UNIVERSITY OF CALIFORNIA, SAN DIEGO

Genome-wide mapping and analysis of chromosome architecture in human tissues

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Anthony Schmitt

Committee in charge:

Professor Bing Ren, Chair
Professor Arshad Desai
Professor Kelly Frazer
Professor Christopher Glass
Professor Kun Zhang

2017

The Dissertation of Anthony Schmitt is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2017

DEDICATION

I would like to dedicate this dissertation to my family. To my parents, Tom and Paula, who have believed in me and taught me humility, love, respect, and work ethic. To my brothers and sister, Nick, Phil, Jay, and Anna, who have been best friends throughout my life and have shaped me into the person I am today. To my amazing wife, Meghan, who, in addition to marrying me, has supported me in numerous ways across my personal and professional endeavors. And to my expecting son, for whom I now strive for excellence. I love you all.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 4

LIST OF TABLES

Chapter 1

Chapter 4

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor, Professor Bing Ren, for his critical guidance and encouragement. He has played an invaluable role in the ideation and refinement of the work. He has supported me undertaking several ambitious projects, and has exemplified great mentorship through his patience, trust, and inspiration.

I would also like to acknowledge our collaborator and friend, Ming Hu. Ming has made significant contributions to the work by way of creative analytical approaches, and countless discussions regarding the interpretation of results. The fruits of this collaboration are described in chapter 1 and chapter 3 of this dissertation.

I would also like to acknowledge our collaborator Yun Li, and especially her post-doc Zheng Xu. Zheng made significant contributions to the work through rigorous analyses of significant 3D interactions in human tissues, which took our study a critical "one step further". Yun made several important contributions to the interpretation of GWAS variants in the context of our study.

I would also like to acknowledge many of the collaborators whom I've worked very closely with throughout my Ph.D., although in works that extend beyond the scope of this dissertation. I would like to acknowledge Kun Zhang, and his student Daniel Jacobsen; Kelly Frazer and her trainees Paola Benaglio, Hurley Li and Naoki Nariai; Christopher Glass and his post-doc Casey Romanoski; and Tom Vondriska and his trainees Manuel Garrido and Doug Chapski.

I would like to thank several people who have helped me immensely during my Ph.D. in both the computational and experimental components of my work. I have been greatly supported by Dr. Siddarth Selvaraj and Dr. Inkyung Jung. In addition to their friendship, I would like to thank Siddarth for his significant contributions to the design and analysis of the capture Hi-C data at the MHC and KIR loci, as well as contributing experimental suggestions extending beyond the scope of this work. Inkyung has been instrumental throughout my PhD, providing critical advice for the interpretation of FIREs in human tissues, and for developing a rigorous analytical framework for analyzing complex promoter Capture-HiC data, and leading the deeper analysis of these datasets.

I would also like to thank Dr. Yan Li, Dr. Jesse Dixon and Catherine Tan for their help in the experimental aspects of this work. Yan first taught me how to perform Hi-C, and provided the initial spark to my interest in studying the 3D genome. Catherine Tan assisted in experiments across several projects throughout my Ph.D., many of which extending beyond the scope of this work. Jesse has played important roles in refining my experimental command of Hi-C and related methods, and has contributed to experimental refinement of the Capture-HiC method, as well as other contributions extending beyond the scope of this work.

I would also like to thank Dr. Jesse Dixon, Dr. Feng Yue, Dr. Uli Wagner, Dr. Gary Hon, Dr. Siddarth Selvaraj, Dr. Inkyung Jung, and Yunjiang Qiu for their critical instruction of how to conduct bioinformatics analysis. When I joined Bing's lab, I had never opened a Terminal window, nor even heard of a single programming language. Jesse was the first to patiently introduce me to the basics of computing infrastructure, while Feng, Uli, and Gary provided useful help during my initial bioinformatics training. Siddarth exhibit much patience and thoroughness in his training of the HalpoSeq algorithm, while Inkyung has given much oversight to my more advanced training. However, I cannot thank Yunjiang Qiu enough for day-to-day assistance with all of my bioinformatics needs. In addition to being a good friend and bay mate, Yunjiang is primarily responsible for my overall progression and proficiency in bioinformatics analysis, and for that, I'm truly grateful.

I would also like to thank the rest of the current and previous Ren lab members, especially David Gorkin. David has exemplified what it takes to be an excellent scientist, and has been a great friend to me throughout my Ph.D. Dave has provided many critical insights on my work, and has inspired me through his hard work and passion for science and the training of other scientists. I'm truly grateful for his friendship, as well as his personal and professional guidance.

Lastly, I would like to thank my committee, Professor Arshad Desai, Professor Kun Zhang, Professor Christopher Glass, and Professor Kelly Frazer for their guidance and helpful suggestions throughout my Ph.D.

Chapter 1, in full, is a reprint of the material as it appears in *Nature Reviews Molecular Cell Biology*. Schmitt, Anthony D.; Hu, Ming; Ren, Bing. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in *BMC Genomics*, volume 16, Nov 5 2015. Selvaraj, Siddarth; Schmitt, Anthony D.; Dixon, Jesse R.; Ren, Bing. The dissertation author was the co-primary investigator and co-primary author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in *Cell Reports*, volume 17, Nov 15 2016. Schmitt, Anthony D.; Hu, Ming; Jung, Inkyung; Xu, Zheng; Qiu, Yunjiang; Tan, Catherine L.; Li, Yun; Barr, Cathy L.; Ren, Bing. The dissertation author was the co-primary investigator and the co-primary author of this material.

Chapter 4, in full, has been submitted for publication of the material as it may appear in *Science*. Inkyung Jung, Anthony Schmitt, Yarui Diao, Dongchan Yang, Zachary Chiang, Marilynn Chan, Catherine Tan, Cathy Barr, Bin Li, Samantha Kuan, Dongsup Kim, Bing Ren. The dissertation author was the co-primary investigator and the co-primary author of this material.

VITA

2009	Bachelor of Science, University of Massachusetts, Amherst
2009-2012	Research Technician, Massachusetts General Hospital
2017	Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Yamada S, Kuroda T, Fuchs BC, He X, Supko JG, **Schmitt A**, McGinn CM, Lanuti M, Tanabe KK (2012) Oncolytic herpes simplex virus expressing yeast cytosine deaminase: relationship between viral replication, transgene expression, prodrug bioactivation. Cancer Gene Therapy. 19 (3): 160-70.

Goodwin JM, **Schmitt AD**, McGinn CM, Fuchs BC, Kuruppu D, Tanabe KK, Lanuti M (2012) Angiogenesis inhibition using an oncolytic herpes simplex virus expressing endostatin in a murine lung cancer model. Cancer Investigation. 30 (3):243-50.

Caravan P, Yang Y, Zachariah R, **Schmitt A**, Mino-Kenudson M, Sosnovik DE, Dai G, Fuchs BC, Lanuti M (2013) Molecular MR Imaging of Pulmonary Fibrosis in Mice. Am J Respir Cell Mol Biol. 49 (6):1120-6.

Jin F*, Li Y*, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, **Schmitt AD**, Espinoza C, Ren B (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 503 (7475):290-4.

Fuchs BC, Hoshida Y, Fujii T, Wei L, Yamada S, Lauwers GY, McGinn CM, DePeralta DK, Chen X, Kuroda T, Lanuti M, **Schmitt AD**, Gupta S, Crenshaw A, Onofrio R, Taylor B, Winckler W, Bardeesy N, Caravan P, Golub TR, Tanabe KK (2014) Epidermal growth factor receptor inhibition attenuates liver fibrosis and development of hepatocellular carcinoma. Hepatology. 59 (4):1577-90.

Leung D*, Jung I*, Rajagopal N*, **Schmitt A**, Selvaraj S, Lee AY, Yen CA, Lin S, Lin Y, Qiu Y, Xie W, Yue F, Hariharan M, Ray P, Kuan S, Edsall L, Yang H, Chi NC, Zhang MQ, Ecker JR, Ren B (2015) Intergrative analyses of haplotype-resolved epigenomes across human tissues. Nature. 518 (7539):350-354

Schultz MD*, He J*, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, Lin S, Lin Y, Jung I, **Schmitt AD**, Selvaraj S, Ren B, Sejnowski TJ, Wang W, Ecker JR (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 523 (7559):212-6.

Selvaraj S*, **Schmitt AD***, Dixon JR, Ren B (2015) Complete haplotype phasing of the MHC and KIR loci with targeted HaploSeq. BMC Genomics. 16 (1):900

Schmitt AD, Hu M, Ren B (2016) Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol Cell Biol. 17 (12): 743-55.

Schmitt AD*, Hu M*, Jung I, Xu, Z, Qui Y, Tan CL, Li, Y, Barr CL, Ren B (2016) A compendium of chromatin contact maps reveal spatially active regions in the human genome. Cell Reports. 17 (8):2042-59.

Jung I*, **Schmitt AD***, Diao Y, Yang D, Chiang Z, Chan M, Tan C, Barr C, Li B, Kuan S, Kim D, Ren B (2016) A compendium of promoter-centered long-range chromatin interactions in 27 human tissues and cell types. Under Revision, Science.

ABSTRACT OF THE DISSERTATION

Genome-wide mapping and analysis of chromosome architecture in human tissues

by

Anthony Schmitt

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2017

Professor Bing Ren, Chair

Gene expression in mammals is regulated by complex networks involving higher order chromatin organization, transcription factor binding, histone and DNA biochemical modifications and other mechanisms. Our understanding of the functional relationship between 3D chromosome architecture and gene regulation has been limited by the technologies to map 3D chromatin looping and the breadth of cell or tissue types analyzed. During my Ph.D. I have addressed technological shortcomings in the field by developing a high-resolution method for mapping chromatin interaction profiles at thousands of loci in a single assay, termed Capture-HiC. We have shown that Capture-HiC is capable of obtaining interaction profiles for contiguous loci, and when used in conjunction with HaploSeq phasing technology, can obtain targeted haplotype phasing information for medically relevant loci such as the MHC and KIR loci. Also

during my Ph.D. I have greatly advanced our understanding of the functional relationship between chromatin organization and gene regulation through Hi-C analysis in 21 human cell lines and primary adult tissues. We have discovered that chromosome architecture in human tissues exhibits distinguishing signatures of local spatially active regions. These regions, termed FIREs, are highly tissue-specific, enriched for active enhancers and GWAS variants, and conserved between human and mouse. We also find that FIREs exhibit promiscuous local interaction behavior and a significant degree of self-interaction. Further, I have developed high-resolution promoter Capture-HiC technology, and used this to map promoter-centered long-range interactomes in 27 human cell and tissue types. We find that promoter-centered interactions in tissues lie within dynamic interaction networks, which cluster by developmental lineage. Most surprisingly, we find widespread promoter-promoter interactions that impact distal gene expression, including hundreds of promoter regions harboring GWAS variants that have functional implications on distal genes. Together, through Hi-C and Capture-HiC analyses in human tissues, we have developed a rich resource for understanding chromatin folding and gene regulation. We anticipate these studies to lay a foundation for future experiments designed to further understand the gene-regulatory function of chromatin folding as well as the future study of how deleterious variants in *cis*-regulatory elements perturb gene regulation.

Chapter 1

THE 3D GENOME

REVIEWS

TECHNOLOGIES AND TECHNIQUES

Genome-wide mapping and analysis of chromosome architecture

Anthony D. Schmitt[1], Ming Hu[2,3] and Bing Ren[4]

Abstract | Chromosomes of eukaryotes adopt highly dynamic and complex hierarchical structures in the nucleus. The three-dimensional (3D) organization of chromosomes profoundly affects DNA replication, transcription and the repair of DNA damage. Thus, a thorough understanding of nuclear architecture is fundamental to the study of nuclear processes in eukaryotic cells. Recent years have seen rapid proliferation of technologies to investigate genome organization and function. Here, we review experimental and computational methodologies for 3D genome analysis, with special focus on recent advances in high-throughput chromatin conformation capture (3C) techniques and data analysis.

[1]*Ludwig Institute for Cancer Research and the University of California, San Diego (UCSD) Biomedical Sciences Graduate Program, 9500 Gilman Drive, La Jolla, California 92093, USA.*
[2]*Department of Population Health, Division of Biostatistics, New York University School of Medicine, 650 First Avenue, Room 540, New York, New York 10016, USA.*
[3]*Present address: Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, Ohio 44195, USA.*
[4]*Ludwig Institute for Cancer Research, Department of Cellular and Molecular Medicine, Moores Cancer Center and Institute of Genomic Medicine, University of California, San Diego (UCSD) School of Medicine, 9500 Gilman Drive, La Jolla, California 92093, USA.*
Correspondence to M.H. and B.R.
hum@ccf.org; biren@ucsd.edu

doi:10.1038/nrm.2016.104
Published online 1 Sep 2016

Recent studies have revealed the existence of millions of potential *cis*-regulatory elements in the human genome, with a great number of them residing in intergenic regions and away from their target gene promoters[1,2]. The distal elements, which largely consist of enhancers, influence the transcription of target genes through looping of chromatin fibres[3–11] during animal development[12–16]. Evidence of chromatin looping has been detected for many enhancers[17–21]. However, the mechanisms by which chromatin interactions are formed and maintained during development remain to be elucidated.

The chromosome conformation capture (3C) method and its derived 3C-based technologies (termed C-technologies) are commonly used for studying chromatin interactions in eukaryotic cells[22–27] (TABLE 1). These techniques have uncovered general features of genome organization, which include the existence of hierarchical chromatin structures, such as compartments[22], topologically associating domains (TADs)[6,10], sub-TADs[11], insulated domains[17] and chromatin loops[27].However, different C-technologies and analysis strategies have produced variable data on chromatin domains and DNA loops; for example, 100-fold differences have been seen in the total number of statistically significant chromatin interactions between studies[19,27]; and different studies have used similarly sounding terminologies to describe different structural features (such as 'loops' versus 'significant interactions' and 'contact domains' versus 'topological domains'), thus clouding our understanding of chromosome topology in cells[19,27]. It is unclear whether the differences in numbers of chromatin domains and loops identified in different studies are due to experimental protocols or data analysis algorithms.

In this Review, we discuss recent experimental and computational advances in C-technologies. We briefly catalogue all C-technologies, and place special emphasis on a few key areas of recent technological advancements regarding methods for chromatin fragmentation, approaches for proximity ligation and the use of a target-enrichment step before performing ultra-high-throughput sequencing. We also thoroughly explore the recent computational advancements that have been developed to analyse data sets produced by C-technologies (termed C-data). We detail the approaches for interrogating various C-data sets, placing special emphasis on methodologies to account for experimental biases, assessment of the resolution of a data set, extraction of global chromosome organization features and identification of chromatin interactions. We also propose key factors for consideration when selecting the appropriate computational methods to analyse C-data. Owing to space limitations, this Review does not cover alternative applications of C-data, such as haplotype phasing[28–30], genome assembly[31–33], metagenomic applications[34–36] and three-dimensional (3D) chromosome modelling[22,24,37,38]. Readers can find excellent reviews on these topics elsewhere[5,39–42]. We conclude by providing perspective on the challenges that remain ahead.

C-technologies: advances and adaptations

3C was invented as a general method to study chromosome organization in eukaryotic cells[43]. It combines protein crosslinking and proximity ligation of DNA to detect long-range chromatin interactions between pairs of genomic loci. Briefly, nuclei are isolated following treatment of cells with formaldehyde, which crosslinks the chromatin proteins to their associated DNA to

Table 1 | **A tabulation of known chromosome conformation capture technologies**

Assay abbreviation	Full assay name	Refs	Related protocols or guidelines
1 versus 1*			
3C	Chromosome conformation capture	43	97–100
1 versus Many/All*			
Multiplexed 3C-seq	Multiplexed chromosome conformation capture sequencing	101	102
Open-ended 3C	Open-ended chromosome conformation capture	103	–
3C-DSL	Chromosome conformation capture combined with DNA selection and ligation	104	–
4C	Circular chromosome conformation capture	45	105
4C	Chromosome conformation capture-on-chip	51	–
4C-seq	Chromosome conformation capture-on-chip combined with high-throughput sequencing	106	46,72, 107,108
TLA	Targeted locus amplification	30	–
e4C	Enhanced chromosome conformation capture-on-chip	109	110
ACT	Associated chromosome trap	111	112
Many versus Many*			
5C	Chromosome conformation capture carbon copy	52	113–116
ChIA–PET	Chromatin interaction analysis paired-end tag sequencing	23	–
Many versus All*			
Capture-3C	Chromosome conformation capture coupled with oligonucleotide capture technology	25	–
Capture-HiC	Hi-C coupled with oligonucleotide capture technology	58	–
All versus All*			
GCC	Genome conformation capture	–	117
Hi-C	Genome-wide chromosome conformation capture	22	69,70,118
ELP	Genome-wide chromosome conformation capture with enrichment of ligation products	119	–
TCC	Tethered conformation capture	24	–
Single-cell Hi-C	Single-cell genome-wide chromosome conformation capture	38	96
In situ Hi-C	Genome-wide chromosome conformation capture with *in situ* ligation	27	–
DNase Hi-C	Genome-wide chromosome conformation capture with DNase I digestion	49	–
Micro-C	Genome-wide chromosome conformation capture with micrococcal nuclease digestion	50	–

*'1', 'Many' and 'All' indicate how many loci are interrogated in a given experiment. For example, '1 versus All' indicates that the experiment probes the interaction profile between 1 locus and all other potential loci in the genome. 'All versus All' means that one can detect the interaction profiles of all loci, genome-wide, and their interactions with all other genomic loci.

fix the chromatin structure. The crosslinked DNA is then digested using restriction enzymes and the ends of the digested DNA fragments are re-ligated in diluted conditions that strongly favour ligation of the juxtaposed DNA fragments. The frequency of ligation between two genomic loci is then assessed using PCR or direct DNA sequencing. Although proximity ligation had earlier been used to detect DNA loops between the rat prolactin promoter and a distal enhancer in uncrosslinked cells[44], the inclusion of formaldehyde crosslinking in 3C enhanced the efficiency and robustness of proximity ligation reactions[43], thereby enabling broad adoption of the 3C technique for high-throughput analyses of chromosome architecture.

Over the years, many additional modifications have been introduced to 3C techniques that further enhanced the scale, resolution and efficiency of chromosome conformation analyses. First, with the rapid advances in DNA sequence analysis technologies, 3C quickly developed into genome-scale methods with the adoption of microarray technology and eventually ultra-high-throughput DNA sequencing as a way to measure the frequency of proximity ligation products (TABLE 1). As only a fraction of DNA fragments generated by the C-technologies are legitimate ligation products between distinct genomic loci, it is necessary both to enrich for ligation junctions and to reduce or eliminate unligated DNA fragments. To achieve this, biotin-labelling with

Hi-C
A high-throughput, genome-wide chromosome conformation capture assay using affinity purification of labelled-DNA ligation junctions to measure pairwise interaction frequencies in cell populations.

biotin-conjugated nucleotides has been used to fill-in the 5′ overhangs left by restriction digestion before proximity ligation. Following proximity ligation, the ligation products are biotin-labelled at the ligation junctions[22,23]. Biotinylated nucleotides at the ends of unligated DNA molecules are conventionally removed by a dedicated T4 DNA polymerase reaction[22] or during the end-repair step of the library preparation procedure[27]. Biotinylated ligation junctions are eventually isolated by affinity purification and subject to ultra-high-throughput DNA sequencing, generating genome-wide chromatin contact maps that reflect chromosome organization in a cell population. The first rendition of this procedure, known as Hi-C, has now been widely used[22].

To increase the resolution of chromosome conformation analyses, modifications have also been made to the restriction digestion step (FIG. 1a). At the very core of C-technologies is the need to first fragment the chromatin of crosslinked nuclei to generate DNA ends capable of re-ligating to other spatially proximal fragmented ends[43]. Until recently, restriction digestion has been generally carried out using '6-cutters' — type II restriction enzymes that recognize a six-base-pair sequence motif. The finest resolution possible using 6-cutter fragmentation would in theory be the size of the restriction fragment generated (termed fragment-level resolution). The closest to achieving this was a recent high-resolution analysis of human fetal lung fibroblast, which achieved nearly

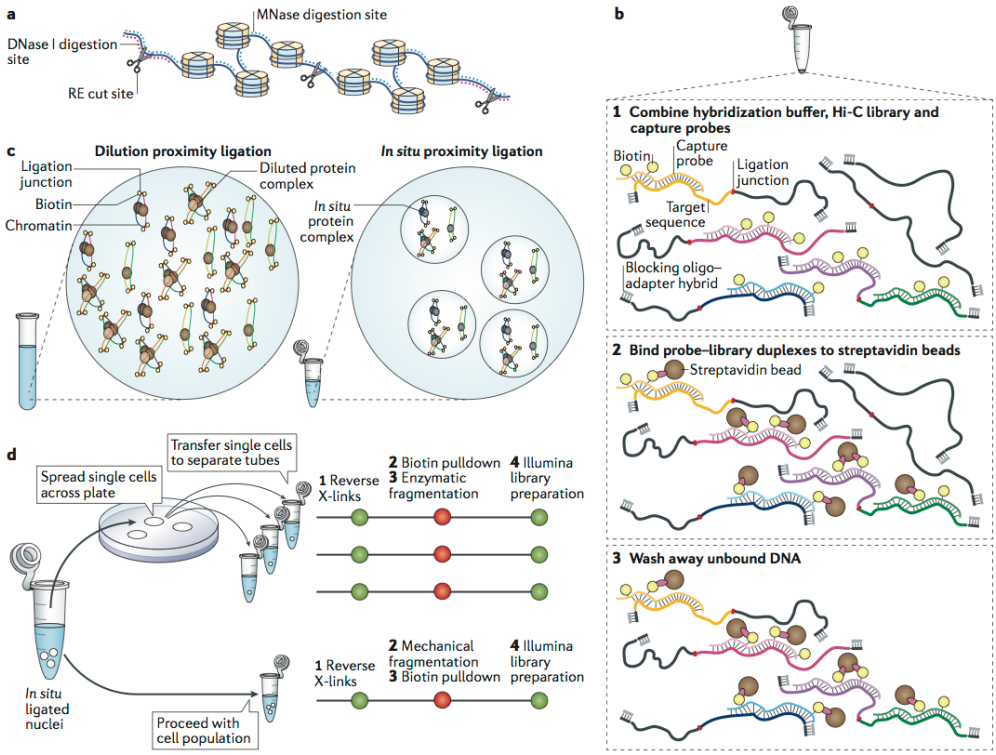


Figure 1 | **Experimental modifications to genome-wide chromosome conformation capture (3C)-based technologies (C-technologies).** **a** | Chromatin fragmentation can be achieved using type II restriction enzymes (REs), which cut at enzyme-specific recognition motifs[22,47], endonucleases such as DNase I, which fragments DNA at sites of open chromatin[48,49], and micrococcal nuclease (MNase), which fragments chromatin in histone linker sequences[50]. **b** | Hi-C includes the sequencing of all biotin-labelled ligation products, which are enriched by biotin-affinity purification and subsequent library preparation[22,69,70]. In Capture-HiC, sequences of interest can be enriched from a Hi-C DNA library to obtain highly multiplexed, targeted interaction profiles[29,53–60]. This involves the hybridization of biotinylated capture-probes to DNA sequences of interest (step 1), the immobilization of this library of probe–target sequence duplexes on streptavidin beads (step 2) and the washing away of unbound DNA, leaving only the captured probe–library duplexes (step 3). **c** | Proximity ligation in Hi-C sample preparation was originally done after nuclei were lysed and chromatin complexes were diluted, to favour intramolecular ligation events[22,69,70] (left). An alternative strategy is to carry out the proximity ligation step within intact nuclei[27,38] before nuclear lysis and DNA–protein crosslink reversal (right). **d** | Single-cell Hi-C[38,96] (top) differs from cell-population Hi-C[22,69,70] (bottom) by the plating of nuclei, the sorting of them individually into tubes and the processing of them using a modified library preparation protocol. X-links, crosslinks.

fragment-level resolution, requiring over 3.4-billion valid chromatin contacts (over 5.6-billion raw read-pairs)[19]. Although 4-cutters potentiate higher-resolution analyses of genome conformation by means of producing smaller restriction-fragment sizes, the total number of restriction fragments genome-wide is ~16-fold higher and the total number of possible pairwise contacts is 256-fold higher. Accordingly, 4-cutter fragmentation was initially applied in targeted chromatin conformation analysis using 4C (circular chromosome conformation capture; also known as chromosome conformation capture-on-chip) technology, as 4C interrogates the chromatin looping landscape of only a single restriction fragment with the rest of the genome, rather than all possible pairwise contacts genome-wide[45,46]. Genome-wide analyses with 4-cutter fragmentation were performed in flies[47], in part owing to their relatively small genome size compared to mouse or human, which significantly reduces the total number of possible pairwise contacts. To date, the finest resolution analysis of mammalian genomes has been carried out using a 4-cutter[27]. In this study, 4.9 billion valid chromatin contacts were required to obtain 1 kb-resolution Hi-C maps in a single cell type ('1 kb resolution' is explained further below). Other methods have now been used for chromatin fragmentation, each offering a unique set of advantages and disadvantages. DNase I has recently been shown to fragment chromatin of crosslinked nuclei for Hi-C applications[48,49] (FIG. 1a). Similarly, micrococcal nuclease (MNase) has been used to fragment chromatin before proximity ligation in yeast nuclei, helping to achieve nucleosome-level resolution of chromatin organization[50]. In addition, mechanical shearing was used to fragment chromatin in a 4C protocol variant and was suggested to be sufficient to fragment chromatin for Hi-C[49], although to our knowledge no Hi-C data from mechanical shearing have yet been published.

Conventional Hi-C requires billions of DNA sequencing reads to achieve truly genome-scale coverage at kilobase-pair resolution[19,27]. By contrast, the first targeted approaches, such as 4C and chromosome conformation capture carbon copy (5C), are PCR-based C-technologies, using PCR enrichment to analyse chromatin contact profiles of a single locus[45,51] or across a continuous locus, respectively[52]. Although these methods are less expensive than Hi-C and are based on relatively straight-forward protocols, they suffer from low throughput (4C) or complex primer design (5C) and, importantly, do not include the key advantage of Hi-C, which is the enrichment of valid ligation products using biotin-labelling of ligation junctions and affinity purification. To gain cost-effectiveness while preserving the efficiency afforded by genome-wide C-techniques, two strategies have been developed that also generate targeted 3C data. First, chromatin immunoprecipitation (ChIP) was introduced before the proximity ligation step to enrich for DNA associated with specific DNA-binding proteins, chromatin modifiers or histone modifications[23]. This method, termed chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), allows for targeted analysis of chromatin conformation at binding sites of transcription factors or at transcriptionally active chromatin domains. It also has the benefit of achieving a higher resolution compared to Hi-C, as only ligation products involving the immunoprecipitated molecule are sequenced.

Second, Hi-C has recently been combined with target enrichment and sequencing (Capture-HiC) to reveal chromatin contacts of mammalian gene promoters[49,53–57] and other specific genomic loci[29,53,58–60]. Unlike 4C and 5C, Capture-HiC involves first generating a library of proximity-ligated DNA fragments using one of several published Hi-C methods. Next, biotinylated RNA or DNA oligonucleotide probes are hybridized to specific sequences of interest (for example, gene promoters) within the Hi-C library, followed by affinity purification of the biotinylated probe–library duplexes, stringent washing of bound DNA and finally ultra-high-throughput DNA sequencing (FIG. 1b). Control over which genomic loci are interrogated in a Capture-HiC experiment is determined by the user when designing the capture probes. Importantly, ligation frequencies of probed regions detected from Capture-HiC experiments are highly similar to ligation frequencies measured by high-resolution, whole-genome Hi-C data[29], yet Capture-HiC data sets are obtained at a small fraction of the cost because only the probed regions are analysed, underscoring both the quality and efficiency of this method. Current Capture-HiC approaches have varied substantially with respect to template Hi-C library preparation procedure, target selection, capture probe design and target enrichment protocol (TABLE 2). Thus, data generated from such experiments vary widely with respect to the quality of target enrichment; for example, the on-target rate differs between studies. One consistent tendency is that Capture-HiC data from studies with larger target size have substantially higher on-target rates than data from studies with smaller captured regions, ranging from ~65% on target in select promoter Capture-HiC studies[54,56,57] to 5–15% when capturing small continuous regions or interspersed loci[29,58,59]. Interestingly, no reports to our knowledge have implemented the 'double-capture' strategy for small target sizes, which uses two consecutive captures to increase the on-target rates for difficult-to-capture templates[61]. Additionally, promoter Capture-HiC data generated using either RNA or DNA probes have reported differing on-target rates, with RNA probes currently outperforming DNA probes[54–57]. However, the first and only report of genome-wide promoter Capture-HiC using DNA probes also used 4-cutter library preparation, rather than 6-cutter, making it challenging to interpret which design approach is superior. Overall, variations in Hi-C library preparation, probe design, target size, number of probes allocated to each target locus and user expertise contribute to the variable quality and depth of coverage across loci in each study, making concrete experimental recommendations premature and creating challenges for downstream data analyses, as discussed in the next section.

A substantial, although variable (~7–50%), proportion of Hi-C contacts detected in mammals using the original Hi-C protocol originate from inter-chromosomal ('*trans*') ligation events[6,19,22,28,62–68]. The reported frequency of *trans* contacts varies tremendously across cell types and even

Chromosome conformation capture carbon copy (5C). A high-throughput chromosome conformation capture assay that examines the spatial proximity of two defined sets of genomic regions, measured using a pair of DNA oligos corresponding to the sequences upstream and downstream of the ligation junction.

Target size
The cumulative length (in base pairs) targeted by capture probes in a Capture-HiC experiment.

Table 2 | **Design and implementation of Capture-HiC experiments**

Oligo array vendor	Probe	Organism	Target (or targets)	Control (or controls)	Hi-C library protocol	Refs
Agilent SureSelect	RNA	Human	Breast cancer risk loci	Size-matched gene dessert regions	Hind III dilution Hi-C	58
Agilent SureSelect	RNA	Human	Colon cancer risk loci	N/A	Hind III dilution Hi-C	59
In-house*	RNA	Human	MHC and KIR loci	N/A	Hind III dilution Hi-C	29
In-house*	RNA	Human	Three ~2-Mb loci	N/A	Mbo I *in situ* Hi-C	60
Roche Nimblegen SeqCap	DNA	Human	LncRNA promoters	β-Globin LCR, *NANOG*, and *SOX2* loci	DNase I dilution Hi-C	49
Agilent SureSelect	RNA	Mouse	Promoters	Random ligation library‡	Hind III dilution Hi-C	56
Agilent SureSelect	RNA	Mouse	Promoters	Random ligation library‡	Hind III Dilution Hi-C	57
Agilent SureSelect	RNA	Human	Promoters	Random ligation library‡	Hind III Dilution Hi-C	54
Agilent SureSelect	RNA	Human	Promoters and autoimmune disease risk loci	*HBA* locus	Hind III dilution Hi-C	53
Roche Nimblegen SeqCap	DNA	Mouse	Promoters	Intergenic and exonic regions	Mbo I dilution Hi-C	55

HBA, haemoglobin subunit alpha; LCR, locus control region; LncRNA, long non-coding RNA; N/A, not applicable. *Single-strand DNA oligonucleotides are obtained from CustomArray and synthesized into RNA probes in-house. ‡In the random ligation library, crosslinks are reversed before the proximity ligation reaction.

biological replicates, and they are much less reproducible than the intra-chromosomal ('*cis*') contacts. This raises the possibility that many of these ligation products result from random inter-molecular ligations occurring during sample preparation in diluted conditions. In the original Hi-C protocol, following restriction digestion and biotin-labelling, nuclei are lysed using sodium dodecyl sulphate (SDS) and crosslinked chromatin complexes are diluted before proximity ligation[22,69,70] (FIG. 1c). Since the inception of Hi-C, 4C protocols have forgone the nuclear lysis step by way of omitting SDS treatment; conducting proximity ligation without intentional lysis of dilution, resulting in fewer observed *trans* contacts[71,72]. A recent study also indicated that in Hi-C, nuclear lysis and dilution of chromatin complexes before proximity ligation can be omitted, corroborating the observation that proximity ligation can occur within intact nuclei[38] (FIG. 1c). By adapting Hi-C with this modified ligation procedure (a process from here termed *in situ* Hi-C), a substantial improvement in the fraction of legitimate, informative ligation products is achieved without affecting the accuracy of conformation capture[27,38,65], with fewer random *trans* contacts, higher reproducibility of contacts across a range of distances and even reduction of previously described experimental bias[65]. Thus, *in situ* Hi-C seems to be the preferred protocol moving forward. However, as Hi-C data can be used only to infer genome organization based on observed contact frequencies, true evaluation of the superior protocol requires comparison to a set of known true interaction frequencies, which does not exist in a comprehensive fashion. Moreover, evaluating data quality based solely on the fraction of observed *trans* or long-range *cis* contacts is not entirely appropriate, as cells may indeed have highly intermingled chromosomes, depending perhaps on cell cycle stage. Instead, additional metrics should be used for assessing data quality, such as estimating random collision frequency[6] and analysis of read orientation as a function of linear genomic distance[19,27].

