mouse embryonic stem cells (mESCs) at 40-kb resolution (chi-square test p value = 0.0052), but not polycomb domains (Dowen et al., 2014; Ji et al., 2016), further underscoring the role of FIREs in active gene regulation (Figure S3D).

Because many FIRE bins were found in clusters, we stitched together adjacent FIRE bins and ranked them by cumulative *Z* score, revealing that a small proportion of FIRE clusters (termed "super-FIREs") contain the majority of bins with the most significant local interaction frequency (Figure S3E). Strikingly, compared to all FIREs (Figure S3F), we observed some tissues, in which nearly 100% of super-FIREs contain either a super-enhancer or typical enhancer (Figure S3G), suggesting that the bins with the highest local interaction frequency almost always mark active enhancer(s). Analysis of super-FIREs not containing an enhancer revealed a moderate enrichment for H3K27me3 across most testable samples, but no other clear trends (Figures S3H–S3M). Given this striking relationship, we wondered to what

Figure 3. FIREs Are Tissue-type Specific and Enriched Near Genes Involved in Tissue Function

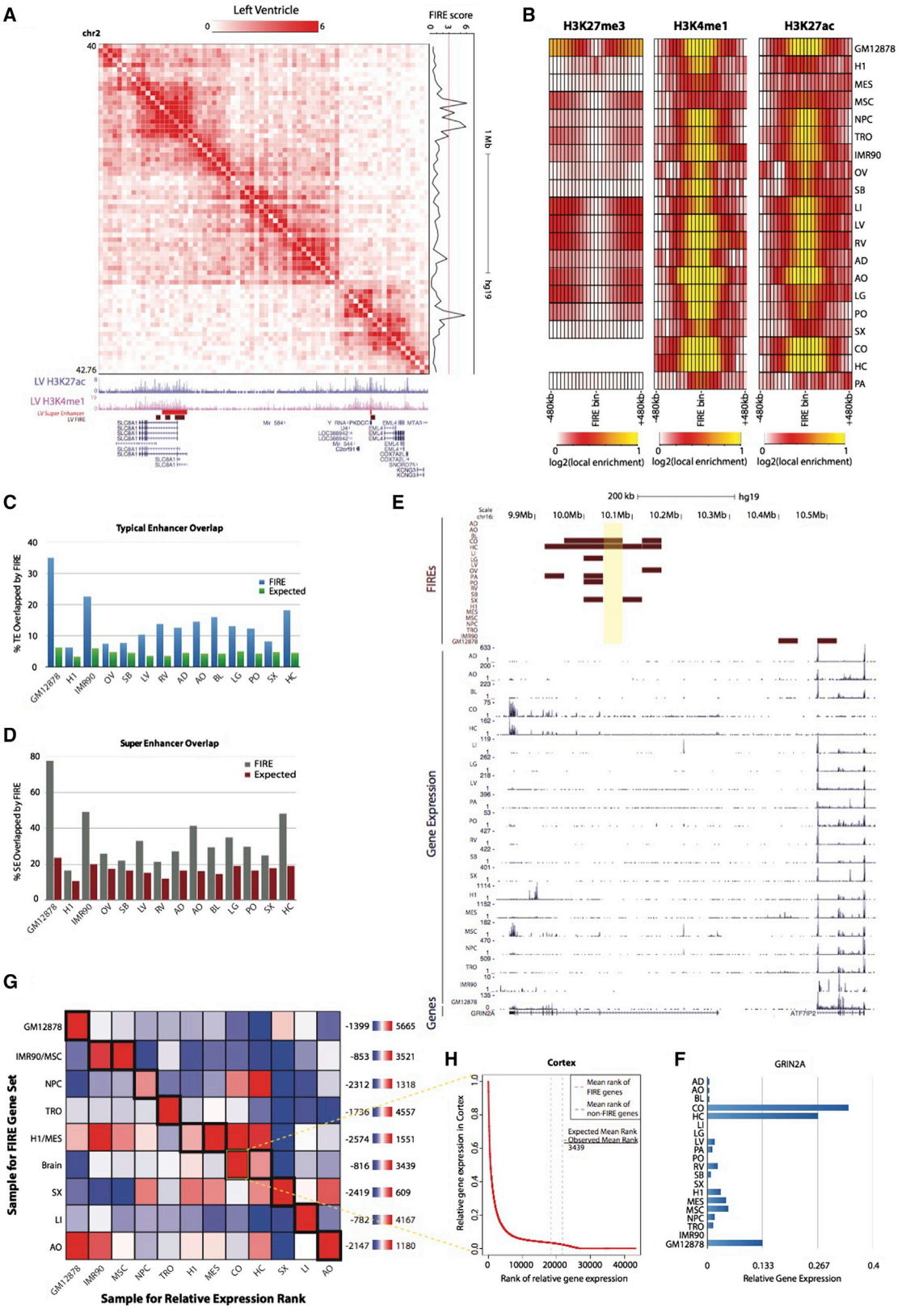
(A) At the top is a dendrogram resulting from a hierarchical clustering analysis using genome-wide FIRE scores for each sample. The y axis is the Euclidean distance between FIRE scores from any two samples. The heat map below shows a subset of FIRE bins (n = 8,371), corresponding to FIRE bins that are called as FIRE in only one or two samples. For ventricle tissues, brain tissues, IMR90/MSC, and H1/MES, FIREs specific to two samples are allowed in the definition of sample specific.

(B) Genome browser snapshot showing a GM12878-specific FIRE region (chr19:6,560,000–6,640,000) (top, maroon) in an 800-kb region around *CD70* (chr19:6,583,193–6,604,114). Below is a line plot of FIRE scores for each sample, showing the GM12878-specific FIRE peak (blue).

(C) Genome browser snapshot showing a brain-specific FIRE region (chr3:78,920,000–78,960,000), shared by CO and HC, in a 760-kb region within *ROBO1* (chr3:78,646,338–79,068,609). Below is a line plot of FIRE scores for each tissue showing CO (yellow) and HC (pea green) FIRE peaks.

(D) GREAT biological process analysis of genes surrounding GM12878-specific FIRE bins (n = 1,464 bins), showing biological processes highly related to immune functions. Plotted values are the −log10 of the Bonferroni-corrected binomial p values.

(E) Same as (D), except using genes surrounding brain (CO and HC) specific FIRE bins (n = 912 FIRE bins) showing several significant processes highly related to brain functionality. Plotted values are the −log10 of the Bonferroni-corrected binomial p values.



(legend on next page)

extent FIRE analysis could be used to predict the locations of typical and super-enhancers in GM12878. By varying the significance thresholds for FIRE calling and performing a receiver operating characteristic (ROC) area under curve (AUC) analysis, we find an impressive predictive power of FIRE analysis to identify typical enhancers and super-enhancers using Hi-C data alone (AUC = 0.813 and AUC = 0.906, respectively) (Figures S3N and S3O). Taken together, the high overlap between super-enhancers and FIREs, as well as the enrichment of tissue identity genes near tissue-specific FIREs, implicates a potential *cis*-regulatory role for FIREs in facilitating tissue-specific gene expression.

FIREs Are Near Tissue-Specifically Expressed Genes

Because super-enhancers are known to be tissue-specific and positioned near cell identity genes, we asked if FIREs are nearby genes that are more transcriptionally active in the corresponding tissue/cell types. By re-analyzing publicly available RNA-seq data (Roadmap Epigenomics Consortium et al., 2015), we indeed found a strong correlation between cell/tissue-specific FIREs and cell/tissue-specific expression of nearby genes. For example, the *GRIN2A* gene, which encodes an important ligand- and voltage-gated N-methyl-D-asparate (NMDA) receptor subunit implicated in epilepsy (Kingwell, 2013) and schizophrenia (Ohi et al., 2016), is predominantly expressed in brain tissues, and the transcription start site (TSS) is ~197 kb from a brain-specific FIRE (Figure 4E). In *GRIN2A*, the relative gene expression in cortex (CO) is the highest among all tissues (Figure 4F; see Supplemental Experimental Procedures). We also calculated the relative gene expression for each gene within 200 kb of a tissue-specific FIRE across all tissues and found significant correlation between tissue-specific FIREs and tissue-specifically expressed genes (Figure S3P). For example, we found that the GM12878-specific FIRE gene set contained genes with significantly higher relative expression in GM12878 compared to any other FIRE gene set (two-sample t test p value < 9.26e−6) (Figure S3P).

Intrigued by these observations in brain tissue and lymphoblast cells, we applied a more systematic mean-rank gene set enrichment test (see Supplemental Experimental Procedures) to further understand the relationship between FIREs and gene expression patterns. For example, in cortex tissue, there is a clear difference between the mean ranks of genes neighboring brain-specific FIREs compared to random FIRE positioning (Figures 4G and 4H). Importantly, this type of analysis can be used to study the extent to which tissue-specific FIRE genes are expressed by testing all combinations of relative expression rank lists and tissue-specific FIRE gene sets (Figure 4G). In other words, if tissue-specific FIRE genes are primarily expressed in that same sample, the enrichment signal should track the diagonal of an all by all comparison (Figure 4G) and generally lower enrichment off the diagonal where the sample for the rank list and FIRE gene set are different. Indeed, we observed this trend, although the neural progenitor cell (NPC)-specific FIRE gene set is ranked higher in the cortex and hippocampus, which may be expected, given that they prominently consist of neural cells or neural progenitors. Taken together, our results suggest that tissue-specific FIREs are likely involved in tissue-specific gene expression.

FIREs Are Conserved in Humans and Mice

If FIREs play a role in gene regulation and developmental programs, one would expect that such chromatin features would be conserved evolutionarily (Dixon et al., 2012, 2015; Vietri Rudan et al., 2015). To test this hypothesis, we compared FIREs between humans and mice in three different sample types (embryonic stem cells, neural progenitor cells, and cortex tissue) (Dixon et al., 2012, 2015; Fraser et al., 2015; Shen et al., 2012). We found that FIREs are significantly conserved in these comparisons (Figure 5A). Specifically, 33.0% of human cortex FIREs

Figure 4. FIREs Are Enriched for Active Enhancers and Positioned Near Tissue-Type-Specific Genes

(A) Normalized Hi-C contact matrix in left ventricle tissue showing a 2.76-Mb locus (chr2:40,000,000–42,760,000). Below are genome browser tracks for previously published (Hnisz et al., 2013) LV super-enhancers (red), LV FIRE bins (brown), and UCSC genes, including isoforms (blue). To the right is the continuous LV FIRE score along this locus.

(B) Heat maps showing the local enrichment (see Supplemental Experimental Procedures) of H3K27me3 (left), H3K4me1 (middle), and H3K27ac (right), centered on FIRE bins for each cell line or adult tissue. H3K27me3 data were not available for CO or HC.

(C) Bar plot showing the observed overlap between actual FIRE bins and previously characterized typical enhancers (blue) (Hnisz et al., 2013) for each available cell line or tissue that has both Hi-C data and typical enhancer calls. Expected values are also shown (green), which are calculated by permuting the location of FIRE bins within each tissue and calculating the overlap with typical enhancers. The y axis shows the percentage of typical enhancers overlapped by FIREs.

(D) Same as (C), except showing the percentage of super-enhancers overlapped by FIRE bins for each testable cell line or tissue.

(E) Genome browser snapshot showing an example of sample-specific gene expression near sample-specific FIREs. Shown here is a 780-kb locus (chr16:9,820,000–10,600,000) around *GRIN2A* (chr16:9,852,375–10,276,611). At the top, FIRE tracks (maroon) for each sample, showing the brain-specific FIRE (chr16:10,040,000–10,080,000, highlighted in yellow) ~197 kb away from *GRIN2A* TSS. Below, RNA-seq data (Roadmap Epigenomics Consortium et al., 2015) for all samples except OV (blue), showing *GRIN2A* is mainly expressed in brain tissues.

(F) Bar plot indicating the relative gene expression (see Supplemental Experimental Procedures) of *GRIN2A* across 20 samples.

(G) All-by-all mean-rank enrichment analysis result showing gene expression specificity of genes within 200 kb of sample-specific FIRE bins (see Supplemental Experimental Procedures). Each row is a different sample type for which the sample-specific FIRE gene set is collected, and columns are the sample type used to calculate the relative expression rank of each gene. IMR90/MSC, M1/MES, and brain tissues were previously shown to have highly overlapped FIRE bins (Figure 3A) and are therefore grouped. The color for each row of the heat map indicates the enrichment. Outlined in thick black boxes along the diagonal are the matrix entries for which the sample for the sample-specific FIRE gene set and expression rank list are the same. Highlighted in a thin yellow box is the analysis portrayed in (H).

(H) Line plot illustrating a single mean-rank enrichment analysis. The plot shows the relative gene expression values (y axis) in the cortex as a function of their numeric ranking (x axis) in the cortex. Vertical dashed lines show the position of the observed mean rank of cortex-specific FIRE genes (red dash), and the expected mean rank based on size-matched randomly selected non-FIRE bins in the cortex (gray dash). The inset is the calculation of the enrichment score.

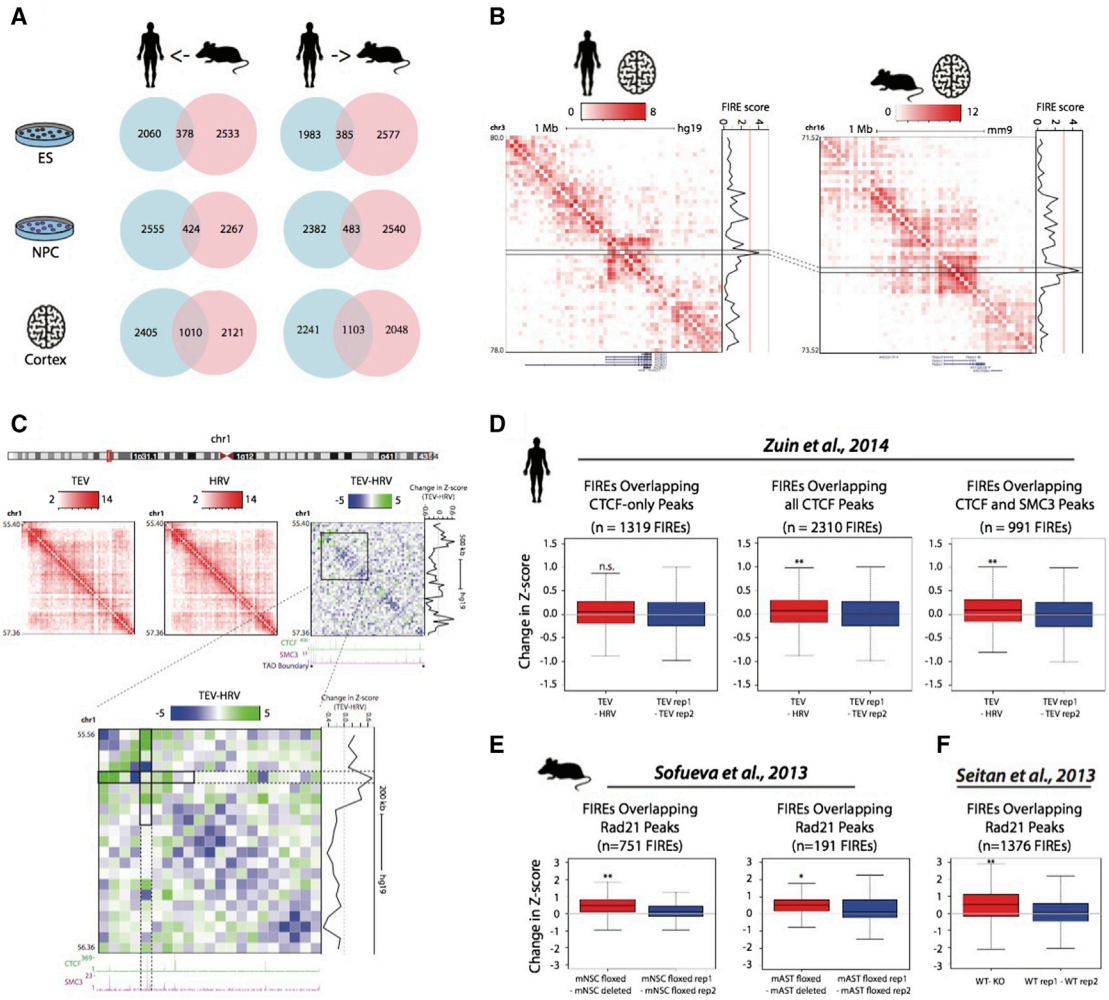


Figure 5. FIREs Are Conserved across Evolution and Mediated by Cohesin

(A) Venn diagrams showing the significant number of conserved FIRE bins when lifting over mouse FIREs onto the human genome (left column) or lifting over human FIREs onto the mouse genome (right column) in either embryonic stem cells (top row, p value < 5.0e−16), neural progenitor cells (middle row, p value < 2.2e−16), and cortex tissue (bottom row, p value < 2.2e−16). Significance evaluated using a Fisher's exact test (see Supplemental Experimental Procedures).