The improved efficiency of *in situ* proximity ligation and Hi-C facilitated the examination of chromatin organization in single mammalian cells using single-cell Hi-C, which provided a deeper understanding of cell-to-cell variability in chromosome architecture[38] (FIG. 1d). In single-cell Hi-C, cell populations are subjected to the initial steps of *in situ* Hi-C but, before crosslink reversal, the intact nuclei are sorted into individual tubes and subjected to a modified Hi-C-library preparation procedure and multiplexed PCR amplification. This strategy was applied in mouse T helper cells and produced genome-wide contact maps for 74 individual cells, with 10 of these maps being of high enough quality for further analysis[38]. The resulting single-cell contact maps, despite being very sparse (at 1 Mb bin size), confirmed the existence of chromosome territories and TADs while highlighting the cell-to-cell variability of chromosome architecture. Merged single-cell maps are similar to Hi-C data generated from millions of cells, supporting the reliability of the single-cell data. A key limitation of the method is that only a small number of unique chromatin contacts, up to 30,000 in the published work[38], were detected. This represents less than 2.5% of the total number of theoretical chromatin contacts in a mouse cell. The sparse data set probably results from inefficient steps in the existing protocol, such as enzymatic chromatin fragmentation, biotin-labelling, proximity ligation and conventional Illumina TruSeq library preparation. Removing the biotin-labelling step and performing sticky-end ligation, as in 3C, may potentiate the detection of more unique ligation junctions, as ligation junction detection will not depend on high efficiency of the enzymatic biotin-labelling reaction or the efficiency of blunt-end ligation. Additionally, more-efficient library preparation methods designed specifically to handle low inputs, such as tagmentation[73], may improve the yield and absolute number of detectable ligation junctions.

Bin size
A measure of Hi-C data resolution. A bin is a fixed, non-overlapping genomic span to which Hi-C reads are grouped to increase the signal of chromatin interaction frequency.

Computational analysis of C-data

The rapid development of C-technologies and fast accumulation of large amounts of data have posed great challenges for data analysis and interpretation, and necessitated the development of sophisticated computational tools that can accurately identify long-range chromatin

interactions and reveal the general principles of chromatin motion and organization. It is important to note that, although the observed frequency of proximity-ligation products has been used to infer the 3D distances between a pair of DNA sequences, procedures including crosslinking, chromatin fragmentation, biotin-labelling and re-ligation can all introduce biases that complicate the interpretation of observed contact frequencies[74–76]. Additionally, the resolution of analysis in the available data sets remains to be rigorously defined. To overcome these challenges, statistically solid and computationally efficient bioinformatics pipelines are essential. Several computational algorithms and tools have been developed in recent years, specifically for analysing C-data. Below, we discuss several key issues that need to be considered.

Accounting for experimental bias. Similarly to analysis of data generated by ChIP followed by sequencing (ChIP-seq) and RNA sequencing (RNA-seq), analysis of C-data can be confounded by multiple layers of bias that originate from different steps of experimental procedures. Accounting for these biases (at times referred to as bias removal or normalization) is the first and arguably the most important step in C-data analysis. Efficient and effective removal of multiple systematic biases is critical for the success of any subsequent analysis of C-data as well as for the proper interpretation of results.

In general, there have been two types of approaches to account for biases in C-data. The first class of bias-removal approaches account for biases in an explicit fashion — by assuming that all sources of systematic biases are known based on biases determined empirically from the observed data (FIG. 2; TABLE 3). The second class of bias-removal approaches account for biases in an implicit way — by assuming no known source (or sources) of bias, and assuming that the cumulative effect of the bias is captured in the sequencing coverage of each locus (or 'bin'). In other words, as Hi-C is a genome-wide assay, the implicit models assume that each locus should receive equal sequence coverage after biases are removed. These implicit models all rely on some implementation of matrix-balancing algorithms, and from here on they are referred to as the matrix-balancing methods (FIG. 2; TABLE 3). Therefore, selecting the appropriate bias-removal methodology depends on whether the sources of the biases in the data are assumed to be known or unknown. In a seminal study, restriction enzyme fragment lengths, GC content and sequence mappability were identified as three major sources of experimental biases in Hi-C data[77]. The key challenge is to estimate the combinatorial bias effect between two interacting loci. To address this challenge, the binary contact status between any two fragment-ends was modelled as the Bernoulli random variable. Next, to estimate the bias effects, the maximal likelihood approach was applied to the joint likelihood function, which is defined as the product of Bernoulli probability mass function for all possible fragment end pairs. In practice, to make such computation feasible, all interacting loci were first grouped into bins based on the percentiles of each bias factor. Next, an empirical distribution was used to estimate such combinational bias effects, leading to a statistically effective but computationally intensive bias-removal method[77]. Later on, HiCNorm, which is a generalized linear regression-based method, was developed to remove the above-mentioned three systematic biases in Hi-C data[78] (FIG. 2b; TABLE 3). Differing from the first explicit model[77], which used a Bernoulli distribution to model the binary contact status between any two fragment-ends, HiCNorm directly models the contact frequency between any two bins as a Poisson distribution or a negative binomial distribution[78]. Noticeably, analysing binned Hi-C data enables HiCNorm to adopt a simple parametric form for the combinatorial bias effect, resulting in much-improved computational efficiency.

In addition to these two explicit approaches, implicit, matrix-balancing approaches have been widely used to account for biases in Hi-C data and rely on two different assumptions. First, the combinatorial-bias effect between two interacting loci can be simplified as the product of the two locus-specific bias effects. Second, if there is no bias effect (that is, when all bias has been accounted for), the total genome-wide contact summation for each locus will be a constant, implying that each locus has 'equal visibility' to the Hi-C assay. Based on these two assumptions, classic matrix-balancing algorithms have been used to account for systemic bias. For example, the first method that described balancing Hi-C contact matrices was termed vanilla coverage[22] (FIG. 2c). To account for bias, the observed contact frequency between locus A and locus B is divided by the product of the total genome-wide contact frequency at locus A and the total genome-wide contact frequency at locus B, and the ratio is used as the normalized contact frequency (FIG. 2c). Later, iterative correction and eigenvector decomposition (ICE) was introduced (FIG. 2d; TABLE 3); this process iterates through the vanilla coverage procedure until there is convergence of the normalized contact frequency, thereby further reducing the coverage variability from locus to locus but greatly increasing the computational cost to achieve bias removal[79]. Since ICE was introduced, several efforts have been made to improve its computational efficiency[80,81]. Meanwhile, a fast version of the matrix-balancing Sinkhorn–Knopp algorithm[82], originally described by Knight and Ruiz[83], has been applied to account for biases in the finest resolution Hi-C data sets[27] (TABLE 3). Matrix-balancing methods may also be preferred when analysing Hi-C data prepared with other chromatin-fragmentation approaches, such as DNase I or mechanical sharing[49], as matrix-balancing methods assume that the source of bias is unknown, and the presence of empirically determined biases from these Hi-C data sets has not yet been thoroughly examined. In practice, both explicit and implicit approaches have been used to account for biases in Hi-C data; therefore, it would be helpful to conduct a comprehensive comparison between the two approaches. To date, only a partial comparison has been made, which highlighted the differences in reproducibility of *cis* and *trans* interaction frequencies at low resolution[84]. A novel computational framework that combines the strengths of the two approaches may enable more accurate bias removal and higher computational efficiency.

Restriction enzyme fragment lengths
The total genomic length in each bin that is within 500 bp of restriction enzyme cut sites used in the Hi-C library preparation.

Mappability
The probability of a read-mapping uniquely to the effective fragment length sequence within each bin.

Poisson distribution
A probability distribution for the discrete random variable in which the variance is the same as the mean.

Negative binomial distribution
A probability distribution for the discrete random variable in which the variance is larger than the mean.

Hi-C contact matrices
Symmetric, two-dimensional matrices (M), for which each matrix entry (M_{ij}) represents the raw or normalized contact frequency between bin i and bin j.

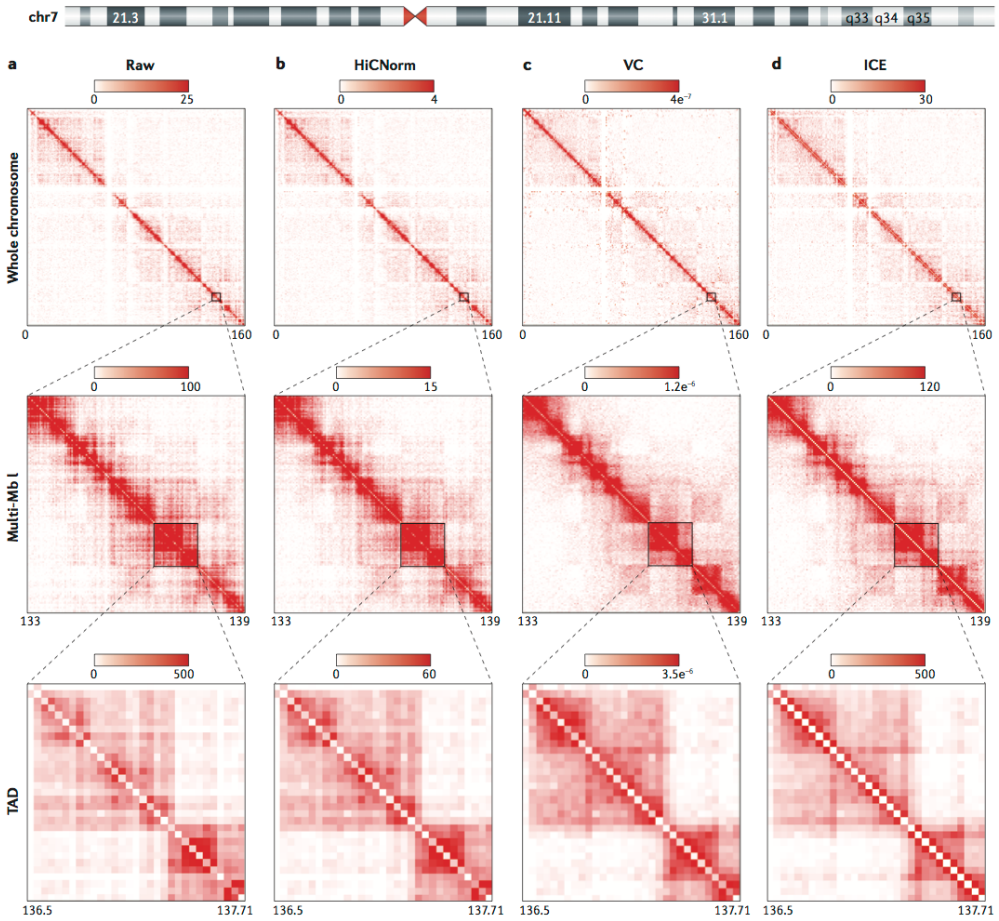


Figure 2 | **Comparison of computational methods to account for bias in Hi-C data.** We reprocessed high-resolution Hi-C data from IMR90 cells[19] uniformly until the bias-removal step, at which point either raw contact matrices were generated or normalization was conducted with one of three methods. Here, we illustrate a semi-quantitative comparison of human chromosome 7 (chr7) for 3 genomic resolutions (whole chromosome, a multi-megabase (multi-Mb) locus and a topologically associating domain (TAD)) at 40 kb bin size for a raw Hi-C contact matrix (part **a**), an explicit model of bias removal (HiCNorm) (part **b**), and two methods of matrix-balancing algorithms for bias removal, namely a fast, rough, single-iteration balancing method, vanilla coverage (VC) (part **c**) and iterative correction and eigenvector decomposition (ICE) (part **d**). It can be visually appreciated that the explicit or implicit assumptions made by each method to account for biases result in quantitative differences in the normalized interaction frequency between loci. The intensity gradient is a linear increase from zero to the maximum noted (units are observed read counts for the raw matrices, and normalized read counts for the normalized matrix columns). Depicted are a series of symmetrical Hi-C contact matrices at various genomic resolutions. The rows (i) and columns (j) of each matrix represent bins along a chromosome, in this case various regions of human chr7. Each matrix entry [i,j] represents the observed or normalized interaction frequency between a pair of genomic loci. Pairwise interactions observed at higher frequency are depicted as a darker red colour along the colour gradient, whereas light red coloration represents very few observed interactions in the Hi-C data. The gradient units for raw matrices (part **a**) are 'observed interaction frequency' and the units for HiCNorm, VC and ICE (parts **b**–**d**) are 'normalized interaction frequency', which become increasingly apparent when analysing more-local Hi-C contacts (closer to the diagonal). Matrix entries near the matrix diagonal represent pairwise interactions between loci that are proximal in linear genomic distance (i~j), whereas matrix entries far off the diagonal (i>>j) represent pairwise interactions between loci that are very distal in linear genomic distance. For whole-chromosome and TAD resolutions, the maximal signal intensity was set to the ninety-ninth percentile for the given matrix. For the multi-Mb resolution, the maximal intensity was set to the ninety-fifth percentile value of the given matrix. Each matrix is a symmetrical matrix, NxN, and the chromosome coordinate information is given below each matrix in megabases.

Table 3 | **Approaches to account for systematic biases in Hi-C data**

Approach	Model assumption*	Implementation‡	Computational speed	Refs
Yaffe and Tanay	Three systematic biases	Perl and R	Slow	77
HiCNorm	Three systematic biases	R	Fast	78
ICE	Equal visibility	Python	Fast	79
Knight and Ruiz	Equal visibility	JAVA	Fast	27
HiC-Pro	Equal visibility	Python and R	Very fast	80

ICE, iterative correction and eigenvector decomposition. *Model assumption refers to the inherent assumptions in the computational model used to account for bias in Hi-C data. These approaches can be classified based on their model assumptions: they are either explicit, assuming that systematic biases are known (three systematic biases), or implicit, assuming systemic biases are unknown and all the bias is captured by the sequencing coverage of each bin (equal visibility). ‡Implementation refers to the programming language in which the normalization programme is written.

As discussed above, Capture-HiC technologies measure chromatin conformation at target loci at high resolution[54–59]. Thus, in addition to the systemic experimental biases already present in Hi-C data, Capture-HiC data contain additional biases, owing to uneven capture efficiency at targeted loci as well as to some capture bias generated when both interacting sequences are targeted by capture probes (compared to when just one end is being probed), which manifests as sequence coverage variability at each locus[54,56,85]. To specifically account for such coverage asymmetry in Capture-HiC data, the CHiCAGO (Capture-HiC analysis of genomic organization) algorithm was developed; this estimates the bait-specific bias and the other-end-specific bias separately[85]. Moreover, it estimates the bait-specific bias by grouping the probed loci with similar local interacting profiles, whereas the other-end-specific bias is estimated by grouping the non-probed loci with similar distal interacting profiles. More studies are needed to fully explore the combinatorial effect of the bait-specific bias and the other-end-specific bias in Capture-HiC data.

Although several methods to account for experimental bias are available (TABLE 3), they should be used with great caution. The validity of each approach depends heavily on its explicit or implicit model assumptions. The explicit approaches assume that the systematic biases are known and taken into account in the statistical model to account for inherent biases in the observed Hi-C contact matrix. These methods can be overly conservative and run the risk of missing additional sample-specific biases whereby the normalized Hi-C data may still be affected by unknown biases, namely biases not taken into account in the explicit model. For example, DNA-circularization bias[86] is not accounted for in the current explicit approaches. By contrast, the matrix-balancing approaches rely on the equal visibility assumption: that each locus throughout the genome has equal likelihood of being engaged in a 3D contact captured by the Hi-C protocol. Therefore, matrix-balancing algorithms assume that, after removing all biases, the normalized Hi-C contact matrix should have constant row (and column) summation. If these row summations are scaled to one, then each matrix entry represents an approximate contact probability between two loci, whereas following bias removal from explicit models, the matrix entries represent normalized contact counts. The equal visibility assumption may seem intuitive, as Hi-C is indeed a genome-wide sequencing technique and approximately equal coverage across the genome may be expected. However, there are many biases that are known to affect read coverage in Hi-C data, such as the restriction cut site position and the mappability and GC content of sequences flanking the restriction enzyme cut sites[77]. Moreover, it has also been appreciated that the restriction enzyme used in library preparation is biased towards cutting at open chromatin regions[49]. These experimental biases, some of which are unique to Hi-C and do not exist for other whole-genome sequencing library preparation methods, will clearly bias the Hi-C sequencing coverage; therefore, matrix-balancing assumes that the cumulative effect of all bias factors is captured in the coverage of each locus. Coverage distribution across bins in Hi-C data is Gaussian (continuous), with several bins having absolutely no coverage, owing to poorly annotated sequence content, lack of restriction enzyme cut sites or other known experimental biases. In general, the bins with no observed coverage are ignored during matrix-balancing. However, bins with very poor coverage can sometimes be corrected by orders of magnitude to have balanced coverage compared with the rest of the genome. Coverage of conventional whole-genome sequencing data is also not perfectly even, so the justification to balance coverage in Hi-C data is imperfect. Finally, as Hi-C data sets seem to be rapidly moving towards high-resolution analyses, it remains unclear which bias assumptions are more appropriate at smaller bin sizes compared with the larger bin sizes that have until recently predominated in the analyses of Hi-C data. Given the limitations of both explicit approaches and matrix-balancing approaches, we recommend that users conduct careful quality control and experimental validation for the normalized Hi-C data sets. In addition, to ensure reproducibility, it is desirable to compare the normalized results from multiple biological replicates and from different computational approaches. It is also good practice to conduct Hi-C data analyses using both types of bias-removal approaches, as this eliminates the possibility of making a discovery that is dependent on the type of bias-removal method.

Resolution of C-data. To study chromosomal spatial organization, the resolution at which to examine the data needs to be determined. As mentioned above, the resolution of a Hi-C experiment is often conveyed as the size of the genomic loci (or bins) used to compute the meaningful chromatin contacts between pairs of genomic loci[19,27]. To determine the correct resolution, it must first be appreciated that the linear increase of resolution requires a quadratic increase in total sequencing depth. For example, the first Hi-C study collected 8.7 million reads to study the human genome at 1 Mb and 100 kb resolutions[22]. The highest resolution Hi-C maps to date collected over 4.9 billion reads to study the human genome at 1 kb resolution[27], demonstrating a 3-orders of magnitude increase in sequencing depth for a 2-orders of magnitude increase in resolution. Noticeably, the

Bait-specific bias
An experimental bias in the Capture-HiC procedure, referring to the unequal probability of probe hybridization to the target sequence as a result of variable sequence content and hybridization properties.

Other-end-specific bias
An experimental bias in the Capture-HiC procedure, referring to the unequal probability of ligation between the bait locus and its interacting restriction fragment as a result of variable local genomic features.

Principal component analysis
(PCA). A statistical approach for multivariate data analysis. PCA converts a set of correlated variables into a set of linearly uncorrelated variables named principal components, each of which is a linear combination of the original correlated variables.

First eigenvector
The coefficients of the linear combination in the first principle component, which has the largest variance among all principal components. In Hi-C data analysis, the sign of the first eigenvector was used to determinate the A and B compartments.

linear genomic distance between two interacting loci is also a key factor required to determine the appropriate resolution. Because Hi-C contact frequency dramatically decreases as the linear genomic distance increases, in practice, only interactions within a certain range of linear genomic distance are considered. For example, a recent study analysed 5–10-kb-resolution Hi-C data for pairwise interactions within a linear genomic distance of 2 Mb[19].

Despite these general principles of resolution, researchers must still arbitrarily select the bin size for which to analyse their Hi-C data, and definitive guidelines for appropriate bin size determination are lacking. Most available approaches for determining bin size are heuristic and difficult to transfer to other experimental settings. For example, resolution has been defined in one study as the smallest bin size for which more than 1,000 valid chromatin contacts can be observed in at least 80% of the bins[27]. Although this lays out a quantitative criterion, it lacks clear theoretical and experimental justification. It may be argued that the resolution of Hi-C data should be determined by the specific biological questions at hand and interpreted from a statistical perspective. For example, suppose the computational task is the detection of enhancer–promoter interactions. First, a set of experimentally validated interacting loci (true positives) and a set of random collisions (true negatives) must be collected; then, the strength (frequency) of chromatin contacts for both must be quantified. The difference in the distribution of chromatin interaction frequency between the true positives and true negatives can then be used to calculate the total sequencing depth that is required to justify the statistical validity of the pre-specified sensitivity and specificity. Such statistically based power analyses and careful experimental design will help to determine the optimal resolution of a specific Hi-C data set and to facilitate appropriate biological interpretation and discovery.

Analyses of features of global chromatin conformation. The development of the Hi-C technique enabled the characterization of global features of chromatin organization (TABLE 4), leading to the discovery of compartmentalization of chromosome folding within the nucleus[22]. Genomic regions at two distinct nuclear compartments, arbitrarily labelled compartment A and compartment B, display high contact frequency within the same compartment and low contact frequency between the compartments. Compartment A roughly corresponds to the euchromatin and features higher gene density, whereas compartment B corresponds to the heterochromatin and is largely made up of gene deserts. Compartment B is also closely correlated with lamina-associated domains (LADs). Interestingly, this large-scale genome compartmentalization is highly dynamic during the differentiation of human embryonic stem cells[62] and between normal and cancer cells[22], suggesting compartmentalization has a crucial role in mediating genome function and cell identity.

Principal component analysis (PCA) on intra- or inter-chromosomal Hi-C contact maps can be applied to designate compartments A and B[22,27]. More specifically, the sign of the first eigenvector determines the compartment label. Although PCA is easy to implement and has straightforward interpretation, it has two major caveats. First, for some chromosomes, the sign of the first eigenvector represents the short and long chromosome arms, rather than the typical A and B patterns observed in most other chromosomes. In this case, the sign of the second eigenvector should be used to determine the compartment designation. Second, the sign of the first eigenvector is an arbitrary identification method. Without additional information, the compartment cannot be determined. In practice, regions with high gene density can be assigned as compartment A, and regions with low gene density as compartment B.

In general, each compartment is continuous and several megabases in size, reflecting relatively large-scale chromatin architecture. In addition, recent Hi-C analysis at high resolution discovered that sub-compartments, which are distinct compartments within the conventional A and B compartments, may exist; these span smaller genomic regions and correlate with the underlying chromatin biochemical activity[27]. Higher resolution Hi-C or 5C studies revealed that compartments consist of TADs[6,10]. In mammals, TADs are approximately 1 Mb in size, conserved across cell types and species, and may serve as the basic unit of genome structure and function. A more comprehensive discussion of the structure and function of TADs can be found in a recent review[87].

Table 4 | **Approaches for the analysis of global chromatin conformation**

Approach	Objective	Pros	Cons	Refs
PCA	Detect nuclear compartments	Easy to implement; straightforward interpretation	First eigenvector may not work; arbitrary compartment assigning	22
DI/HMM	Detect TADs	Model the change of upstream and downstream interaction bias	Heuristic tuning parameters	6
Arrowhead	Detect TADs	High computational efficiency with dynamic programming	Heuristic tuning parameters	27
Insulation score	Detect TADs	Robust to different sequencing depth; can detect dynamics of TAD boundaries	Heuristic tuning parameters	90
Armatus	Detect TADs	TAD calling robust in different resolutions	Fails to provide uncertainty in TAD calling	88
HiCseg	Detect TADs	Models the uncertainty in Hi-C data	Fails to detect multi-level TADs	89

DI, directionality index; HMM, hidden Markov model; PCA, principle component analysis; TAD, topologically associating domain.

Table 5 | **Approaches for chromatin contact peak-calling**

Approach	Assumption on background model	Pros	Cons	Refs
Jin *et al.*	Global background	Models contact-frequency uncertainty as a negative binomial distribution	Variability of local chromatin organization may introduce biases	19
Fit-Hi-C	Global background	Accurate background model using non-parametric spline	Variability of local chromatin organization may introduce biases	91
GOTHiC	Global background	Models contact-frequency uncertainty as binomial distribution	Variability of local chromatin organization may introduce biases	54
HiCCUPS	Local background	Designed for high-resolution Hi-C data	Deep sequencing is required	27
HMRF	Global or local background	Models spatial dependency among adjacent, interacting loci	High computation cost	93

GOTHiC, genome organisation through HiC; HiCCUPS, Hi-C computational unbiased peak search; HMRF, hidden Markov random field.

Developing computational approaches for detecting TADs is an active research area (TABLE 4). The first published approach was based on a hidden Markov model (HMM)[6]. For each given bin, the total number of interactions located 2 Mb upstream and 2 Mb downstream were calculated and quantified in a metric termed the directionality index. It was assumed that the total number of upstream and downstream interactions are comparable at the centre of TADs but are highly imbalanced at bins adjacent to TAD boundary regions. Based on such an assumption, an HMM was used to capture the sharp transition from the upstream interaction bias to the downstream interaction bias at the TAD boundary regions, which is a distinctive signature of two spatially separate, self-interacting domains. Later on, the Arrowhead algorithm was used to annotate contact domains genome-wide[27]. Dynamic programming was used to ensure efficient implementation of the Arrowhead algorithm to the high-resolution Hi-C data. Meanwhile, the Armatus algorithm was developed for detecting consistent TAD patterns at different resolutions[88]. In addition, the HiCseg algorithm can narrow down the problem of annotating TADs from 2D image segmentation to linear (1D) segmentation[89]. Similarly, a sliding insulation score approach was recently introduced that also transforms the Hi-C contact matrix into an intuitive 1D insulation score vector[90]. This approach has been demonstrated to detect dynamics of TAD boundary strength in different experimental conditions[90]. Importantly, most of these approaches rely on heuristic tuning parameters, such as the threshold on the maximal linear genomic distance between two interacting loci when computing the directionality index, which is a measure of orientation biases in chromatin interactions originating from a genomic locus, or the window size for computing insulation, which is a measure of interaction permissibility across a genomic locus. Currently, we suggest researchers try different tuning parameters and visually check the TAD coordinates alongside the Hi-C contact matrix to ensure the validity and reproducibility of TAD-calling results. It is also likely that the hierarchical level of genome organization that can be detected is affected by the tuning parameters. For example, smaller insulation windows or small directionality index windows are more capable of detecting smaller scale chromatin folding structures compared with larger windows.

A key challenge in the analysis of global chromatin conformation lies in the fact that the genome is folded into multiple hierarchical structures, from compartments to TADs, nested sub-TAD structures and individual chromatin loops. Understanding the principles underlying this hierarchical chromosome organization requires the development of novel computational approaches. An excellent review[41] highlights the recent computational advance in the analysis of global chromosome organization.

Analyses of local features of chromosome conformation. As a result of ever-increasing DNA sequencing throughput and decreasing sequencing cost, high-resolution Hi-C data sets are attainable and have enabled the analysis of chromatin contacts at nearly kilobase resolution. As this resolution is nearly the size of individual *cis*-regulatory elements, high-resolution Hi-C data sets can be interrogated for fine-mapping of long-range *cis*-regulatory interactions and provide novel insights on transcription regulation mechanisms. To that end, many computational approaches have been developed for detecting biologically meaningful long-range chromatin contacts, which is a process termed peak-calling (TABLE 5). In pioneering work, chromatin contact frequencies obtained from Hi-C data were modelled as a negative binomial distribution and a global background model was devised that consists of both systematic bias factors and the linear genomic distance factor[19]. The Fit-HiC algorithm uses a non-parametric spline approach to model the background-chromatin contact frequency[91]. Both methods take advantage of a global background model in which the expected interaction frequency of a given pair of loci follows the trend derived from genome-wide contact frequencies at a given linear genomic distance. In both methods, peak-calling led to millions of statistically significant chromatin contacts; however, by using the global background model, this approach may over-estimate chromatin interactions, leading to false positives. Meanwhile, the GOTHiC (genome organisation through HiC) algorithm uses a simple binomial distribution model to simultaneously remove biases in Hi-C data and detect significant interactions by assuming that the global background interaction frequency of two loci depends also on the relative genome-wide

Hidden Markov model (HMM). A statistical model assuming that the observed data are determined by a set of unobserved (hidden) states with the Markov property: the future state depends on only the current state and is independent of all the previous states.

Heuristic tuning parameters The parameters in the statistical models and computational pipelines that are not estimated from the observed data but are determined based on prior knowledge and expectation.

Global background model The statistical model for the expected chromatin contact frequency estimated from genome-wide measurements. It is used to systematically identify significant pairwise Hi-C interactions throughout the genome. All interacting loci pairs at a given linear distance share the same global background model.

Non-parametric spline A statistical approach to fit the observed data using a piecewise-defined polynomial function.

coverage[54,92]. Another feature of GOTHiC is that it implemented the Benjamini–Hochberg multiple-testing correction to control for the false discovery rate. By applying this method to a Hi-C data set from mouse cells, ~90,000 statistically significant interactions could be identified[92]. By contrast, HiCCUPS (Hi-C computational unbiased peak search) uses a local background model and has been applied to detect chromatin loops in several human and mouse cell lines at 1 kb or 5 kb resolution from *in situ* Hi-C data[27]. HiCCUPS identified around ~2,500–10,000 chromatin loops, depending on the resolution of the data set. Recently, the computational problem of detecting significant chromatin interactions was tackled from a different angle, by assuming that the background model (either a global background or a local background) is known and by developing a hidden Markov random field (HMRF) algorithm to model the spatial dependency among neighbourhood interacting loci[93] (TABLE 5). In other words, the dependency implies that, if two loci are inferred to be spatially proximal based on Hi-C data, then all the neighbouring loci will have a higher probability of interacting. The HMRF algorithm can achieve higher reproducibility and improves statistical power, especially for the analysis of pairwise contacts in high-resolution Hi-C data. In the future, it would be of great interest to compare the interaction frequency at these identified peaks, as well as other loci, among different experimental conditions and biological contexts. A software package named diffHiC[94] was recently developed to detect dynamic chromatin interactions across experimental conditions or cell types. Using the same statistical framework of the edgeR (empirical analysis of DGE in R) package[95], which has achieved great success in detecting differentially expressed genes in RNA-seq data, diffHiC has the potential to become a powerful tool for differential-interaction analysis.

Capture-HiC shows great promise in the detection of chromatin interactions at loci of interest[29,49,54–60]. The computational methods for the analysis of Capture-HiC data are still under development. One study used a heuristic observed read-count cut-off in identifying significant interactions, but this lacks solid statistical justification[55]. Later on, a statistical model based on a convolution of negative binomial and Poisson distributions was proposed to account for background distribution in the Capture-HiC data[85]. As Capture-HiC technology becomes more popular, novel computational methods will be developed to better-characterize its data.

Several key issues need to be considered with the above peak-calling approaches. First, whether to use a global background model or a local background model is still under debate. Unlike peak-calling in ChIP-seq data analysis, in which input DNA is frequently used as control, it is unclear how to characterize the random collision frequency between chromatin loci. Second, to detect biologically meaningful chromatin interactions, such as those between individual *cis*-regulatory elements, a great number of candidate loci needs to be considered when statistically determining if any two loci of interest are interacting more frequently than expected. In practice, this imposes a challenging multiple-comparisons problem, which requires highly intensive computation and rigorous statistical justification. Third, biologically meaningful, long-range chromatin contacts are spatially and temporally dynamic. Without a 'gold standard' of true-positive and true-negative chromatin contacts, it is difficult to fully evaluate the sensitivity and specificity of each approach. Moreover, to address biological hypotheses, it is important to conduct targeted analyses across different cell types to identify cell-type-specific chromatin contacts[94]. It is just as important to closely examine cell-type-common chromatin contacts where cell-type-specific enhancer activation is observed, as these may be controlled by different transcription-factor-binding events, rather than by differential chromatin looping[27]. However, the careful evaluation of technical variability and biological variability of chromatin interaction frequency as well as the comprehensive experimental validation of cell-type-specific chromatin interactions are still lacking. We envision that further advancement in both experimental technologies and computational algorithms for the targeted analysis of chromatin conformation will occur in the near future.

Future perspectives

Although C-technologies have been increasingly used, current experimental protocols have some significant limitations that could prevent the uncovering of additional chromatin organization features. First, common methods produce only static molecular interaction maps that overlook the temporal dynamics of chromatin in live cells and disregard cell-to-cell variability in a population, potentially leading to incorrect models of chromatin organization. Second, current maps of chromatin interaction still lack the fine resolution needed to resolve interactions between individual *cis*-regulatory elements, greatly limiting our ability to interrogate the functional roles of chromatin structure in gene regulation. Third, current methods for mapping chromatin interactions permit the efficient mapping of only pairwise interactions, thus failing to detect potential multi-way interaction hubs that are suspected to exist in the nucleus. Last, with various different techniques for mapping and analysing chromatin topology, a critical comparison of these methods is greatly needed.

The recently launched 4D Nucleome Project will address these challenges through a multi-pronged approach. In particular, new data standards for assessing different experimental protocols and data analysis methods will be developed. Such standards could include pairs of DNA loci for which chromatin interactions have been rigorously assessed genetically, biochemically and by using microscopy imaging. This US National Institutes of Health (NIH) common fund initiative is also expected to develop improved methods for generating high-resolution chromatin-interaction maps, through a combination of substantial optimization and improvement of experimental protocols, innovative algorithms for data analysis and structural modelling. New methods for determining chromosome organization in small numbers of cells or even

Benjamini–Hochberg multiple-testing correction
A statistical procedure that uses stringent statistical significance thresholds to control the false discovery rate when performing multiple comparisons.

Local background model
The statistical model for the expected chromatin contact frequency estimated from local chromatin interaction properties. Each pair of interacting loci has a unique local background model, which depends on the definition of its local neighbouring regions.

single cells will be developed, along with methods that generate complementary views of genome organization without fixation, restriction digestion or ligation.

Live-cell imaging tools and analysis approaches are needed that can accurately inform on dynamic chromatin organization both within and between TADs. Multicolour live-cell 3D imaging tools will be particularly useful for studying chromatin motion in live cells. The results of such experiments could uncover the basic principles governing dynamic chromatin organization at various scales in mammalian cells and help to interpret the contact probability data obtained from C-technologies.

Finally, to achieve a thorough understanding of the structural and functional role of chromatin organization in transcription regulation, 3D chromatin organization data sets will need to be integrated with other genomic and epigenomic data sets in a wide range of cell types and tissues, such as those produced by large-scale consortia like the NIH Encyclopedia of DNA Elements (ENCODE) project, Roadmap Epigenome project and the International Human Epigenome Consortium. This will result in improved knowledge of the functional relationships between chromatin organization and genome function.

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
3. Bickmore, W. A. & van Steensel, B. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–1284 (2013).
4. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).
5. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, 11–24 (2012).
6. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
The original study to describe TADs from Hi-C analysis, using novel computation approaches. It discovered that TADs are conserved between cell types and species, and demarcated by CCCTC-binding factor (CTCF) binding at TAD boundaries.
7. Gorkin, D. U., Leung, D. & Ren, B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell* **14**, 762–775 (2014).
8. Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13–25 (2014).
9. Nora, E. P., Dekker, J. & Heard, E. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? *Bioessays* **35**, 818–828 (2013).
10. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
11. Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).
12. Deng, W. *et al.* Reactivation of developmentally silenced globin genes by forced chromatin looping. *Cell* **158**, 849–860 (2014).
13. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233–1244 (2012).
14. Kim, A. & Dean, A. Chromatin loop formation in the β-globin locus and its role in globin gene transcription. *Mol. Cells* **34**, 1–5 (2012).
15. Krivega, I. & Dean, A. Enhancer and promoter interactions-long distance calls. *Curr. Opin. Genet. Dev.* **22**, 79–85 (2012).
16. Plank, J. L. & Dean, A. Enhancer function: mechanistic and genome-wide insights come together. *Mol. Cell* **55**, 5–14 (2014).
17. Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374–387 (2014).
18. Heidari, N. *et al.* Genome-wide map of regulatory interactions in the human genome. *Genome Res.* **24**, 1905–1917 (2014).
19. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
The first paper to report Hi-C interaction maps at the resolution of individual restriction fragments in mammals. This study also introduced the global background model.
20. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).
21. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
22. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
The original study describing Hi-C technology. This study was also the first to describe the genome compartments A and B, which respectively mark colocalizing active and repressed regions of the genome.
23. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
24. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2012).
25. Hughes, J. R. *et al.* Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
26. Kolovos, P. *et al.* Targeted chromatin capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements. *Epigenetics Chromatin* **7**, 10 (2014).
27. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
The highest-resolution Hi-C analysis to date, at 1–5kb resolution in 9 human and mouse cell types. This study reports that the genome is organized globally into 6 sub-compartments, within which the genome is organized into ~10,000 chromatin loops, many of which are conserved across species and cell types, and are anchored by CTCF binding in convergent orientation.
28. Selvaraj, S., R. Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.* **31**, 1111–1118 (2013).
29. Selvaraj, S., Schmitt, A. D., Dixon, J. R. & Ren, B. Complete haplotype phasing of the MHC and KIR loci with targeted HaploSeq. *BMC Genomics* **16**, 900 (2015).
30. de Vree, P. J. *et al.* Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat. Biotechnol.* **32**, 1019–1025 (2014).
31. Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
32. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).
33. Marie-Nelly, H. *et al.* High-quality genome (re)assembly using chromosomal contact data. *Nat. Commun.* **5**, 5695 (2014).
34. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).
35. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)* **4**, 1339–1346 (2014).
36. Marbouty, M. *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife* **3**, e03318 (2014).
37. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
38. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59–64 (2013).
39. Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16**, 344–358 (2015).
40. Flot, J. F., Marie-Nelly, H. & Koszul, R. Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. *FEBS Lett.* **589**, 2966–2974 (2015).
41. Imakaev, M. V., Fudenberg, G. & Mirny, L. A. Modeling chromosomes: beyond pretty pictures. *FEBS Lett.* **589**, 3031–3036 (2015).
42. Serra, F. *et al.* Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.* **589**, 2987–2995 (2015).
43. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
The original study describing 3C technology.
44. Cullen, K. E., Kladde, M. P. & Seyfred, M. A. Interaction between transcription regulatory regions of prolactin chromatin. *Science* **261**, 203–206 (1993).
45. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).
A study reporting chromosome conformation capture-on-chip (4C), which explores the genome-wide interactions of individual loci at high resolution.
46. van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods* **9**, 969–972 (2012).
47. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
48. Deng, X. *et al.* Bipartite structure of the inactive mouse X chromosome. *Genome Biol.* **16**, 152 (2015).
49. Ma, W. *et al.* Fine-scale chromatin interaction maps reveal the *cis*-regulatory landscape of human lincRNA genes. *Nat. Methods* **12**, 71–78 (2015).
The first study to report the use of DNase Hi-C and DNase Capture-HiC, and the first application of Capture-HiC to specifically enrich for gene promoters.
50. Hsieh, T. H. *et al.* Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* **162**, 108–119 (2015).
51. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
Another study reporting chromosome conformation capture-on-chip (4C), which explores the genome-wide interactions of individual loci at high resolution.
52. Dostie, J. *et al.* Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
The original study describing 5C, which explores the interaction profiles of several contiguous loci with each other at high resolution.

53. Martin, P. *et al.* Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* **6**, 10069 (2015).
54. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
55. Sahlen, P. *et al.* Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol.* **16**, 156 (2015).
56. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
The first application of Capture-HiC to capture all promoters in the genome, demonstrating the feasibility and quality of obtaining high-resolution promoter interaction profiles for >20,000 loci in a single assay.
57. Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat. Genet.* **47**, 1179–1186 (2015).
58. Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.* **24**, 1854–1868 (2014).
The original study describing Capture-HiC technology and its use to interrogate the interaction landscapes of several disease-associated risk loci.
59. Jager, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6**, 6178 (2015).
60. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl Acad. Sci. USA* **112**, E6456–E6465 (2015).
61. Schmitt, M. W. *et al.* Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat. Methods* **12**, 423–425 (2015).
62. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
A high-resolution Hi-C analysis in human embryonic stem cells and four derived cell types, revealing a relationship between dynamic chromatin organization and gene expression, as well as haplotype-resolved dynamics in chromatin organization patterns.
63. Fraser, J. *et al.* Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* **11**, 852 (2015).
64. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350–354 (2015).
65. Nagano, T. *et al.* Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biol.* **16**, 175 (2015).
66. Seitan, V. C. *et al.* Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res.* **23**, 2066–2077 (2013).
67. Sofueva, S. *et al.* Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J.* **32**, 3119–3129 (2013).
68. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl Acad. Sci. USA* **111**, 996–1001 (2014).
69. Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
70. van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **39**, 1869 (2010).
71. Comet, I., Schuettengruber, B., Sexton, T. & Cavalli, G. A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. *Proc. Natl Acad. Sci. USA* **108**, 2294–2299 (2011).
72. van de Werken, H. J. *et al.* 4C technology: protocols and data analysis. *Methods Enzymol.* **513**, 89–112 (2012).
73. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density *in vitro* transposition. *Genome Biol.* **11**, R119 (2010).
74. Williamson, I. *et al.* Anterior-posterior differences in HoxD chromatin topology in limb development. *Development* **139**, 3157–3167 (2012).
75. Bickmore, W. A. The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.* **14**, 67–84 (2013).
76. Williamson, I. *et al.* Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence *in situ* hybridization. *Genes Dev.* **28**, 2778–2791 (2014).
77. Yaffe, E. & Tanay, A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* **43**, 1059–1065 (2011).
78. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics* **28**, 3131–3133 (2012).
79. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
80. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
81. Li, W., Gong, K., Li, Q., Alber, F. & Zhou, X. J. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics* **31**, 960–962 (2015).
82. Knopp, P. & Sinkhorn, R. Concerning nonnegative matrices and doubly stochastic matrices. *Pacif. J. Math.* **21**, 343–348 (1967).
83. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Analysis* **33**, 1029–1047 (2012).
84. Shavit, Y. & Lio, P. Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. Biosyst.* **10**, 1576–1585 (2014).
85. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in capture Hi-C data. *Genome Biol.* **17**, 127 (2015).
86. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).
87. Dekker, J. & Heard, E. Structural and functional diversity of topologically associating domains. *FEBS Lett.* **589**, 2877–2884 (2015).
88. Filippova, D., Patro, R., Duggal, G. & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.* **9**, 14 (2014).
89. Levy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**, i386–i392 (2014).
90. Crane, E. *et al.* Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240–244 (2015).
91. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* **24**, 999–1011 (2014).
92. Mifsud, B. *et al.* GOTHiC, a simple probabilistic model to resolve complex biases and to identify real interactions in Hi-C data. Preprint at *bioRxiv* http://dx.doi.org/10.1101/023317 (2015).
93. Xu, Z. *et al.* A hidden Markov random field based Bayesian method for the detection of long-range chromosomal intereactions in Hi-C data. *Bioinformatics* **32**, 650–656 (2015).
94. Lun, A. T. & Smyth, G. K. diffHic: a bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**, 258 (2015).
95. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
96. Nagano, T. *et al.* Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. *Nat. Protoc.* **10**, 1986–2003 (2015).
97. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat. Methods* **3**, 17–21 (2006).
98. Hagege, H. *et al.* Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.* **2**, 1722–1733 (2007).
99. Louwers, M., Splinter, E., van Driel, R., de Laat, W. & Stam, M. Studying physical chromatin interactions in plants using chromosome conformation capture (3C). *Nat. Protoc.* **4**, 1216–1229 (2009).
100. Naumova, N., Smith, E. M., Zhan, Y. & Dekker, J. Analysis of long-range chromatin interactions using chromosome conformation capture. *Methods* **58**, 192–203 (2012).
101. Ribeiro de Almeida, C. *et al.* The DNA-binding protein CTCF limits proximal Vκ recombination and restricts κ enhancer interactions to the immunoglobulin κ light chain locus. *Immunity* **35**, 501–513 (2011).
102. Stadhouders, R. *et al.* Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat. Protoc.* **8**, 509–524 (2013).
103. Wurtele, H. & Chartrand, P. Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended chromosome conformation capture methodology. *Chromosome Res.* **14**, 477–495 (2006).
104. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* **470**, 264–268 (2011).
105. Gondor, A., Rougier, C. & Ohlsson, R. High-resolution circular chromosome conformation capture assay. *Nat. Protoc.* **3**, 303–313 (2008).
106. Splinter, E. *et al.* The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on *Xist* RNA. *Genes Dev.* **25**, 1371–1383 (2011).
107. Gheldof, N., Leleu, M., Noordermeer, D., Rougemont, J. & Reymond, A. Detecting long-range chromatin interactions using the chromosome conformation capture sequencing (4C-seq) method. *Methods Mol. Biol.* **786**, 211–225 (2012).
108. Splinter, E., de Wit, E., van de Werken, H. J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods* **58**, 221–230 (2012).
109. Schoenfelder, S. *et al.* Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.* **42**, 53–61 (2010).
110. Sexton, T. *et al.* Sensitive detection of chromatin coassociations using enhanced chromosome conformation capture on chip. *Nat. Protoc.* **7**, 1335–1350 (2012).
111. Ling, J. Q. *et al.* CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. *Science* **312**, 269–272 (2006).
112. Ling, J. & Hoffman, A. R. Associated chromosome trap for identifying long-range DNA interactions. *J. Vis. Exp.* **50**, 2621 (2011).
113. Dostie, J., Zhan, Y. & Dekker, J. Chromosome conformation capture carbon copy technology. *Curr. Protoc. Mol. Biol.* http://dx.doi.org/10.1002/0471142727.mb2114s80 (2007).
114. Ferraiuolo, M. A., Sanyal, A., Naumova, N., Dekker, J. & Dostie, J. From cells to chromatin: capturing snapshots of genome organization with 5C technology. *Methods* **58**, 255–267 (2012).
115. Fraser, J., Ethier, S. D., Miura, H. & Dostie, J. A. Torrent of data: mapping chromatin organization using 5C and high-throughput sequencing. *Methods Enzymol.* **513**, 113–141 (2012).
116. Umbarger, M. A. Chromosome conformation capture assays in bacteria. *Methods* **58**, 212–220 (2012).
117. Rodley, C. D., Bertels, F., Jones, B. & O'Sullivan, J. M. Global identification of yeast chromosome interactions using genome conformation capture. *Fungal Genet. Biol.* **46**, 879–886 (2009).
118. Duan, Z. *et al.* A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes. *Methods* **58**, 277–288 (2012).
119. Tanizawa, H. *et al.* Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res.* **38**, 8164–8177 (2010).

Acknowledgements
The authors dedicate this manuscript in loving memory of Joseph Schmitt. They would like to give special thanks to members of the Ren laboratory for their suggestions. This work is supported by the Ludwig Institute for Cancer Research, La Jolla, California, USA, and grants from US National Institutes of Health (NIH; grant U54DK107977 to B.R. and M.H., and grants U54 HG006997 and R01 ES024984 to B.R.). A.D.S. is supported by NIH genetics training grant T32 GM008666.

Competing interests statement
The authors declare competing interests: see Web version for details.

FURTHER INFORMATION
4C protocol variant: http://www.nature.com/protocolexchange/protocols/1979
4D Nucleome Project: http://www.4dnucl.org; https://commonfund.nih.gov/4dnucleome/index
ALL LINKS ARE ACTIVE IN THE ONLINE PDF

Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in *Nature Reviews Molecular Cell Biology*. Schmitt, Anthony D.; Hu, Ming; Ren, Bing. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Selvaraj et al. BMC Genomics (2015) 16:900
DOI 10.1186/s12864-015-1949-7

BMC Genomics

METHODOLOGY ARTICLE Open Access

Complete haplotype phasing of the MHC and KIR loci with targeted HaploSeq

Siddarth Selvaraj[1†], Anthony D. Schmitt[1,2†], Jesse R. Dixon[1,3] and Bing Ren[1,4,5*]

Abstract

Background: The MHC and KIR loci are clinically relevant regions of the genome. Typing the sequence of these loci has a wide range of applications including organ transplantation, drug discovery, pharmacogenomics and furthering fundamental research in immune genetics. Rapid advances in biochemical and next-generation sequencing (NGS) technologies have enabled several strategies for precise genotyping and phasing of candidate HLA alleles. Nonetheless, as typing of candidate HLA alleles alone reveals limited aspects of the genetics of MHC region, it is insufficient for the comprehensive utility of the aforementioned applications. For this reason, we believe phasing the entire MHC and KIR locus onto a single locus-spanning haplotype can be a critical improvement for better understanding transplantation biology.

Results: Generating long-range (>1 Mb) phase information is traditionally very challenging. As proximity-ligation based methods of DNA sequencing preserves chromosome-span phase information, we have utilized this principle to demonstrate its utility towards generating full-length phasing of MHC and KIR loci in human samples. We accurately (~99 %) reconstruct the complete haplotypes for over 90 % of sequence variants (coding and non-coding) within these two loci that collectively span 4-megabases.

Conclusions: By haplotyping a majority of coding and non-coding alleles at the MHC and KIR loci in a single assay, this method has the potential to assist transplantation matching and facilitate investigation of the genetic basis of human immunity and disease.

Keywords: HaploSeq, MHC, HLA-Typing, KIR, Phasing

Background

The major histocompatibility complex (MHC) and the killer cell immunoglobulin-like receptor (KIR) are important regulators of human immune responses and are involved in many human diseases [1, 2]. These loci are highly polymorphic, allowing an extensive antigen-presenting repertoire that enables strong immunity against a wide range of foreign antigens, pathogens and tumor cells [1–3]. At the same time, its immunogenic heterogeneity can also create incompatibility in allotransplantation procedures, causing graft rejections and graft-versus-host disease (GVHD) [4, 5]. Furthermore, many of the hundreds of genes within these immunogenic loci are increasingly recognized as major susceptibility genes for drug hypersensitivity reactions and appear to play a significant role in numerous diseases, including cancer [6–8]. Taken together, the clinical implications of these loci make it useful to determine the sequence type of these molecules.

Typing of human leukocyte antigen (HLA) genes, located within the MHC locus, has traditionally been achieved in low resolution using serotyping techniques [9]. With advancements in technologies including PCR and more recently, next generation DNA sequencing (NGS), molecular-based methods have now enabled more clinically significant high-resolution HLA typing [10–12]. Notably, single-molecule NGS-based DNA sequencing has been demonstrated to resolve allele ambiguity by generating haplotypes of entire genes, resulting in super high-resolution (8-digit) haplotyping of HLA genes [13, 14]. However, even precise gene-level haplotyping may not be sufficient for many

* Correspondence: biren@ucsd.edu
†Equal contributors
[1]Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA
[4]Department of Cellular and Molecular Medicine, and UCSD Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA
Full list of author information is available at the end of the article

applications. For example, while gene-level haplotyping for several candidate HLA genes can reduce risk of graft failure in transplantation matching, recipients could still be susceptible to graft-versus-host disease, as the totality of transplantation associated genes have not been fully understood. In particular, reports suggest that non-HLA gene families such as inflammatory genes, immune receptors, or others across the MHC or KIR haplotype can contribute to transplantation biology [15–17]. In addition, the strong linkage disequilibrium (LD) patterns across the MHC and KIR loci can allow coordinated functional activities of alleles on the same haplotype, complicating our understanding of transplantation biology [4, 5, 9, 18, 19]. Indeed, knowledge of haplotypes across several HLA genes has been shown to generate improved transplantation outcome predictions [19, 20] and can therefore facilitate determination of novel haplotype patterns for drug discovery and genome-wide association studies [21]. In summary, it appears useful to haplotype the entirety of the MHC and KIR loci to enable better understanding of immune genetics through analyses of compound heterozygous alleles.

Several experimental protocols have been developed to construct long-range haplotypes. Specifically, methods have been developed to generate mega-base-sized haplotypes [22–25], while others can phase the entire chromosome [26–29]. However, the adaptability of these methods to generate user-defined targeted haplotypes is unclear. More recently, Targeted Locus Amplification (TLA) has been developed to accomplish targeted phasing [30], but as the haplotypes from TLA are limited to a few-hundred kilobases, they may not be amenable for phasing large mega-base scale loci such as the MHC. Here, we develop a method, referred to as targeted HaploSeq, to generate full-length complete haplotypes of MHC and KIR loci from a single assay. Specifically, targeted HaploSeq combines the previously published HaploSeq [26] method developed for genome-wide haplotype phasing, with oligo capture and sequencing. As a proof of principle, we have applied targeted HaploSeq to the MHC and KIR loci in human lymphoblastoid cells. We phased over 90 % of the alleles in MHC and KIR loci at an estimated accuracy of ~99 %. To our knowledge, targeted HaploSeq is the first method to phase the MHC and KIR loci into a single haplotype structure. These results establish the utility of targeted HaploSeq for MHC and KIR typing in biomedical research as well as clinical settings.

Results and discussion

Experimental design

In the targeted HaploSeq method, a conventional Hi-C library [31] is generated using HindIII restriction digestion and amplified to obtain suitable material for oligonucleotide probe-based enrichment of the target loci (Fig. 1a). Briefly, based on simulation results (Additional file 1: Fig. S1), we computationally generated the probe sequences, at 4X tiling density, using the SureDesign Software (Agilent Technologies) and targeted the non-repetitive +/− 400 bp regions adjacent to HindIII cut sites over the MHC and KIR loci (Fig. 1b, Additional file 2: Fig. S2a). In addition, to facilitate better phasing of genic regions, we designed probes across the exons within the MHC locus (Fig. 1a).

Next, by performing capture-sequencing [32, 33], we generated targeted HaploSeq data in GM12878 lymphoblastoid cells at 2× whole-genome sequencing depth with 30–50 fold target enrichment across the MHC and KIR loci (Fig. 2a, Additional file 2: Fig. S2b). More than 90 % of probes had at least 5-fold sequence coverage compared to data from virtual probes with an average of ~100 fold enrichment. This highlights the sensitivity of the probes from our targeted HaploSeq protocol. Next, to validate the quality of our targeted HaploSeq data, we compared it to a previously published HaploSeq dataset [26] generated from the same cell line. As HaploSeq utilizes chromatin interaction patterns to reconstruct haplotypes, we compared these between the two datasets and observed a high concordance ($r^2 = 0.8$, Fig. 2b, Additional file 3: Fig. S3a, b). By using haplotype inference from the parent–child trio whole-genome sequencing (WGS) data [34], we examined the fraction of chromatin interactions between the homologous chromosomes (h-trans interactions), whose rarity is critical for accurate *de novo* haplotyping. Similar to HaploSeq, targeted HaploSeq data rarely exhibit h-trans interactions (Additional file 4: Fig. S4a).

Of note, the MHC locus appears to have a higher h-trans ratio in both HaploSeq and targeted HaploSeq datasets, but several lines of evidence suggest that these might be systematic errors from sequencing and analysis protocols. First, reads supporting h-trans interactions are primarily observed in complex regions with high variant density (Additional file 4: Fig. S4b). Second, >85 % of h-trans interactions from targeted HaploSeq dataset originate from the same end of a given paired-end fragment. Lastly, about 95 % of these same-end h-trans interactions are also observed in long-fragment reads (LFR) in previously published Moleculo datasets [25] from the same individual, indicating that a significant fraction of these h-trans interactions could have arisen from incorrect local haplotype inferences from the parent-child trio WGS data (Fig. 2c, d, Additional file 5). Taken together, our targeted HaploSeq data is of high quality and therefore enables accurate analyses of haplotype structures across the MHC and KIR loci.

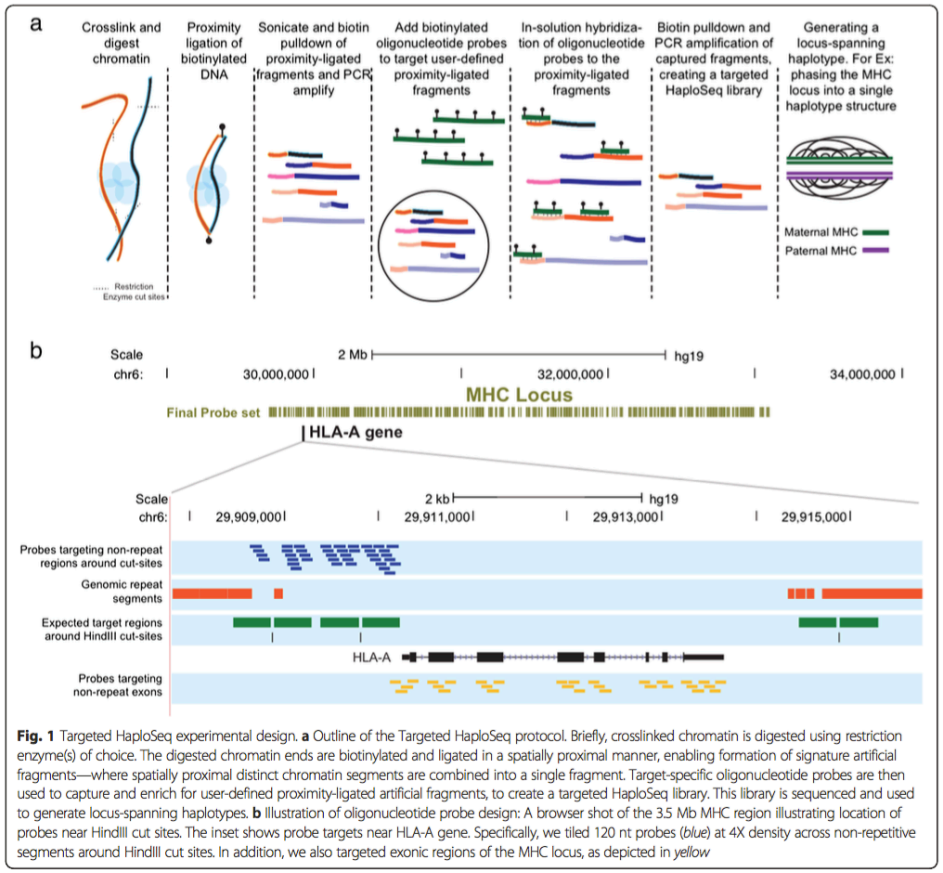


Fig. 1 Targeted HaploSeq experimental design. **a** Outline of the Targeted HaploSeq protocol. Briefly, crosslinked chromatin is digested using restriction enzyme(s) of choice. The digested chromatin ends are biotinylated and ligated in a spatially proximal manner, enabling formation of signature artificial fragments—where spatially proximal distinct chromatin segments are combined into a single fragment. Target-specific oligonucleotide probes are then used to capture and enrich for user-defined proximity-ligated artificial fragments, to create a targeted HaploSeq library. This library is sequenced and used to generate locus-spanning haplotypes. **b** Illustration of oligonucleotide probe design: A browser shot of the 3.5 Mb MHC region illustrating location of probes near HindIII cut sites. The inset shows probe targets near HLA-A gene. Specifically, we tiled 120 nt probes (*blue*) at 4X density across non-repetitive segments around HindIII cut sites. In addition, we also targeted exonic regions of the MHC locus, as depicted in *yellow*

High-resolution and accurate phasing of MHC and KIR loci

By utilizing heterozygous genotype identifications (SNVs) from the trio-based WGS data [34], we used the HaploSeq and LCP protocols to perform *de novo* haplotyping. We generated a single haplotype structure over the MHC locus resolving over 90 % of ~9,400 heterozygous alleles and we used the trio-based haplotype structure to estimate the accuracy of our approach to be ~97.7 % (Additional file 6: Fig. S5). However, as the parent-child trio data could have accumulated incorrect phasing at regions with high variant density, we repeated the *de novo* haplotyping protocol after ignoring variants that we found to be h-trans in both our and LFR datasets. Consequently, our phasing accuracy improved to 98.94 % (Additional file 6: Fig. S5). Despite reducing the phasing error by over 50 %, from 2.3 to 1.06 %, we still observe a majority of phasing errors occurring in the high variant density regions (Fig. 2e). This suggests that the accuracy can potentially be further improved by using long-read or single molecule technologies that may be more suitable for mapping such complex regions. Of note, unlike switch errors—the standard method to calculate phasing error rates where an incorrect haplotype block is penalized only once, we estimate error by testing each variant independently and therefore our error rate represents worst-case scenario. To this end, as the density of variants affects the resolution of HaploSeq-based haplotyping, we observed a relatively lower resolution phasing for the KIR locus (Additional file 1: Fig. S1b). Regardless, we obtained accurate phasing of 348 out of 353 variants resolved at the KIR

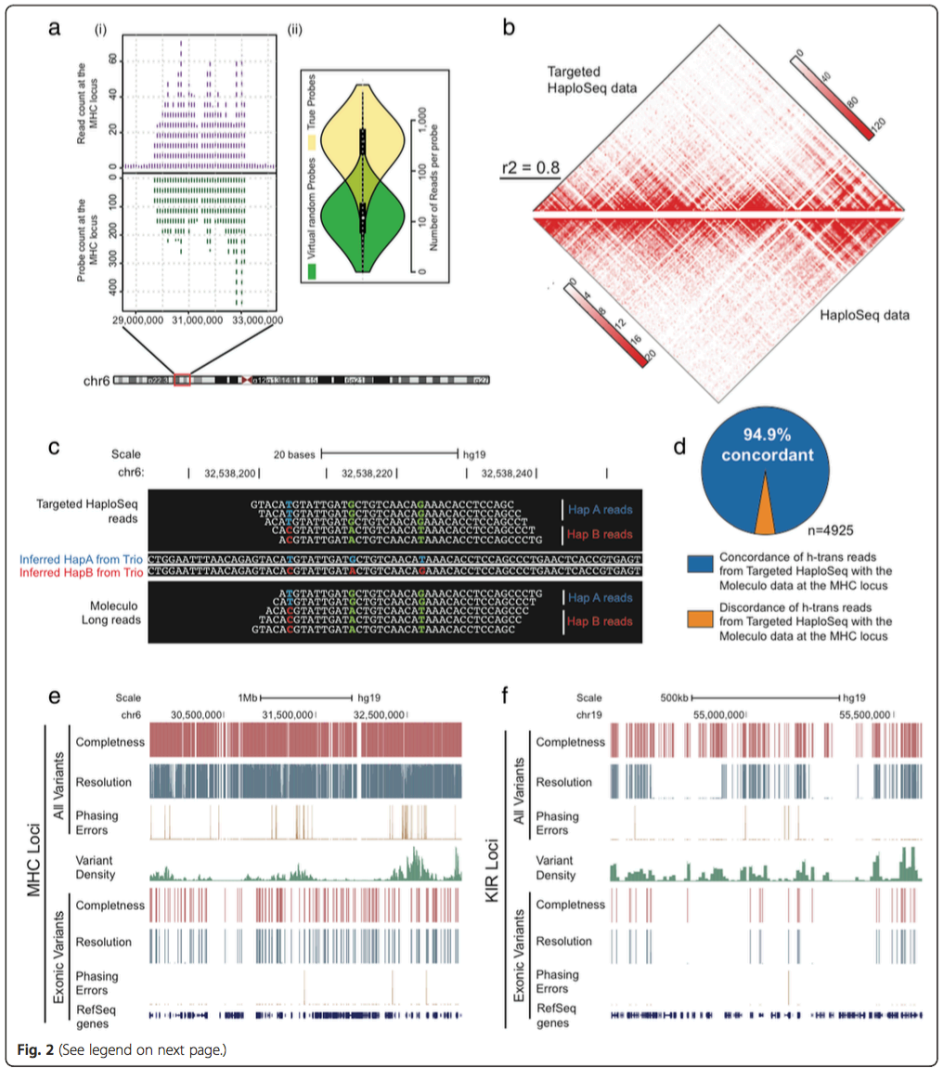


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 High-resolution and accurate phasing of MHC and KIR loci. **a** (i) Top chart demonstrates enrichment of targeted HaploSeq reads at the 100 kb binned MHC locus and the bottom plot shows number of probes in 100 kb bins used across the MHC locus. Visually, we can observe a high correlation between these plots, demonstrating the expected relationship between density of probes and the sequencing depth of targeted HaploSeq reads. (ii) To illustrate the sensitivity of probes, we virtually created random probes flanking HindIII cut sites and compared the enrichment in targeted HaploSeq data from these regions to the data from regions containing true probes. We observe ~100 fold more reads from true regions (on target, *yellow*) than the random regions (off target, *green*) and this fold-enrichment suggests high-sensitivity of our probes. **b** High correlation of targeted HaploSeq and the previously published HaploSeq datasets from GM12878 cells at the MHC locus ($r^2 = 0.8$). **c** An example of haplotype inconsistency in the parent-child trio WGS data. Specifically, HapA (TGT-blue) and HapB (CAG-red) represent two haplotypes inferred from the trio dataset. Single-end reads from targeted HaploSeq (*top*) and Moleculo long-fragment reads (*bottom*) support a case of an inter-haplotype adjacent SNP-pair (*green*) and therefore raises an inconsistency with the parent-child trio haplotype inference. **d** Overall, ~95 % of the targeted HaploSeq reads representing homologous-trans (h-trans) interacting SNVs are concordant with the Moleculo LFR data. **e** High-resolution phasing capabilities of targeted HaploSeq method at the MHC locus. Completeness represents the collection of all heterozygous SNVs (*red*) within the MHC locus. Resolution represents the set of phased or resolved heterozygous SNVs in a single haplotype structure. While we observe ~1 % error, these errors are highly concentrated in the high variant density regions. The bottom section represents phasing of only exonic variants. **f** Similar figure as e) for the KIR locus

loci (Fig. 2f). Together, we resolved ~90 % of alleles among the MHC and KIR loci at ~99 % accuracy (Additional file 4: Fig. S4), demonstrating that our approach can generate complete, high-resolution and accurate haplotypes.