(B) Normalized Hi-C contact matrix in human cortex (left) and mouse cortex (right) for a 2-Mb syntenic region (human chr3:78,000,000–80,000,000; mouse chr16:71,520,000–73,520,000) showing a conserved FIRE (connected black lines) within the same tissue type but across species. Below is a UCSC gene track, and to the right of the contact matrix is the continuous FIRE score across the locus. For the human data, the Hi-C contact matrix, gene track, and FIRE score plot have been inverted to show synteny with the mouse data.

(C) Normalized Hi-C contact matrices (red and white) or delta matrix (green and blue) for the 1.96-Mb locus (chr1:55,400,000–57,360,000) illustrating the change of interaction frequency between TEV and HRV. Directly below the delta matrix are binding profiles of CTCF and the Cohesin subunit SMC3 in wild-type HEK cells (Zuin et al., 2014) as well as TAD boundary annotations. To the right of the Hi-C delta matrices is the continuous FIRE *Z* score difference between TEV and HRV. Below is a delta matrix at a zoomed-in 800-kb region (chr1:55,560,000–56,360,000) for TEV-HRV, showing the greatest reduction of FIRE score occurs at the bin with co-binding of CTCF and SMC3. The FIRE *Z* score difference is plotted to the right of the subtraction matrices.

(D) Box plots showing the change in *Z* score at FIREs overlapping bins bound by CTCF but not SMC3 "CTCF-only" (left plot), all CTCF peaks (middle plot), and CTCF and SMC3 co-binding (right plot) for the comparison of TEV and HRV. The red boxes show distributions of FIRE score change at FIRE bins called in wild-type cells minus the mutant cells, whereas the blue boxes are distributions for FIRE score change at FIRE bins called in wild-type cells but between biological replicates of wild-type cells. These comparisons show the significant reduction of FIRE score at all CTCF peaks, and especially at CTCF SMC3 co-bound peaks overlapping FIRE bins (*p = 1.0e−4; **p = 4.04e−5; two-sample t test).

(E) Similar to (D), except analysis of *Z* score change was done considering FIREs overlapping the Cohesin subunit Rad21 peaks using previously published Hi-C data and Rad21 ChIP-seq data in mouse neural stem cells (left plot) and mouse post-mitotic astrocytes (middle plot) (Sofueva et al., 2013). Comparison of *Z* score change upon deletion of Rad21 shows a significant decrease compared to changes observed between biological replicates (*p < 0.01; **p < 2.2e−16; two-sample t test).

(F) Similar to (E), except analysis of *Z* score change was conducted on previously published Hi-C data and Rad21 ChIP-seq data in mouse thymocytes (Seitan et al., 2013). Comparing the distributions of *Z* score changes at FIRE bins bound by Rad21 shows a significant reduction in *Z* score between the wild-type and Rad21 knockout cells compared to changes between wild-type biological replicates (**p < 2.2e−16; two-sample t test).

are also FIREs in the mouse cortex, whereas only 8.7% is expected by chance (Fisher's exact test p value < 2.2e−16). For example, returning to the *ROBO1* locus, we found that both the mouse and human cortex have only one FIRE bin in the 2-Mb region around *ROBO1*, and the single FIRE position is conserved across species (Figure 5B). Interestingly, the degree of FIRE conservation between a human and mouse is the highest in cortex tissue and less, although statistically significant, in embryonic stem cells and neural progenitor cells (ESC p value < 5.0e−16; NPC p value < 2.2e−16, Fisher's exact test) (Figure 5A). More generally, by randomly sampling syntenic bins across a range of FIRE scores, we find a modest yet significant correlation of FIRE score between a human and a mouse in each cell type (Pearson correlation coefficient = 0.20–0.42; p value < 2.2e−16) (Figures S4A–S4F). These data indicate a tendency for the local contact frequency to be conserved in syntenic regions throughout the human and mouse genome as well as conservation of the strongest locally interacting hotspots.

CTCF and Cohesin Complex Contribute to Establishment of FIREs

We posited that FIREs might be mediated by the Cohesin complex, which has been previously shown to modulate enhancer/promoter interactions in mammalian cells (Kagey et al., 2010). To test this hypothesis, we re-analyzed three previously published Hi-C datasets, in which a Cohesin subunit was experimentally depleted in human or mouse cells (Seitan et al., 2013; Sofueva et al., 2013; Zuin et al., 2014), and investigated FIRE scores upon loss of a Cohesin subunit. We began by systematically examining the Hi-C datasets generated in HEK293 cells before and after depletion of the Cohesin subunit SMC3 (Figure 5C). Because the Cohesin complex is frequently bound together with CTCF throughout the genome, we focused our analysis to CTCF-only binding sites and CTCF/SMC3 co-bound peaks. SMC3-only peaks were ignored because only ~0.7% of SMC3 peaks overlapping FIREs were not co-occupied with CTCF (Figure S4G). We then compared FIRE score changes at FIRE bins upon loss of SMC3. We observed a significant decrease of the FIRE score at CTCF/SMC3 co-bound sites (two-sample t test p value = 6.78e−6 for TEV-HRV) (Figures 5C and 5D). By contrast, there is no statistically significant FIRE score decrease at FIRE bins that had CTCF binding *without* binding of SMC3 (Figure 5D). Quantitatively similar results were seen in mouse neural stem cells, post-mitotic astrocytes, and thymocytes in the case of Rad21 deletion (two-sample t test p value = 0.0011 for post-mitotic astrocytes; two-sample t test p value < 2.2e−16 for both neural stem cells and thymocytes) (Figures 5E and 5F) (Seitan et al., 2013; Sofueva et al., 2013). Importantly, the significant decrease of the FIRE score was only observed at FIRE bins. Cohesin loss did not systemically affect FIRE scores at randomly selected and size-matched (5% of the genome) control regions (Figures S4H and S4I). We also re-analyzed Hi-C data in HEK293 cells, in which CTCF had been experimental knocked down (Zuin et al., 2014), and again observed that FIRE score is most significantly reduced at FIRE bins occupied by CTCF/SMC co-binding in wild-type cells (Figure S4J). Collectively, these results, as well as the significant enrichment of Cohesin at FIRE bins (Figure S4K), suggest that both CTCF and the Cohesin complex contribute to the formation of FIREs, and such a mechanism is likely conserved across the human and mouse.

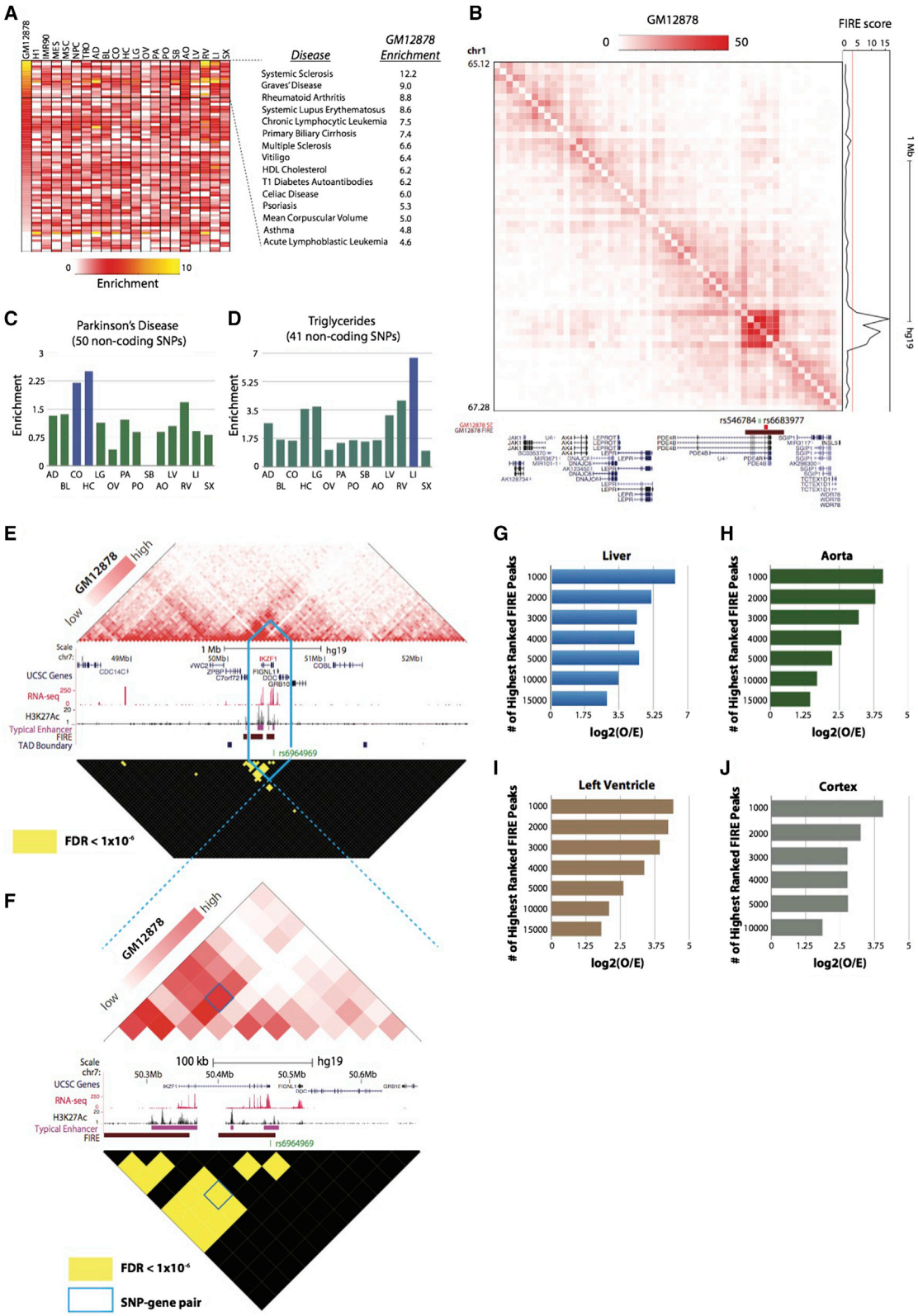
FIREs Are Enriched for Disease-Associated SNPs

Our analyses have indicated that FIREs are enriched for active enhancers and super-enhancers (Figures 4A–4D; Figures S3B, S3C, S3F, S3G, S3N, and S3O). Because typical and super-enhancers contain a significant proportion of disease-associated SNPs (Hnisz et al., 2013), we further investigated the overlap between FIREs and disease-associated SNPs. First, we mapped 4,327 previously annotated disease-associated non-coding SNPs to FIREs defined in each cell line and tissue (see Supplemental Experimental Procedures) (Hnisz et al., 2013). Consistent with previous results (Hnisz et al., 2013), we observed 7.06 and 3.76 SNPs per megabase, and among 354 GM12878 FIREs overlapped with super-enhancers and 2,800 GM12878 FIREs overlapped with typical enhancers, respectively (Figure S5A). Surprisingly, among 1,615 GM12878 FIREs that do *not* overlap an annotated enhancer, we also observed 3.33 SNPs per megabase, which is ~2.3-fold higher than the genome-wide SNP density (1.42 SNPs per megabase) (Figure S5A). Importantly, these SNPs would not be captured by directly overlapping super-enhancers or typical enhancers with disease-associated SNPs (Hnisz et al., 2013).

Next, we examined the overlap between disease-associated SNPs and FIREs for 456 diseases and quantitative traits (Hnisz et al., 2013). We defined the enrichment score for each disease as the ratio between the proportion of SNPs overlapped with FIREs and the proportion of FIRE bins in the genome. Strikingly, numerous immune-related diseases exhibit strong SNP enrichment in GM12878, but mild or weak enrichment in the other cell lines or tissues (Figure 6A). In fact, the vast majority of the top enrichment scores come from diseases previously implicated with immune pathology (Jostins et al., 2012) (Figure 6A). Motivated by these observations, we closely examined genes near FIREs harboring disease-associated SNPs, and found many genes associated with that type of disease. For example, two SNPs associated with acute lymphoblastic leukemia (ALL), rs6683977 and rs546784, are within a GM12878-specific super-FIRE (Figure 6B) and within *PDE4B*, a gene associated with ALL (Yang et al., 2011).

We then conducted an SNP enrichment analysis for the tissue datasets and observed similar results for some diseases and quantitative traits, with the most striking findings in the brain and liver (Figures 6C and 6D; Figures S5C and S5D). A careful examination of SNP and FIRE overlap also revealed disease candidate genes. For example, two Alzheimer's disease-associated SNPs, rs3851179 and rs536841, are within a brain FIRE (Figure S5B). Here, rs3851179 is within a brain-specific super-enhancer, whereas rs536841 is outside the super-enhancer. Interestingly, this brain-specific FIRE overlaps with *PICALM*, which contains the SNP (rs3851179) previously related to the incidence of late-onset Alzheimer's disease (Liu et al., 2016).

The presence of deleterious variants has been shown to mediate the expression of distal genes and confer pathology through DNA looping (Smemo et al., 2014). Therefore, we posited that significantly interacting bin pairs (i.e., "peaks")



(legend on next page)

anchored at SNP-bearing FIREs (termed "FIRE peaks") may be enriched for SNP-gene pairs, relative to peaks anchored at non-FIRE bins (termed "non-FIRE peaks"). To explore this, we first used Fit-Hi-C (Ay et al., 2014) (see Supplemental Experimental Procedures) and a stringent statistical significance (FDR < 1e−6) cutoff to obtain the most confident peak calls within a 2-Mb genomic distance for all samples in our primary cohort (Supplemental Information). We found that this significance cutoff corresponds well to previously published total peak counts (Jin et al., 2013) and can also be used to link disease-associated SNPs to genes previously implicated in a particular disease. For example, Fit-Hi-C peak-calling analysis in GM12878 lymphoblasts reveals a highly significant (FDR = 6.29e−83) pairwise Hi-C contact between a bin containing a SNP associated with ALL (rs6964969) and a distal (~130 kb) TSS of *IKZF1*, a gene previously implicated in ALL (Mullighan et al., 2009) (Figures 6E and 6F). To further explore SNP-gene-pair linkages in our tissue datasets, we collected statistically associated SNP-gene pairs from the GTEx eQTL database in tissues matching our Hi-C datasets (GTEx Consortium, 2015; Lonsdale et al., 2013). We then selected six of our higher resolution tissue Hi-C datasets that were also present in GTEx for further analysis and found that FIRE peaks were indeed significantly enriched for SNP-gene pairs compared to non-FIRE peaks (Table S4). However, this may be expected because FIREs are enriched for disease-associated SNPs, and FIREs are likely to have more local peaks than non-FIREs based on the definition of FIRE. Therefore, we analyzed the enrichment of GTEx SNP-gene pairs in subsets of the most significant FIRE peaks (i.e., the lowest FDR bin pairs). We found that the most statistically significant FIRE peaks exhibited the strongest enrichment of SNP-gene pairs, and relaxing the FDR for peak calling results in statistically significant, but less enriched, SNP-gene pairs (Figures 6G–6J; Table S4).

FIREs Display Promiscuous Local Chromatin Interactions

Although FIREs are identified on the basis of their cumulative local contact frequency, this could result from FIREs either having a single local target with exquisitely high contact frequency or numerous local targets with moderate to high contact frequency. Because FIREs and super-FIREs are highly enriched for active enhancers, exploring the interaction patterns of FIRE regions may provide further insight into the interaction behavior of active *cis*-regulatory loci. First, as expected, we find that FIREs are highly enriched for local interactions compared to non-FIREs, but, unexpectedly, this contact enrichment extends in many cases to an ~500-kb genomic distance (Figure 7A). Because FIREs tend to be positioned near the TAD center, it's likely that FIREs are highly interactive with all loci within the confines of their respective TADs. Next, using the most statistically confident (FDR < 1e−6) Hi-C contacts determined by Fit-Hi-C, we find that FIREs have significantly more local (≤200 kb) peaks compared to non-FIREs (Figures 7B and 7C; Table S4) (two-sample t test p value < 0.01 for ovary (OV) and small bowel (SB); < 2.2e−16 for remaining samples), with an average of three to seven local peaks per FIRE bin, depending on the sample and sequencing depth (Figures 7B and 7C; Table S4). One example is the *BCL11A* locus in GM12878 lymphoblast cells, where numerous enhancer-bearing FIRE bins significantly interact with each other and with the bin containing the promoter for *BCL11A* (Figure 7D). Interesting, *BCL11A* is also known to be involved in numerous lymphoid pathologies (Satterwhite et al., 2001).