As current HLA typing protocols primarily type candidate genes across the MHC loci, we analyzed our method's phasing capabilities across heterozygous genes from MHC and KIR loci. In total, we resolve ~92 % of heterozygous variants, representing over 92 % of heterozygous genes, at an accuracy of 99.34 % (Fig. 2e, f, Additional file 7: Fig. S6). In this regard, we generate highly accurate phasing for several "classical" genes used in conventional HLA typing protocols. For example, in the case of genes such as HLA-B, HLA-C, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1 and HLA-DPB1, we resolve phasing of >99.5 % of the heterozygous variants at 100 % accuracy. Similarly at the KIR loci, we accurately predict all but one exonic variant (Additional file 7: Fig. S6). To our knowledge, our method is the first to demonstrate high-resolution and accurate haplotyping across the entire MHC and KIR loci, phasing not only the highly diverse major and minor alleles, but also other important immunological genes and variants at non-genic regions across the locus together in a single haplotype structure.

Conclusions

Here, we describe the targeted HaploSeq method to generate large mega-base scale haplotypes in human cells. Using this technology, we reconstruct complete phase information of MHC and KIR loci. In principle, targeted HaploSeq is blind to genotyping and can be used to identify genetic variants *de novo* within the targeted loci. For example at the MHC locus, our method identified ~27 % of variants at an accuracy of 99.76 and 89.21 % for heterozygous and homozygous genotypes, respectively. This performance can be further improved with the use of multiple 4-base or 6-base cutters during Hi-C library preparation [35], instead of a single 6-base recognizing restriction enzyme as demonstrated in this manuscript. Alternatively, computational strategies such as population-based imputation can be also be used to generate comprehensive genotyping [36].

High-resolution genotyping and phasing of immunogenic loci such as MHC and KIR has several applications. First, it has the potential to greatly improve the practice of HLA typing/matching for clinical transplantation procedures [13, 15, 20, 37], as this method provides access to alleles that are otherwise un-typed using current methods. In addition, with population-scale MHC and KIR haplotyping, our method can help to elucidate a refined set of minimal alleles that confer the highest risk for GVHD, thereby informing follow-up cost-effective selective typing of these most informative alleles. Second, as our method phases coding and non-coding cis-regulatory sequences together, one can study patterns of compound heterozygosity and linkage of human immune variation [7, 16, 17]. Finally, several studies have uncovered numerous disease-associated HLA and KIR alleles and by understanding long-range haplotypes, we can now start to unravel mechanistic underpinnings of human immune disorders [21, 38, 39].

Recently, proximity-ligation methods such as Hi-C have been demonstrated to be useful in assembling genomes *de novo* [40, 41]. As targeted HaploSeq obtains high-quality chromatin interaction datasets, similar to Hi-C [31], this methodology can potentially be used to generate diploid assembly of complex regions, such as the MHC or T-cell receptor beta (Tcrb) locus [42], of human and other large genomes. Similarly, Hi-C has also recently been used in metagenomics studies to deconvolute the species present in complex microbiome mixtures [43, 44]. With the advent of targeted HaploSeq, it is now possible to capture distinct loci that are informative and discriminative enough to delineate species mixtures based on the captured proximity-ligation fragments.

Taken together, we present targeted HaploSeq and demonstrate its application for targeted phasing of HLA and KIR loci in the human genome. We believe that this method will lead to new avenues in biomedical research and in personalized clinical genomics.

Data access

All sequencing data have been submitted to the Gene Expression Omnibus (GEO) database and will be publically available upon publication. Data has been made available under the accession number GSE65726.

Ethics

Not applicable, non-human subjects.

Additional files

Additional file 1: Figure S1. Targeting regions around HindIII cut sites allows complete and high-resolution haplotyping of MHC and KIR loci. a) (i) and (ii) depict completeness and resolution at MHC locus, respectively. We simulated reads across +/− 400 bp from HindIII cut sites in the MHC region to study our ability to obtain complete and high-resolution haplotypes. As the MHC region has a high-density of het. variants (a het. variance every ~300 bases), 2X sequencing coverage is enough to generate complete haplotypes, regardless of read length. On the same lines, we obtain high-resolution seed haplotypes at low sequencing coverage. b) (i) and (ii) depict completeness and resolution at KIR locus respectively. On the contrary, as the KIR locus has a lower density of variants, high sequencing coverage is required to obtain complete haplotypes. In particular, 40 bp reads are not enough to obtain complete phasing even at 50X coverage and therefore is omitted in the resolution plot. Similary, even at high sequencing coverage, resolution is very limited regardingless of read length. (TIFF 8219 kb)

Additional file 2: Figure S2. Targeted enrichment at the KIR genomic locus. a) Genome browser shot of the ~1 Mb KIR region. The inset shows targets near KIR3DL2 gene, depicting target regions (green) around HindIII cut sites and repeat segments (red). We tiled 120-bp probes (blue) at 4X density accross these non-repeat target regions. b) (i) Top Plot demonstrates enrichment of GM12878 Targeted-HaploSeq reads at the 100 kb binned KIR locus while the bottom plot shows number of probes used across the KIR locus. Together, these plots show a high correlation among probes and read enrichment. (ii) Plot demonstrating sensitivity of capture probes—the true probes capture reads ~100 fold than random probes created virtually near HindIII cut sites (TIFF 8219 kb)

Additional file 3: Figure S3. Targeted HaploSeq data has large pool of long insert fragments. a) Insert-size distribution of targeted Haploseq (green) and b) HaploSeq (purple) in GM12878 LCLs. Both these datasets have similar amount of long-insert fragments which is critical for long range haplotyping. (TIFF 8219 kb)

Additional file 4: Figure S4. Homologous chromosomal interactions are rare and most of them are enriched in high variant density regions of the MHC loci. Using haplotypes indentified from the parent-trio whole genome sequencing data, we define homologous trans (h-trans) interactions in the Targeted Haploseq (green) and HaploSeq—from our previous publication (purple). a) h-trans interactions are rare —< 1 % in whole genome (i), about 5–6 % in the MHC locus (ii) and <0.5 % in KIR locus (iii). While h-trans interactions are <1 % whole-genome, we see them in significantly higher fractions at the MHC locus (~5 %). Interestingly, majority of these are found at regions with very high variant density (b), suggeting that the haplotype predictions from parent-trio data at these regions could be error-prone, which in-turn results in higher h-trans in HaploSeq datasets. (TIFF 8219 kb)

Additional file 5: Online Methods. (DOCX 149 kb)

Additional file 6: Figure S6. Targeted HaploSeq generates a single (complete) haplotype structure across MHC/KIR locus. The performance metric of the Targeted HaploSeq protocol, measured by completeness (span of the haplotype bloc), resolution (fraction of het. alleles resolved), and accuracy. While each of these metrics were defined after performing read-based as well as population based haplotyping, seed resolution is estimated only based on read-based haplotyping. The overall resolution is defined as the weighted average among all alleles accross the MHC and KIR loci together. We observe over 50 % decrease in error rate from 2.3 to 1.06 % after correcting for potential incorrect local haplotypes from parent-trio data. (TIFF 8219 kb)

Additional file 7: Figure S7. Targeted HaploSeq generates high quality phasing of heterozygous genes. Over 92 % of exonic het. variants are phased at an accuracy of 99 %. (TIFF 8219 kb)

Competing interests

S.S., A.D.S, J.R.D., and B.R. are named inventors on a patent application on the technology described in this manuscript. S.S., J.R.D. and B.R. are co-founders of Arima Genomics, Inc.

Authors' contributions

B.R., S.S. and A.D.S conceived the strategy. A.D.S performed the experiments and optimized the targeted aspects of HaploSeq. J.R.D assisted in the experiments. S.S. conducted the analysis. S.S. prepared the manuscript with assistance from A.D.S and B.R. All authors read and approved the final manuscript.

Authors' information

Not applicable.

Availability of data and materials

Not applicable.

Acknowledgements

We thank members of the Ren laboratory for helpful suggestions throughout the course of this work.

Funding

Research is supported by funds from NIH (R01ES024984), LICR and UCSD provided to B. R. A.D.S is supported in part by the UCSD Genetics Training Grant (T32 GM008666).

Author details

[1]Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA. [2]Biomedical Sciences Graduate Program, University of California San Diego, La Jolla, CA 92093, USA. [3]Medical Scientist Training Program, University of California San Diego, La Jolla, CA 92093, USA. [4]Department of Cellular and Molecular Medicine, and UCSD Moores Cancer Center, University of California San Diego, La Jolla, CA 92093, USA. [5]Institute of Genomic Medicine, University of California San Diego, La Jolla, CA 92093, USA.

Received: 16 February 2015 Accepted: 16 September 2015
Published online: 05 November 2015

References

1. Jin P, Wang E. Polymorphism in clinical immunology - From HLA typing to immunogenetic profiling. J Transl Med. 2003;1:8. doi:10.1186/1479-5876-1-8.
2. Middleton D, Gonzelez F. The extensive polymorphism of KIR genes. Immunology. 2010;129:8–19. doi:10.1111/j.1365-2567.2009.03208.x.
3. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, Almeida J, et al. Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics. 2008;60:1–18. doi:10.1007/s00251-007-0262-2.
4. Petersdorf EW. The major histocompatibility complex: a model for understanding graft-versus-host disease. Blood. 2013;122:1863–72. doi:10.1182/blood-2013-05-355982.
5. Proll J, Danzer M, Stabentheiner S, Niklas N, Hackl C, Hofer K, et al. Sequence capture and next generation resequencing of the MHC region highlights potential transplantation determinants in HLA identical haematopoietic stem cell transplantation. DNA Res. 2011;18:201–10. doi:10.1093/dnares/dsr008.

6. Chung WH, Hung SI, Chen YT. Human leukocyte antigens and drug hypersensitivity. Curr Opin Allergy Clin Immunol. 2007;7:317–23. doi:10.1097/ACI.0b013e3282370c5f.
7. Rizzo R, Bortolotti D, Baricordi OR, Fainardi E. New insights into HLA-G and inflammatory diseases. Inflamm Allergy Drug Targets. 2012;11:448–63.
8. Zeestraten EC, Reimers MS, Saadatmand S, Dekker JW, Liefers GJ, van den Elsen PJ, et al. Combined analysis of HLA class I, HLA-E and HLA-G predicts prognosis in colon cancer patients. Br J Cancer. 2014;110:459–68. doi:10.1038/bjc.2013.696.
9. Mahdi BM. A glow of HLA typing in organ transplantation. Clin Transl Med. 2013;2:6. doi:10.1186/2001-1326-2-6.
10. Chang CJ, Chen PL, Yang WS, Chao KM. A fault-tolerant method for HLA typing with PacBio data. BMC bioinformatics. 2014;15:296. doi:10.1186/1471-2105-15-296.
11. Boegel S, Lower M, Schafer M, Bukur T, de Graaf J, Boisguerin V, et al. HLA typing from RNA-Seq sequence reads. Genome medicine. 2012;4:102. doi:10.1186/gm403.
12. Bai Y, Ni M, Cooper B, Wei Y, Fury W. Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads. BMC Genomics. 2014;15:325. doi:10.1186/1471-2164-15-325.
13. Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics. 2013;14:355. doi:10.1186/1471-2164-14-355.
14. Shiina T, Suzuki S, Ozaki Y, Taira H, Kikkawa E, Shigenari A, et al. Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens. 2012;80:305–16. doi:10.1111/j.1399-0039.2012.01941.x.
15. Furst D, Muller C, Vucinic V, Bunjes D, Herr W, Gramatzki M, et al. High-resolution HLA matching in hematopoietic stem cell transplantation: a retrospective collaborative analysis. Blood. 2013;122:3220–9. doi:10.1182/blood-2013-02-482547.
16. Mullighan C, Heatley S, Doherty K, Szabo F, Grigg A, Hughes T, et al. Non-HLA immunogenetic polymorphisms and the risk of complications after allogeneic hemopoietic stem-cell transplantation. Transplantation. 2004;77:587–96.
17. Guo Z, Hood L, Malkki M, Petersdorf EW. Long-range multilocus haplotype phasing of the MHC. Proc Natl Acad Sci U S A. 2006;103:6964–9. doi:10.1073/pnas.0602286103.
18. Traherne JA. Human MHC architecture and evolution: implications for disease association studies. Int J Immunogenet. 2008;35:179–92. doi:10.1111/j.1744-313X.2008.00765.x.
19. Petersdorf EW, Malkki M, Horowitz MM, Spellman SR, Haagenson MD, Wang T. Mapping MHC haplotype effects in unrelated donor hematopoietic cell transplantation. Blood. 2013;121:1896–905. doi:10.1182/blood-2012-11-465161.
20. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. PLoS Med. 2007;4, e8. doi:10.1371/journal.pmed.0040008.
21. Larsen CE, Alford DR, Trautwein MR, Jalloh YK, Tarnacki JL, Kunnenkeri SK, et al. Dominant sequences of human major histocompatibility complex conserved extended haplotypes from HLA-DQA2 to DAXX. PLoS Genet. 2014;10, e1004637. doi:10.1371/journal.pgen.1004637.
22. Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, et al. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature. 2012;487:190–5. doi:10.1038/nature11236.
23. Kaper F, Swamy S, Klotzle B, Munchel S, Cottrell J, Bibikova M, et al. Whole-genome haplotyping by dilution, amplificaiton, and sequencing. Proc Natl Acad Sci. 2013;110:5552–7.
24. Kitzman JO, Mackenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, et al. Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol. 2011;29:59–63. doi:10.1038/nbt.1740.
25. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, Blauwkamp T, et al. Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol. 2014;32:261–6. doi:10.1038/nbt.2833.
26. Selvaraj S, DixonJ R, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol. 2013;31:1111–8. doi:10.1038/nbt.2728.
27. Kirkness EF, Grindberg RV, Yee-Greenbaum J, Marshall CR, Scherer SW, Lasken RS, et al. Sequencing of isolated sperm cells for direct haplotyping of a human genome. Genome Res. 2013;23:826–32. doi:10.1101/gr.144600.112.
28. Fan HC, Wang J, Potanina A, Quake SR. Whole-genome molecular haplotyping of single cells. Nat Biotechnol. 2011;29:51–7. doi:10.1038/nbt.1739.
29. Yang H, Chen X, Wong H. Completely phased genome sequencing through chromosome sorting. Proc Natl Acad Sci. 2012;109:3190–3190. doi:10.1073/pnas.1200309109.
30. de Vree PJ, de Wit E, Yilmaz M, van de Heijning M, Klous P, Verstegen MJ, et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. Nat Biotechnol. 2014;32:1019–25. doi:10.1038/nbt.2959.
31. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93. doi:10.1126/science.1181369.
32. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009;27:182–9. doi:10.1038/nbt.1523.
33. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461:272–6. doi:10.1038/nature08250.
34. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73. doi:10.1038/nature09534.
35. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014;159:1665–80. doi:10.1016/j.cell.2014.11.021.
36. Browning BL, Browning SR. Improving the Accuracy and Efficiency of Identity-by-Descent Detection in Population Data. Genetics. 2013;194:459–71. doi:10.1534/genetics.113.150029.
37. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood. 2007;110:4576–83. doi:10.1182/blood-2007-06-097386.
38. Traherne JA, Horton R, Roberts AN, Miretti MM, Hurles ME, Stewart CA, et al. Genetic analysis of completely sequenced disease-associated MHC haplotypes identifies shuffling of segments in recent human history. PLoS Genet. 2006;2, e9. doi:10.1371/journal.pgen.0020009.
39. Romero V, Larsen CE, Duke-Cohan JS, Fox EA, Romero T, Clavijo OP, et al. Genetic fixity in the human major histocompatibility complex and block size diversity in the class I region including HLA-E. BMC Genet. 2007;8:14. doi:10.1186/1471-2156-8-14.
40. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31:1119–25. doi:10.1038/nbt.2727.
41. Kaplan N, Dekker J. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat Biotechnol. 2013;31:1143–7. doi:10.1038/nbt.2768.
42. Spicuglia S, Pekowska A, Zacarias-Cabeza J, Ferrier P. Epigenetic control of Tcrb gene rearrangement. Semin Immunol. 2010;22:330–6. doi:10.1016/j.smim.2010.07.002.
43. Beitel CW, Froenicke L, Lang JM, Korf IF, Michelmore RW, Eisen JA, et al. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ. 2014;2, e415. doi:10.7717/peerj.415.
44. Burton JN, Liachko I, Dunham MJ, Shendure J. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. G3. 2014;4:1339–46. doi:10.1534/g3.114.011825.

Generation of Hi-C libraries. GM12878 (Coriell) cells were cultured in suspension in 85% RPMI media supplemented with 15% FBS and 1X penicillin/streptomycin. Cells were harvested, fixed in 1% formaldehyde, and 5 million cells were subject to the Hi-C protocol as previously described[1]. Prior to target enrichment, Hi-C libraries were amplified by 14 cycles of PCR using a high-fidelity (Fusion) polymerase. The number of pre-enrichment PCR cycles can be tailored depending on how much DNA is required for downstream target hybridization reactions. In this case, we performed several parallel PCR reactions using small amounts of bead-bound Hi-C library input at 14 cycles to obtain sufficient material.

Generation of RNA baits. Probes were computationally designed using the SureDesign software suite (Agilent Technologies). The custom design targeted the upstream and downstream 400bp adjacent to HindIII cut sites spanning the MHC (chr6:29689001-33098550) and KIR (chr19:54538900-55596120) loci using the hg19 genome build. SureDesign parameters were set to 4X tiling density, maximum probe boosting, and maximum repetitive sequence masking. We also targeted MHC exons at 2X tiling density, balanced boosting, and maximum repetitive element masking. In sum, 8,702 unique probes sequences were computationally generated, of which 6,413 correspond to regions flanking HindIII cut sites across the MHC locus, 765 correspond to HLA exons, and 1,524 probes were allocated to regions flanking HindIII cut sites across the KIR locus. In total, 12,298 probes were synthesized after considering some probes are duplicated in

the design due to the boosting parameter of SureDesign. Using the probe sequences output by SureDesign, single-stranded DNA (ssDNA) oligos were synthesized by CustomArray Inc. ssDNA oligos contained universal forward and reverse priming sequences. Universal forward priming sequences contained a truncated SP6 RNA polymerase recognition sequence that is engineering to be completed upon PCR amplification of the ssDNA oligos. The reverse universal priming sequence contained a BsrDI recognition sequence for 3' dsDNA probe cleavage prior to *in vitro* transcription. To convert ssDNA oligos into biotinylated RNA baits, ssDNA oligos were first diluted to 200pg/ul and then PCR-amplified using high-fidelity DNA polymerase (KAPA) and purified using Ampure XP beads (Beckman Coulter). As mentioned, the PCR reaction also serves the purpose to complete the remainder of the 5' SP6 recognition sequence. Next, reverse priming sequences were removed by digesting the dsDNA with BsrDI (New England Biosciences) and purified again using Ampure XP beads to remove the digested fragments. Lastly, we performed *in vitro* transcription (IVT) according to manufacturers protocol (Ambion) in the presence of biotinylated UTP (Epicentre). RNA was then column-purified (Qiagen), diluted to working concentration (500ng/ul) and stored at -80 until use.

Generation of targeted HaploSeq libraries. To enrich our Hi-C libraries for proximity ligation fragments mapping to the MHC and KIR loci, we performed target enrichment using our custom RNA baits, followed by PCR amplification and sequencing. Briefly, 500ng of Hi-C library was incubated overnight at 65

degrees with 500ng of biotinylated RNA probe along with 2.5ug of human Cot-1 DNA (Life Technologies), 2.5ug Salmon Sperm DNA (Life Technologies) and blocking primers at a final concentration of 6.67uM. Next, RNA:DNA hybrids were isolated using T1 streptavidin-coated beads (Invitrogen) while DNA molecules not bound by RNA baits were washed away. Finally, captured products were resuspended in water and PCR amplified (Fusion) on-bead using 10-11 cycles. Lastly, PCR products were purified using AMPure XP beads (Beckman Coulter) and then subject to next-generation sequencing on Illumina HiSeq2500 to obtain approximately 51M reads pairs.

Genotyping. Variant calls and genotypes for GM12878 were downloaded from 1000 genomes project[2] and these were used for *de novo* haplotype reconstruction from targeted HaploSeq data. Predicted haplotypes were compared with phasing Information for GM12878 from the 1000 genomes project to estimate accuracy.

Read Alignment. We mapped targeted HaploSeq reads to the hg19 genome. Reads were aligned using BWA Mem[3] (Version: 0.7.5a-r405) as single end and these reads were manually paired using in-house scripts. Unmapped reads were removed and PCR duplicate reads were removed using Picard (Version: 1.49). The aligned datasets were then finally processed with GATK[4] (Version: March 2013) for indel realignment and variant recalibration. The alignment process

resulted in ~86% of reads mapping to genome in a non-duplicative fashion, of which about 7% of the reads mapped to the MHC/KIR loci.

Simulations of haplotype completeness and resolution over MHC and KIR loci. Here, we used different combinations of read length and coverage to obtain predictions at various haplotyping resolutions (Supplementary Figure 1). In particular, we used read lengths of 40, 50, 75, 100, 150 and 250 bases at coverage up to 50X of the MHC and KIR loci. To maintain the Hi-C insert distributions, we used human H1 Hi-C[5] intra-chromosomal read starting positions at the MHC/KIR locus where at least one end of the pair-end is within 400bp from the HindIII cut-sites. We constructed graphs with nodes representing heterozygous variants in GM12878 (MHC/KIR) and edges corresponding to reads that cover multiple variants. These graphs allowed to us to predict completeness and resolution of the haplotypes through simulated data. Specifically, if the edge covers from "first" to the "last" SNP, the graph is 100% complete and the fraction of SNPs covered or phased in the longest graph is termed resolution. From various simulations (Supplementary Figure 1), the high density of heterozygous variants in the MHC loci allows complete and high-resolution haplotypes with low sequencing coverage. On the other hand, longer reads are necessary to obtain complete haplotypes in the KIR loci.

Target Enrichment violin plots. The violin plots (Figure 2aii and Supplementary Figure 2bii) represent reads per probe in two cases. In the first case, we take the

actual probe locations in the MHC and KIR locus and estimate the read count per probe. In the second case, we designed "virtual" probes randomly across the genome (excluding the MHC and KIR loci) but confined to regions adjacent to HindIII cut sites similar to actual probes. We then counted reads per virtual probe. A 100-fold difference between read counts in these two cases demonstrates that our oligonucleotide based capture is sensitive in obtaining data from the targeted loci.

Homologous Trans interaction estimations. We used haplotype predictions from parent-child trio dataset[2] to estimate the fraction of *cis* and fraction of homologous *trans* interaction (read1 and read2 map between the two homologous chromosomes) from the targeted HaploSeq data and the HaploSeq data. As reported earlier in the HaploSeq method[6], we observe rare homologous interactions genome-wide (excluding MHC locus, Supplementary Figure 4ai). However, the fraction of homologous *trans* (or h-trans) is notably higher in the MHC locus (Supplementary Figure 4aii). Interestingly, the vast majority of these h-trans interactions from the MHC locus occur in regions with very high heterozygous variant density (Supplementary Figure 4b).

Haplotyping. Using variant calls and genotypes from the 1000 genomes project[2], we used HaploSeq protocol[6] to generate locus-spanning haplotypes across the MHC and KIR locus.

Using Moleculo LFR data to decipher discordancy in haplotype calls. We downloaded raw reads from Moleculo LFR data from GM12878 cells[7]. We adopted the Hi-C mapping strategy for Moleculo data. A snapshot of the data is described in Figure 2c. In order to estimate an overall rate for concordance among targeted HaploSeq and Moleculo data regarding the inter-haplotypes from parent-trio data, we did the following; Using the haplotype predictions from parent-child trio data, we first grouped all the SNP-pairs that represented inter-haplotype interactions from targeted HaploSeq. Then, for all these SNP-pairs, we asked what fraction of them is also inter-haplotype from Moleculo data. This fraction was approximately 67%. However, it turns out that these 67% of SNP-pairs are the well-supported inter-haplotype junctures, as they represent ~95% the inter-haplotype reads from targeted HaploSeq. In other words, 95% of inter-haplotype reads from the targeted HaploSeq data are concordant with molecule data (Figure 2d). Consequently, these SNPs from these SNP-pairs were removed from the genotype list and the haplotyping analysis was repeated, resulting in more than a 50% reduction in haplotyping error (Supplementary Figure 5).

Exon Phasing. We downloaded RefSeq gene list for hg19 build and kept only the longest transcript for a given gene in both the MHC and KIR locus. Therefore, each gene is represented by only one canonical transcript and the heterozygous variants from the exons of these canonical transcripts were used for the haplotyping analysis (Figure 2e,f and Supplementary Figure 6).

Genotype Predictions using targeted HaploSeq data. We used GATK pipeline and UnifiedGenotyper to call variants. We retained genotype predictions that had a sequencing depth of at least 10 and had a PASS filter after the GATK variant recalibration step.

References

1 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

2 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

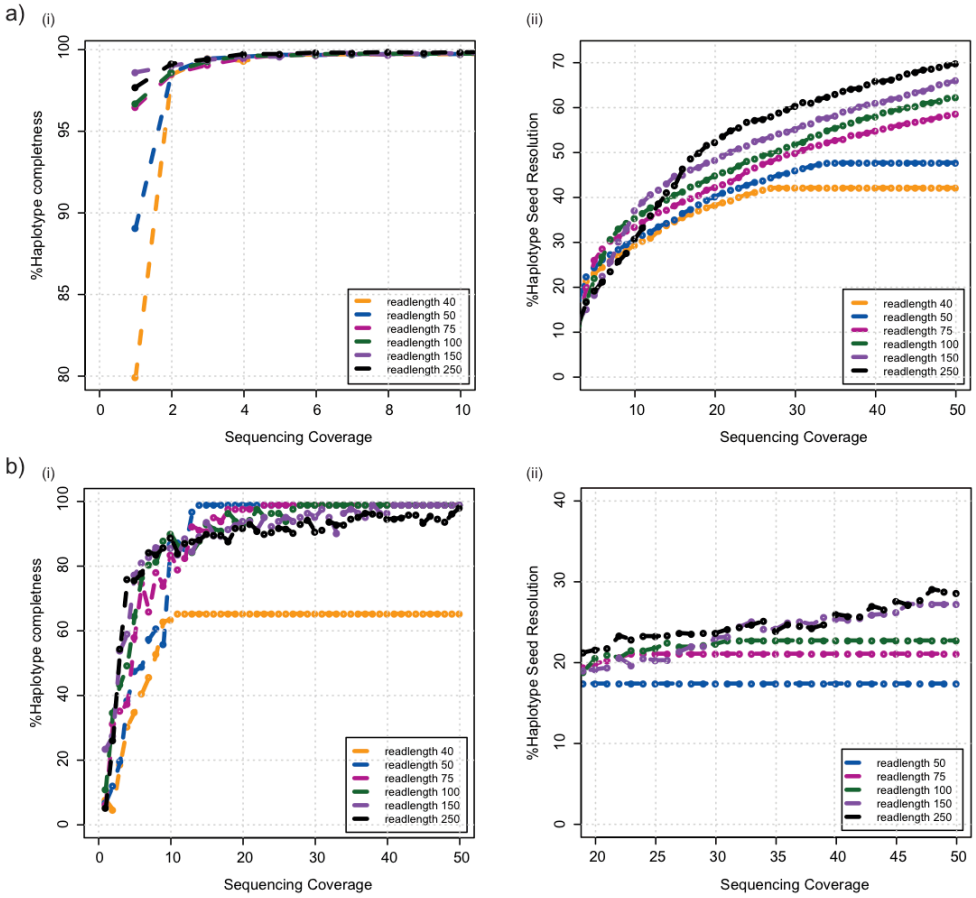
3 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Preprint at arXiV:1303.3997v2 [q-bio.GN]* (2013).

4 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498, doi:10.1038/ng.806 (2011).

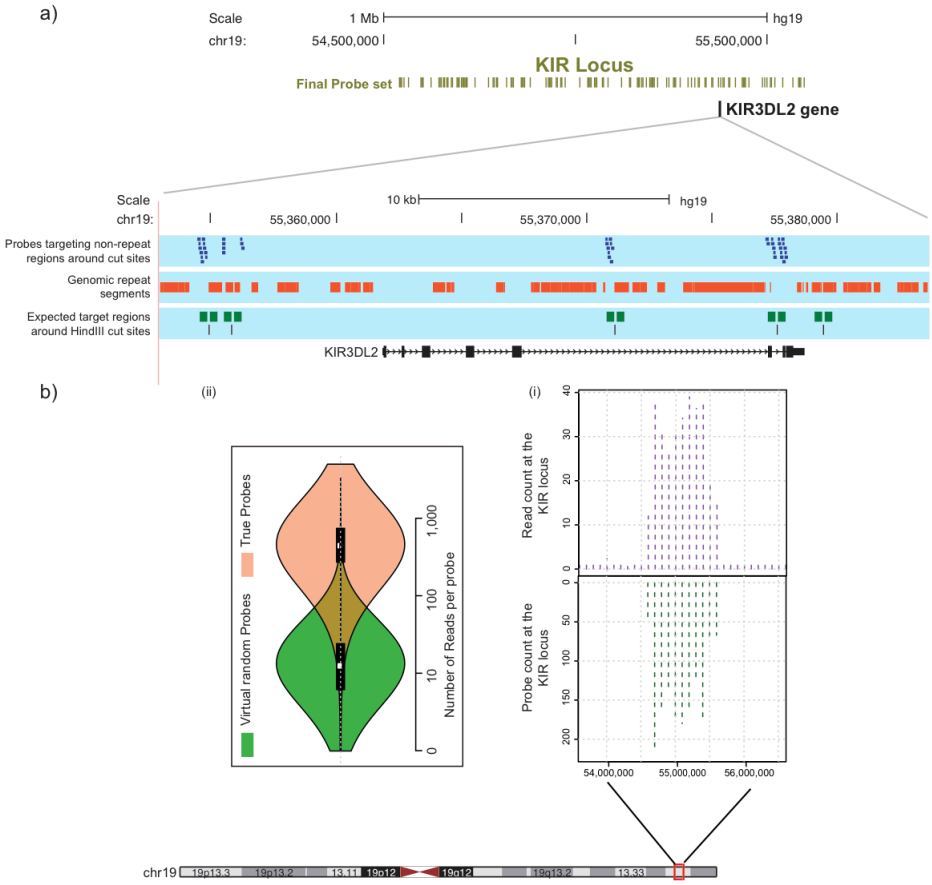
5 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

6 Selvaraj, S., J, R. D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **31**, 1111-1118, doi:10.1038/nbt.2728 (2013).

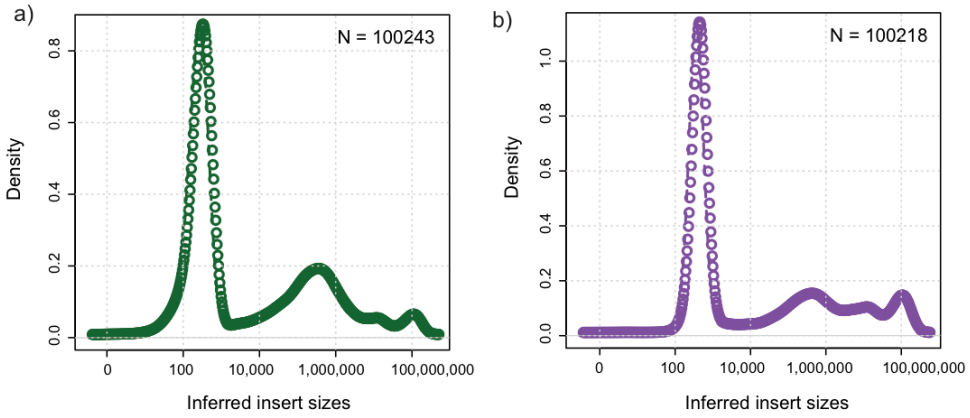
7 Kuleshov, V. *et al.* Whole-genome haplotyping using long reads and statistical methods. *Nature biotechnology* **32**, 261-266, doi:10.1038/nbt.2833 (2014).



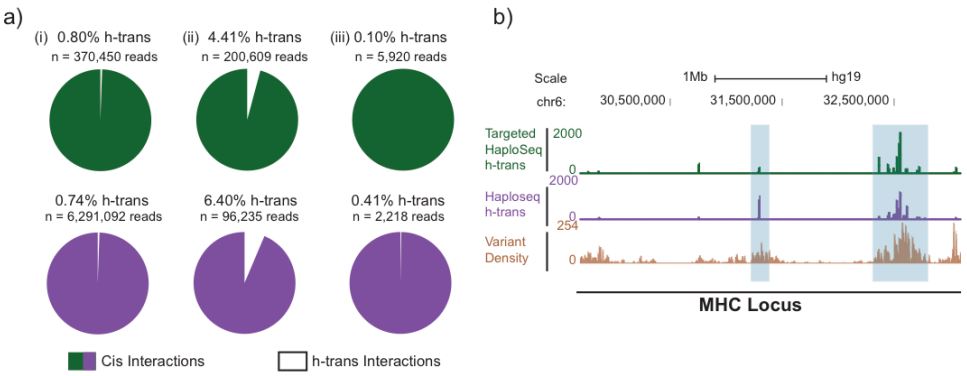
Supplementary Figure 1: Targeting regions around HindIII cut sites allows complete and high-resolution haplotyping of MHC and KIR loci.
a) (i) and (ii) depict completness and resolution at MHC locus, respectively. We simulated reads across +/- 400bp from HindIII cut sites in the MHC region to study our ability to obtain complete and high-resolution haplotypes. As the MHC region has a high-density of het. variants (a het. variant every ~300 bases), 2X sequencing coverage is enough to generate complete haplotypes, regardless of read length. On the same lines, we obtain high-resolution seed haplotypes at low sequencing coverage. b) (i) and (ii) depict completness and resolution at KIR locus respectively. On the contrary, as the KIR locus has a lower density of variants, high sequencing coverage is required to obtain compelete haplotypes. In particular, 40bp reads are not enough to obtain complete phasing even at 50X coverage and therefore is omitted in the resolution plot. Similarly, even at high sequencing coverage, resolution is very limited regardless of read length.



Supplementary Figure 2: Targeted enrichment at the KIR genomic locus.
a) Genome browser shot of the ~1Mb KIR region. The inset shows targets near KIR3DL2 gene, depicting target regions (green) around HindIII cut sites and repeat segments (red). We tiled 120-bp probes (blue) at 4X density accross these non-repeat target regions. b) (i) Top Plot demonstrates enrichment of GM12878 Targeted-HaploSeq reads at the 100kb binned KIR locus while the bottom plot shows number of probes used accross the KIR locus. Together, these plots show a high correlation among probes and read enrichment. (ii) Plot demonstrating sensitivity of capture probes - the true probes capture reads ~100 fold than random probes created virtually near HindIII cut sites.



Supplementary Figure 3: Targeted HaploSeq data has large pool of long insert fragments. a) Insert-size distribution of targeted Haploseq (green) and b) HaploSeq (purple) in GM12878 LCLs. Both these datasets have similar amount of long-insert fragments which is critical for long range haplotyping.



Supplementary Figure 4: Homologous chromosomal interactions are rare and most of them are enriched in high variant density regions of the MHC loci.
Using haplotypes identified from the parent-trio whole genome sequencing data, we define homologous trans (h-trans) interactions in the Targeted Haploseq (green) and HaploSeq - from our previous publication (purple). a) h-trans interactions are rare - <1% in whole genome (i), about 5-6% in the MHC locus (ii) and <0.5% in KIR locus (iii). While h-trans interactions are <1% whole-genome, we see them in significantly higher fraction at the MHC locus (~5%). Interestingly, majority of these are found at regions with very high variant density (b), suggesting that the haplotype predictions from parent-trio data at these regions could be error-prone, which in-turn results in higher h-trans in HaploSeq datasets.

	Locus	Seed Resolution	Complet-ness	Reso-lution	Accuracy
GM12878 uncorrected	MHC	39.5	100	92.3	97.7
	KIR	17.2	100	71.6	98.6
GM12878 Corrected	MHC	37.1	100	90.9	98.9
	KIR	17.1	100	71.6	98.6
	Uncorrected Overall			91.2	97.7
	Corrected Overall			89.9	98.9

Supplementary Figure 5: Targeted HaploSeq generates a single (complete) haplotype structure across MHC/KIR locus.
The performance metric of the Targeted HaploSeq protocol, measured by completness (span of the haplotype block), resolution (fraction of het. alleles resolved), and accuracy. While each of these metrics were defined after performing read-based as well as populaiton based haplotyping, seed resolution is estimated only based on read-based haplotyping. The overall resolution is defined as the weighted average among all alleles accross the MHC and KIR loci together. We observe over 50% decrease in error rate from 2.3% to 1.06% after correcting for potential incorrect local haplotypes from parent-trio data.

Locus	Number of het. genes	Number of het. genes resolved	Number of exonic het. variants	Number of exonic het. variants resolved	Number of exonic het. variants incorrecty phased
MHC	103	96	612	572	3
KIR	16	14	45	37	1
Overall		92.43		92.69	99.34

Supplementary Figure 6: Targeted HaploSeq generates high quality phasing of heterozygous genes
Over 92% of exonic het. variants are phased at an accurcay of 99%.

Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in *BMC Genomics*, volume 16, Nov 5 2015. Selvaraj, Siddarth; Schmitt, Anthony D.; Dixon, Jesse R.; Ren, Bing. The dissertation author was the co-primary investigator and co-primary author of this paper.

Chapter 3

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,*]

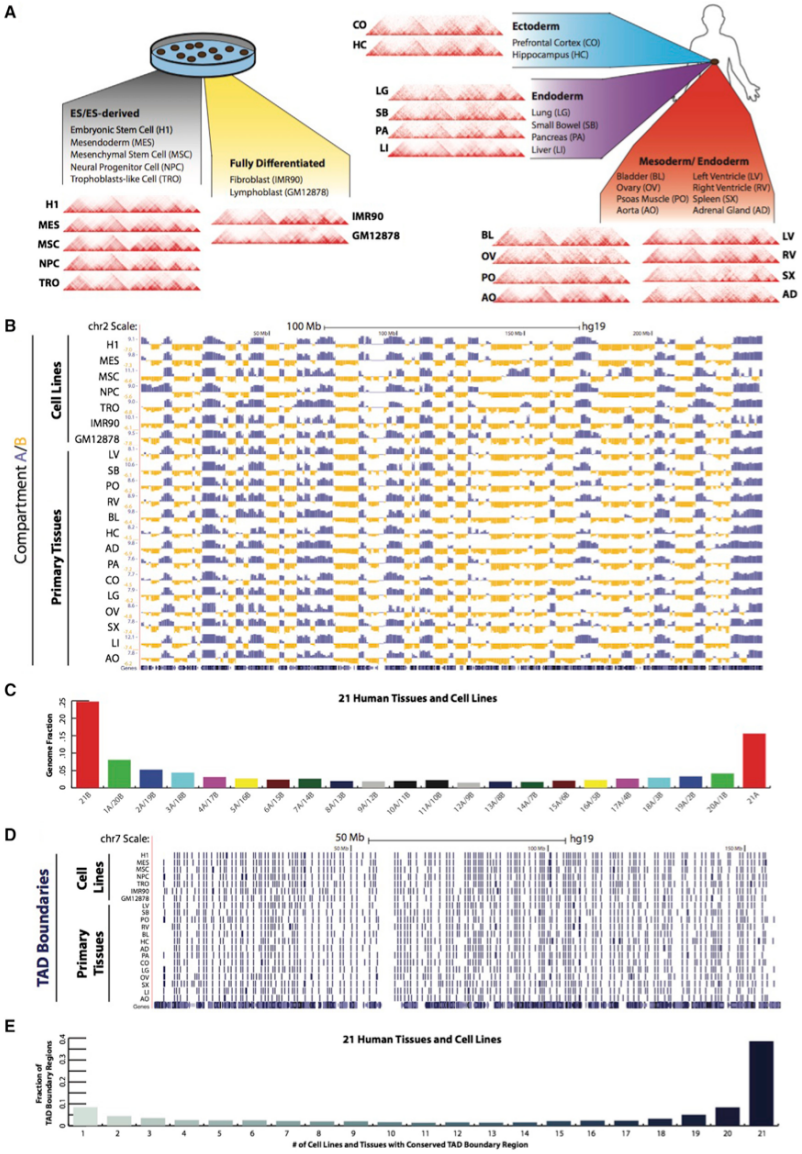
[1]Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA
[2]UCSD Biomedical Sciences Graduate Program, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[3]Division of Biostatistics, Department of Population Health, New York University School of Medicine, 650 First 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)

OPEN ACCESS
CellPress

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

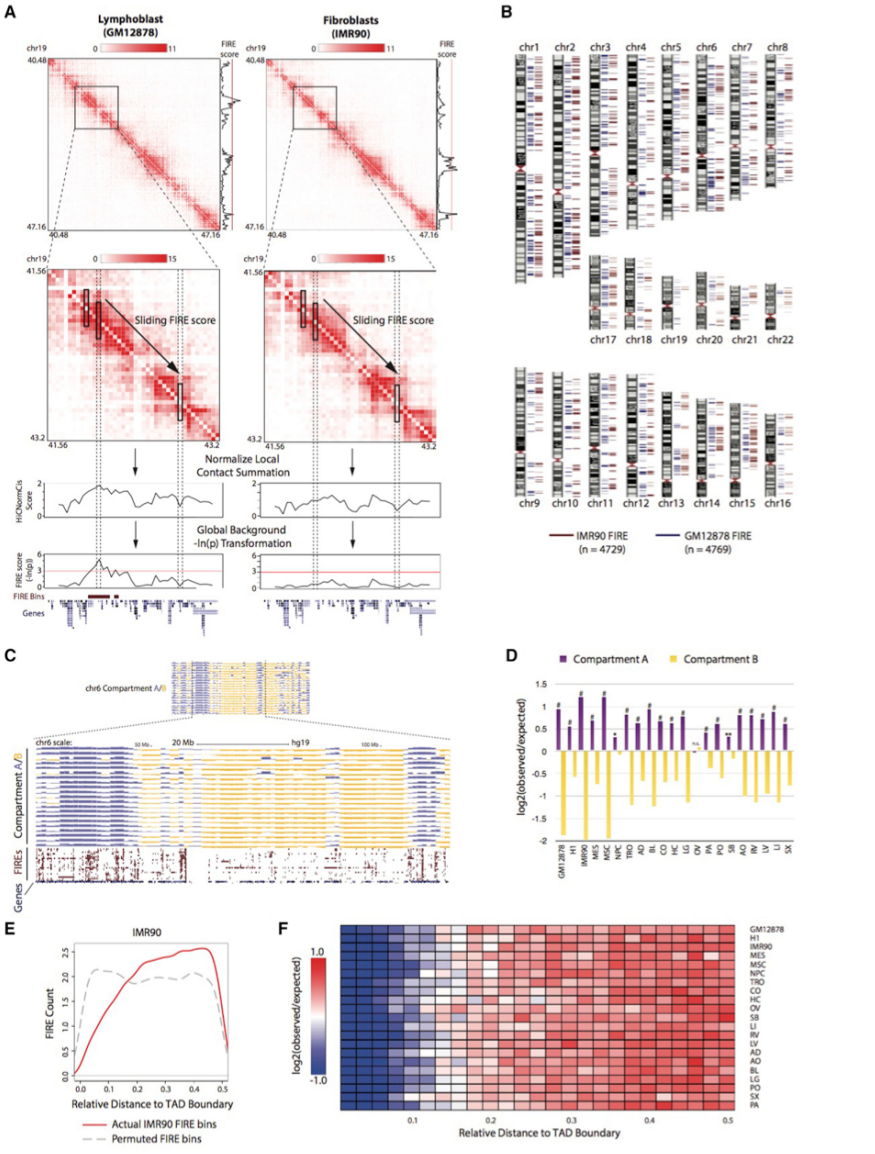
(A) Illustration of the primary 21 Hi-C datasets analyzed, depicting the cell (left panel) or tissue (right panel) origin of the samples as well as the germ layer origin for tissues (right panel). Hi-C interaction patterns across an 11.68-Mb region (chr12:82,840,000–94,520,000) are shown for all 7 cell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

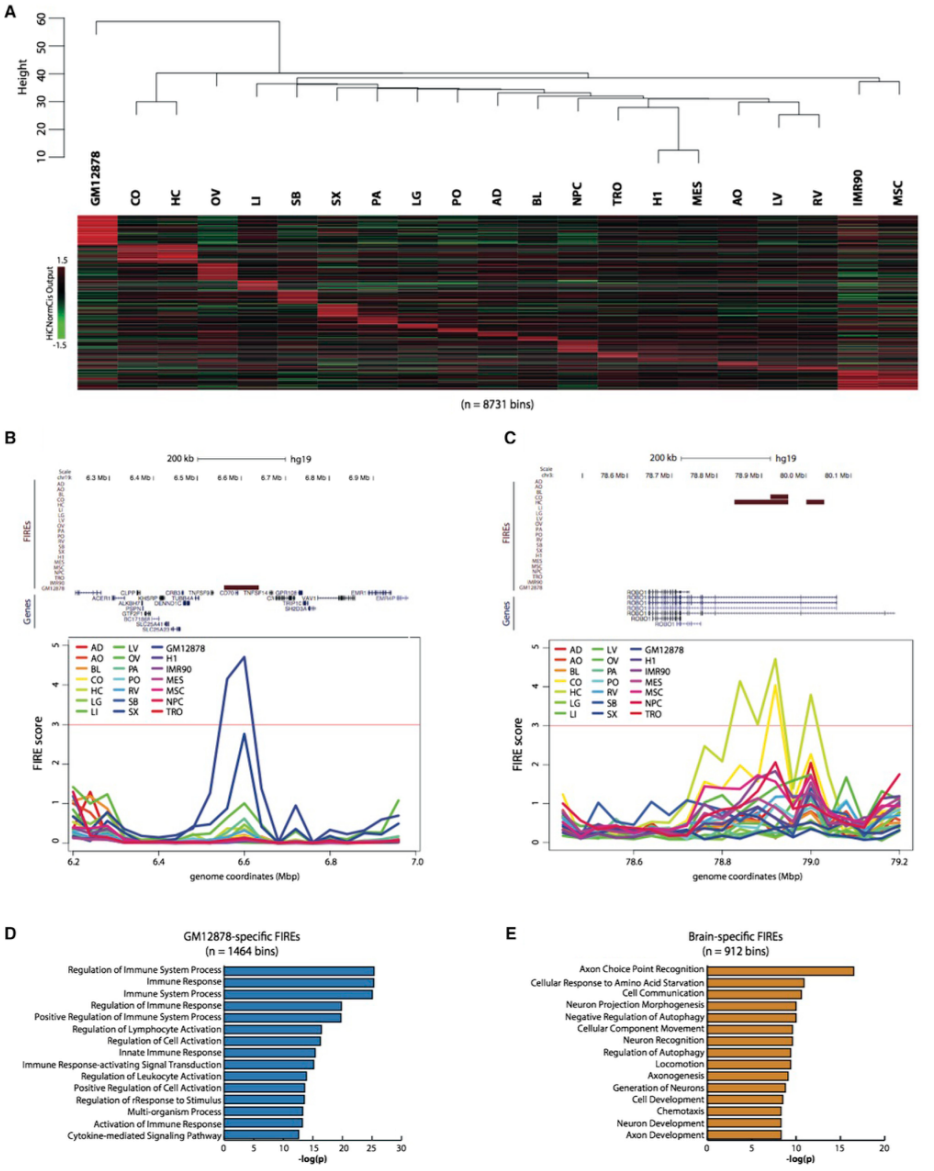
(B) Chromosome ideograms showing the genome-wide positional distribution of FIRE bins in GM12878 (blue, n = 4,769) and IMR90 (maroon, n = 4,729). Genome-wide visualization captures both conserved and specific FIRE bins. Only autosomes are depicted.

(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)

OPEN ACCESS
CellPress

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

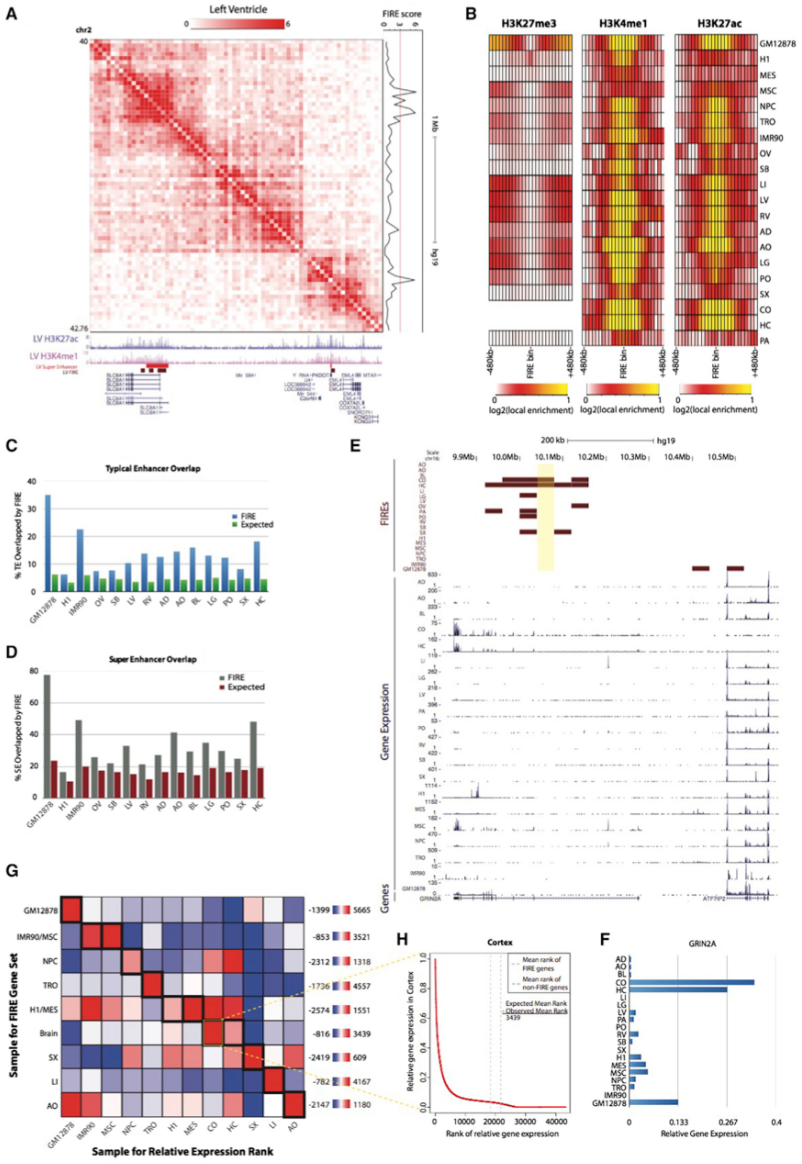
(A) At the top is a dendrogram resulting from a hierarchical clustering analysis using genome-wide FIRE scores for each sample. The y axis is the Euclidean distance between FIRE scores from any two samples. The heat map below shows a subset of FIRE bins (n = 8,371), corresponding to FIRE bins that are 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.

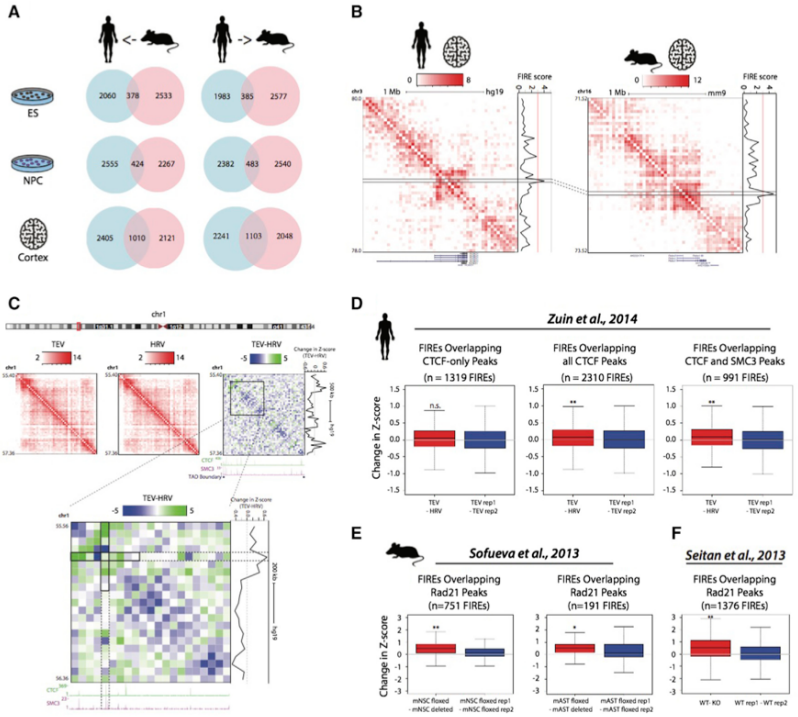


Figure 5. FIREs Are Conserved across Evolution and Mediated by Cohesin

(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.

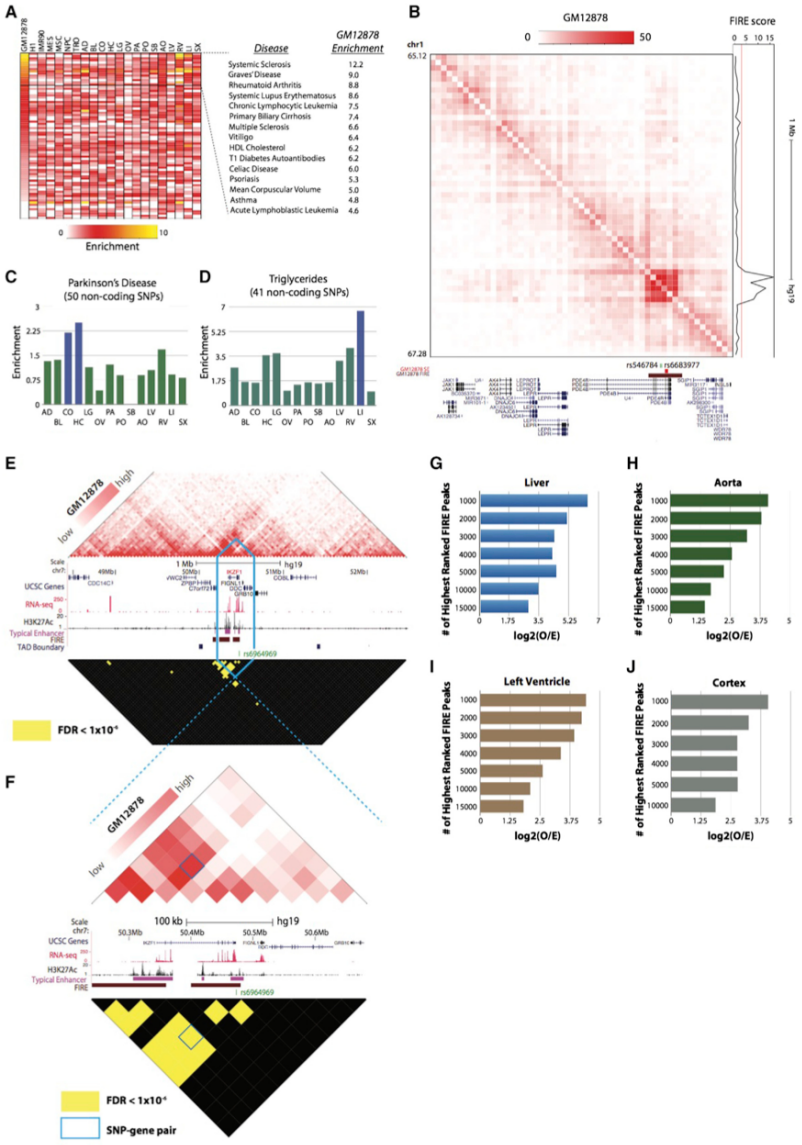
FIREs Are Enriched for Disease-Associated SNPs

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.

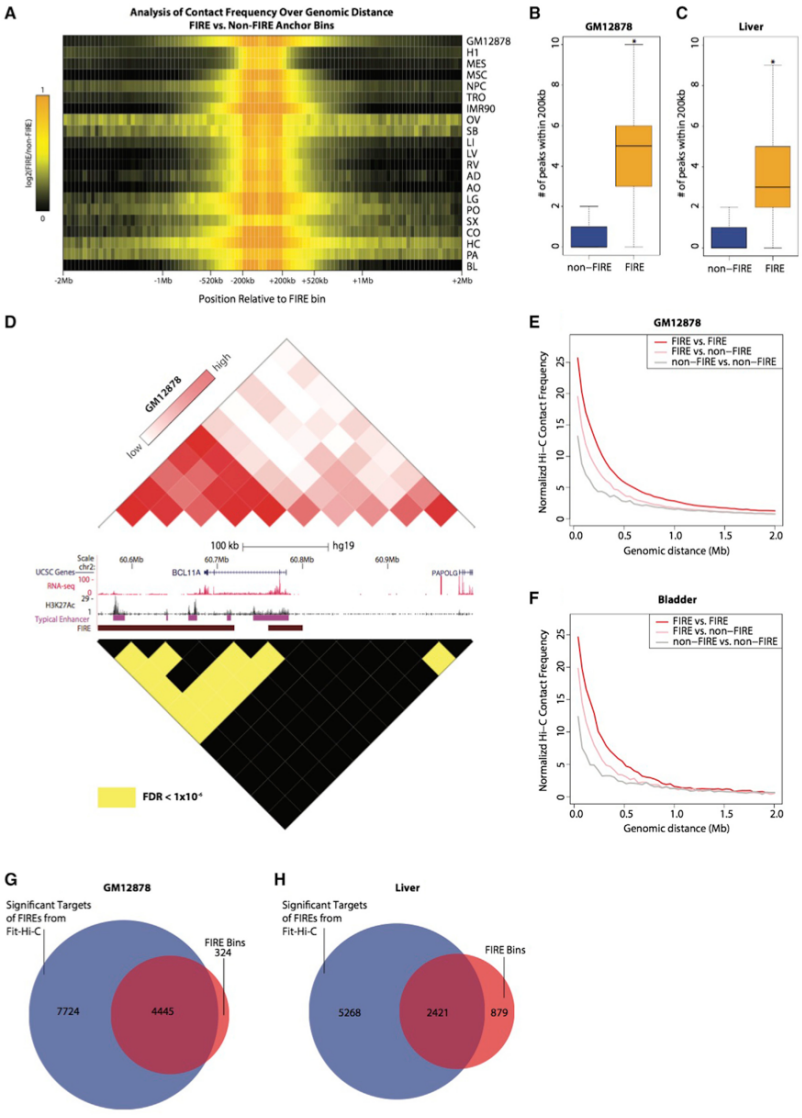


Figure 7. FIREs Have Several Targets and Are Self-Interactive

(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)

OPEN ACCESS
CellPress

(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 $R_i = Y_i / \exp(\hat{\beta}_0 + \hat{\beta}_F F_i + \hat{\beta}_{GC} GC_i + \hat{\beta}_M M_i)$ as the normalized total *cis* interaction. Noticeably, $\exp(\hat{\beta}_0)$ 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.