To further quantify the contacts between FIREs, we examined the contact frequencies of FIREs and non-FIRE bins across a spectrum of genomic distances within 2 Mb. We find a significantly high contact frequency between FIREs beyond 200 kb

Figure 6. FIREs Are Enriched with Disease-Associated GWAS SNPs

(A) Heat map showing the enrichment of disease-associated GWAS SNPs (see Supplemental Experimental Procedures) in FIRE bins for each cell line or tissue (columns). Rows represent the enrichment of disease-associated SNPs for one disease, and all rows in the presented heat map are sorted from high to low based on enrichment score in GM12878 (lymphoblast cell line). Only diseases with >15 SNPs are shown. Noted to the right are the top 15 diseases for which disease-associated SNPs are most enriched in GM12878 FIREs, showing the high enrichment of several diseases (all except mean corpuscular volume) with previously noted immune-mediated pathology (Jostins et al., 2012).

(B) Normalized Hi-C contact matrix of a 2.16-Mb locus (chr1:65,120,000–67,280,000) in GM12878 cells. The tracks below depict the presence of two SNPs associated with acute lymphoblastic leukemia (rs546784 and rs6683977) located within a FIRE bin (brown, chr1:66,760,000–66,800,000), ~30 kb outside of a GM12878-specific super-enhancer (red) and also within the *PDE4B* gene sequence. To the right of the Hi-C contact matrix is the FIRE score.

(C) Bar plots showing the enrichment of Parkinson's disease-associated SNPs across 14 primary adult tissue FIRE annotations, also highlighting the highest enrichment in FIREs from both brain tissues (CO and HC).

(D) Bar plots showing the enrichment of SNPs associated with the quantitative triglycerides trait across 14 primary adult tissue FIRE annotations, also highlighting the highest enrichment in liver FIREs.

(E) Normalized Hi-C contact matrix (top) in GM12878 for a 4.04-Mb locus (chr7:48,440,000–52,480,000) centered on *IKZF1* (red text). The Hi-C color scale ranges from the 15th to 99th percentile normalized contact frequencies within this locus. The reflected matrix shows the statistically significant (FDR < 1e−6) bin-pairs within 2-Mb genomic distance across the locus. Only bin pairs with FDR < 1e−6 are yellow; the rest are black. Between the matrices are a UCSC gene annotations (blue, top), RNA-seq data (red), H3K27Ac data (black), typical enhancer annotations (Hnisz et al., 2013) (purple), FIRE annotations (brown), TAD boundary calls (blue), and an SNP that is statistically linked to the IKZF1 TSS (green). The blue lines outline the 440-kb locus (chr7:50,240,000–50,680,000) that is shown in (F).

(F) Same as (E), except a zoomed-in snapshot of a 440-kb locus (chr7:50,240,000–50,680,000) centered on a SNP-bearing FIRE bin (chr7:50,440,000–50,480,000) containing the 3′ UTR of *IKZF1* and the SNP rs6964969. The blue box outlines the bin pair that is the significant interaction between previously known SNP-gene pairs.

(G) Bar plots showing the enrichment of liver GTEx eQTLs in FIRE peak bin pairs as a function of the subset of top liver FIRE peaks (based on the lowest false discovery rate) determined by Fit-Hi-C.

(H) Same as (G), except using aorta GTEx eQTLs, FIREs, and FIRE peaks.

(I) Same as (G), except using left ventricle GTEx eQTLs, FIREs, and FIRE peaks.

(J) Same as (G), except using cortex GTEx eQTLs, FIREs, and FIRE peaks.

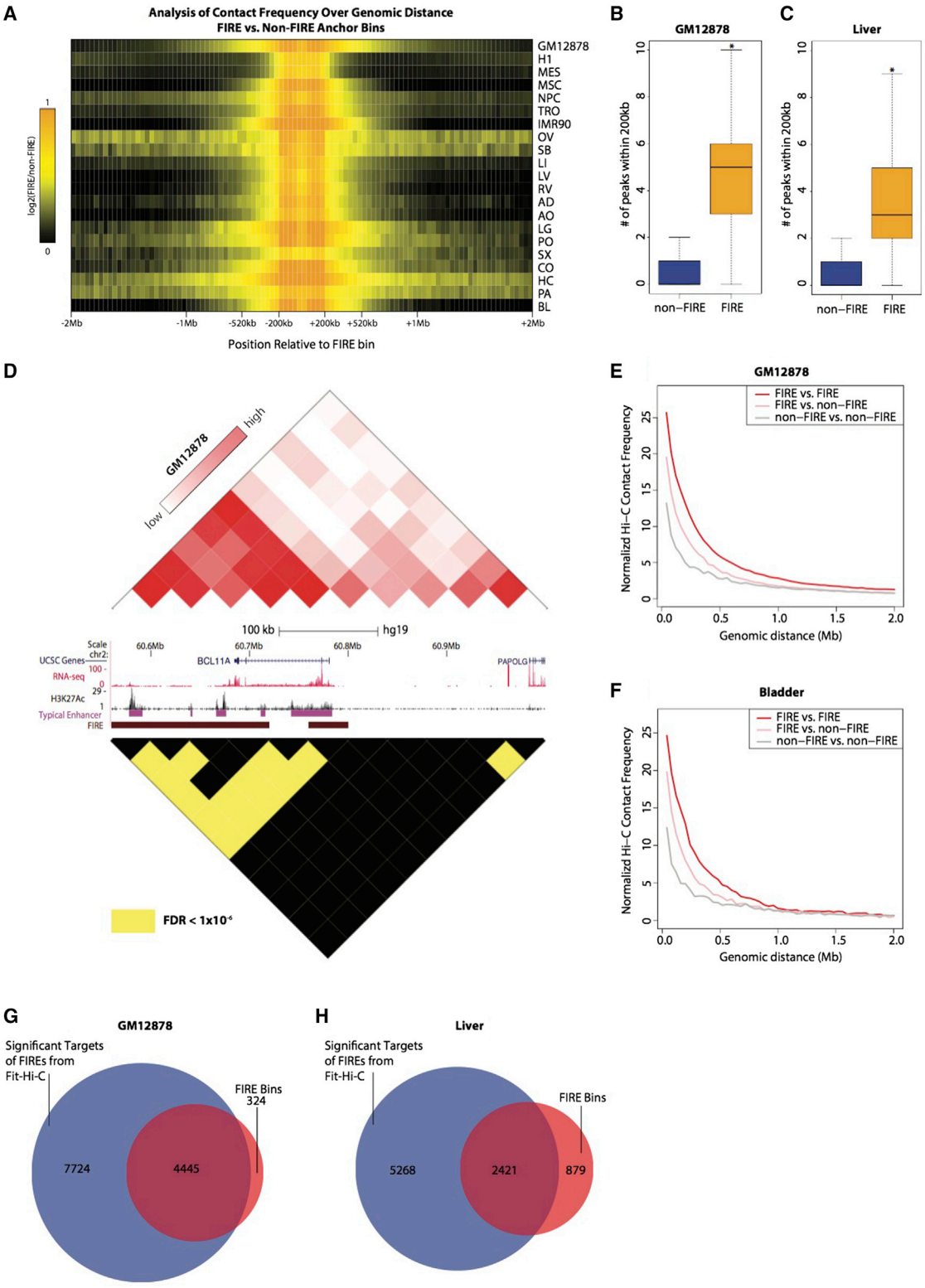


Figure 7. FIREs Have Several Targets and Are Self-Interactive

(A) Heat map showing the relationship between the mean observed contact frequencies at FIREs compared to the mean observed contact frequency at non-FIREs. Enrichment is shown as the ratio between the two contact observed mean contact frequencies (FIRE:non-FIRE) per unit genomic distance, from ± 40 kb to ± 2 Mb, centered on FIRE bins. Each row represents the analysis of a different sample, and the color intensity corresponds to the enrichment value.

(legend continued on next page)

(Figures 7E and 7F), often up to ~500 kb and even up to 2 Mb in some cell lines and tissues (Figure S5E; Table S4). Furthermore, we find a significant proportion of FIREs are targets of other FIREs (chi-square test p value < 1e−5 for OV and < 2.2e−16 for the rest of the samples) (Figures 6, 7E, 7G, 7H, and S5E; Table S4). Taken together, these data support the notion that FIREs represent spatially active regions in the genome.

DISCUSSION

3C and related technologies have been instrumental for understanding the hierarchical organization of mammalian genomes. Comparative analyses across cell types or species have thus far revealed a number of organizational features, including dynamic chromosomal compartments (Dixon et al., 2015; Lieberman-Aiden et al., 2009), TADs (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012), sub-TADs (Phillips-Cremins et al., 2013), insulated neighborhoods (Dowen et al., 2014), and chromatin loops (Rao et al., 2014). Here, through a comprehensive survey of chromatin organization in 21 human tissues and cell types, we report the finding of a previously under-appreciated feature of chromatin organization, FIRE, defined as regions that show substantial levels of local chromatin interactions. FIREs are distinct structural features compared to the previously described 3D genome features, such as TADs, chromatin loops, and compartments. FIREs are enriched in compartment A and display strong tissue-type specificity, with nearly 60% of the FIREs found in two or fewer tissues and cell types out of 21 surveyed. Perhaps most surprisingly, FIREs appear to engage in promiscuous chromatin interactions within their local chromatin neighborhood. The majority of the FIREs identified interact with multiple partners, while the reported chromatin loops typically connect two genomic regions together. Thus, FIREs are hotspots of local chromatin interactions. Finally, FIREs likely represent genomic regions actively engaged in gene regulation. Indeed, they reside near cell-identity genes, harbor significant levels of active chromatin marks, and are enriched for active enhancers, especially super-enhancers.

Further analysis reveals FIREs are closely related to previously reported super-enhancers (Hnisz et al., 2013). In GM12878 cells, in which deeply sequenced Hi-C data were available, nearly 100% of the super-enhancers are FIREs. Such an observation sheds light on the spatial architecture of super-enhancers and other active enhancers. Specifically, our results suggest that in addition to the high density of transcription factor binding and active chromatin modification, these long-range control elements also share a unique spatial feature: a high level of local chromatin interactions. Three additional properties about FIREs carry implications for the understanding of chromatin organization of enhancers. First, FIREs are not only highly interactive within 200 kb, but also highly interactive beyond 200 kb. Because FIREs are often positioned toward the TAD center, this likely means these FIREs are free to explore and interact with a substantial fraction of the TAD structure. Second, we find that FIREs often have numerous significant local interaction partners. Coupled with the observation that FIREs and super-FIREs are highly enriched for enhancers, this uncovers the promiscuously interactive behavior of active enhancer sequences. This could mean that enhancers are likely to explore and physically engage with several loci in their local neighborhood in search for compatible targets. Lastly, we find that FIREs are highly self-interactive, even beyond the local (±200 kb) neighborhood. This underscores the significant degree of active *cis*-regulatory element spatial clustering occurring within the topological framework of larger domains. These observations, in conjunction with the notion that FIREs exhibit a high degree of tissue-specificity, reveal the degree to which tissues contain unique chromatin folding signatures at their active *cis*-regulatory elements. Through their heightened local contact frequency, FIREs are likely to engage with several *cis*-regulatory elements in their TADs and cooperatively regulate gene expression.

By analyzing the effects of Cohesin depletion in three independent studies involving both mouse and human cells, we found that the Cohesin complex is a key mediator of FIREs, and this mechanism is conserved across species. Previous analyses of chromatin architecture in mammalian cells indicated that loss of Cohesin results in a reduction of interaction frequency within TADs ("intra-TAD"), whereas knockdown of CTCF results in both loss of intra-TAD contact frequency and an increase in inter-TAD contact frequency (Zuin et al., 2014). Our re-analysis of these data in the context of very local chromatin interaction frequency indicates that upon loss of Cohesin or CTCF, the most dramatic reduction in FIRE score at FIRE bins was observed at loci containing CTCF/Cohesin co-bound peaks but not CTCF-only sites. We further demonstrate the Cohesin dependence of FIREs in murine neural progenitor cells, astrocytes, and thymocytes, supporting a conserved mechanism of FIRE establishment.

In sum, by generating a rich resource of chromatin contact maps across 21 human tissues and cell types and exploring

(B) Box plot for GM12878 showing the distributions of a number of statistically significant (FDR < 1e−6) Hi-C contacts within 200 kb emanating from non-FIRE (blue box) or FIRE (yellow box) bins (two-sample t test p value < 2.2e−16).
(C) Same as (B), except analysis of liver data.
(D) Comparison of the normalized contact matrix (top triangle) to statistically confident (FDR < 1e−6) pairwise contacts (bottom triangle) in GM12878 across a 440-kb locus centered on *BLC11A*. Between the matrices are the UCSC gene annotations (blue), RNA-seq (red), H3K27Ac (black), typical enhancer annotations (purple) (Hnisz et al., 2013), and FIRE annotations (brown). Color bar values of the Hi-C contact matrix correspond to the 15th and 99th percentiles, respectively, across this locus. In the lower triangle matrix, only the most confident bin pairs (FDR < 1e−6) are colored yellow.
(E) Line plots in GM12878 showing the normalized Hi-C contact frequency (y axis) as a function of genomic distance (x axis) for three categories of pairwise interactions: FIRE-FIRE interactions (red line), FIRE-non-FIRE interactions (pink line), and non-FIRE-non-FIRE interactions (gray line).
(F) Same as (E), except analysis is in bladder tissue.
(G) Venn diagram showing the overlap between all annotated FIRE bins (red circle) in GM12878 and all bins that are involved in statistically significant (FDR < 1e−6) pairwise contacts (blue circle).
(H) Same as (G), except analysis is in liver tissue.

with integrative analytic methods, we have cataloged 3D genome interactions at various hierarchical levels and uncovered the highly dynamic nature of local interaction hotspots. These results provide insights into the chromatin organization in mammalian cells.

EXPERIMENTAL PROCEDURES

Hi-C

Hi-C experiments on all human tissues were performed as previously described using the HindIII restriction enzyme (Lieberman-Aiden et al., 2009), with minor modifications pertaining to handling flash frozen primary tissues (Leung et al., 2015). All previously published Hi-C datasets analyzed in this study were generated using the original "dilution" Hi-C protocol (Lieberman-Aiden et al., 2009) and HindIII, unless otherwise noted (Table S1).

Hi-C Data Processing

Newly generated Hi-C datasets were sequenced on either the Illumina HiSeq2000 or HiSeq2500 instrument. Published datasets were obtained from the SRA and converted to fastq files. Data were then processed using a custom pipeline, beginning with aligning each read end to the mm9 or hg19 reference genomes using BWA –mem. Chimeric read ends were filtered to keep only 5′ alignments with MAPQ > 10, and then read ends were paired and de-duplicated. Raw contact matrices were constructed using in-house scripts, and then further processed using HiCNormCis (described below) or using HiCNorm (Hu et al., 2012), Vanilla Coverage (Rao et al., 2014), or ICE (Imakaev et al., 2012), where indicated.

Compartment A/B Identification

Compartment A/B analysis was performed at 1-Mb resolution, as previously described (Lieberman-Aiden et al., 2009), using the "prcomp" function in R on the Pearson correlation matrix.

Identification of Topological Domains

Topological domain boundaries were identified at 40-kb bin resolution using the previously described insulation score analysis approach, with two minor modifications (Crane et al., 2015). Because mammalian TAD have been previously identified to be ~1 Mb, a 1-Mb genomic region was used rather than 500 kb. Additionally, a 200-kb window, rather than 100 kb, was used for calculation of the delta vector.