OPEN ACCESS
CellPress

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*

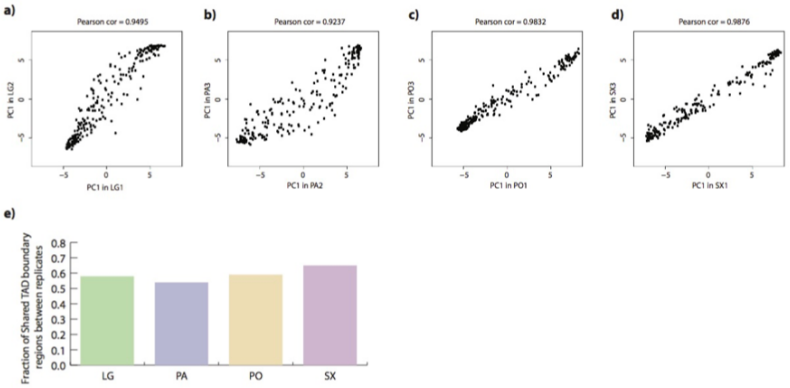
Supplemental Information

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%

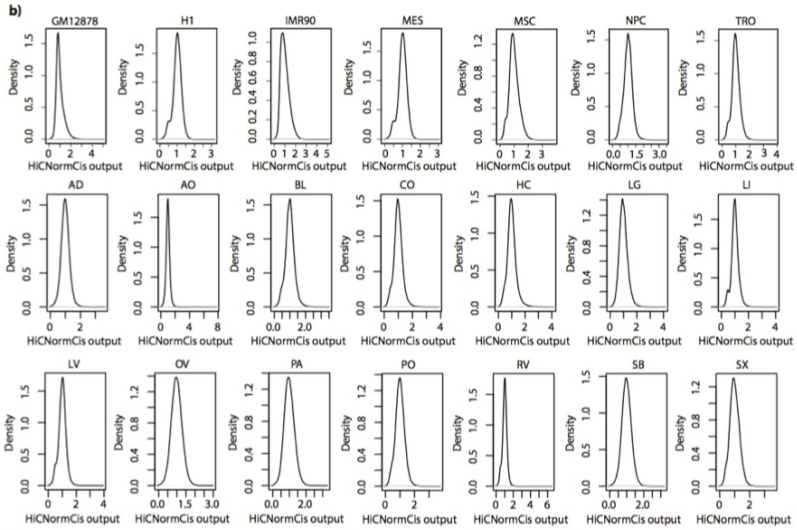
Figure S1. Hi-C data reproducibility and compartment A/B conservation, related to Figure 1.

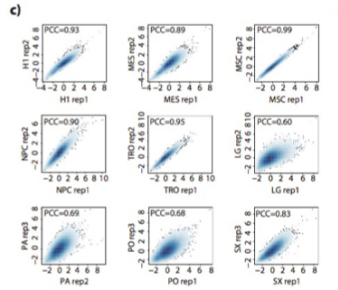
A) Scatter plots from replicates of LG, showing the genome-wide of the PC1 values used for the Compartment A/B analysis. The plot title contains the Pearson correlation coefficient of all 1Mb bin-pairs. The x- and y-axes are labeled according to their tissue type and donor. For example, LG2 corresponds 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.

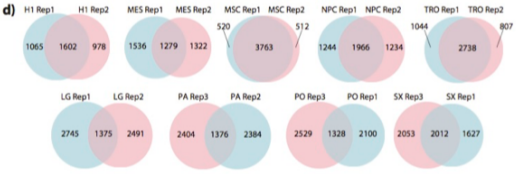
Figure S2

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



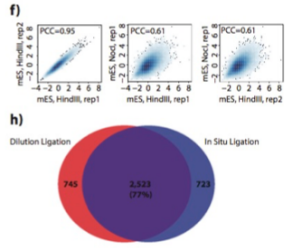


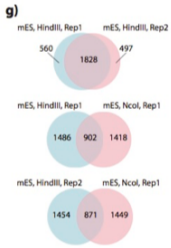


e)

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





i)

Sample	before HiCNormCis			After HiCNormCis		
	F	GC	M	F	GC	M
GM12878	0.140	-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	0.704	-0.230	0.174	0.126	-0.070	-0.003
CO	0.675	-0.257	0.214	0.113	-0.059	-0.009
HC	0.663	-0.291	0.221	0.100	-0.053	-0.010
OV	0.729	-0.332	0.177	0.149	-0.100	0.001
SB	0.757	-0.407	0.188	0.150	-0.091	-0.002
LG	0.624	0.181	0.206	0.061	0.073	-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	0.309	0.136	-0.031	0.128	-0.007
LI	0.735	-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	-0.338	0.131	0.177	-0.118	0.004
BL	0.721	-0.297	0.135	0.124	-0.072	-0.003

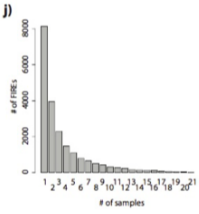
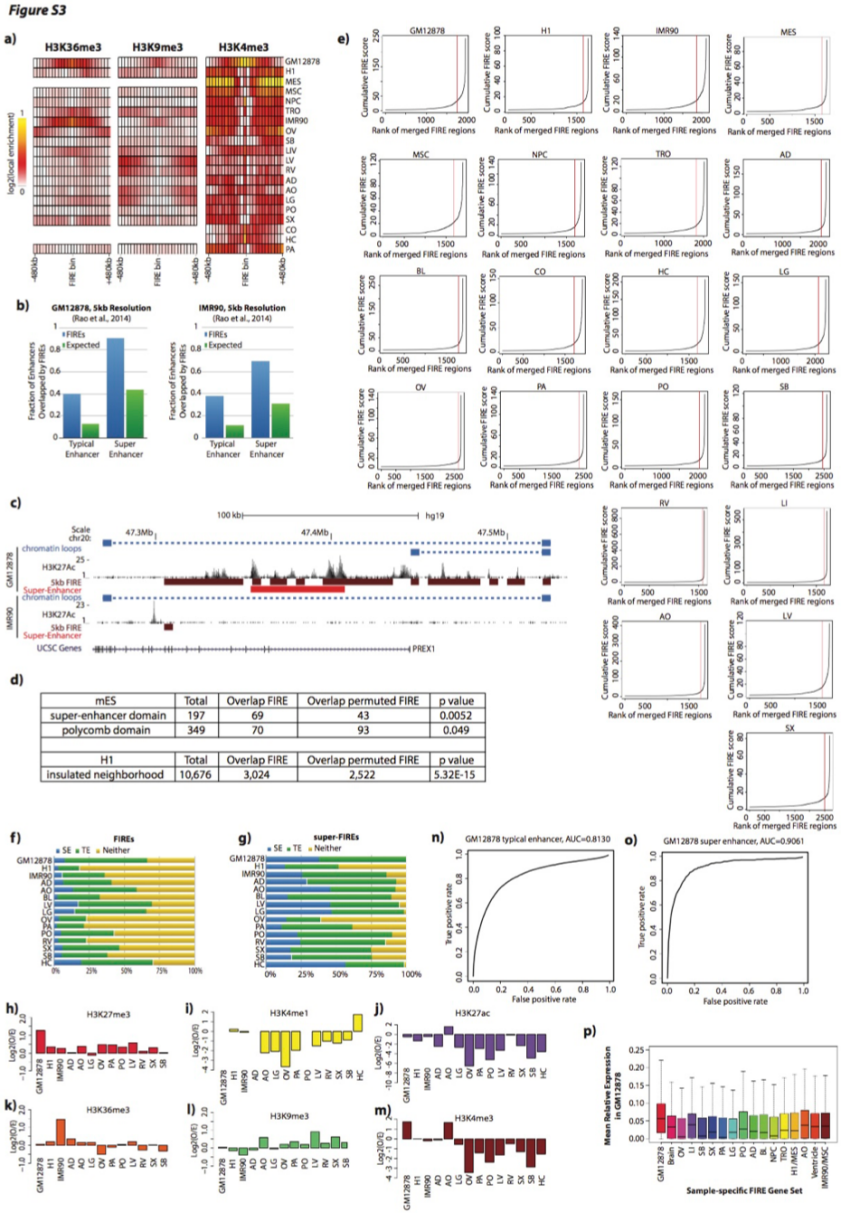


Figure S2. FIRE calling methodology, related to Figure 2.

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.

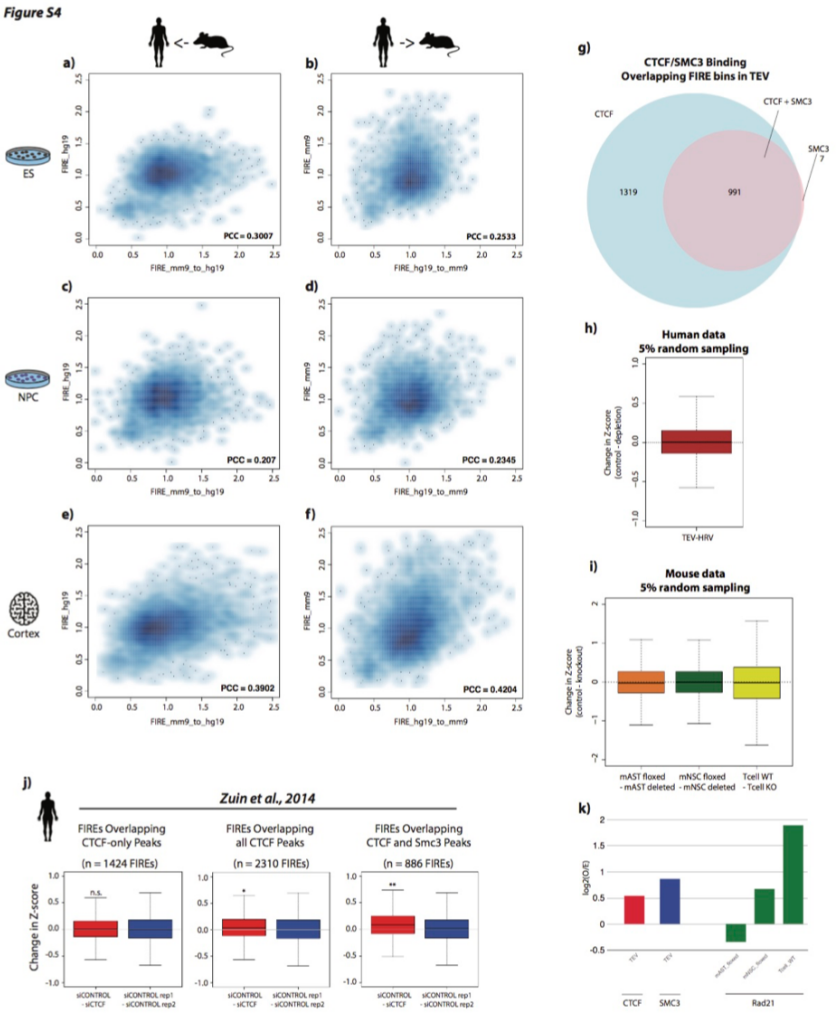


Figure S4
a)
b)
c)
d)
e)
f)
ES
NPC
Cortex
PCC = 0.3007
PCC = 0.2533
PCC = 0.207
PCC = 0.2345
PCC = 0.3902
PCC = 0.4204
FIRE_mm9_to_hg19
FIRE_hg19_to_mm9
g)
CTCF/SMC3 Binding
Overlapping FIRE bins in TEV
CTCF
CTCF + SMC3
SMC3
1319
991
7
h)
Human data
5% random sampling
Change in Z-score
(control - depletion)
TEV-HIV
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(TAE)
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

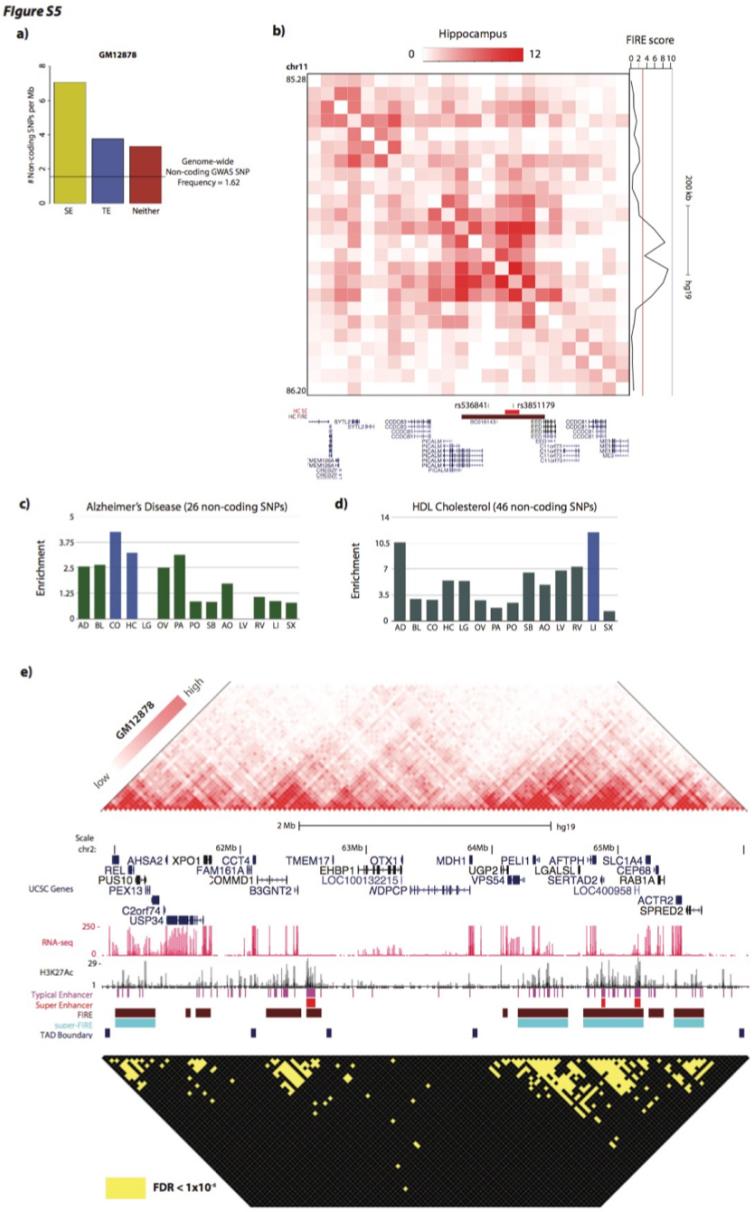

a)
GM12878
Non-coding SNPs per Mb
Genome-wide
Non-coding GWAS SNP
Frequency = 1.62
SE
TE
Neither
b)
Hippocampus
0
12
FIRE score
chr11
85.28
86.20
200 kb
hg19
rs536841
rs3851179
c)
Alzheimer's Disease (26 non-coding SNPs)
Enrichment
AD BL CO HC LG OV PA PO SB AO LV RV LI SX
d)
HDL Cholesterol (46 non-coding 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
REL
PUS10
PEX13
FAM161A
COMMD1
B3GNT2
TMEM17
OTX1
EHBP1
LOC100132215
NDPCP
MDH1
PELI1
AFTPH
SLC1A4
UGP2
VPS54
LGALSL
CEP68
SERTAD2
RAB1A
LOC400958
ACTR2
SPRED2
C2orf74
USP34
RNA-seq
H3K27Ac
Typical Enhancer
Super Enhancer
FIRE
super-FIRE
TAD Boundary
FDR < 1x10⁻⁴

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.

Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in *Cell Reports*. Schmitt, Anthony D.; Hu, Ming; Jung, Inkyung; Xu, Zheng; Qiu, Yunjiang; Tan, Catherine L.; Li, Yun; Barr, Cathy L.; Ren, Bing. The dissertation author was the co-primary investigator and the co-primary author of this material.

Chapter 4

A compendium of promoter-centered long-range chromatin interactions in 27 human tissues and cell types

Introduction

Genome-Wide Association Studies (GWAS) have uncovered thousands of genetic variants that are associated with human diseases and phenotypic traits (*1*). The fact that these variants are generally located in non-coding sequences and are enriched for distal *cis*-regulatory elements (*c*RE) suggests that a substantial fraction of them may contribute to pathogenesis of disease by affecting transcriptional regulation of specific genes (*2, 3*). However, to formally test this hypothesis, it is critical to first identify the target genes of *cis*-regulatory elements. *c*REs modulate transcription of their target genes from a distance through long-range chromatin interactions (*4-7*). Mapping of chromatin architecture by chromosome conformation capture (3C) techniques such as 4C-seq, ChIA-PET and Hi-C (*8-12*) (*13, 14*) (*11, 15*) could uncover long-range interactions between cREs and therefore may reveal promoter-enhancer targeting relationships. Recently, Hi-C combined with capture sequencing has provided a cost-effective way to map chromatin interactions at specific regions at high-resolution (*12, 16-19*). In the current study, we used the capture Hi-C strategy to generate high-resolution maps of promoter-centered chromatin interactions across 27 human tissue/cell types.

Results

We performed Promoter Capture Hi-C (pcHi-C) (Fig. S1) using 280,445 custom-made RNA capture probes to interrogate chromatin interactions centered at 19,539 well annotated human gene promoters across 27 different human tissues or cell types representing a wide spectrum of cell lineages (Fig. S2A-C, Table S1, 2) (*20, 21*). The capture probes synthesis efficiency was highly reproducible between two replicates (Fig. S2D) and covered nearly all targeted promoters (99%) (Fig. S2E). On the other hand, the coverage of capture probes across different target promoters was highly variable (Fig. S2F-G), which can introduce experimental bias to the pcHi-C data. To remove such experimental biases

together with intrinsic sequence biases, we first normalized observed interaction frequencies based on the "capturability" of each DNA fragment using a β-spline regression model (Fig. S3A, see Methods). We then defined significant pcHi-C interactions after removing distance dependent background signals (Weibull p value < 0.01, Table S3) (see Methods).

Analysis of the pcHi-C data resulted in promoter-centered long-range interaction map at DNA fragment resolution (Fig. 1A, see Methods). Based on HindIII restriction sites, 514,738 DNA fragments were defined where 8,698 fragments contain at least one promoter and 126,604 fragments contain at least one putative *c*RE based on H3K27ac signals (see Methods). In total, we identified 561,574 significant pcHi-C interactions across 27 human cell/tissue types (Fig. 1A, Fig. S4A, B). The majority of significant pcHi-C interactions were within 500kb (89%, Fig. S4B) and were significantly enriched for promoter-promoter (P-P, 7.3%, Fisher Exact p value < 2.2e-16) and promoter-*c*RE (P-*c*RE, 36.7%, Fisher Exact p value < 2.2e-16) interactions (Fig. S4C, see Methods) compared to random expectations. We noted that many non-annotated distal regions that interact with promoters were actually marked by diverse transcription factors (*22*) (Fig. S4D, E), suggesting that most promoter-centered long-range interactions are associated with functional elements in the human genome. Interestingly, P-P interactions tend to show shorter interaction distances compared to other interactions (Fig. S4F).

Two independent lines of evidence support the reliability of the identified chromatin interactions. We first compared the results of IMR90 pcHi-C and a previous high-resolution Hi-C dataset from the same cell line (*11*). We found that 90% of promoters showed statistically significant similarity in their long-range interaction profiles between the two datasets (Fig. S5A-C, see Methods). Second, we compared the significant pcHi-C interactions with previous 4C-seq datasets at six loci in the human H1 embryonic stem cells and H1-derived Mesenchynal Stem Cells (MSC) (*10*) and promoter-centered "loops" from IMR90 *in situ* Hi-C results and lymphoblast cells (LCL) (*15*) (see Methods). The pcHi-C results showed high concordance with these orthogonal datasets (Fig. S5D-F). Taken together, our pcHi-C approach is a highly efficient and accurate means to detect to identify significant promoter-centered long-range interactions with low sequencing cost.

Taking advantage of the chromatin and transcriptome datasets collected for these tissue/cell types analyzed by the ENCODE (*22*) and Roadmap Epigenome consortiums (*21*), we next carried out integrative analysis to examine the relationship between the long-range chromatin interactions and chromatin states (*10, 11*). In this analysis we excluded 6 tissue types due to the comparatively low sequencing coverage. Consistent with previous reports (*9, 12*), pcHi-C interactions are often found at active chromatin regions (Fig. 1B). Notably, certain DNA fragments showed extensive long-range interactions with multiple promoters (Fig. S6A). We systematically defined these promiscuously interacting regions from promoter-promoter interaction maps as P-P interaction hotspot (iHS) or from promoter-other interaction maps as P-O iHS (Poisson p value < 0.01, see Methods). For each cell/tissue type we identified around 700~1400 such interaction hotspots (Table S4). According to the classic enhancer-promoter communication model, physical interactions between transcription factors (TFs) bounds at enhancer and promoter regions facilitate enhancer/promoter communication (*23, 24*). As we also observed that long-range promoter-centered interactions are associated with TFs (Fig. S5D, E), we first sought to explore the relationship between the interaction hotspots and TF binding patterns. We examined the TF ChIP-seq data from H1 and GM12878 generated by the ENCODE consortium (*22*) (Table S5-6, see Methods) and found that both P-P and P-O iHS significantly overlap with the TF clusters (Fig. 1C, Fig. 6B-E), which were often found in super enhancer regions and cell-type specific (*25*). As expected, both P-P and P-O iHS are cell/tissue type specific (Fig. 1D and Fig. S6F) and P-O iHS cluster along the germ layers of each tissue/cell type (Fig. 1D, Fig. S6G). Super-enhancers are known to be key regulatory sequences for regulating important cellular identity genes (*25, 26*), and we found that super-enhancers significantly overlap with P-O iHS (Fig. S6H) and also are highly associated with corresponding cell/tissue types (Fig. S6I). For example, a P-O iHS in left ventricle tissue consists of a super-enhancer interacting with multiple promoters, each with high transcriptional activity compared to non-interacting promoters (Fig. 1E). Taken together, our results suggest a strong association between TF clusters/super-enhancers and long-range interaction hotspots and their functional implication on gene regulation.

Identification of functional long-range promoter-cRE interactions is critical to dissect gene regulatory mechanisms. Historically, correlation-based approaches using chromatin state information at

promoters and distal *c*RE have been widely used for this purpose (*27, 28*). Although we observed statistically significant correlation of H3K27ac signals between significant promoter-*c*RE interaction pairs from pcHi-C across cell/tissue types (Fig. S7A, KS-test p value < 2.2e-16) many of them showed modest correlation. Thus, we sought to examine the similarity of promoter-*c*RE pairs identified by either pcHi-C or correlation-based approaches, and also examine the enrichment of functional relationships in pairs defined by each method. When we defined the same number of promoter-*c*RE pairs using H3K27ac correlation we found that only 6% of promoter-*c*RE pairs overlapped the significant promoter-*c*RE interaction pairs based on pcHi-C (Fig. S7B, see Methods). To test which model is more accurate to detect regulatory relationships, we utilized eQTL information obtained from GTEx database (*29*) (see Methods). Several examples illustrate the consistent promoter-*c*RE pairs detected by both eQTL relationships and significant pcHi-C interactions (Fig. 2A, Fig. S7C-E). We systematically assessed the enrichment of eQTL relationships with the matched tissue types between our pcHi-C and GTEx database (*29*), which tissue-type specific information is not available for correlation-based methods, and found that the significant interaction pairs based on pcHi-C are highly enriched in eQTL relationships (Fig. 2B, Fig. S7F). Next, we aggregated all significant interactions from pcHi-C and eQTL relationships (*29*) to compare with the correlation-based method. eQTLs were 6-fold more enriched in promoter-*c*RE interaction pairs based solely on pcHi-C, which is much higher than eQTL enrichment in promoter-*c*RE pairs identified solely based on the correlation-based method (Fig. S8A), indicating that DNA looping information is critical to identify regulatory long-range promoter-*c*RE relationships.

Formation of chromatin interactions is a critical step during transcriptional activation of a gene by distal enhancers (*30, 31*). Since we observed dynamic long-range promoter-*c*RE interactions across cell/tissue types (Fig. 1A, Fig. S8B), we asked to what extent are long-range promoter-*c*RE interactions correlated with variations in gene expression. We focused this analysis on 3,454 testable promoters (Fig. S8C-D, see Methods), which are those covered by at least 4,000 pcHi-C reads in more than 10 cell/tissue types. We found that transcription levels of 66% of gene are positively correlated (PCC > 0.5) with the chromatin interactions profiles between the gene promoter and at least one distal element (Fig. S8E, see Methods). For example, the interaction between *POU3F3* promoter and one cRE showed highly correlated

dynamic patterns between pcHi-C interaction strength (left-hand side Fig. 2C) and *POU3F3* gene expression (right-hand side Fig. 2C). Our results provide support for using long-range chromatin interactions as a tool to infer target genes for distal regulatory elements. We provided a list of significant promoter-cRE interaction pairs identified from pcHi-C (Table S7).

We also found extensive long-range promoter-promoter interactions (21,479 unique P-P interactions) in our datasets (Fig. S4C). Widespread P-P interactions have been observed before in culture mammalian cells, and our results extended the observation to diverse primary human tissues and cell types (*14, 32*). These promoter pairs that exhibit strong interactions also display striking correlative chromatin activities across diverse cell/tissue types (average PCC is 0.41, 0.59, 0.52, and 0.04 for H3K27ac, H3K4me1, H3K4me3, and random permutation) (Fig. 3A-B). For example, dynamic H3K27ac signals at *TMED4* promoter are highly correlated with those at significantly interacting promoters based on our pcHi-C result (Fig. 3A) We also calculated correlation coefficients of transcription levels across 27 cell/tissue types for promoter-promoter pairs defined by our pcHi-C, ChIA-PET (*14*), adjacent promoters, and randomly selected pairs (see Methods). To our surprise, transcription levels between interacting promoter pairs are only weakly correlated (average PCC is 0.15, Fig. 3C,), even lower than adjacent promoter pairs (average PCC is 0.26). Indeed, we often found that non-expressed gene promoters interact with highly expressed gene promoters (Fig. S9A).

We hypothesized that promoter regions can function as enhancer and thus able to regulate distal genes through long-range promoter-promoter interactions. This is in part based on the observation of widespread enhancer signatures or activities at promoter regions (*21, 33*) and enhancer-like function of lncRNA promoters (*34*). We termed these promoter regions as Enhancer-like TSS Proximal eLement (EPL). In support of the functional significance of the EPL, we found that ~11000 eQTLs collected from GTEx database for all available tissue types are in fact P-P interaction pairs (Fig. 3F, Fig. S9B, see Methods). For instance, a significant pcHi-C interaction is found between BLCAP and GHRH gene promoter regions in aorta, and one significant eQTL (rs55705839) of BLCAP gene is located in the *GHRH* promoter in the same tissue type (Fig. 3D). Interestingly this eQTL did not show any meaningful associations with an adjacent downstream gene (*GHRH*) or nearby genes except *BLCAP* (Fig. 3E). There is

no significant association between rs55705839 and *GHRH* in other tissue types according to GTEx database (*29*). Another example is P-P interactions between *POU5F1* (Oct4) gene and two promoters (*CCHCR1* and *TCF19*), which were found to regulate *POU5F1* using a functional screening approach (in submission).

To further test whether EPLs act their function in *cis* to their target genes, we investigated allele-biased chromatin activities in terms of H3K27ac ChIP-seq at EPLs and their interacting promoters. If EPLs act as regulatory sequences, we would expect concordant allelic biases of chromatin activities at EPLs and their interacting promoters. Since the hapolotyes of the genomes analyzed in this study were previously phased (*21*), we defined allelically biased promoter activities using H3K27ac ChIP-seq datasets (see Methods). Around 70% of significant P-P pairs are biased in the same allele, which is very significant concordance rate compared to random expectation (Fig. S9C, *** Empirical P value <0.001, see Methods). For example, in sigmoid colon transcriptionally repressed *VSTM1* gene promoter showed P1 allele biased activity and interacted with active *OSCAR* gene promoter that biased in the same allele (Fig. 3G). Our results provide further support for the enhancer-like function of promoter regions or very promoter proximal regulatory sequences. Our results also provide insight into functional roles of EPLs in regulation of transcription of both the immediate adjacent genes and spatially proximal distal genes.

The promoter-centered long-range interaction maps generated in this study could serves as a resource to infer the target genes of sequences harboring disease-associated sequence variants. For example, a genetic determinant of human obesity is located in the first intron of *FTO* gene, and this genetic determinant affects the distal genes *IRX3* and *IRX5* instead of *FTO* (*6, 7*). The pcHi-C interactions captured these known functional relationships (Fig. 4A). Uncovering disease-disease relationships can greatly advance our understanding of mechanisms underlying human disease (*35, 36*), yet the lack of GWAS-SNP target gene information prevents the estimation of disease associations. However, as our promoter-centered long-range interaction maps can provide putative target genes we sought to explore the relationships between diseases by linking GWAS variants and their target genes. We obtained GWAS SNP information from the GWAS Catalog database (*1*) and expanded the list based on Linkage Disequilibrium (LD) information (>0.8 r^2, see Methods), resulting in 87,433 putative disease-associated genetic variants. Based

on significant pcHi-C interactions in each cell/tissue type we predicted putative target genes of these variants. We found that frequently targeted genes were enriched amongst a set of reported disease-associated genes (Fig. S10A-B, see Methods), supporting the capability of our promoter-centered long-range interaction maps to detect GWAS-SNP target genes.

We further grouped GWAS-SNPs in terms of their 'mapped trait' annotation as GWAS sub-categories (see Methods). After that we identified putative target genes of GWAS-SNPs in each GWAS sub-category. Based on the similarity of these putative target genes between GWAS sub-categories we performed K-means clustering and grouped them into 30 clusters (see Methods). This analysis uncovers several clusters of GWAS sub-categories with interesting biologically relevant features (Fig. 4B). For example, autoimmune related disease (C1 and C28), brain disorders (C2), obesity related phenotypes (C19), and eyes related phenotypes (C22) were grouped together respectively. The recent striking report on the immune basis of Alzheimer's disease (*37*) was also well characterized in our approach (blue boxed region in Fig. 4B). We also revealed a novel association between autoimmune related diseases and cancers (C1 in Fig. 4B), which need to be investigated further. Interestingly, several genes related to immune system such as *HLA-DRB5* and *BTNL2* were frequently recognized as putative GWAS-SNP target genes regardless of disease types, may suggesting the importance of immune system in various diseases. To further understand the related biological functions of GWAS-SNP target genes we carried out gene ontology (GO) analysis (see Methods). For example, putative target genes in C1 were enriched by immune system related signaling pathways and biological functions that are biologically relevant to autoimmune diseases (Fig. 4C). The summarized GO biological functional enrichment result in each cluster defined in Fig. 4B provides both relevant and novel biological insights to corresponding disease and trait types (Fig. 4D).

Lastly, we identified 859 GWAS SNPs that reside promoter regions but interact with distal genes both as eQTLs (*29*) and long-range chromatin interactions. For example, the genetic variant rs12691307 located in *KCTD13* gene promoter regions has been identified as a schizophrenia-associated genetic locus (*38*) (Fig. S10C). However this gene has no specific functional association with schizophrenia so far. In contrast, the distal target gene both as eQTLs and pcHi-C, *DOCA2* functions on calcium-dependent

spontaneous release of neurotransmitter. We found many similar examples (Fig. S10D, E). Traditionally, the GWAS-SNPs located at promoters are assumed to target immediate downstream genes, but our results provide an additional view to interpret GWAS-SNPs with a novel mechanistic insight by considering long-range promoter-promoter interactions.

Discussion

In summary we generated high-resolution promoter-centered long-range interaction maps across diverse human cell/tissue types and provided a resource to understand human disease associated variants. The maps would enable further investigation into the role of distal elements in target gene expression and uncover mechanisms of long-range gene regulation exhibited by both enhancers and promoters. This resource provides a new tool to interpret the function of GWAS-SNPs and dissect gene regulatory networks in human cells.

Methods

Obtaining human tissue samples

Esophagus, lung, liver, pancreas, small bowel, sigmoid colon, thymus, bladder, adrenal gland, aorta, gastric, heart, ovary, psoas, spleen, and fat tissues were obtained from deceased donors at the time of organ procurement at the Barnes-Jewish Hospital (St. Louis, USA) as part of the Epigenome Roadmap Consortium collection (*21*). Samples were flash frozen with liquid nitrogen. The same tissue types from different donors were combined together during downstream data analysis. Human dorsolateral prefrontal cortex (DLPFC) and hippocampus (HC) tissues were obtained from the National Institute of Child Health and Human Development (NICHD) Brain Bank for Developmental Disorders. These two samples were from a healthy single male donor, age 31. Ethics approval was obtained from the University Health Network and The Hospital for Sick Children for use of the tissues.

Hi-C library on human tissue samples and early embryonic cell types

Human tissue samples were flash frozen and pulverized prior to formaldehyde cross-linking. Fibroblasts (IMR90) and lymphoblast (GM12878 and GM19240) cells were cultured and formaldehyde cross-linked with 5 million cells for each Hi-C library. Hi-C was then conducted on the samples as previously described (*40*). Previously constructed Hi-C libraries (*10*) were used for hESC (H1) and early embryonic cell types including mesendoderm, mesenchymal stem cell, neural progenitor cells, and trophoblast-like cells.

Generation of capture RNA probes

In order to perform promoter capture Hi-C we computationally designed RNA probes that capture promoter regions of previously annotated human protein coding genes. Capture regions were selected for 19,704 protein coding genes across 22 autosomes and X chromosome according to GENCODE v19 annotation. For each transcription start site, the two nearest left hand- and right hand-side HindIII restriction sites were selected. Six capture oligos were designed at 120 nucleotide (nt) length and 30nt tiling overhang. Oligos were designed upstream and downstream 300bp adjacent to each restriction site. As two restriction sites were chosen for each transcription start site, in total 12 capture oligos were designed to target each promoter region. Capture sequences overlapped with directly adjacent HindIII restriction sites were removed. GC contents of 94% capture sequences were ranged from 25% to 65%. Since some HindIII fragments contain multiple TSS, 14,508 promoter regions (73%) were uniquely targeted by RNA probes, while the remaining promoters are shared by at least one other promoter in a HindIII fragment. In total, our capture oligo design generated 280,445 unique probe sequences including randomly selected capture regions (i.e. gene deserts). Single-stranded DNA oligos were then synthesized by CustomArray Inc. Single-stranded DNA oligos contained universal forward and reverse primer sequences (total length 31nt), whereby the forward priming sequence contained a truncated SP6 recognition sequence that was completed by the overhanging forward primer during PCR amplification of the oligos. After PCR, double-stranded DNA was converted into biotinylated RNA probes through *in vitro* transcription with the SP6 Megascript kit and in the presence of a biotinylated UTP, as previously described (*41*).

Promoter Capture Hi-C library construction

Promoter Capture Hi-C library was constructed by performing a target enrichment protocol (enriching target promoter centered proximity ligation fragments from Hi-C library using capture RNA probes). Briefly, incubated 500ng Hi-C library 24h at 65 °C within the humidified hybridization chamber with 2.5ug human Cot-1 DNA (Life Technologies), 2.5ug salmon sperm DNA (Life Technologies), and p5/p7 blocking oligos with hybridization buffer mix (10X SSPE, 10mM EDTA, 10X Denhardts solution, and 0.26% SDS) and 500ng RNA probes. Enriched RNA probe hybridized proximity ligation fragments using 50ul T1 streptavidin beads (Invitrogen) with 30min incubation at RT followed by additional 15min incubation in wash buffer1 (1X SSC and 0.1% SDS). Washed three times beads bound DNA fragments with 500ul of pre-warmed (65 °C) wash buffer2 (0.1X SSC and 0.1% SDS), and resuspended in nuclease-free water. Performed qPCR and amplified capture Hi-C library on beads. Purified PCR products with AMPure XP beads followed by sequencing.

Promoter Capture Hi-C library sequencing, read alignment, and off-target read filtering

Promoter Capture Hi-C library sequencing procedures were carried out as described previously according to Illumina HiSeq2500 or HiSeq4000 protocols with minor modifications (Illumina, San Diego, CA). Read pairs from Promoter Capture Hi-C library were independently mapped human genome hg19 using BWA-mem and manually paired with in house script. Unmapped, non-uniquely mapped, and PCR duplicates were removed. Trans-chromosomal read pairs and putative self-ligated products (<15kb read pairs) were removed. Off-target reads were removed when both read pairs did not match to capture probe sequences. The on-target rates in Promoter Capture Hi-C library were ranged from 17% to 44%.

Promoter Capture Hi-C normalization

Interaction frequencies obtained from Promoter Capture Hi-C were normalized in terms of DNA fragment level restricted by HindIII. We defined DNA fragment that spans each HindIII restriction site. The start and end of DNA fragment was defined by taking midpoint of adjacent upstream and downstream restriction site, respectively. We merged adjacent DNA fragments within 3kb. As a result, 514,738 DNA fragments were defined. Median length of DNA fragments was 4.8kb. After that, we calculated raw interaction

frequencies at DNA fragment resolution and performed normalization to remove experimental biases caused by intrinsic DNA sequence biases such as GC contents, mappability, and effective fragment lengths, RNA probe synthesis efficiency bias, and RNA probe hybridization efficiency bias. Although RNA probe synthesis efficiency was highly reproducible between replicates (0.98 Pearson correlation coefficient, Fig. S2D) the coverage of capture probes was highly variable across target regions (Fig. S2F, G). Due to the high complexity of different types of experimental biases, we defined a new term named "Capturability" that means the probability of the region being captured. We assumed that "Capturability" represents all combined experimental biases and can be estimated as a total number of capture reads spanning a given DNA fragment divided by a total number of capture reads. We found that "Capturability" in each DNA fragment is highly reproducible across samples. Therefore, we defined universal "Capturability" as summation of all "Capturability" defined in each sample. The basic idea of our normalization approach is correcting raw interaction frequency using "Capturability" of two DNA fragments. During normalization we processed promoter-promoter interactions and promoter-other interactions, separately because promoter regions tend to show very strong "Capturability" as our capture probes were designed to target promoter regions. Also, we only considered promoter-centered long-range interactions within 2Mb from TSS. Let Y_{ij} represents raw interaction frequency between DNA fragment *i* and *j*. Let C_{i} represents "capturability" defined in DNA fragment *i*. Assume Y_{ij} follows a negative binomial distribution with mean μ and variance $\mu + \alpha\mu^2$. We fitted a negative binomial regression model as follows: $\log \mu_i = \beta_0 + \beta_1 \mathrm{BS}(\mathrm{C_i}) + \beta_2 BS(\mathrm{C_i})$, and defined the residual $\mathrm{R_{ij}} = \mathrm{Y_i} / \exp\left(\hat{\beta}_0 + \hat{\beta}_1 BS(C_{\mathrm{i}}) + \hat{\beta}_2 \mathrm{BS}(\mathrm{C_j})\right)$ as a normalized interaction frequency between DNA fragment *i* and *j*. *BS* represents a basis vector obtained from *B*-spline regression function. The purpose of *B*-spline regression function is dimension reduction during fitting a negative binomial regression model.

Identification of significant chromatin interactions

In order to identify significant pcHi-C chromatin interactions we removed distance dependent background signals from normalized interaction frequencies. Again, we assumed that normalized interaction frequencies $\mathrm{R_{ij}}$ follow a negative binomial distribution with mean μ and variance $\mu + \alpha\mu^2$. As similar to

above interaction frequency normalization step, we calculated expected interaction frequency at a given distance by fitting to a negative binomial regression model with basis vectors obtained from *B*-spline regression of distance between two DNA fragments. Let D_{ij} represents expected interaction frequency at a given distance *d* calculated from a negative binomial regression model. Distance dependent background signals were removed by taking signal to background ratio as follow: $(R_{ij} + avg(R_{ij})) / (D_{ij} + avg(R_{ij}))$. Significant pcHi-C interactions were defined in terms of 0.01 p-value thresholds by fitting normalized interaction frequencies (after removing distance dependent background signals) with 3-paramters Weibull distribution.

Validation of significant pcHi-C interactions in IMR90

The visual inspection of normalized interaction frequencies from IMR90 Promoter Capture Hi-C suggests highly reproducible results compared to high resolution IMR90 Hi-C with only 10% sequencing depth (Fig. S5A). The average correlation coefficient of normalized interaction frequency at upstream and downstream 2Mbp regions of each promoter was 0.57 between IMR90 Hi-C and Promoter Capture Hi-C (Fig. S5B), which is statistically significant compared to randomly permutated data (KS-test p value <2.2e-16). Next, we compared the significant Promoter Capture Hi-C interactions with "loops" identified from *in situ* IMR90 or GM12878 Hi-C experiments (*15*). We only considered "loops" emanating from promoter containing DNA fragments defined in our Promoter Capture Hi-C result.

Functional annotation of DNA fragment

We annotated functional elements to each DNA fragment. If DNA fragment contains at least one annotated protein coding TSS we assigned the fragment as promoter containing DNA fragment. Next, we defined putative distal *cis*-regulatory elements using H3K27ac peaks across all 27 cell and tissue types. We combined these peaks and merged if the peaks are within 3kb each other, resulting in 126,604 putative distal *cis*-regulatory elements. We assigned the *c*RE containing DNA fragment if the DNA fragment contains at least one *cis*-regulatory element. When a DNA fragment contains both TSS and *c*RE we defined

the fragment as a promoter-containing DNA fragment because our data is highly biased to capture promoter regions.

Enrichment of functional elements in promoter interacting regions.

In order to test the enrichment of functional elements in promoter interacting regions, we first calculated the expected coverage by each type of element. 8,698 DNA fragments contain at least one promoter (2.6%), 8,217 DNA fragments contain both promoter and distal cRE (4.2%), 126,604 DNA fragments contain at least one cRE (28.3%) and 371,219 DNA fragments (28.3%) contain neither promoter nor *c*REs. Both promoter and distal *c*RE containing DNA fragments were considered as promoter containing DNA fragment. Fisher-exact test was performed for the statistical test.

Identification of interaction hot spots

We observed that a certain DNA fragment frequently interacts with multiple promoters. To systematically identify such highly interacting regions (i.e interaction hot spots) we first investigate the distribution of the number of interaction frequencies with promoters for each DNA fragment. To minimize experimental biases caused by capturing promoter regions, we conducted our analysis by separating promoter-promoter interactions and promoter-other interactions. For each cell or tissue-type, we selected highly interaction regions in terms of 0.01 p value cutoff after fitting the number of interacting promoters with Poisson distribution. We termed these highly interacting regions as promoter interaction hotspot (P-P iHS) from promoter-promoter interactions and other interaction hotspot (P-O iHS) from promoter-other interactions.

Identification of TF clusters for H1-hESC and GM12878.

Transcription factor ChIP-seq experiments on human lymphoblast (GM12878) and human embryonic stem cell (H1-hESC) by ENCODE were collected. These ChIP-seq reads were aligned against human genome hg19 using BWA-mem with default parameters. We collected only uniquely mapped reads with 10 or greater alignment quality score. Samtools version 1.3 sorted these bam files by coordinate and we removed duplicated reads by Picard. Peak calling of individual ChIP-seq experiments was performed with MACS2

callpeak with default parameters (*42*). We defined TF clusters by calling peaks from combined bed files of TF peaked regions using MACS2 bdgpeakcall. The regions occupied by multiple TF peaks were recognized as TF clusters. To minimize parameter dependent bias, we retrieved TF cluster regions 40 times with various parameter sets as following; minimum # of TFs within cluster (5 or10), minimum length of cluster (2-fold increase from 100bp to 1600bp), maximum gap length within cluster (2-fold increase from 100bp to 51.2kb). Final TF clusters were defined when the region is dected as TF clusters more than 20 times from 40 different parameter set.

Permutation test of TF clusters and super-enhancers

Permutation test was performed to measure how TF clusters are enriched near interaction hotspot. Bedtools shuffleBed generate genomic locations that resemble actual TF clusters with the same size but different genomic coordinate as shuffled random clusters. Bedtools intersectBed identified overlap between interaction hotspots and TF clusters or shuffled random clusters. Average and standard deviation of shuffled random clusters were calculated from 10,000 sets of random data sets. Similarly, enrichment of super-enhancers was conducted by generating random data sets with the same size but different genomic coordinate. The list of super-enhancers were obtained from the author's website (*25*).

k-medoids clustering of interaction hotspot and hierarchical clustering of cell/tissue types based on interaction hotspot

We first collect all P-P or P-O interaction hotspots in each cell/tissue types, respectively as putative interaction hotspot DNA fragments. Then, assigned –log10 (interaction hotspot p value) for each putative interaction hotspot DNA fragment in each cell/tissue type, resulting in an interaction hotspot p value profile where each entry indicates –log10 (p value). Using the interaction hotspot p value profiles, we carried out hierarchical clustering between 21 samples with Pearson correlation metric. To test cell/tissue type specificity of interaction hotspots we conducted k-medoids (k=50) clustering of all putative interaction hotspots using JuliaStats package. After generating 50 k-medoids clusters we manually reorder the clusters in terms of hierarchical clustering result in above.

Correlation-based H3K27ac promoter-enhancer pairs

In order to define promoter-enhancer pairs by using correlation based approaches, we first calculate input normalized RPKM values at enhancer and promoter regions. We take log2(H3K27ac RPKM +1)-log2(input RPKM +1) as the input normalized H3K27ac RPKM values. By collecting H3K27ac ChIP-seq results for all 27 cell/tissue types, each enhancer or promoter region has a 1 by 27 H3K27ac RPKM vector. We computed Pearson correlation coefficient of these vectors between all pairs of enhancer and promoter within 1Mbp. We selected top ranked enhancer-promoter pairs with the same number of significant enhancer-promoter interaction pairs defined by using pcHi-C as a correlation-based H3K27ac promoter-enhancer pairs.

Comparison between eQTL relationships and promoter-other significant chromatin interactions

In order to validate functional enrichment for significant promoter-other pcHi-C interactions we compared significant eQTL relationship for all matched tissue types (n=13, AD, AO, DLPFC, EG, HC, LF, LI, LV, OV, PA, SB, SG, and SX) from GTEx database. We only considered eQTLs that are located in the fragment without a gene and target a gene located in the other fragment. After that, we counted the number of eQTLs that match the significant P-O pcHi-C interactions. For random expectation values, we downloaded all tested eQTLs and randomly selected the same number of significant eQTLs and counted the number of matched eQTLs. Standard deviation of error bars were obtained from 1,000 iteration of random eQTLs.

To compare functional enrichment of promoter-cRE pairs between pcHi-C and correlation-based methods, we combined all eQTL relationships from 13 tissue types used in the above analysis. After that, we calculate the number of matched eQTLs in promoter-cRE pairs by pcHi-C and by a correlation-based method. 6,381 and 2,354 eQTLs were matched to promoter-cRE pairs in pcHi-C and the correlation-based method, respectively. For the random expectation values, we randomly selected the same number of

promoter-cRE pairs and calculate standard deviation of error bars from 100 iterations of random promoter-cRE pairs.

Comparison between eQTL relationships and promoter-promoter significant chromatin interactions

We collected all eQTL relationships from GTEx database with 44 tissue types as putative eQTL relationships. After that, we counted the number of eQTLs that match the significant P-P pcHi-C interactions where we considered eQTLs that resided within 2.5kb from TSS. For random expectation values, we randomly selected P-P pairs within 1Mbp as the same number of significant P-P chromatin interactions and then counted the number of eQTLs that match the random P-P pairs. Average and standard deviation values were calculated after performing random selection of P-P pairs 1,000 times. To ensure the result is not biased depending on distance between P-P pairs we also tested randomly selected P-P pairs within 500kb and 100kbp , but found that the significant P-P pcHi-C pairs are always significantly matched with eQTL relationships compared to randomly selected P-P pairs.

Linking between dynamic long-range P-O interactions and variations in gene expression

We linked dynamic long-range P-O interaction to variations in gene expression. We first collected testable promoter regions because of different sequencing depth across samples or capture probe density across promoter regions. Based on P-O pcHi-C interaction profiles between GM12878 two biological replicates, we found that the reproducibility of the pcHi-C interaction profiles is affected by the coverage of reads at promoter regions. We found that 4,000 is a minimum number of reads spanning at promoters to fairly compare P-O interaction profiles between samples. We collected testable promoters when the promoter is covered by more than 4,000 reads at least 10 cell/tissue types, resulting in 3,454 testable promoters. For those testable promoter regions, we filtered again based on gene expression variations. We defined the gene expression is variable when the maximum and minimum FPKM values of the gene show more than 2-fold difference, resulting in 2,903 testable promoters. After that we computed Pearson correlation coefficient (PCC) between pcHi-C normalized interaction frequencies and variations in gene expressions (FPKM) for a given promoter. The pcHi-C normalized interaction frequencies represent chromatin interactions between

the promoter and another DNA fragment within 2Mbp from the TSS. We only considered P-O pairs showing over 0.5 absolute PCC value. If the variations of gene expressions are only positively correlated (>0.5 PCC) with other fragments, we defined the gene is linked by dynamic long-range interactions positively (n=1,250). If the variations of gene expression is only negatively correlated (>0.5 PCC) with other fragments, we defined the gene is linked by dynamic long-range interactions negatively (n=276). If the variations of gene expressions show both positively and negatively correlation with dynamic long-range interactions, we defined the gene is linked by dynamic long-range interactions both positively and negatively (n=661).

Linking between allele biased promoter activities and significant P-P chromatin interactions

In order to support *cis*-regulatory function of promoters on distal gene regulation, we utilized allelically biased promoter activity information. We first defined allelelically biased promoters using very deeply sequenced paired-end H3K27ac ChIP-seq data for matched 14 tissue types (AD, AO, EG, GA, LG, LV, OV, PA, PO, RA, RV, SG, SX, and TH). We denoted one allele as P1 and another allele as P2. We obtained haplotype-resolved these ChIP-seq data from our previous study (*21*) and collected testable promoter regions when the promoter was covered more than 15 allele specific reads, resulting in 131,535 testable promoters. Then we calculated binomial p value between P1 and P2 alleles and defined allele biased promoter activity in terms of 0.05 FDR cutoff (n=10,844). For each tissue type, we collected testable significant P-P pcHi-C interactions where both promoters are allelically biased. Without any restriction of distance information we found 46 testable significant P-P pcHi-C interaction pairs and among them 61% of pairs were concordantly biased in the same allele. When we restrict the distance between P-P pairs as less than 100kb, we found 32 testable significant P-P pcHi-C interaction pairs and among them 72% of pairs were concordantly biased in the same allele. For the random expectation, we randomly assigned the biased allele for testable promoters and iterated this procedure 1,000 times.

Extended GWAS-SNPs list with LD information.

As GWAS-SNPs obtained from GWAS catalogue database contain tag SNP information only, we extended the GWAS-SNP information using linkage disequilibrium (LD) structure. LD scores were calculated using PLINK for five different populations obtained from 1000 genome phase 3 data. For each tag SNP we included all associated SNPs showing tight LD score (>0.8) for all five populations (AFR, AMR, EAS, EUR, and SAS).

Enrichment test of disease genes in putative GWAS-SNP target genes.

In order to test the enrichment of disease associated genes in GWAS-SNP putative target genes we first downloaded the list of putative disease associated genes from GeneCard database, resulting in 9,989 disease associated genes. Then, for each cell/tissue type we defined putative target genes of GWAS-SNPs based on significant pcHi-C interactions in each cell/tissue type. To remove false putative target genes, we only considered putative target genes that frequently interact with multiple GWAS-SNPs. We used top 20 frequently targeted genes as high confident putative target genes and calculated the ratio between disease-associated genes and other genes as a measure of disease gene enrichment.

Clustering of GWAS sub categories based on putative target genes

Based on "mapped traits" from GWAS catalog database, we first grouped GWAS-SNPs into sub categories. For each sub categories we defined putative target genes by aggregating all unique significant pcHi-C interactions. We only considered sub categories containing more than 5 putative target genes, resulting in 907 sub categories. Based on the frequency of putative target genes in each sub category we calculated PCC between sub categories, resulting in 907 by 907 symmetric PCC matrix. We performed K-mean clustering (n=30) for this matrix. 367 sub categories were grouped into 29 clusters, but rest of them was not grouped well. We only focused on these 29 well-clustered groups during downstream analysis.

Analysis of functional enrichment using DAVID

We use DAVID 6.8 Beta version to perform the functional enrichment test. We use all human genes as background and we select UP_TISSUE, KEGG_PATHWAY, and GO_BP as functional annotations.

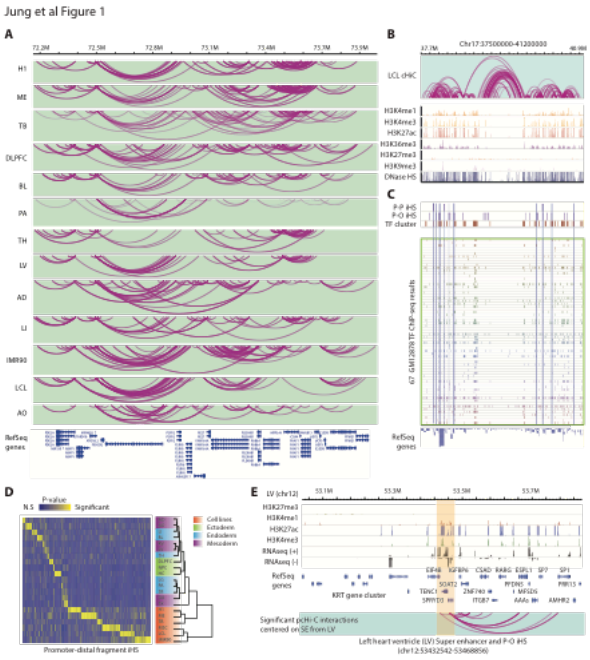


Figure 4.1. Mapping long-range promoter-centered chromatin interactions on 27 human tissue and cell types.

(**A**) Significant interactions identified from pcHi-C across multiple cell/tissue types, with the darkness of the purple corresponding to the strength of the interactions. RefSeq genes are presented together. (**B**) Depiction of identified long-range promoter-centered interactions across a 3.7 Mb locus in lymphoblast cells (LCL) (top). Shown below are histone modification signals obtained from ChIP-seq analyses (*39*) as well as accessible chromatin regions measured from DNaseI hypersensitivity assay (**C**) Depiction of Promoter-Promoter interaction hotspots (P-P iHS), Promoter-Other interaction hotspots (P-O iHS), and transcription factor (TF) clusters identified in GM12878 cells for the same region shown in Fig. 1B. Below are 67 TF binding profiles obtained form ChIP-seq analysis, and RefSeq genes. Highlighted in translucent blue are overlapping iHS and TF clusters. (**D**) Heatmap showing cell/tissue-type specific P-O iHS. Each row represents a distinct cell or tissue type (n=21), and each column is a unique P-O iHS (n=3,951). The color bar ranges from non-statistical significance (N.S) to high significance. The dendrogram (right-hand side) is based on hierarchical clustering of P-O iHS similarities between cell and tissue types. Cell and tissue types are colored based on their developmental origin, or cell line status as indicated. (**E**) Snapshot illustrating the promiscuous interaction profile of a Left Ventricle (LV) super-enhancer, overlapping with P-O iHS. The top rows depict histone modification signals as measure by ChIP-seq, followed by transcriptional levels measure RNA-seq in LV. Below are RefSeq genes, and then observed interactions emanating from the LV super-enhancer (highlighted in translucent orange). Interactions detected by pcHi-C are colored in purple, with the darkness of the purple corresponding to the strength of the interactions.

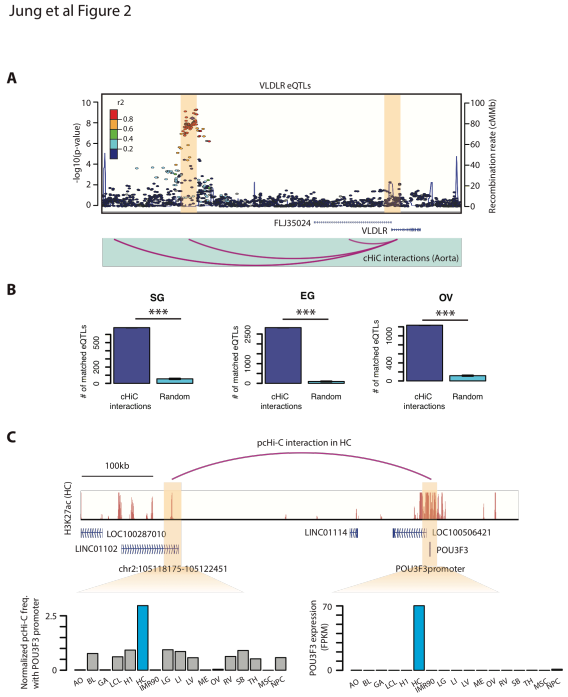


Figure 4.2. Long-range promoter-distal cRE interactions are enriched for functional relationships.

(A) Illustrative Locus zoom plot of eQTLs for *VLDLR* (top) and significant pcHi-C interactions emanating from the *VLDLR* promoter region in Aorta tissue (bottom). Dots along the locus zoom plot represent SNPs, and their significance effect on *VLDLR* gene expression levels is plotted along the left y-axis. Dots are also color-coded based on their Linkage Disequilibrium (LD) scores with a tag SNP. The blue line traveling across the scatterplot indicates the recombination rate, as plotted along the right y-axis. (**B**) Barplots showing the number of matched eQTLs with significant P-O pcHi-C interactions and the number of matched randomly selected eQTLs with significant P-O pcHi-C interactions for sigmoid colon (SG), esophagus (EG), and ovary (OV). Fisher-exact test was performed for statistical significant (*** < p value 10e-3). (**C**) Illustrative example of dynamic gene expression showing positive correlation with changes in long-range promoter-*c*RE interaction frequency. The top interaction shows the significant interaction between the *POU3F3* promoter and a distal *c*RE ~350kb upstream in hippocampus tissue (HC). The bar plot below and left shows the normalized pcHi-C interaction frequencies between *POU3F3* promoter and the distal *c*RE and the right shows the gene expression levels of *POU3F3*, highlighted in blue.

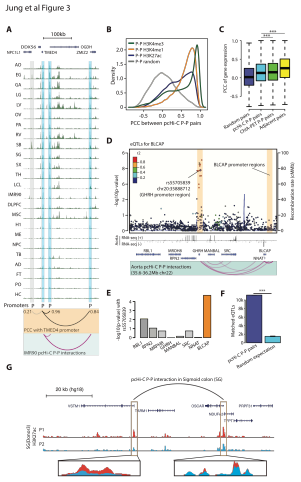


Figure 4.3. Widespread promoter-promoter interactions in distal gene regulation.

(**A**) Browser snapshot of the *TMED4* locus, showing promoter-promoter interactions that correspond to correlated H3K27ac signals at promoters. Shown and the top are RefSeq genes, and below are H3K27ac histone modification signals as measured by ChIP-seq. Highlighted by blue boxes are promoters who are linked both by highly correlated H3K27ac signal and by significant pcHi-C interactions. A highlighted by gray box is an adjacent promoter of TMED4. Below, are PCC values and links based on H3K27ac signal and links based on pcHi-C. (**B**) Density plots showing the PCC distributions of H3K27ac (blue), H3K4me1 (orange), and H3K4me3 (green) signals for promoter-promoter pairs exhibiting significant DNA looping interactions detected by pcHi-C. A density plot showing the PCC distributions of H3K27ac signals for randomly selected promoter-promoter pairs (gray). X-axis indicates PCC of histone modification signals between promoter-promoter pairs across 27 cell/tissue types. (**C**) Box plot showing the PCC distributions of gene FPKM signals for randomly selected promoter-promoter pairs (dark blue) compared to promoter-promoter pairs exhibiting significant pcHi-C interactions (blue), promoter-promoter pairs defined by ChIA-PET (green), and adjacent promoter-promoter pairs within 20kb (yellow). KS-test was performed for statistical significance (*** < p value 0.005). (**D**) Illustrative LocusZoom plot of eQTLs for *BLCAP* gene expression in Aorta. Both *BLACP* gene promoter region and *GHRH* promoter that contain significant eQTLs are highlighted in translucent orange, Dots along the LocusZoom plot represent SNPs, and their significance of association with *BLCAP* gene expression is plotted along the left y-axis. Dots are also color-coded based on their LD score with a tag SNP (rs55705839). The blue line traveling across the LocusZoom plot indicates the estimated recombination rate, as plotted along the right y-axis. Gene expression levels detected by RNA-seq and RefSeq genes position are plotted below the LocusZoom plot. (**E**) Bar plots showing the eQTL association of genes with the SNP rs55705839, with the most significant association with the distal gene, *BLCAP*. Y-axis indicates –log10 eQTL p values. (**F**) Bar plots showing the absolute number of eQTLs that are matched with promoter-promoter interactions identified from pcHi-C data (darkblue) and randomly selected promoter-promoter pairs (blue). Error bar indicates standard deviation from 1,000 random data sets. Fisher-exact test was performed for statistical significance (*** p value < 2.2e-16). (**G**) Illustrative example of concordant allelically-biased H3K27ac signals at promoters which are linked through a significant pcHi-C interactions in SG tissue. RefSeq genes are shown at the top, followed by allelically mapped H2K27ac signal (P1 and P2). Highlighted in orange boxes are the promoters for *VSTM1* and *OSCAR*, which show allelically biased H3K27ac signals towards the P1 allele (red).

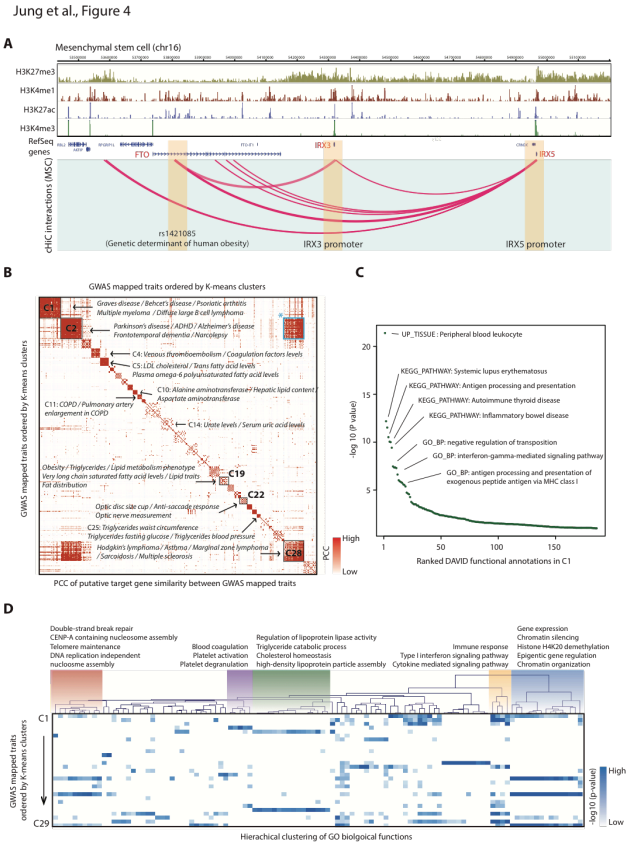


Figure 4.4. Putative target genes of GWAS SNPs linked by promoter-centered long-range chromatin interactions.

(**A**) Browser snapshot of the *FTO/IRX3/IRX5* locus in mesenchymal stem cells (MSC). Highlighted in orange boxes is the intronic enhancer in *FTO* bearing the genetic determinant for human obesity, rs1421085, as well as the promoters for *IRX3* and *IRX5*. The top tracks show histone modification signals obtained from ChIP-seq analysis. Below are RefSeq genes, follow by all significant long-range interactions originating from the promoters of *IRX3* or *IRX5*. (**B**) Shown are K-means clustering (n = 29) results of disease-disease associations using putative GWAS-SNP target gene similarities. Each dot indicates PCC of the target gene similarities between GWAS sub-categories. Clusters are shown along the arbitrary order of K-means clusters from cluster 1 to cluster 29 as from top to bottom and from left to right. Representative diseases or traits are shown together for several clusters. (**C**) Gene ontology analysis of putative target genes in cluster1 using DAVID. GO terms are presented according to p-values (green dots). UP_TISSUE is for up-regulated tissue type, GO_BP is for GO biological process. P values of corresponding GO terms using nearest gene information are shown as gray dots. (**D**) Hierarchical clustering of GO biological processes (each column, n=109) across the K-means clusters (each row, n=29). Each entry indicate – log10(p) value of GO biological processes in the corresponding cluster. Several representative biological processes are described together.

References:

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42, D1001-6 (2014).

2. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R. & Stamatoyannopoulos, J.A. in Science 1190-1195 (2012).

3. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. & Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America* 106, 9362-7 (2009).