Identifying Frequently Interacting Regions

We developed a Poisson-regression-based normalization approach, named "HiCNormCis," to identify FIRE bins. Specifically, we first partitioned the entire genome into bins, and calculated the total number of intra-chromosomal ("*cis*") interactions in the contact distance range of 15–200 kb for each bin. Bins with low mappability (<0.9) around HindIII cut sites were removed. HiCNormCis then takes into account biases from three known factors known to bias observed Hi-C contact counts, including effective fragment length, GC content, and mappability (Yaffe and Tanay, 2011) (related to Figures 2 and S2). Let Y_i represent the total *cis* interactions (15–200 kb) for the *i*th bin. Additionally, let F_i, GC_i, and M_i represent the effective fragment length and GC represent content and mappability in the *i*th bin, respectively. The detailed calculation of F_i, GC_i, and M_i is described in our previous work (Hu et al., 2012). Assume Y_i follows a Poisson distribution, with a mean of θ_i. We fitted a Poisson regression model as follows: $\log \theta_i = \beta_0 + \beta_F F_i + \beta_{GC} GC_i + \beta_M M_i$, and defined the residual $R_i = Y_i / \exp(\widehat{\beta}_0 + \widehat{\beta}_F F_i + \widehat{\beta}_{GC} GC_i + \widehat{\beta}_M M_i)$ as the normalized total *cis* interaction. Noticeably, $\exp(\widehat{\beta}_0)$ is proportional to the overall sequencing depth, and the residual R_i has a mean of 1. Therefore, the normalized total *cis* interactions are robust to different sequencing depths, and are directly comparable among different samples. Visual inspection revealed that R_i follows a Gaussian distribution (related to Figure S2). Therefore, we converted R_i to the corresponding *Z* score and −ln(p value). The same approach can theoretically be applied to any Hi-C dataset generated using a restriction enzyme and at any bin size.

Identification of Significant Hi-C Contacts

Statistically significant contacts in Hi-C data were identified at 40-kb resolution using Fit-Hi-C, as previously described (Ay et al., 2014) (see Supplemental Experimental Procedures). We used the default Fit-Hi-C code to calculate a p value and q value for each bin pair within a 2-Mb genomic distance. For all analyses in this study, we used a conservative peak-calling threshold of FDR < 1e−6.

ACCESSION NUMBERS

The accession number for the Hi-C and re-analyzed RNA-seq data reported in this paper is GEO: GSE87112.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and nine tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2016.10.061.

AUTHOR CONTRIBUTIONS

Conceptualization, A.D.S, M.H., and B.R; Formal Analysis, M.H., A.D.S., Z.X., I.J., Y.Q., and Y. Li; Investigation, A.D.S. and C.L.T.; Resources, C.L.B., S.L., and Y. Lin; Writing – Original Draft, A.D.S., M.H., and B.R.; Writing – Review and Editing, A.D.S., M.H., and B.R.

ACKNOWLEDGMENTS

We would like to dedicate this manuscript in loving memory of Joseph Schmitt. We would like to give special thanks to Samantha Kuan and Bin Li for operation of the sequencing instruments and data processing. We'd like to acknowledge the help of Michael Yu from Trey Ideker's laboratory (UCSD), Doug Chapski from Tom Vondriska's laboratory (UCLA), and Jesse Dixon (Salk Institute) for sharing helpful files or codes to facilitate this study. We would also like to give special thanks to David Gorkin for numerous helpful discussions throughout the project, as well as the additional members of the Ren laboratory. This work is supported by the Ludwig Institute for Cancer Research and grants from NIH (U54DK107977 to B.R. and M.H. and R01 ES024984 to B.R.). A.D.S. is supported by an NIH genetics training grant T32 GM008666. C.L.B. is supported by funding from The Ontario Mental Health Foundation, The Krembil Foundation, and The Hospital for Sick Children Psychiatric Endowment Fund. Y. Li and Z.X. are partially supported by NIH R01HG006292 and R01HL129132 (awarded to Y. Li).

Received: July 14, 2016
Revised: September 2, 2016
Accepted: October 18, 2016
Published: November 15, 2016

REFERENCES

Arens, R., Nolte, M.A., Tesselaar, K., Heemskerk, B., Reedquist, K.A., van Lier, R.A.W., and van Oers, M.H.J. (2004). Signaling through CD70 regulates B cell activation and IgG production. J. Immunol. *173*, 3901–3908.

Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res. *24*, 999–1011.

Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., and Meyer, B.J. (2015). Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature *523*, 240–244.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin domains: the unit of chromosome organization. Mol. Cell *62*, 668–680.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell *159*, 374–387.

Dryden, N.H., Broome, L.R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., Nagano, T., Andrews, S., Wingett, S., Kozarewa, I., et al. (2014). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. *24*, 1854–1868.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., et al.; FANTOM Consortium (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol. Syst. Biol. *11*, 852.

GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J.S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics *28*, 3131–3133.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods *9*, 999–1003.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I., et al. (2016). 3D chromosome regulatory landscape of human pluripotent cells. Cell Stem Cell *18*, 262–275.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C.A., et al.; International IBD Genetics Consortium (IIBDGC) (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kingwell, K. (2013). Epilepsy: GRIN2A mutations identified as key genetic drivers of epilepsy-aphasia spectrum disorders. Nat. Rev. Neurol. *9*, 541.

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A.Y., Yen, C.A., Lin, S., Lin, Y., Qiu, Y., et al. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature *518*, 350–354.

Leyva-Díaz, E., del Toro, D., Menal, M.J., Cambray, S., Susín, R., Tessier-Lavigne, M., Klein, R., Egea, J., and López-Bendito, G. (2014). FLRT3 is a Robo1-interacting protein that determines Netrin-1 attraction in developing axons. Curr. Biol. *24*, 494–508.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Liu, G., Xu, Y., Jiang, Y., Zhang, L., Feng, R., and Jiang, Q. (2016). PICALM rs3851179 variant confers susceptibility to Alzheimer's disease in Chinese population. Mol. Neurobiol, Published online April 5, 2016. http://dx.doi.org/10.1007/s12035-016-9886-2.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Meaburn, K.J., and Misteli, T. (2007). Cell biology: chromosome territories. Nature *445*, 379–781.

Montavon, T., and Duboule, D. (2013). Chromatin organization and global regulation of Hox gene clusters. Philos. Trans. R. Soc. Lond. B Biol. Sci. *368*, 20120367.

Mullighan, C.G., Su, X., Zhang, J., Radtke, I., Phillips, L.A.A., Miller, C.B., Ma, J., Liu, W., Cheng, C., Schulman, B.A., et al. (2009). Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. N. Engl. J. Med. *360*, 470–480.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Ohi, K., Shimada, T., Nitta, Y., Kihara, H., Okubo, H., Uehara, T., and Kawasaki, Y. (2016). Specific gene expression patterns of 108 schizophrenia-associated loci in cortex. Schizophr. Res. *174*, 35–38.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Satterwhite, E., Sonoki, T., Willis, T.G., Harder, L., Nowak, R., Arriola, E.L., Liu, H., Price, H.P., Gesk, S., Steinemann, D., et al. (2001). The BCL11 gene family: involvement of BCL11A in lymphoid malignancies. Blood *98*, 3413–3420.

Seitan, V.C., Faure, A.J., Zhan, Y., McCord, R.P., Lajoie, B.R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A.G., et al. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. *23*, 2066–2077.

Selvaraj, S., R Dixon, J., Bansal, V., and Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat. Biotechnol. *31*, 1111–1118.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. *38*, 1348–1354.

Smemo, S., Tena, J.J., Kim, K.-H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371–375.

Sofueva, S., Yaffe, E., Chan, W.-C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A., and Hadjur, S. (2013).

Cohesin-mediated interactions organize chromosomal domain architecture. EMBO J. *32*, 3119–3129.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell *163*, 1611–1627.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. *10*, 1297–1309.

Xu, Z., Zhang, G., Jin, F., Chen, M., Furey, T.S., Sullivan, P.F., Qin, Z., Hu, M., and Li, Y. (2015). A hidden Markov random field based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data. Bioinformatics *32*, 650–656.

Xu, Z., Zhang, G., Wu, C., Li, Y., and Hu, M. (2016). FastHiC: a fast and accurate algorithm to detect long-range chromosomal interactions from Hi-C data. Bioinformatics *32*, 2692–2695.

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat. Genet. *43*, 1059–1065.

Yang, J.J., Cheng, C., Devidas, M., Cao, X., Fan, Y., Campana, D., Yang, W., Neale, G., Cox, N.J., Scheet, P., et al. (2011). Ancestry and pharmacogenomics of relapse in acute lymphoblastic leukemia. Nat. Genet. *43*, 237–241.

Zuin, J., Dixon, J.R., van der Reijden, M.I., Ye, Z., Kolovos, P., Brouwer, R.W., van de Corput, M.P., van de Werken, H.J., Knoch, T.A., van IJcken, W.F., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. U S A *111*, 996–1001.

Cell Reports, Volume *17*

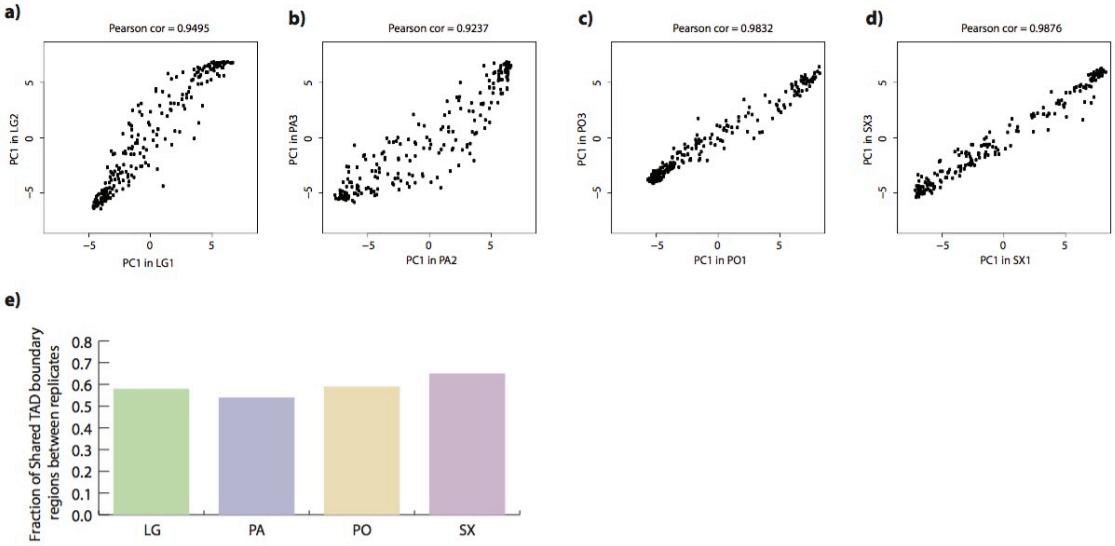
Supplemental Information

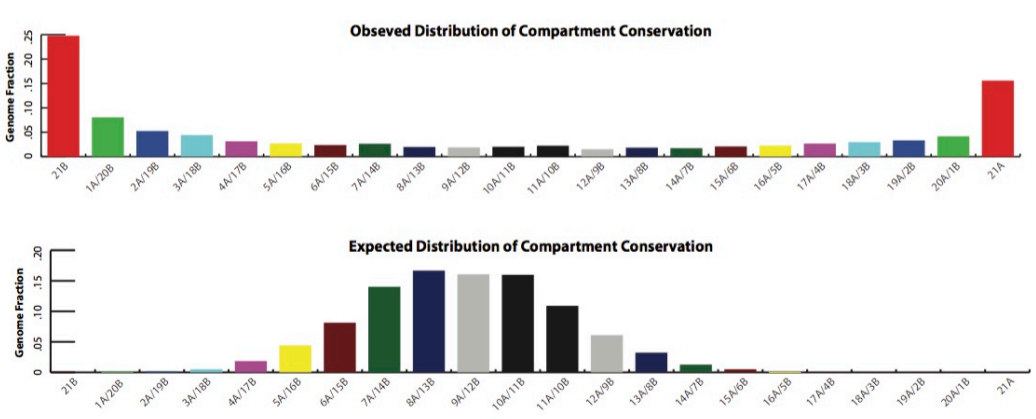
A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome

Anthony D. Schmitt, Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L. Tan, Yun Li, Shin Lin, Yiing Lin, Cathy L. Barr, and Bing Ren

1. Supplemental figures.

Figure S1





Compartment	Observed	Expected
21B	682	0
1A+20B	222	1
2A+19B	145	4
3A+18B	122	14
4A+17B	87	51
5A+16B	75	122
6A+15B	66	224
7A+14B	73	386
8A+13B	55	459
9A+12B	52	442
10A+11B	56	440
11A+10B	62	300
12A+9B	42	168
13A+8B	51	89
14A+7B	48	35
15A+6B	58	14
16A+5B	63	2
17A+4B	74	0
18A+3B	82	0
19A+2B	92	0
20A+B	115	0
21A	429	0

Chi-square test p-value < 2.2e-16
Binomial test p-value: P(# of 7B>=682) < 2.2e-16
Binomial test p-value: P(# of 7A>=429) < 2.2e-16

g)

Sample	# of TAD boundaries
0.05	2,534
0.1	2,511
0.2	2,517
0.4	2,509
0.6	2,519
0.8	2,508
All	2,516

h)

Sample	# of overlap with TAD boundary regions	% of TAD boundary regions
0.05	2,497	92.34%
0.1	2,479	91.68%
0.2	2,478	91.64%
0.4	2,479	91.68%
0.6	2,484	91.86%
0.8	2,474	91.49%
All	2,479	91.68%

i)

Sample-specificity	# of TAD boundary regions	% of TAD boundary regions
Only in one sample	128	4.73%
Shared by 2 samples	58	2.14%
Shared by 3 samples	41	1.52%
Shared by 4 samples	48	1.78%
Shared by 5 samples	51	1.89%
Shared by 6 samples	90	3.33%
Shared by all 7 samples	2,288	84.62%

Figure S1. Hi-C data reproducibility and compartment A/B conservation, related to Figure 1.

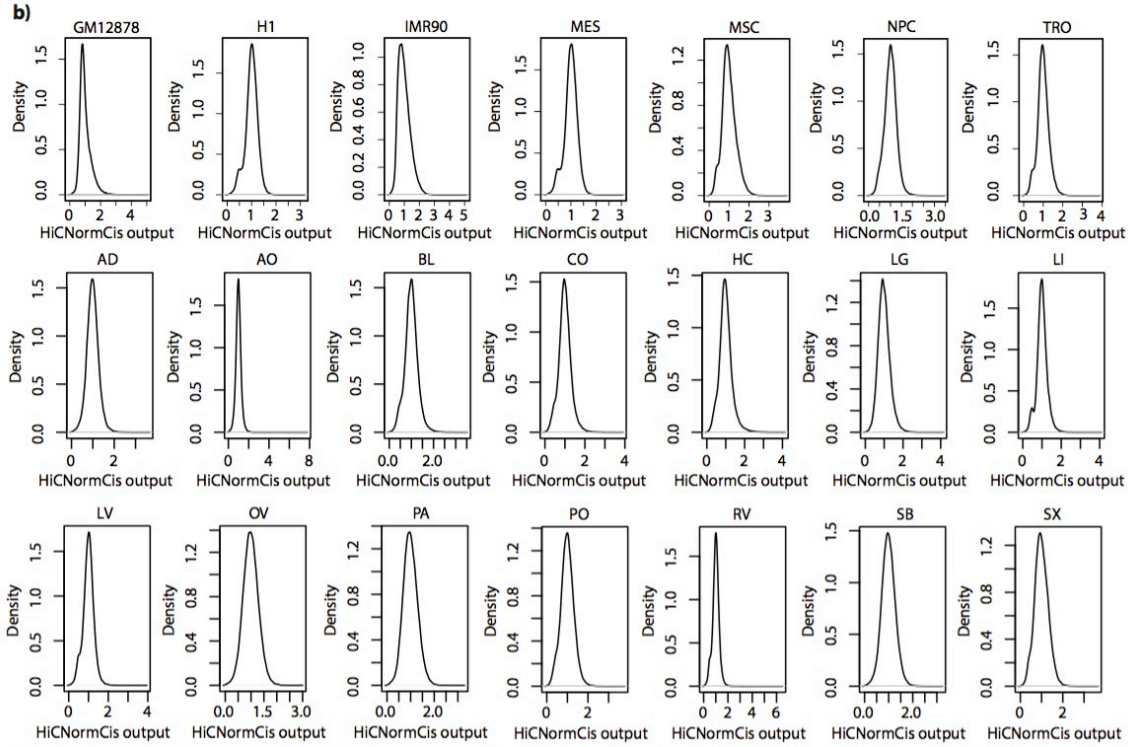
A) Scatter plots from replicates of LG, showing the genome-wide of the PC1 values used for the Compartment A/B analysis. The plot title contains the Pearson correlation coefficient of all 1Mb bin-pairs. The x- and y-axes are labeled according to their tissue type and donor. For example, LG2 corresponds to Lung tissue from donor 2.
B) Same as Panel A, except analysis of biological replicates for PO.
C) Same as Panel A, except analysis of biological replicates for PA.
D) Same as Panel A, except analysis of biological replicates for SX.
E) Bar plots showing the statistically significant fraction of overlapping TAD boundaries in LG, PO, PA, and SX (Chi squarest test p value < 2.2e-16).
F) Bar plots showing the observed (top) and expected (bottom) distributions of compartment A/B conservation. Labels on the x-axis indicate the number of samples and compartment label for which there is conservation and the Y-axis indicates the total genome fraction that corresponds to that compartment label. For example, 16A/5B indicates the total number of 1Mb bins for which 16 human cell lines or tissues had an A compartment label and 5 samples had a B compartment label. In the bottom table, the 'Compartment' column indicates the how many samples are shared for each compartment label, while the 'Expected' and 'Observed' columns indicate how many 1Mb bins fall into 'Compartment' category. Statistical analysis comparing the observed and expected distributions are done with Chi-square test, and statistical analysis of having complete conservation across all samples (i.e. 21A or 21B) was done with a binomial test.
G) Table showing the total number of topological domain boundaries detected using the insulation square method (Crane et al., Nature, 2015) applied to downsampled Hi-C from H1 cells. The left column indicates what fraction of the full H1 dataset was obtained from downsampling, and the right column indicates the total number of TAD boundaries detected.
H) Table showing the absolute number of TAD boundary regions overlapping all putative boundaries identified across all downsampling samples (middle column). The right column indicates the corresponding fraction out of all putative boundaries identified across all downsampling samples. The left column indicates what fraction of the full H1 dataset was obtained from downsampling.
I) Table showing the percentages of TAD boundaries that were unique to subsets of the downsampled H1 datasets. The left column indicates how many of the 7 degrees of sampling share a particular TAD boundary region. The middle column indicates how many TAD boundaries regions were common to a particular subset denoted in the left column. The right column is the corresponding fraction the common TAD boundary regions are of the total putative boundaries in downsampled H1.

Figure S2

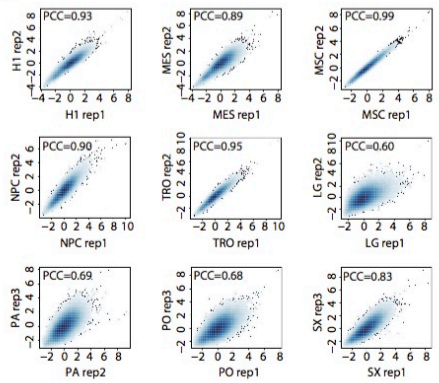
a)

	Data Source				
Bias Factor	*Raw*	*ICE*	*VC*	*HiCNorm*	*HiCNormCis*
Effective Fragment Length	0.47	-0.39	-0.30	0.25	0.01
GC Content	0.15	0.60	0.60	0.08	0.05
Mappability	0.18	0.02	-0.03	0.10	-0.01

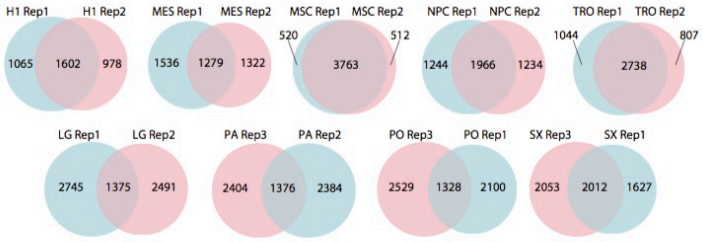
b)



c)



d)

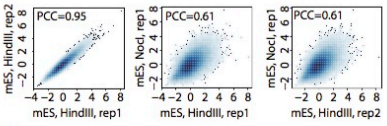


e)

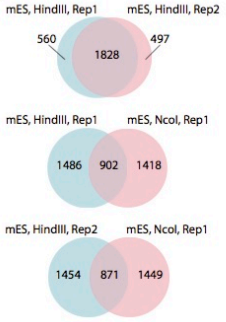
Sample	cis>15kb
80%	153,016,660
60%	114,766,901
40%	76,509,297
20%	38,255,359
10%	19,130,174

Sample	80%	60%	40%	20%	10%
80%	1.00	0.97	0.95	0.94	0.94
60%	0.97	1.00	0.98	0.98	0.97
40%	0.95	0.98	1.00	0.99	0.99
20%	0.94	0.98	0.99	1.00	0.99
10%	0.94	0.97	0.99	0.99	1.00

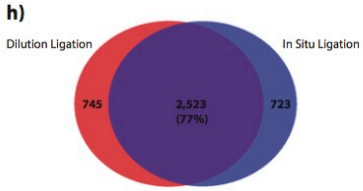
f)



g)



h)



i)

Sample	before HiCNormCis			After HiCNormCis		
	F	GC	M	F	GC	M
GM12878	0.540	-0.162	0.069	0.049	-0.023	0.000
H1	0.768	-0.441	0.175	0.175	-0.117	0.000
IMR90	0.528	0.015	0.170	0.066	-0.021	-0.004
MES	0.774	-0.372	0.171	0.169	-0.113	0.000
MSC	0.696	-0.299	0.163	0.105	-0.065	-0.003
NPC	0.700	-0.451	0.189	0.105	-0.056	0.003
TRO	0.708	-0.230	0.179	0.126	-0.070	-0.003
CO	0.675	-0.257	0.214	0.113	-0.059	-0.009
HC	0.661	-0.291	0.221	0.100	-0.053	-0.010
OV	0.729	-0.552	0.177	0.149	-0.100	0.001
SB	0.757	-0.407	0.188	0.150	-0.091	-0.002
LG	0.524	0.181	0.206	0.061	0.015	-0.010
PO	0.682	-0.151	0.188	0.114	-0.054	-0.010
PA	0.627	0.121	0.181	0.105	0.000	-0.013
SX	0.211	0.509	0.136	-0.031	0.128	-0.007
LI	0.720	-0.396	0.115	0.108	-0.069	-0.001
LV	0.770	-0.474	0.099	0.152	-0.095	0.001
RV	0.758	-0.432	0.100	0.142	-0.091	0.000
AD	0.771	-0.406	0.185	0.154	-0.099	-0.004
AO	0.799	-0.538	0.121	0.177	-0.118	0.004
BL	0.721	-0.297	0.135	0.124	-0.072	-0.003

j)

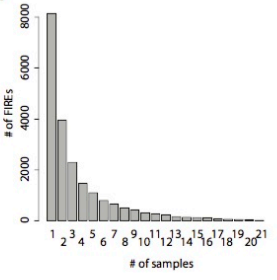
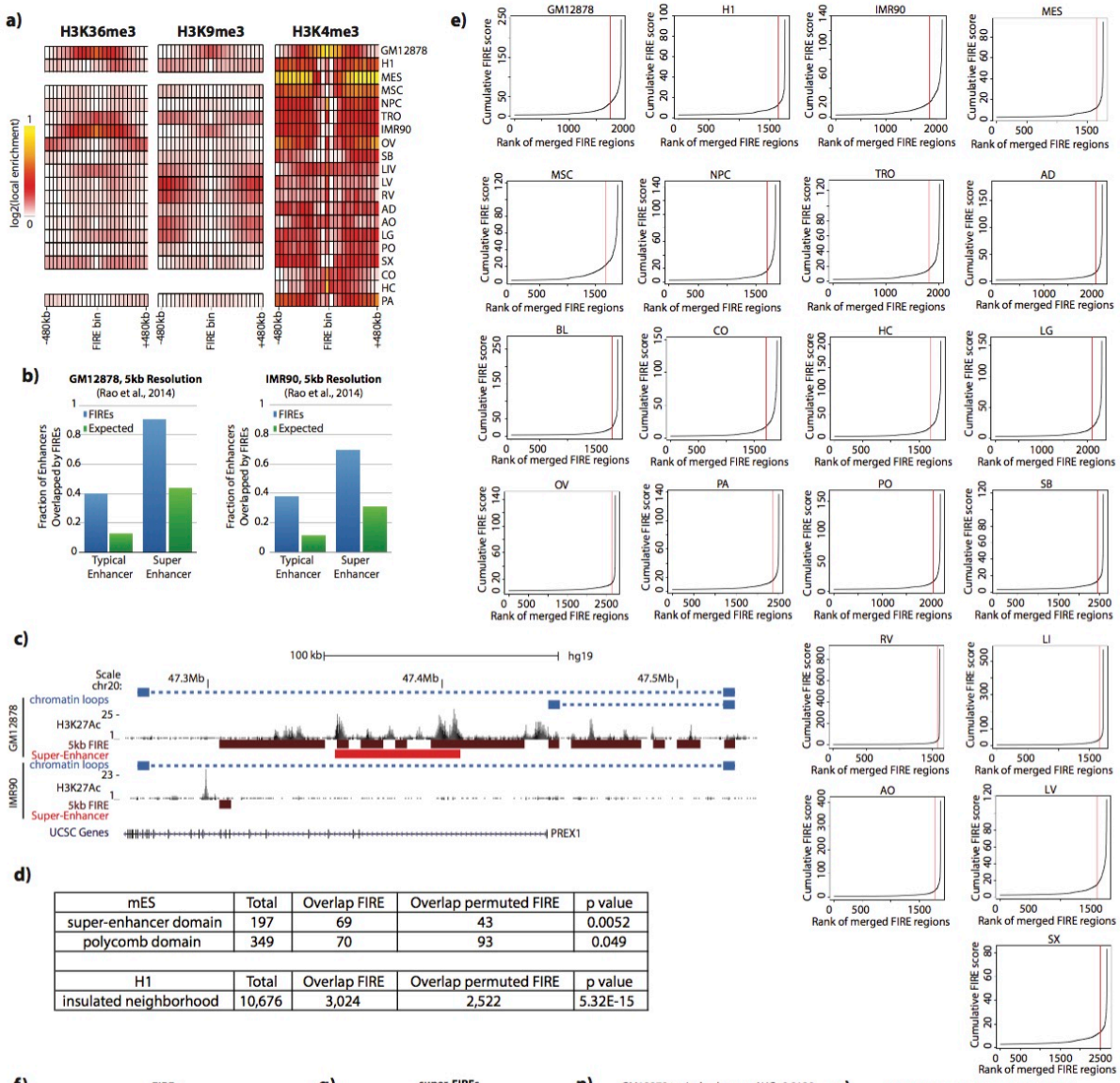


Figure S2. FIRE calling methodology, related to Figure 2.

A) Box plot showing the distribution of the total raw 15-200kb cis interactions per bin in each sample. Box plots for various degrees of downsampling of H1 rep2 (Dixon .., 2015) are shown in green, cell lines are shown in blue, and primary tissue Hi-C data is shown in yellow.
B) Table showing the Pearson correlation coefficient (PCC) between local contact summation of each bin with their respective effective restriction fragment length, GC content, and mappability (as rows). The normalization method (or lack thereof for raw matrix) to prepare the Hi-C contact data is listed as column headers. PCC values are rounded to the nearest hundredth.
C) Density plots showing the distribution of HiCNormCis outputs for each sample. The y-axes show the density and the x-axes are the HiCNormCis output values. The sample name is indicated in the title of each plot.
D) Scatterplots showing the genome-wide pairwise correlation of FIRE score between two biological replicates for H1, MES, MSC, NPC, TRO, LG, PA, PO, and SX. Inset is the Pearson correlation coefficient.
E) Pie charts showing the overlapping FIRE calls in 9 pairs of biological replicates from cell lines or tissues. Same 9 samples as Panel D. (Chi-square test p value < 2.2e-16).
F) Left, table showing the number of long-range cis interactions in a downsampled replicate of H1 (H1 rep2 from Dixon et al., 2015) Hi-C data. The 'Sample' column indicates what fraction of the full dataset was extracted during downsampling, and 'cis>15kb' is the total number of long-range cis interactions from the downsampled data. To the right, a table showing the Pearson correlation coefficient (PCC) of the genome-wide FIRE scores for downsampled H1 data. Each row/column corresponds to what downsampled fraction of the Hi-C data was used for the correlation analysis. Each table entry is the PCC.
G) Scatter plots showing the genome-wide Pearson correlation coefficient (PCC) between 3 different samples, including two biological replicates of mES cells prepared using HindIII and 1 sample of mES cells prepared using NcoI (data from Dixon et al., 2012). Inset is the genome-wide PCC value.
H) Pie chart showing the significant FIRE bin overlap between two biological replicates of mES cells prepared with HindIII (left), or mES HindIII rep1 and mES NcoI (middle), or mES HindIII rep2 and mES NcoI (right). (Chi-square test p value < 2.2e-16).
I) Pie charts showing the significant FIRE bin overlap between samples either prepared using the in situ ligation procedure (right) or the "dilution ligation" procedure (left). (Chi-square test p value < 2.2e-16).
J) Table showing the Pearson correlation coefficient (PCC) for total cis interactions counts (within 15-200kb distance) and fragment length of a given bin (column 'F'), GC content (column 'GC'), and mappability (column 'M'), either before (group 'Before HiCNormCis'), or after normalization (group 'After HiCNormCis'), and for each sample (rows).

Figure S3



mES	Total	Overlap FIRE	Overlap permuted FIRE	p value
super-enhancer domain	197	69	43	0.0052
polycomb domain	349	70	93	0.049
H1	Total	Overlap FIRE	Overlap permuted FIRE	p value
insulated neighborhood	10,676	3,024	2,522	5.32E-15

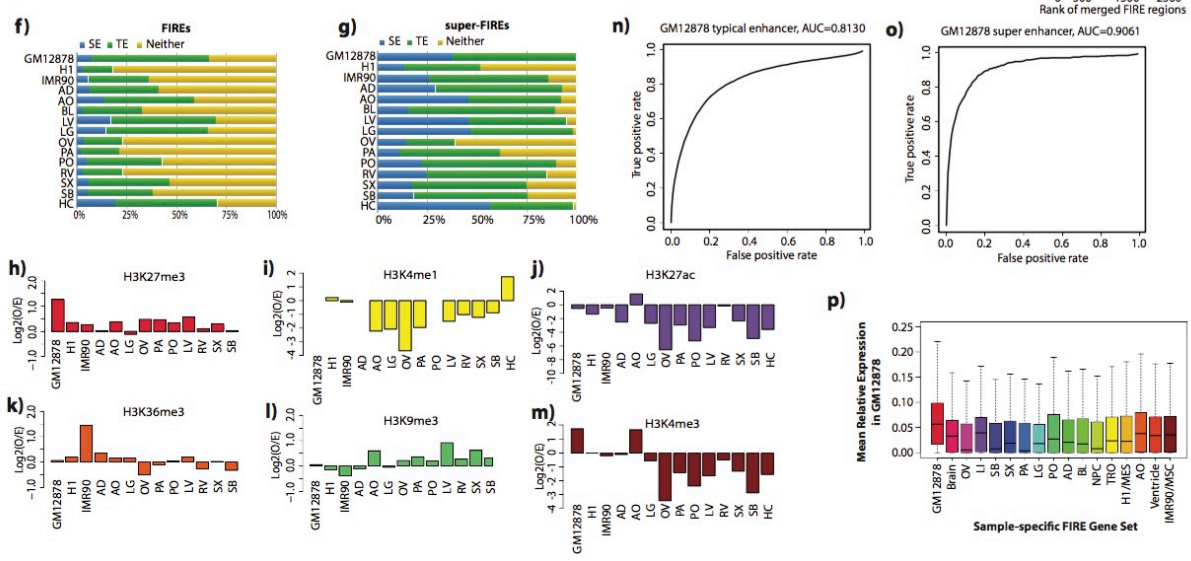


Figure S3. Analysis of chromatin biochemical features at FIREs and super-FIREs, related to Figure 4.

A) Heatmaps showing the local enrichment (see Supplemental Methods) of H3K36me3 (left), H3K9me3 (middle), and H3K4me3 (right), centered on FIRE bins for each cell line or tissue. Local enrichment is calculated relative to the peaks per bin for H3K4me3, and RPKM values for H3K36me3 and H3K9me3. H3K36me3 and H3K9me3 data were not available for CO or HC.
B) Bar plots showing the fraction of typical or super-enhancers overlapped by observed FIRE calls (blue bars) in GM12878 (left plot) and IMR90 (right plot) at 5kb resolution (Rao et al., 2014), or size-matched randomly permuted FIRE calls (green bars). Within each plot, analysis of typical enhancers is on the left, analysis of super-enhancers is on the right.
C) Genome browser snapshot of the PREX1 locus (chr20:47,263,536-47,534,527) in GM12878 (top set of tracks) and IMR90 (bottom set of tracks). Shown for each cell line are previously annotated (Rao et al., 2014) chromatin loops (blue; square is loop anchor, dash to loop), H3K27Ac signal (black), FIREs defined at 5kb resolution (brown), and previously annotated (Hnisz et al., 2013) super-enhancers (red). The bottom of the snapshot shows the positioning of UCSC genes at this locus.
D) Table showing the overlap between FIREs, super-enhancer domains, polycomb domains in mESCs (Dowen et al., 2014) (top section) and insulated neighborhoods in H1 cells (Ji et al., 2016) (bottom section). Tabulated are the total number of domains or insulated neighborhoods, how many are overlapped by a FIRE, and how many are expected to overlap based on random permutation of FIRE positioning in that respective cell type. The Chi-square test p-value is reported in the right column.
E) Line plots showing the cumulative FIRE scores (y-axis) of ranked stitched FIRE bins (x-axis) from the FIREs with the lowest cumulative FIRE scores (left side) to the highest FIRE scores (right side). The red vertical line indicates the inflection point, whereby stitched FIRE bins to the right of this line are called as super-FIREs.
F) Stacked bar plots showing the fraction of FIREs containing at least 1 super-enhancer (SE, blue bars), typical enhancer (TE, green bars), or no SE or TE (yellow bars). Each row is the analysis of a different cell or tissue type.
G) Same as Panel F, except analysis of super-FIREs.
H) Bar plots showing the enrichment (y-axis) of H3K27me3 at super-FIREs that do not contain any annotated typical enhancer or super-enhancers. Each bar represents the analysis of a different tissue, which has been previously annotated for super-enhancers (Hnisz et al., 2013). Hippocampus (HC) tissue is not shown because there is no H3K27me3 ChIP-seq data in HC.
I) Same as Panel H, except analysis of H3K4me1.
J) Same as Panel H, except analysis of H3K27ac.
K) Same as Panel H, except analysis of H3K36me3. No ChIP-seq data available for HC.
L) Same as Panel H, except analysis of H3K9me3. No ChIP-seq data available for HC.
M) Same as Panel H, except analysis of H3K4me3.
N) Line plot showing the relationship between the True Positive rate, defined as the fraction of FIRE bins overlapping typical enhancers (Hnisz et al., 2013), and the False Positive rate, defined as the fraction of FIRE bins not overlapping a typical enhancer, as a function of the significance threshold using to define FIREs in GM12878 cells. (AUC=0.813).
O) Same as Panel N, except for super-enhancers (Hnisz et al., 2013). (AUC=0.906).
P) Genome-wide analysis showing the relative gene expression levels for genes within 200kb of GM12878-specific FIREs. Genes within 200kb of GM12878-specific FIREs were collected, and then for each sample, the relative gene expression levels are calculated. Shown are the box plots of the distribution of relative gene expression levels for each sample indicating that GM12878 relative gene expression levels are higher than any other sample (Two-sample t-test p-value < 2.2e-16 compared to brain, OV, LI, SB, SX, PA, LG, AD, NPC, ventricle, and IMR90/MSC; p-value < 5.66e-7 compared to PO; p-value < 2.93e-8 compared to BL; p-value < 1.04e-9 compared to TRO; p-value < 4.84e-10 compared to H1/MES; p-value < 9.26e-6 compared to AO). Boxplots show the median (black line) and interquartile range.

Figure S4

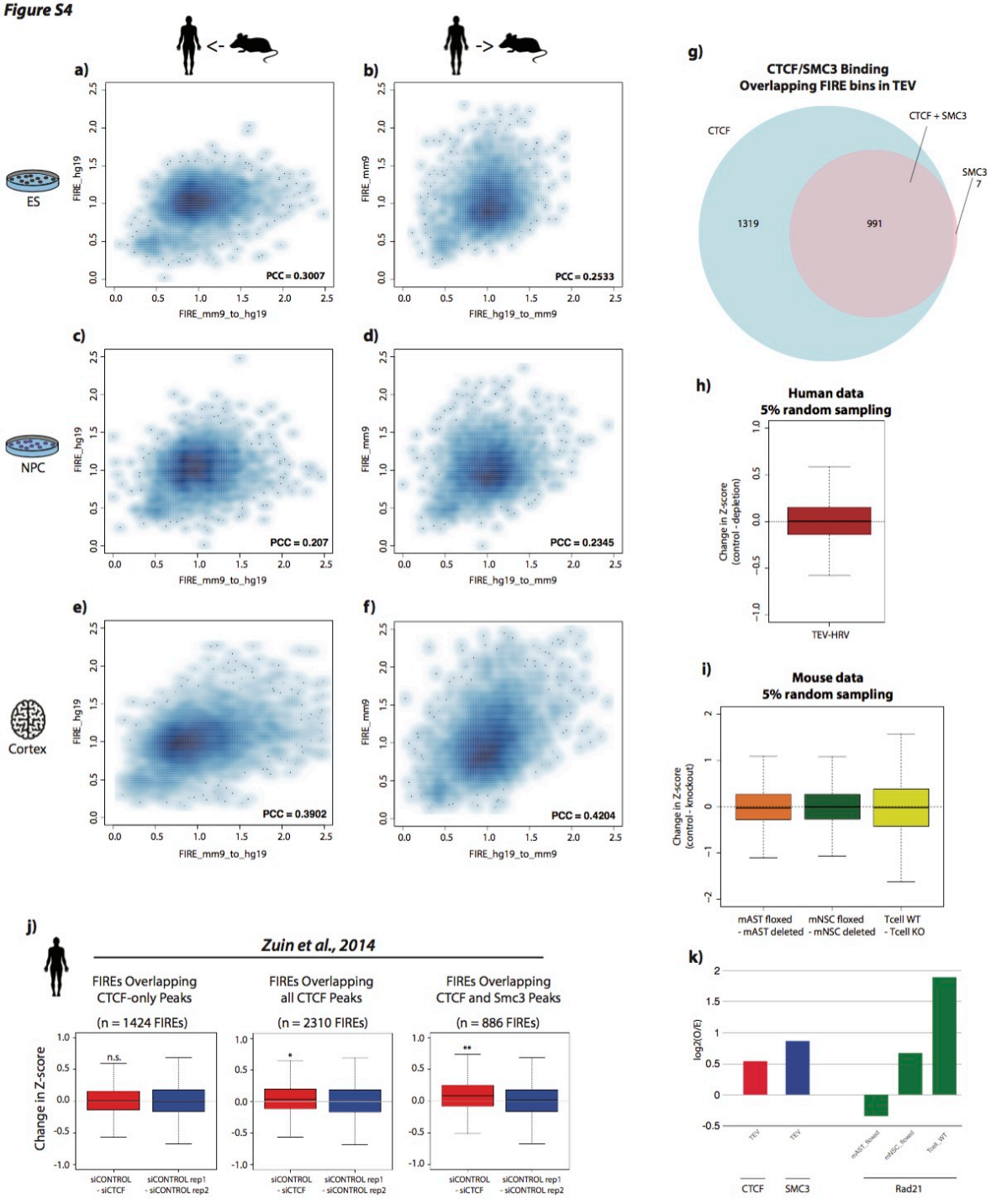

a)
b)
c)
d)
e)
f)
ES
NPC
Cortex
FIRE_hg19
FIRE_mm9
FIRE_mm9_to_hg19
FIRE_hg19_to_mm9
PCC = 0.3007
PCC = 0.2533
PCC = 0.207
PCC = 0.2345
PCC = 0.3902
PCC = 0.4204
g)
CTCF/SMC3 Binding
Overlapping FIRE bins in TEV
CTCF
CTCF + SMC3
SMC3
7
1319
991
h)
Human data
5% random sampling
Change in Z-score
(control - depletion)
TEV-HRV
i)
Mouse data
5% random sampling
Change in Z-score
(control - knockout)
mAST floxed
- mAST deleted
mNSC floxed
- mNSC deleted
Tcell WT
- Tcell KO
j)
Zuin et al., 2014
FIREs Overlapping
CTCF-only Peaks
(n = 1424 FIREs)
FIREs Overlapping
all CTCF Peaks
(n = 2310 FIREs)
FIREs Overlapping
CTCF and Smc3 Peaks
(n = 886 FIREs)
Change in Z-score
n.s.
*
**
siCONTROL
- siCTCF
siCONTROL rep1
- siCONTROL rep2
k)
log2(O/E)
TEV
TEV
mAST_floxed
mNSC_floxed
Tcell_WT
CTCF
SMC3
Rad21

Figure S4. FIRE score species conservation and reduction upon loss of Cohesin, related to Figure 5.

A) Scatterplot showing the correlation between randomly selected non-FIRE bins in mouse ES cells that liftover to the hg19 reference genome. Shown on the x-axis are the FIRE scores from the randomly selected mouse bins that can be liftover to hg19. Shown on the y-axis are the FIRE scores in the corresponding human bins. The PCC value is shown in the bottom right corner.
B) Scatterplot showing the correlation between randomly selected non-FIRE bins in human ES cells that liftover to the mm9 reference genome. Shown on the x-axis are the FIRE scores from the randomly selected human bins that can be liftover to mm9. Shown on the y-axis are the FIRE scores in the corresponding mouse bins. The PCC value is shown in the bottom right corner.
C) Same as Panel A, expect using NPC cell data.
D) Same as Panel B, except using NPC cell data.
E) Same as Panel A and C, except using cortex tissue data.
F) Same as Panel B and D, except using cortex tissue data.
G) Pie charts showing the overlap between FIRE bins called in the TEV sample and bins bound by CTCF only (blue shading, left), SMC3 only (pink shading, right), or co-bound peaks (blue+pink overlap, center).
H) Box plots depicting the change in Z-score in a random sampling of 5% of bins in TEV and HRV cells. There is no significant change in FIRE score in either comparison. Change in Z-score is used for comparison, rather than change in FIRE score (-ln(p-value)), since Z-score has approximate Gaussian distribution.
I) Same as Panel H, except for comparing mAST (floxed – deleted, left boxplot), mNSC (floxed-deleted, middle boxplot), and T-cells (WT-Knockout). In all cases, there is not significant change in FIRE score at a random sampling of FIRE bins. Change in Z-score is used for comparison, rather than change in FIRE score (-ln(p-value)), since Z-score has approximate Gaussian distribution.

J) Box plots showing the change in Z-score at FIREs overlapping bins bound by CTCF but not SMC3 “CTCF-only” (left column), all CTCF peaks (middle column), and CTCF and SMC3 co-binding (right column) for the comparison of siCONTROL and siCTCF samples. The red boxes show distributions of FIRE score change at FIRE bins called in wild type cells minus the mutant cells, while the blue boxes are distributions for FIRE score change at FIRE bins called in wild type cells but between biological replicates of wild type cells. These comparisons show the significant reduction of FIRE score at all CTCF peaks, and especially at CTCF SMC3 co-bound peaks overlapping FIRE bins (*p=4.88e-5, **p=3.89e-9; two sample t-test). Change in Z-score is used for comparison, rather than change in FIRE score (-ln(p-value)), since Z-score has approximate Gaussian distribution.

K) Bar plots showing the significant enrichment of CTCF, SMC3, or Rad21 in FIREs from control samples in 3 different studies (From left to right - One-sample t-test p value < 1.11e-15, < 6.54e-14, < 1.71e-10, < 1.33e-13, and < 2.2e-16). The sample name is indicated across the x-axis, and the log2(O/E) values are plotted on the y-axis.

Figure S5

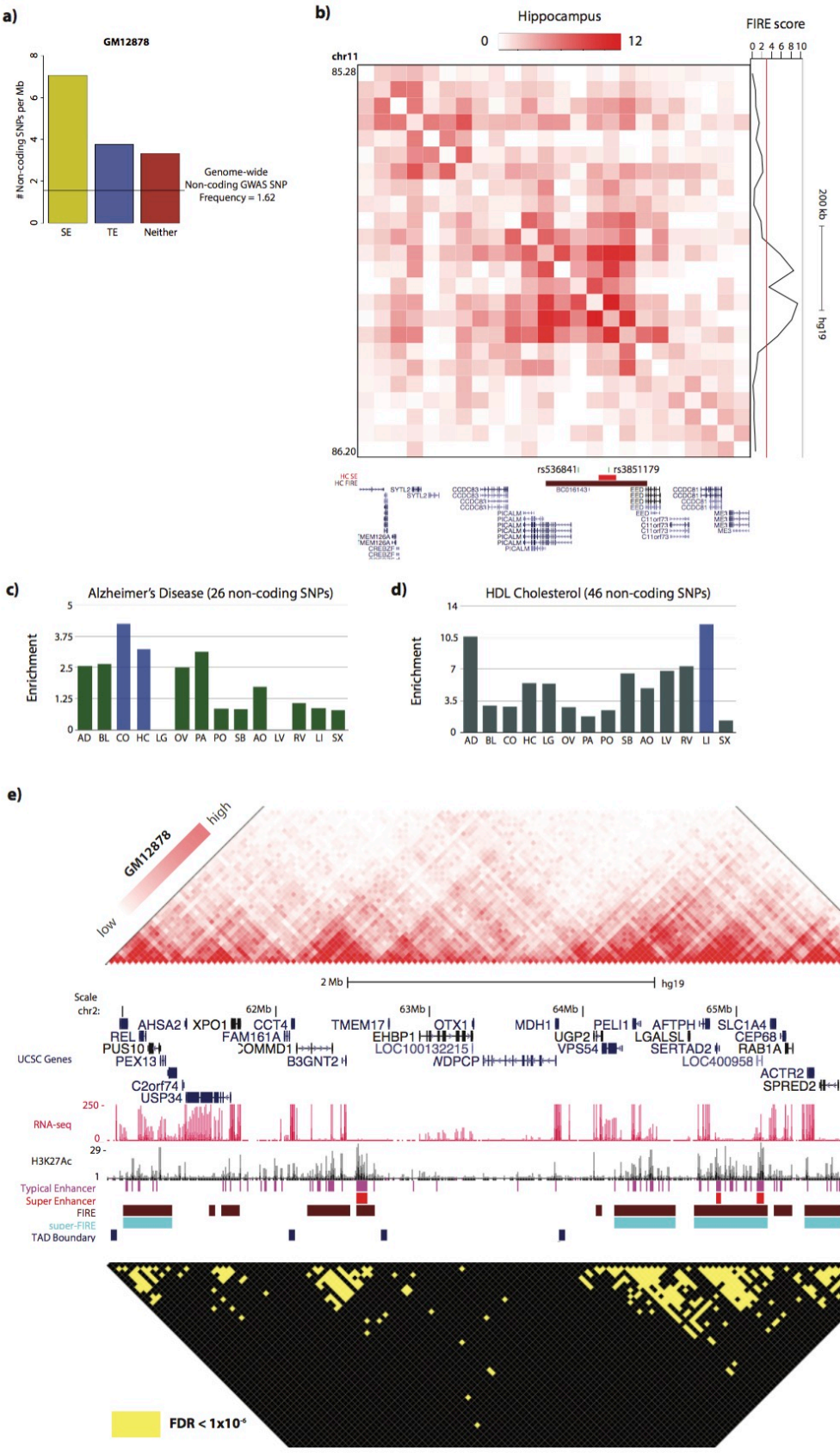

a)
GM12878
Non-coding SNPs per Mb
Genome-wide
Non-coding GWAS SNP
Frequency = 1.62
SE
TE
Neither
b)
Hippocampus
0
12
FIRE score
chr11
85.28
86.20
200 kb
hg19
rs536841
rs3851179
c)
Alzheimer's Disease (26 non-coding SNPs)
Enrichment
AD BL CO HC LG OV PA PO SB AO LV RV LI SX
d)
HDL Cholesterol (46 non-coding SNPs)
Enrichment
AD BL CO HC LG OV PA PO SB AO LV RV LI SX
e)
GM12878
high
low
2 Mb
hg19
Scale
chr2:
62Mb
63Mb
64Mb
65Mb
UCSC Genes
AHSA2
XPO1
CCT4
TMEM17
OTX1
MDH1
PELI1
AFTPH
SLC1A4
REL
FAM161A
EHBP1
UGP2
LGALSL
CEP68
PUS10
COMMD1
LOC100132215
VPS54
SERTAD2
RAB1A
PEX13
B3GNT2
NDPCP
LOC400958
ACTR2
C2orf74
SPRED2
USP34
RNA-seq
250
0
H3K27Ac
29
1
Typical Enhancer
Super Enhancer
FIRE
super-FIRE
TAD Boundary
FDR < 1x10⁻⁶

Figure S5. Analysis of non-coding disease-associated SNPs in FIREs and FIRE-FIRE contacts, related to Figure 6.

A) Bar plot showing the number of non-coding GWAS SNPs per megabase in FIRE overlapping super-enhancers (SE), FIREs overlapping typical enhancers (TE), and FIREs not overlapping either TE or SE. The horizontal line indicates the genome-wide SNP frequency. All analysis was done using GM12878 FIRE data.

B) Normalized Hi-C contact matrix of a 920kb locus (chr11:85,280,000-86,200,000) in human hippocampus tissue (HC). The tracks below show the presence of two Alzheimer's disease associated SNPs (rs536841 and rs3851179) located within a broad FIRE region (brown, chr11:85,840,000-85,880,000). One SNP resides within a HC super-enhancer (red) and the other SNP resides outside of the super-enhancer but within the FIRE region. Both SNPs reside in close proximity to *PICALM*, as shown in the bottom UCSC gene track. Right of the Hi-C contact matrix is the continuous FIRE score across this locus.

C) Enrichment of Alzheimer's disease-associated SNPs across 14 primary tissue FIRE annotations, showing the highest enrichment in FIREs from both brain tissues (CO and HC).

D) Enrichment of SNPs associated with quantitative HDL cholesterol metrics across 14 primary tissue FIRE annotations, showing the highest enrichment in liver FIREs.

E) Normalized Hi-C contact matrix (top) in GM12878 for a 5.14Mb locus (chr2:60,900,000-66,040,000) illustrating the extent of statistically significant FIRE-FIRE interactions. Hi-C color scale ranges from low to high, corresponding to the 15^{th} and 99^{th} percentile contact frequencies within this locus. The reflected matrix shows the statistically significant (FDR<1e-6) Hi-C contacts within 2Mb genomic distance across the locus. Only bin-pairs with FDR<1e-6 are yellow, and the rest are black. Between the matrices are UCSC gene annotations (blue, top), RNA-seq data (red), H3K27Ac data (black), typical enhancer annotations (Hnisz et al., 2013) (purple), super-enhancer annotations (Hnisz et al., 2013) (red), FIRE annotations (brown), super-FIRE annotations (cyan), and TAD boundary calls (blue).

2. Supplemental tables.

Table S1. Hi-C Data Manifest and Quality Metrics, Related to Figure 1

Table S2. Compartment A/B Patterns and PC1 values, Related to Figure 1, 2

Table S3. TAD boundary annotations, Related to Figure 1, 2

Table S4. Fit-Hi-C peak calling summary, and related analyses, Related to Figure 6, 7

Table S5. Genome-wide FIRE scores, Related to Figure 2

Table S6. FIRE calls and sample-specific FIRE calls in the primary cohort, Related to Figure 2, 3

Table S7. Observed and Expected Values of FIREs in Compartment A and B, Related to Figure 2

Table S8. Gene Ontology (GO) analysis of genes near sample-specific FIREs; top biological process terms, Related to Figure 3

Table S9. Gene Ontology (GO) analysis of genes near sample-specific FIREs; top disease ontologies, related to Figure 3

3. Supplemental experimental protocols.

Tissue Collection

For all human tissues except for dorsolateral prefrontal cortex (CO) and hippocampus (HC), samples were collected as previous described as part of the Epigenome Roadmap Consortium collection (The Roadmap Epigenomics Consortium, 2015). Human dorsolateral prefrontal cortex (CO) and hippocampus (HC) tissue were obtained from the National Institute of Child Health and Human Development (NICHD) Brain Bank for Developmental Disorders. Ethics approval was obtained from the University Health Network and The Hospital for Sick Children for use of the tissues. The two specimens used here were from a single male donor, age 31, who was classified as healthy.

4. Computational methods.

Histone ChIP-Seq data processing and peak-calling

Published single- or paired-end ChIP-Seq raw data were downloaded for H3K4me1, H3K4me3, H3K27ac, H3K9me3, H3K27me3, and H3K36me3 from GEO database under accession number GSE16256 and from SRA database under accession umber SRP000941 (Roadmap Epigenomics Consortium et al., 2015). The raw data were aligned to hg19 human reference genome using BWA-mem. Unmapped, non-uniquely mapped, and low quality (less than 10 quality score) reads were removed. We also removed PCR duplicate reads with PicardTools. ChIP-seq peaks were identified using MACS2 with the following parameters (--format=BAM -g mm -m 5 50 -p 1e-5) with corresponding input ChIP-Seq data as a background model. We also calculated input normalized RPKM values for H3K9me3, H3K27me3, and H3K36me3 in each 40kb bin.

RNA-Seq data processing

Published RNA-Seq raw sequencing data were downloaded from GEO database under accession number GSE16256 SRP000941 (Roadmap Epigenomics Consortium et al., 2015). RNA-Seq raw reads were aligned to hg19 human reference genome using BWA-mem. Unmapped and non-uniquely mapped reads were removed. Transcription levels were obtained based on GENCODE annotation v19 and normalized to FPKM values using Cufflinks. FPKM values from multiple replicates or multiple donors were combined together and the mean FPKM value was calculated for each gene.

Hi-C data processing

Unpublished Hi-C libraries described in this manuscript were sequenced on either Illumina HiSeq2000 or HiSeq2500 instrument. All other published Hi-C data were downloaded from SRA and converted to paired-end FASTQ files. Paired-end reads were then aligned independently to either the hg19 human reference genome or mm9 mouse reference genome using BWA-mem. As BWA-mem retains multiple alignments for a single read-end if it maps in two locations (i.e. a chimeric read), we kept only the 5' alignments for each read-end. Read-pairs in which both read-ends had mapping quality greater than 10 were paired using in-house scripts and converted into BAM files using Samtools. PCR duplicates were then removed using PicardTools. If downsampling was performed, we then used PicardTools 'DownsampleSam' function to downsample this final processed BAM file. Then, raw contact matrices were constructed using in-house scripts, and then further processed using HiCNormCis (described below) for the FIRE analysis. For all other Hi-C analyses not pertaining to FIRE scores, Hi-C data were normalized using HiCNorm (Hu et al., 2012), Vanilla Coverage (Rao et al., 2014), or ICE (Imakaev et al., 2012), where indicated. For all datasets of similar nature [such as the main cell lines in this study (GM12878, IMR90, H1, H1-derived) or the primary tissue collection, or the samples from each respective publication], we performed quantile normalization on HiCNorm matrices to normalize for differences in sequencing depth between samples within each group. This was done prior to any downstream comparative analyses.

Compartment A/B Calling

Compartment A/B analysis was performed at 1Mb resolution as previously described (Lieberman-Aiden et al., 2009). First of all, we calculated the average read count for each 1Mb bin in each sample. For cell line data, we removed 1Mb bins with average read count <=100. For tissue data, we removed 1Mb bins with average read count <=10. We used different thresholds for cell line data and tissue data, since tissue data have generally lower sequencing depth than the cell line datasets. Such filtering step has removed around 10% low coverage regions in the entire genome. Due to varying sequencing depths, the filtered regions are slightly different in each sample, and only bins which had a numeric value across all samples were used for downstream compartment analysis. After generating the first three principle components using the 'prcomp' function in R on the Pearson correlation matrix, we visually examined the first principle component (PC1) in each of 7 cell lines and 14 tissues, and found that for a few tissues the PC1 vectors of chr3 and chrX correspond to two chromosome arms, instead of A/B compartment. In specific, these outliers are PC1 vector of chr3 in bladder (BL), dorsolateral prefrontal cortex (CO), hippocampus (HC), lung (LG), psoas muscle (PO), aorta (AO), left ventricle (LV), right ventricle (RV), and PC1 vector of chrX in adrenal gland (AD), dorsolateral prefrontal cortex (CO), hippocampus (HC), pancreas (PA), psoas muscle (PO), left ventricle (LV), right ventricle (RV). For those outliers, the second principle component (PC2) was used to call A/B compartment. Visual examination of those PC2 vectors confirmed they match to the plaid-pattern observed in the normalized Hi-C contact matrices, instead of two chromosome arms.

Compartment A/B Conservation Analysis

To estimate the degree of compartment label conservation (related to Figure 1b, c; Figure S1f), we first scanned every 1Mb bin across the genome and counted the number of cell lines or tissue types that shared the same compartment label, and recorded which label was shared. By performing this at genome-wide scale, we obtained an observed distribution of A/B compartment conservation (Figure 1c, Figure S1f). To statistically determine if this distribution deviates from expectation, or to statistically test the significance of ubiquitous conservation (same label in all cell lines and tissue types), we first created an expected distribution of compartment conservation. First, for each cell line or tissue type, we randomly permuted the compartment label for each bin, while preserving the total number of A or B compartments on each chromosome. We then conducted the same conservation enumeration described for the observed data, and obtained an expected distribution of conservation (Figure S1f). This distribution was compared to the observed distribution using a Chi-square test. Testing the significance of observing the same compartment label ("ubiquitous conservation") across all cell lines or tissue types was done by comparing to the expected values using a binomial test.

TAD Boundary Reproducibility and Conservation Analyses

To estimate the degree of TAD boundary region conservation across samples in the primary cohort (related to Figure 1d, e; Figure S1e), we first identified TAD boundaries at 40Kb bin resolution for each sample independently, and then concatenated unique boundary bins across all samples into a single putative boundary region reference file. Consecutive TAD boundaries within 200Kb distance were also merged into a TAD boundary "region". Merging of adjacent boundary bins was performed because often times larger TAD boundaries (up to 400Kb) may result in slightly shifted (by a few bins) boundary calls between samples, and though they do not directly overlap, then both are a bin within the same boundary region. Moreover, in previous reports, TAD boundaries have been defined as 40-400Kb (Dixon et al., 2012) while regions >400kb are characterized as regions of "disorganized chromatin". Given this, and after defining boundary "regions" using our approach, the final list of unique TAD boundary regions ranged in size from 40-400Kb, consistent with previous definitions (Dixon et al., 2012). Using the cumulative list of TAD boundary regions, we evaluated the fraction of the total number of cell lines and tissues that had a boundary bin overlap with the given boundary region. To evaluate the overlap of TAD boundaries between tissue Hi-C biological replicates (LG, PA, PO, SX), boundaries within 80kb of each other were considered overlapping, which may underestimate the true boundary overlap since TAD boundaries have been previously defined as up to 400kb, and large boundaries regions are subject to technical variation in TAD calling at 40kb resolution. A chi-square test was used to evaluate statistical significance of TAD boundary overlap between replicates.

TAD Boundary Reproducibility and Conservation Analyses

To understand if our TAD identification method is robust across the sequencing depths used in this manuscript, we downsampled H1 rep2 Hi-C data (Dixon et al., 2015) as described above, and constructed HiCNorm contact maps. We then applied the insulation square method (Crane et al., 2015) to identify TAD boundaries. To determine what fraction of TAD boundaries within a given downsampled dataset overlap other putative TAD boundaries in H1 downsampled data, we first collected all putative TAD boundary regions from each of the 7 samples and made a reference putative boundary file (approximately 2,700 putative TAD boundary regions). For each downsampled dataset, we then asked what fraction of TAD boundary regions overlaps the boundaries in the reference putative boundary list (related to Figure S2h). To understand what fraction of TAD boundary regions are shared across all downsampled datasets we calculated the percentage of TAD boundaries that were unique to subsets of the downsampled files, including TAD boundaries that were shared across all downsampling datasets (related to Figure S2i).

Comparison of FIREs and chromatin loops and insulated neighborhoods

To explore the relationship between FIREs and chromatin loops, we called FIREs using the methods described in this manuscript, except at 5kb resolution using in situ Hi-C data in GM12878 and IMR90 (Rao et al., 2014). To compute the enrichment of chromatin loops in FIREs, we first assigned each chromatin loop anchor to a 5kb bin using the previously published loop annotations. We then computed the observed overlap between 5kb FIREs and 5kb loop anchors, and the expected overlap by permuting the FIRE positioning. Statistical significance was computed using Chi-square test. Conversely, to analyze the enrichment for FIREs at chromatin loop anchors, we conducted the same type of analysis, except asking what fraction of loop anchors are overlapped by a FIRE.

To explore the relationship between FIREs and insulated neighborhoods, super-enhancer domains and polycomb domains, we computed the enrichment (observed overlap / expected overlap) of 40kb FIREs at insulated neighborhoods defined in H1 cells (Ji et al., 2016), and the enrichment of 40kb FIREs at super-enhancer domains and polycomb domains in mESCs (Dowen et al., 2014). Statistical significance was computed using Chi-square test.

Identifying super-FIREs

To identify super-FIREs, we used a similar approach of that used to identify super-enhancers (Hnisz et al., 2013). First we merged all book-ended FIRE bins into large continuous FIRE regions. We then ranked the merged FIRE regions by their cumulative Z-score, and plotted the ranked FIRE regions as a function of their cumulative Z-score (related to Figure S3c). We then found the inflection point of the line plot, and defined the FIRE regions to the right of the inflection point as super-FIREs. The same procedure can be done for 5kb bin resolution FIREs, but by stitching FIRE bins within 15kb of one another.

Enrichment of FIRE in compartment A or compartment B

Using the compartment A/B calls at 1Mb resolution for each sample, observed FIRE bins were categorized into either compartment A or compartment B, depending on which compartment the FIRE bin resided. For all observed FIRE calls, the total compartment A overlap and compartment B overlap were enumerated ($O_{FIRE(A)}$ or $O_{FIRE(B)}$). To generate expected values, FIRE bins were randomly permuted while preserving the total number of FIREs per sample and per chromosome, and then re-categorized into either compartment A or compartment B ($E_{FIRE(A)}$ or $E_{FIRE(B)}$). Enrichment for compartment A or compartment B was calculated as either $\log2(O_{FIRE(A)}/ E_{FIRE(A)})$ and $\log2(O_{FIRE(B)}/ E_{FIRE(B)})$, respectively. To statistically evaluate the significance of enrichment of FIREs in compartment A or compartment B, for we created a two by two table using total compartment A overlap and compartment B overlap in observed FIRE calls ($O_{FIRE(A)}$ or $O_{FIRE(B)}$) and expected FIRE calls ($E_{FIRE(A)}$ or $E_{FIRE(B)}$), respectively. Chi-square test was performed to access the statistical significance (related to Table S7) and the process was performed independently for each sample.

FIRE positioning relative to TAD

For each sample and each FIRE bin, we found the TAD for which the FIRE bin resides using TAD calls for that given sample (related to Figure 2e, f). For each FIRE bin within a given TAD, we set the center position of the TAD

to 0.5 relative distance units, corresponding to 'halfway' between each adjacent TAD boundary. We then computed the distance from the TAD center to the boundary (D_{center}), as well as the distance of the FIRE bin to the nearest boundary (D_{FIRE}). Selecting the nearest boundary ensures the D_{FIRE} will always be less than or equal to D_{Center}. The relative distance units of the FIRE within a TAD are then computed as $(D_{FIRE}/D_{Center})/2$.

FIRE clustering analysis

We performed hierarchical clustering analysis using all samples in our primary cohort. Specifically, we first used the normalized total cis interaction (HiCNormCis) value for each 40Kb bin, and calculated the Euclidean distance of two genome-wide FIRE score vectors between any two samples, using the R function "dist". We then used the R function "hclust" with option "single linkage" to perform the hierarchical clustering analysis (related to Figure 3a). Next, we selected 40Kb bins which are cell line or tissue specific FIREs, and visualized their HiCNormCis scores using software JAVA TreeView (Saldanha, 2004).

Genomic Regions Enrichment of Annotations Tool (GREAT) analysis

We performed the GREAT analysis (McLean et al., 2010) to investigate the biological processes and disease ontologies for genes in the neighborhood of cell line or tissue specific FIRE bins (related to Figure 3d, e; Table S8-9). Specifically, we input our list of cell- or tissue-specific FIRE bins for each sample into the GREAT software (http://bejerano.stanford.edu/great/public/html/), and allowed the software to test neighboring genes for biological process and disease ontology enrichment. GREAT then evaluates the statistical significance of enrichment for each biological process, compared to the whole genome background. A Bonferroni-corrected Binomial test was used to obtain the p-value. Reported are the top fifteen biological processes ranked by the most significant p-values, in GM12878-specific FIREs and brain-specific FIREs, respectively (related to Figure 3e, f) and top terms for all samples as well as top disease ontologies are found in Tables S8-9.

Histone Local Enrichment Analysis

For each 40Kb FIRE bin in each sample, we calculated either the number of peaks per bin (for narrow peaks H3K27ac, H3K4me1 and H3K4me3) or the RPKM values per bin (for broad peaks H3K27me3, H3K9me3 and H3K36me3) and then calculated these values for each of the 12 bins upstream and 12 bins downstream of the FIRE bin, creating a vector of 25 values, centered on the FIRE bin (related to Figure 4b; Figure S3a). Those 25 values represent the histone mark profile in 1Mb region centered at each FIRE bin. As a control, to generate an expected histone mark profile, we randomly permuted the location of FIRE bins ten times within each sample, and calculated the averaged peak count or RPKM value at each position across ten random permutations. To calculate the local enrichment, we first calculated the ratio between observed value and expected value for each of the 25 positions around a FIRE bin, creating an enrichment score profile. Then, to assess the magnitude of local enrichment, we normalized each enrichment score relative to the local minima, by taking the log2 of the position enrichment divided by the minimum local enrichment. This converts the data to have a local enrichment of 0 at the local minima and specifically allows one to appreciate the enrichment of FIRE bins relative to the local neighboring bins, rather than relative to genome-wide levels.

Mean-rank Gene Set Test

To determine if genes near sample-specific FIREs tend to be expressed predominantly in the same tissue, we adapted the Mean-rank Gene Set Test concept, originally described in the 'Limma' R package (Ritchie et al., 2015) (https://bioconductor.org/packages/release/bioc/html/limma.html). Conceptually, the mean-rank gene set test evaluates whether a particular subset of genes is highly ranked relative to other genes in terms of a given statistic. Then using the Wilcoxon test, evaluates the null hypothesis that the mean rank of a subset of genes is not different than the expected mean ranking. A 'p-value' is generated by using the 'WilcoxGST' function in the Limma R package whereby the statistic parameter is a ranked list of relative gene expression values (with 1 being the gene with the highest relative expression, defined more below), and the index parameter is the positional indices of the genes within 200kb of a sample-specific FIRE set. However, the Wilcoxon test only evaluates if the mean rank of the test genes are different from the expected ranking, therefore not specifically addressing whether the mean rank is

more towards 1 compared to the expected ranking. Therefore, we present the results as the difference between the expected rank and actual mean rank, whereby a positive value indicates that the mean ranking is closer to 1 than the expected ranking.

In more detail, for each cell line or tissue, we first collected genes whose transcription start site (TSS) is within 200kb of a sample-specific FIRE. The collection of these genes within 200kb of sample-specific FIREs make up the sample-specific FIRE gene set, termed "FIRE genes". To prepare the Relative Expression rank file for each cell line or tissue, we used RNA-Seq data to first filter out genes with zero FPKM in all 21 samples, and then transformed the expression values into Log2(FPKM+1) values. Next, we divided each gene expression value by its cumulative gene expression sum across all 21 samples, to create the relative gene expression value (related to Figure 4f). For each sample, we then sorted all genes by their relative gene expression to assign each gene an expression rank, with 1 being the gene with the highest relative gene expression in that sample. Using these ranks for each sample, we calculated the mean expression rank for genes from a sample-specific FIRE gene set (related to Figure 4h), and then across all sample-specific FIRE gene sets (related to Figure 4g). A gene set enriched for sample-specific expression is expected to have a lower numeric mean rank (towards 1). By random chance, the mean rank will be approximately half of the total number of expressed genes. Therefore, we defined the enrichment score as the expected mean rank – observed mean rank. A large positive enrichment score indicates that genes within 200kb of sample-specific FIREs are primarily expressed in that sample relative to others, whereas a large negative enrichment score indicates that genes within 200kb of sample-specific FIREs are lowly expressed in that sample relative to other samples.

FIRE bin conservation

To investigate the degree of conservation of FIRE bins between human and mouse in three difference cell types (related to Figure 5a, b), we first identified FIRE bins using our HiCNormCis approach in the human and mouse samples. Next we identified breakpoints of major genomic rearrangements between human and mouse based on UCSC "net" alignments (Chiaromonte et al., 2001; Kent et al., 2003; Schwartz et al., 2003). To identify breakpoints in hg19, we used the alignment where hg19 is the target genome and mm9 is the query genome (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/vsMm9/hg19.mm9.net.gz). To identify breakpoints in mm9, we used the alignment where mm9 is the target and hg19 is the query (http://hgdownload.soe.ucsc.edu/goldenPath/mm9/vsHg19/mm9.hg19.net.gz). From each alignment, we calculated the genomic coordinates of the boundaries of all "fill" and "gap" blocks of size >50kb. We sorted these coordinates and then recursively merged those that are separated within 25kb into a single genomic interval. The resulting set of merged intervals defined our breakpoints. Any FIRE bins containing human<->mouse synteny breakpoints as defined above were removed from downstream analyses. UCSC liftover tool was then used to convert the genomic location of FIREs between hg19 human reference genome and mm9 mouse reference genome. Since in many cases the a 40kb bin in one species lifts over to a region that is part of 2 40kb bins in the other species, we considered a "conserved FIRE" if 1 of the 2 bins was a FIRE call. As a control, we also lifted over the genomic location of randomly permuted FIREs (that don't contain a breakpoint) between human and mouse, and calculated the number of FIREs that are conserved. For each of the six comparisons in Figure 6a, we also obtained the expected level of conservation. A Chi-square test was used to evaluate the statistical significance of FIRE conservation between human and mouse.

FIRE score conservation

To estimate the FIRE score conservation between human and mouse across a range of FIRE scores (related to Figure S4a-f), we randomly selected 4,000 40Kb bins, and used UCSC liftover tool to convert the genomic location of the randomly selected 40Kb bins between hg19 human reference genome and mm9 mouse reference genome. Since in many cases the a 40kb in one species lifts over to a region that is part of 2 40kb bins in the other species, we took the average FIRE score of the 2 40kb bins when conducting the correlation analysis. We then made a scatter plot of FIRE scores between the paired human and mouse datasets at the syntenic 40Kb bins, and calculated the Pearson correlation coefficient.

Change in FIRE score upon loss of Cohesin or CTCF

To investigate the impact of Cohesin loss on local interaction frequency (i.e. on FIRE tendency), we evaluated the change in local interaction frequency (as 'Change in Z-score') upon loss of Cohesin (related to Figure 5c-e) or CTCF (related to Figure S4j). In these analysis, we used the Z-score for each FIRE bin, instead of negative -ln(p-value), since Z-scores has approximate Gaussian distribution. For comparison of Z-score change between "control cells" (defined within each experiment as the condition without Cohesin manipulation or CTCF knockdown) and experimental cells (defined within each experiment as the condition with Cohesin depletion or knockout, or CTCF knockdown), we first identified the most confident FIRE bins in control cells, defined as FIRE bins in both control biological replicates. Next, we calculated the change of Z-score between control and experimental, at those selected most confident FIRE bins. As an analysis control, we also calculated the change of Z-score between two control biological replicates at the same set of high confidence FIRE bins. A two sample t-test was used to evaluate the statistical significance of the difference in Z-scores between control vs. experimental, as well as between two biological replicates of control samples. Since two WT biological replicates are symmetric, we took the absolute value of the difference in Z-score between the biological replicates. Therefore, the Z-score difference between two control biological replicates is always positive, and is a fair comparison to the Z-score difference between control and experimental.

CTCF and SMC3/Rad21 Enrichment Analysis

To determine if FIREs are enriched for CTCF or SMC3 (in TEV sample) or Rad21 (in mAST_floxed mNSC_floxed or Tcell_WT samples), we calculated how many CTCF or Cohesin subunit peaks are present in FIREs. We also permuted FIRE positioning 10 times, and asked the same question to obtain a distribution of expected values. To determine statistical significance, we compared this observed value to the expected distribution using a one-sample t-test.

FIRE and disease-associated SNP analyses

We collected the 4,378 non-coding disease associated GWAS SNPs (referred to hereafter as "SNPs") used in a previous study (Hnisz et al., 2013), and converted each SNP ID to its genomic location in hg19 human reference genome, using NCBI dbSNP online tool (http://www.ncbi.nlm.nih.gov/projects/SNP/dbSNP.cgi?list=rslist), resulting in 4,327 SNPs. Next, we mapped each SNP to FIRE bins identified from each of 7 cell lines and 14 tissues, and calculated the SNP density, defined as the number of mapped SNPs per 1Mb of FIRE bins. We further divided FIRE bins based on their overlap with typical enhancers and super-enhancers, and calculated the SNP density within each sub FIRE groups. Additionally, we performed disease-based FIRE SNP overlap analysis. For each of 456 diseases, we defined the enrichment score as the ratio between the proportion of SNPs overlapped with FIRE bins and the proportion of FIRE bins in the genome. Higher enrichment score indicates stronger overlap between SNPs and FIRE bins.

Calling Significant Interaction Pairs in Hi-C data

Statistically significant contacts in Hi-C data were identified using Fit-Hi-C, as previously described (Ay et al., 2014). First, Fit-Hi-C assumes that the expected contact frequency is a function of genomic distance. Fit-Hi-C also assumes the observed contact counts follow a Poisson model for non-peak Hi-C bin-pairs, (i.e. $O_{ij} \sim Poisson(\lambda(d_{ij}))$), and assumes an observed contact count is significantly higher than this Poisson variable for a peak bin-pairs (i.e. a statistically significant Hi-C contact). Fit-Hi-C conducts fitting and removing outliers iteratively. Fit-Hi-C requires the user to specify the range of genomic distance to assess for statistical significance. Based on this genomic distance input and for each iteration, Fit-Hi-C first bins the specific genomic distance into B bins (by default B=100), then estimates the mean observed contact count of currently labeled non-peak bin-pairs from each bin and then fits a spline curve $\lambda(d_{ij})$ based on average observed count at each distance determined by B and the user-input distance cutoff. For example, if one were to input B=50 and 2Mb genomic distance, then the spline curve will fit the mean contact count across 50 distance data points. Then, Fit-Hi-C tests each observed count O_{ij} against the calibrated Poisson distribution $Poisson(\lambda(d_{ij}))$. Fit-Hi-C rejects the null hypothesis when p value is small and labels this observation as a significant bin-pair "peak" (a significant Hi-C contact). In the next iteration, Fit-Hi-C conducts the same processes of calibrating the background distribution and significance testing. After converting our Hi-C contact matrix into the correct input format for Fit-Hi-C, we used the default Fit-Hi-C code to

calculate a p value and q value (a false discovery rate, FDR) for each bin-pair within 2Mb genomic distance. The generic example code for Fit-Hi-C can be found here: (https://noble.gs.washington.edu/proj/Fit-Hi-C/). For all analyses in this study (except where noted) we used a conservative peak-calling threshold of FDR<1e-6. This is based on the observation that more relaxed peak calls (FDR<0.05, the Fit-Hi-C default parameter) seemed to overcall peaks, and, FDR<1e-6 corresponds to ~1 million total peaks in IMR90, very similar to previous reports (Jin et al., 2013).

eQTL Enrichment Analyses

Statistically significant SNP-gene pairs were downloaded from the GTEx Portal (http://www.gtexportal.org/home/), using Version 6 (filed called GTEx_Analysis_V6_eQTLs.zip). Since only a subset of our tissue types can be found in the GTEx dataset, we extracted 6 GTEx datasets corresponding to 6 of our higher depth tissue Hi-C datasets. The following files were used from the GTEx datasets: Adrenal_Gland_Analysis.snpgenes, Liver_Analysis.snpgenes, Brain_Frontal_Cortex_BA9_Analysis.snpgenes, Artery_Aorta_Analysis.snpgenes, Heart_Left_Ventricle_Analysis.snpgenes, Heart_Left_Ventricle_Analysis.snpgenes.

To evaluate whether statistically significant contacts emanating from FIRE bins are enriched for SNP-gene pairs, and also to address whether the most significant Hi-C peaks are further enriched for SNP-gene pairs compared to less significant Hi-C peaks, we first used Fit-Hi-C to generate q values (i.e. FDRs) for all bin-pairs within 2Mb genomic distance for each tissue type and sub-selected higher depth tissue datasets in which we also obtained GTEx information (i.e. 6 tissues listed above). For the analysis of each sample, we first ranked significant bin-pairs by their FDR, from most significant pairwise contact to contacts with FDR approaching 0.05 (default Fit-Hi-C significance cutoff). This generates a genome-wide ranked list of significant pairwise contacts. We then divided significant bin-pairs into two groups depending on whether the anchor bin is a FIRE bin or non-FIRE bin, creating two groups termed “FIRE bin peaks” and “non-FIRE bin peaks”. In order to evaluate whether there is a difference in the presence of known SNP-gene pairs emanating from FIRE bins compared to non-FIRE bins, we selected the top 1K-20K significant FIRE peaks at 1K step size. As a control, we randomly selected a size-matched statistically significant bin-pairs emanating from non-FIRE bins. To evaluate whether FIRE peaks contained more SNP-gene pairs than non-FIRE bin peaks, we tested whether the average number of SNP-gene pairs captured by the top set of FIRE peaks is significantly higher than the size-matched control set (from non-FIRE bin peaks), using a one-side two-sample t test. Due to the random nature of selecting the size-matched control set, we generated 10 control datasets for each comparison (i.e. 1k, 2k…20k). To assess if the most significant FIRE bin peaks are more enriched for SNP-gene pairs than less significant FIRE bin peaks, we have plotted the log2(O/E) values for the top 1k,2k,3,4k,5k,10k,15k FDR groups (related to Figure g-j). Using a p value here is not entirely appropriate to address this analysis since p values for two-sample t tests are sensitive to sample size.

FIRE peak analyses

To evaluate whether FIREs have more local peaks than non-FIREs, we used Fit-Hi-C peak-calling results at stringent statistical significance (FDR<1e-6) to obtain distributions of the number of peaks emanating from FIRE bins or size-matched randomly permuted non-FIRE bins. To determine if the observed number of peaks from FIREs is greater than non-FIREs, we used a two-sample t-test.

To determine if FIREs self-interact at higher frequency than FIREs with non-FIREs or non-FIREs with non-FIREs, we first collected all FIRE bins, and then for each distance (d) from 40kb to 2Mb, we calculated the mean interaction frequency in which a FIRE bin was contacting another FIRE bin. Therefore, for each distance increment, we obtain a mean FIRE-FIRE interaction frequency. We then repeated the same procedure, but this time calculating the interaction frequency of FIREs with non-FIREs at each distance increment. Lastly, we randomly permuted FIRE bin locations to obtain a set of random non-FIRE bins and then calculated the interaction frequency with other non-FIRE bins for each distance increment. Then, for each genomic distance increment, we compared the FIRE-FIRE frequency with either the FIRE-nonFIRE or nonFIRE-nonFIRE using a two-sample t-test (related to Figure 7e; Table S4). This process was done independently for each sample.

To evaluate if FIREs are often the significant contact target of other FIREs we first collected all significant (FDR<1e-6) FIRE target bins determined by Fit-Hi-C, as well as all FIRE bins. We then intersected the FIRE target bins and FIRE bin annotations, creating three groups: FIRE targets that are non FIREs, FIRE targets that are FIREs,

and FIRE bins that are not targets of other FIREs (related to Figure 7g, h). The statistical significance of whether a FIRE bin is more likely a target of another FIRE bin was evaluated using a chi-square test.

Supplemental References:

Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res 24, 999–1011.

Chiaromonte, F., Yap, V., and Miller, W. (2001). Scoring pairwise genomic sequence alignments. Pacific Symp.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331–336.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. Cell 159, 374–387.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell 155, 934–947.

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J.S. (2012). HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics 28, 3131–3133.

Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods 9, 999–1003.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I., et al. (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell 18, 262–275.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290–294.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. U. S. A. 100, 11484–11489.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (80-.). 326, 289–293.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501.

Rao, S.S.P., Huntley, M.H., Durand, N.C., and Stamenova, E.K. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 1–16.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.

Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics 20, 3246–3248.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res 13, 103–107.

The Roadmap Epigenomics Consortium (2015). Integrative analysis of 111 reference human epigenomes.