4. Lettice, L.A., Horikoshi, T., Heaney, S.J., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., Shibata, M., Suzuki, M., Takahashi, E., Shinka, T., Nakahori, Y., Ayusawa, D., Nakabayashi, K., Scherer, S.W., Heutink, P., Hill, R.E. & Noji, S. Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci U S A* 99, 7548-53 (2002).

5. Uslu, V.V., Petretich, M., Ruf, S., Langenfeld, K., Fonseca, N.A., Marioni, J.C. & Spitz, F. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nat Genet* 46, 753-8 (2014).

6. Claussnitzer, M., Dankel, S.N., Kim, K.H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., Abdennur, N.A., Liu, J., Svensson, P.A., Hsu, Y.H., Drucker, D.J., Mellgren, G., Hui, C.C., Hauner, H. & Kellis, M. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* 373, 895-907 (2015).

7. Smemo, S., Tena, J.J., Kim, K.H., Gamazon, E.R., Sakabe, N.J., Gomez-Marin, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., Lee, J.H., Puviindran, V., Tam, D., Shen, M., Son, J.E., Vakili, N.A., Sung, H.K., Naranjo, S., Acemel, R.D., Manzanares, M., Nagy, A., Cox, N.J., Hui, C.C., Gomez-Skarmeta, J.L. & Nobrega, M.A. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507, 371-5 (2014).

8. de Wit, E., Bouwman, B.A., Zhu, Y., Klous, P., Splinter, E., Verstegen, M.J., Krijger, P.H., Festuccia, N., Nora, E.P., Welling, M., Heard, E., Geijsen, N., Poot, R.A., Chambers, I. & de Laat, W. The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* 501, 227-31 (2013).

9. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* 489, 109-13 (2012).

10. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V.V. & Ecker, J.R. Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331-336 (2015).

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

12. Sahlen, P., Abdullayev, I., Ramskold, D., Matskova, L., Rilakovic, N., Lotstedt, B., Albert, T.J., Lundeberg, J. & Sandberg, R. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol* 16, 156 (2015).

13. Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S.Z., Penrad-Mobayed, M., Sachs, L.M., Ruan, X., Wei, C.L., Liu, E.T., Wilczynski, G.M., Plewczynski, D., Li, G. & Ruan, Y. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* 163, 1611-27 (2015).

14. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., Sim, H.S., Peh, S.Q., Mulawadi, F.H., Ong, C.T., Orlov, Y.L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K.I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M.J., Cheung, E., Liu, E., Sung, W.K., Snyder, M. & Ruan, Y. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148, 84-98 (2012).

15. 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. & Aiden, E.L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665-80 (2014).

16. Jager, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H.E., Heindl, A., Whiffin, N., Carnicer, M.J., Broome, L., Dryden, N., Nagano, T., Schoenfelder, S., Enge, M., Yuan, Y., Taipale, J., Fraser, P., Fletcher, O. & Houlston, R.S. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* 6, 6178 (2015).

17. Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., Herman, B., Happe, S., Higgs, A., LeProust, E., Follows, G.A., Fraser, P., Luscombe, N.M. & Osborne, C.S. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* 47, 598-606 (2015).

18. Dryden, N.H., Broome, L.R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., Nagano, T., Andrews, S., Wingett, S., Kozarewa, I., Assiotis, I., Fenwick, K., Maguire, S.L., Campbell, J., Natrajan, R., Lambros, M., Perrakis, E., Ashworth, A., Fraser, P. & Fletcher, O. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res* (2014).

19. Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S., Cooper, N.J., Barton, A., Wallace, C., Fraser, P., Worthington, J. & Eyre, S. Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat Commun* 6, 10069 (2015).

20. Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., Yang, H., Wang, T., Lee, A.Y., Swanson, S.a., Zhang, J., Zhu, Y., Kim, A., Nery, J.R., Urich, M.a., Kuan, S., Yen, C.A., Klugman, S., Yu, P., Suknuntha, K., Propson, N.E., Chen, H., Edsall, L.E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.Y., Chi, N.C., Antosiewicz-Bourget, J.E., Slukvin, I., Stewart, R., Zhang, M.Q., Wang, W., Thomson, J.a., Ecker, J.R. & Ren, B. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* 153, 1134-1148 (2013).

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

22. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74 (2012).

23. Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E.M. & Pennacchio, L.A. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457, 854-858 (2009).

24. Gorkin, D.U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell* 14, 762-775 (2014).

25. Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I. & Young, R.A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153, 307-19 (2013).

26. Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A. & Young, R.A. Super-enhancers in the control of cell identity and disease. *Cell* 155, 934-47 (2013).

27. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E. & Stamatoyannopoulos, J.A. The accessible chromatin landscape of the human genome. *Nature* 489, 75-82 (2012).

28. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. *Nature* 488, 116-20 (2012).

29. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648-60 (2015).

30. Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A. & Blobel, G.A. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149, 1233-44 (2012).

31. Xu, J., Sankaran, V.G., Ni, M., Menne, T.F., Puram, R.V., Kim, W. & Orkin, S.H. Transcriptional silencing of {gamma}-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes Dev* 24, 783-98 (2010).

32. Zhang, Y., Wong, C.H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., Mulawadi, F.H., Sung, W.K., Nicolis, S., Ahituv, N., Ruan, Y. & Wei, C.L. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* 504, 306-10 (2013).

33. Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M. & Stark, A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074-7 (2013).

34. Paralkar, V.R., Taborda, C.C., Huang, P., Yao, Y., Kossenkov, A.V., Prasad, R., Luan, J., Davies, J.O., Hughes, J.R., Hardison, R.C., Blobel, G.A. & Weiss, M.J. Unlinking an lncRNA from Its Associated cis Element. *Mol Cell* 62, 104-10 (2016).

35. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.R., ReproGen, C., Psychiatric Genomics, C., Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case

Control, C., Duncan, L., Perry, J.R., Patterson, N., Robinson, E.B., Daly, M.J., Price, A.L. & Neale, B.M. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47, 1236-41 (2015).

36. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J. & Barabasi, A.L. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347, 1257601 (2015).

37. Gjoneska, E., Pfenning, A.R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.H. & Kellis, M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* 518, 365-9 (2015).

38. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421-7 (2014).

39. Consortium, R.E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Wang, J., Ward, L.D., Sarkar, A., Quon, G., Kheradpour, P., Coarfa, C., Harris, R.A., Ziller, M.J., Schultz, M., Eaton, M.L., Pfenning, A., Wang, X., Polak, P., Karlic, R., Amin, V., Wu, Y.-c., Sandstrom, R.S., Ray, P., Wu, J., Kulkarni, A., Lister, R., Hong, C., Gascard, P., Carles, A., Mungall, A.J., Moore, R., Chau, E., Tam, A., Zhou, X., Li, D., Zhang, B., Mercer, T.R., Neph, S.J., Siebenthall, K.T., Thurman, R.E., Canfield, T., Hansen, R.S., Kaul, R., Sabo, P.J., Beaudet, A.E., Boyer, L., Jager, P.D. & Peggy, J. Integrative analysis of 111 reference human epigenomes. *Nature* (2015).

40. 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., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S. & Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289-93 (2009).

41. Selvaraj, S., Schmitt, A.D., Dixon, J.R. & Ren, B. Complete haplotype phasing of the MHC and KIR loci with targeted HaploSeq. *BMC Genomics* 16, 900 (2015).

42. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., Nussbaum, C., Myers, R.M., Brown, M., Li, W. & Liu, X.S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137 (2008).

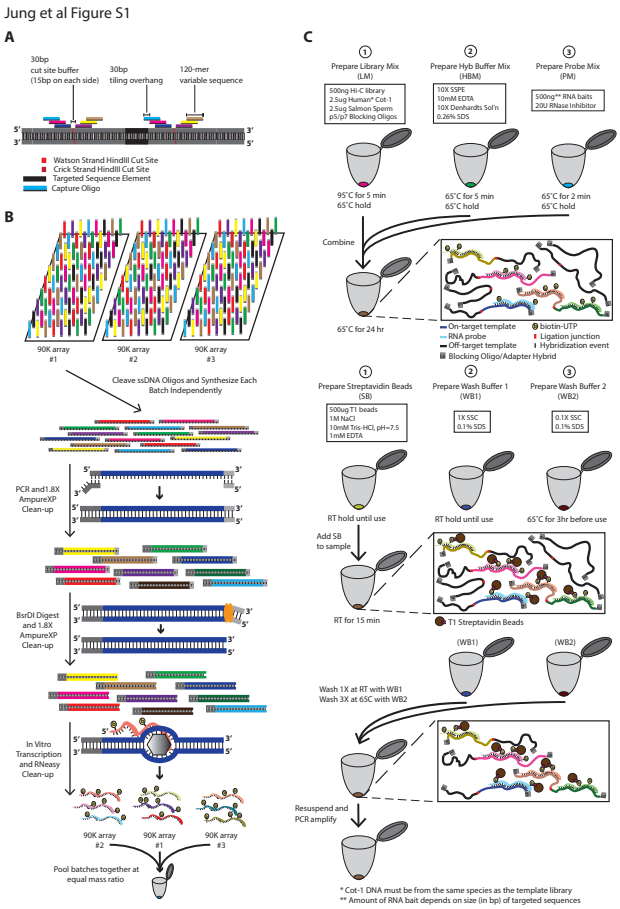


Figure S4.1. Capture Hi-C design, probe synthesis, and target enrichment workflow.

(A) Schematic of probe design for promoter Capture-HiC experiments. For each promoter (black rectangle), two flanking HindIII cut sites were identified. A 15bp buffer was then added to each side of the HindIII cut site, and then 3 120-mer capture probes were allocated to each side of the HindIII cut site, with a 30bp shift between adjacent probes. In total, 12 capture probes were assigned to each promoter, and all probes were targeted towards the Watson Strand cut site. (**B**) Schematic workflow of custom RNA probe synthesis. From top to bottom, ssDNA probe synthesis by CustomArray, Inc, PCR amplification with SP6 recognition sequence completion and purification, BsrDI digest and purification, *in vitro* transcription in the presence of biotinylated UTP and purification, and pooling of probe batches using equal mass ratios. (**C**) Schematic workflow of target enrichment of Hi-C libraries (Promoter Capture-Hi-C). From top to bottom, preparation of library mix, hybridization buffer, and probe mix, following by combining the mixes and overnight incubation to bind probes to Hi-C template. Then, preparation of streptavidin beads and wash buffers. Then, binding RNA:DNA duplexes to streptavidin beads and rigorous washing to remove off-target binding. And lastly, PCR amplification of the resulting Promoter Capture-Hi-C library.

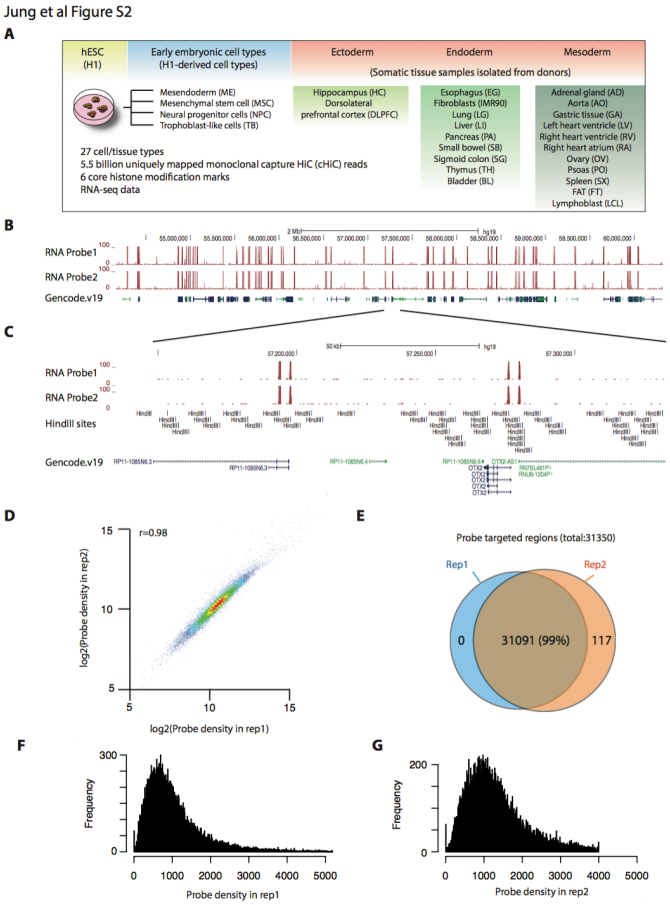


Figure S4.2. Overview of samples and datasets and capture probe quality control.

(**A**) Schematic overview of the cell and tissue types analyzed by Promoter Capture-Hi-C, and note of other datasets available for these samples. Embryonic or embryonic-derived cell types are on the left, and tissues are tabled on the right according to their developmental origin. (**B**) Snapshot of Promoter Capture-Hi-C probe density from RNA-seq analysis of the capture probes. Two replicates of probe synthesis and subsequent RNA-seq are shown, followed by Gencode gene annotations. (**C**) Zoomed-in snapshot of Promoter Capture-Hi-C probe density from RNA-seq analysis of the capture probes. Below the replicate RNA-seq datasets are the HindIII cut sites and Gencode gene annotations, illustrating that the vast majority of probe density is only found around HindIII cut sites flanking promoters. (**D**) Scatter plot showing the reproducibility of probe density from RNA-seq data across two probe synthesis experiments. Each dot on the scatter plot represents a single promoter, and the value is the aggregate probe density from all probes assigned to that given promoter. (**E**) Venn diagram showing the number of targeted regions that contain detectable probe density based on RNA-sequencing of the capture probes from each replicate of probe synthesis. (**F-G**) Histogram of the probe densities measured by RNA-seq (x-axis) in each promoter from replicate 1 (F) and replicate 2 (G) of probe synthesis.

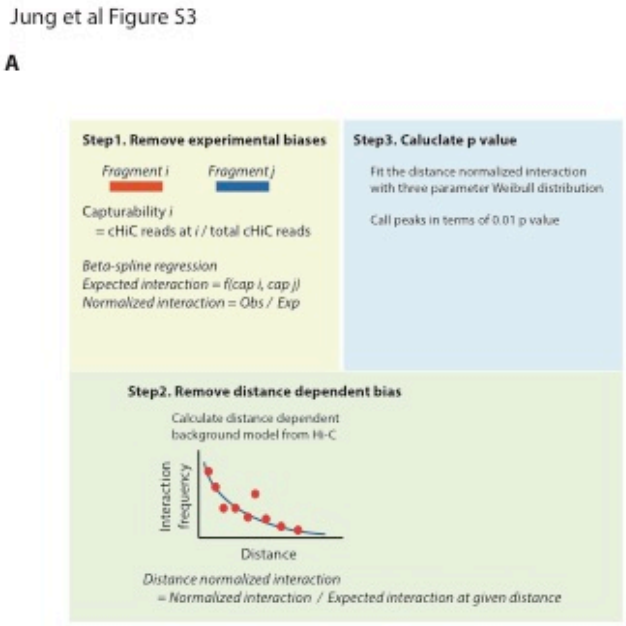


Figure S4.3. Identification of significant Promoter Capture-Hi-C interactions.

(**A**) 3-step procedure for identifying significant interactions in Promoter Capture-Hi-C data.

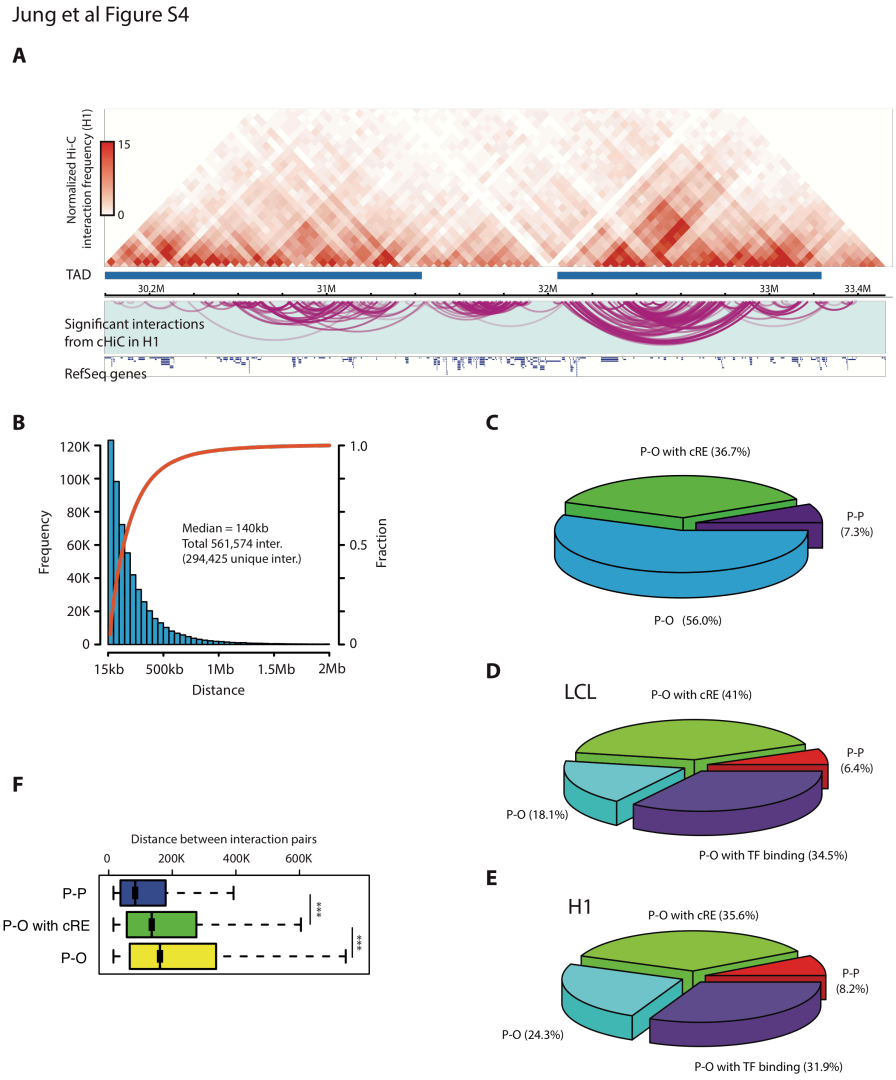


Figure S4.4. General characterization of promoter-centered long-range interactions.

(**A**) Snapshot of a locus showing promoter-centered long-range interactions from Promoter Capture-Hi-C data in H1 cells (bottom, purple loops) in the context of TADs (blue rectangles) detected from Hi-C data (top, red) in H1 cells. RefSeq genes are shown at the bottom. (**B**) Histogram showing the distribution of genomic distances of promoter-centered long-range interactions. The cumulative fraction of promoter-centered long-range interactions is plotted as a red line, and corresponds to values plotted along the right y-axis. (**C**) Pie chart showing the classification of all unique significant pcHi-C interactions obtained from all tissues and cell types. P-P corresponds to promoter-promoter interactions; P-O corresponds to promoter interactions with non-promoters. P-O class of interactions has been sub-divided to P-O with cRE and rest of P-O. (**D-E**) Pie chart showing the classification of all unique significant pcHi-C interactions from Promoter Capture-Hi-C in LCL (D) and H1 (E). The rest of P-O class of interactions has been sub-divided again to show P-O interactions that are also TF binding sites.

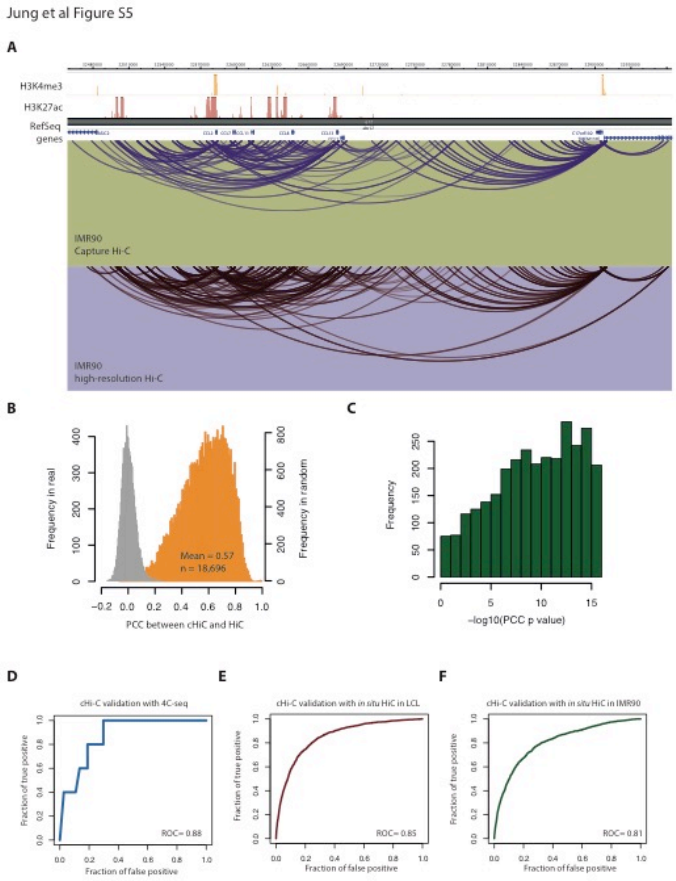


Figure S4.5. Validation of Promoter Capture-Hi-C.

(**A**) Browser snapshot of the CCL gene cluster, highlighting the similarity of promoter-centered interactions from Promoter Capture-Hi-C and high resolution Hi-C data in IMR90 cells. The top two tracks show histone modification signals for H3K4me3 and H3K27ac, obtained from ChIP-seq data, and RefSeq genes. Below, are promoter-centered DNA looping interactions obtained from Promoter Capture-Hi-C in IMR90 cells (blue loops), and promoter-centered DNA looping interactions from high-resolution Hi-C data in IMR90 (purple loops). (**B**) Overlapping histograms, showing the PCC of promoter centered interaction profiles between Promoter Capture-Hi-C and high resolution Hi-C data in IMR90 cells. Each data point in orange color represent the PCC of the interaction profile of a single promoter between the two datasets. In gray, each data point represents the PCC of interaction profiles of randomly selected promoter-pairs between the two datasets. (**C**) Histogram showing distribution of PCC p values of interaction profiles of a single promoter between the two datasets. (**D-F**) ROC plots showing the prediction performance of Promoter Capture Hi-C result compared to 4C-seq (**D**), *in situ* Hi-C loops anchored at promoters lymphoblast cells (**E**), and *in situ* Hi-C loops anchored at promoters in IMR90 (**F**).

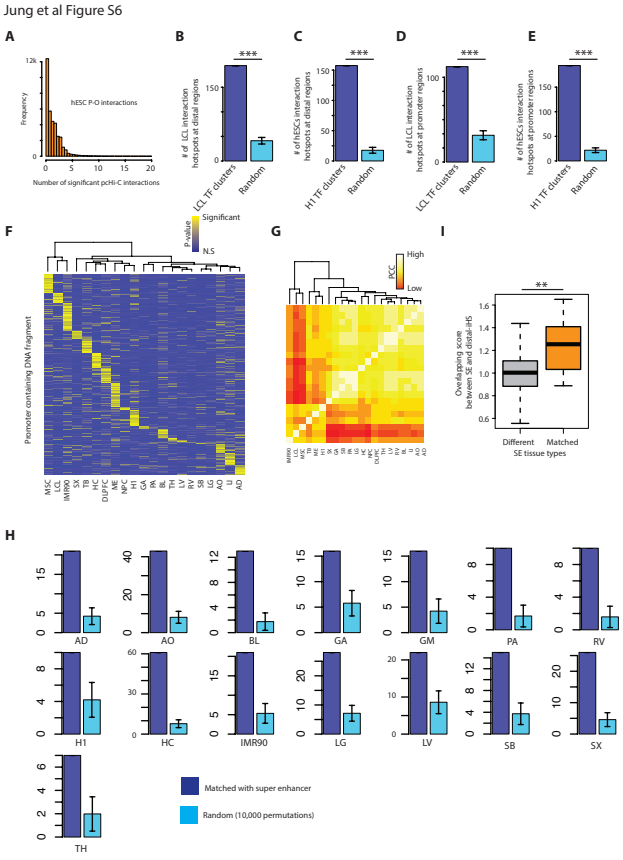


Figure S4.6. Characterization of interaction hotspots (iHS).

(**A**) Histogram showing the distribution of number of interacting promoters for each DNA fragment in H1. (**B-C**) Bar plots showing the number of P-O iHS in lymphoblast cells (LCL) (B) or H1 (C) overlapping with TF clusters compared to random expectation. Fisher-exact test was performed for statistical significance (*** p value < 2.2e-16). (**D-E**) Bar plots showing the number of P-P iHS in lymphoblast cells (LCL) (D) or H1 (E) overlapping with TF clusters compared to random expectation. Fisher-exact test was performed for statistical significance (*** p value < 2.2e-16). (**F**) Heatmap showing cell/tissue-type specific P-P iHS. Each column represents a distinct cell or tissue type, and each row is a putative P-P iHS. The color bar ranges from low statistical significance, to high significance of P-P iHS p-value. The above dendrogram is clustered using the hierarchal clustering based on PCC of P-P iHS similarity between samples. (**G**) Heatmap showing PCC of P-O iHS similarities. The above dendrogram is clustered using the hierarchal clustering based on the PCC between samples. (**H**) An array of bar plots showing the number of P-O iHS overlapping with super-enhancers (left, purple), compared to random expectation (right, blue). Each bar plots represents an analysis of a different cell or tissue, depending on which cells/tissues have super-enhancer annotations. (**I**) Box plots showing the overlapping score between super-enhancers and P-O iHS when the super-enhancer annotation set and iHS set are from the same tissue ('matched', orange), or from different tissues ('different', gray). KS-test was performed for statistical significance (** p value < 0.01).

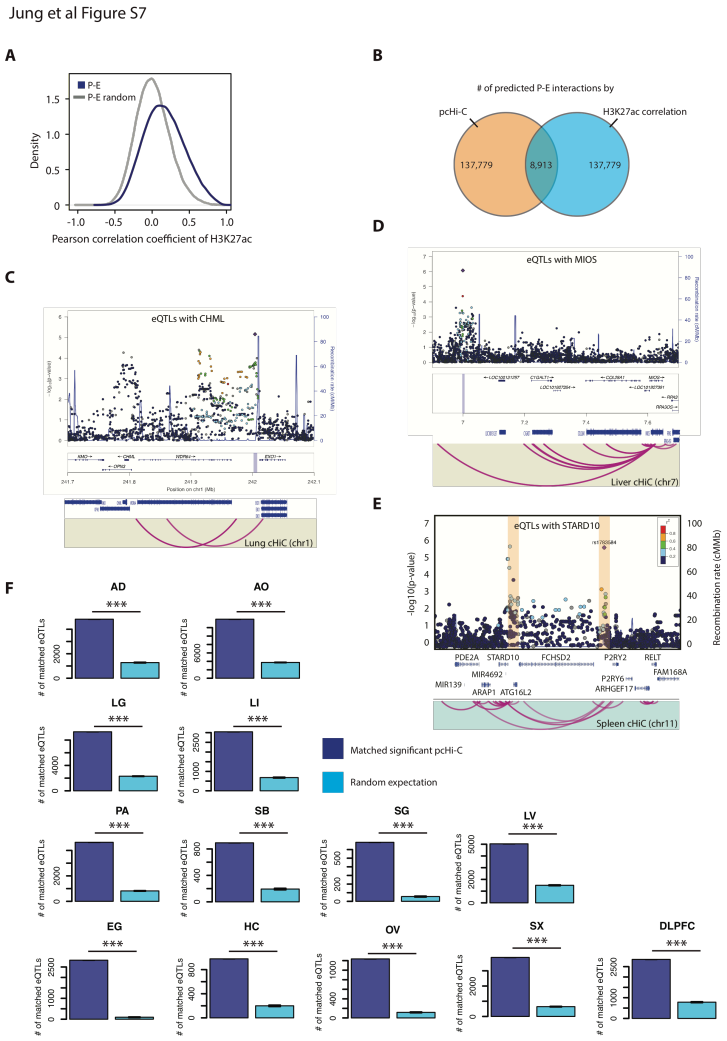


Figure S4.7. Enrichment of eQTL relationships in significant P-cRE interactions.

(**A**) Density plots showing the distribution of PCC for H3K27ac signals at true P-cRE pairs identified from Promoter Capture-Hi-C (blue), compared to expected distributions of PCC values (gray). (**B**) Venn diagram illustrating the amount of overlapping of P-cRE pairs identified from Promoter Capture-Hi-C data (orange) and P-cRE pairs identified using correlation-based approaches (blue). **(C-E)** Illustrative locus zoom plots of eQTLs for *CHML* (C), *MIOS* (D), and *STARD10* (E) gene expression in lung, liver, and spleen, respectively. RefSeq genes position is plotted below the locus zoom plot. Significant Promoter Capture Hi-C are shown as purple in the bottom. (**F**) Array of bar plots showing number of matched eQTL relationships between significant P-O pcHi-C interactions compared to random expectation across 10 matched tissue/cell types from GTEx database. Significant P-O pcHi-C interactions highly enriched by eQTL relationships compared to random expectation in all 10 matched tissue/cell types based on Fisher-exact test p-values (<0.001).

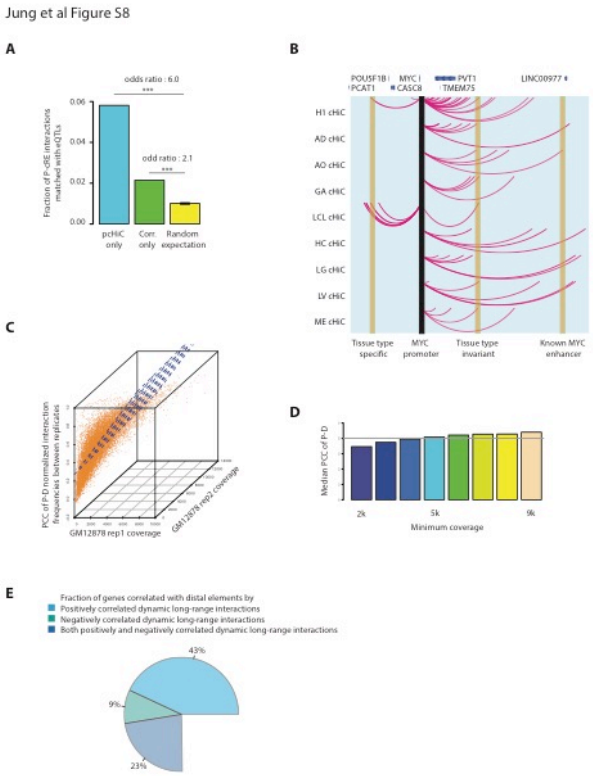


Figure S4.8. Dynamic long-range promoter-cRE interactions.

(A) Bar plots showing fraction of matched eQTL relationships in uniquely identified significant Promoter Capture Hi-C interactions after aggregate all 27 cell/tissue types, uniquely identified by correlation-based method, and random expectation. Standard deviation of error bars is shown for random expectation (n=100). Fisher-exact test was performed for statistical significance (*** p value < 2.2e-16). **(B)** Illustrative dynamic long-range promoter-centered interactions for *MYC* promoter across multiple cell and tissue types. Orange boxes highlighted tissue type specific, tissue type invariant, and known MYC enhancer interactions. *MYC* promoter region is highlighted as black. **(C)** A 3D-scatter plot between numbers of Promoter Capture Hi-C read coverage of promoters between GM12878 rep1, rep2, and PCC of normalized interaction frequencies between two replicates for each promoter. Each dot indicates individual promoter. The hyperplane is shown for regression result. **(D)** Bar plots showing median PCC of promoter and other normalized interaction frequencies between two biological replicates GM12878 in terms of minimum coverage of Promoter Capture Hi-C reads at promoters. **(E)** Pie chart showing the percentage of genes whose gene expression levels are correlated by dynamic promoter-other interactions, positively, negatively, or both.

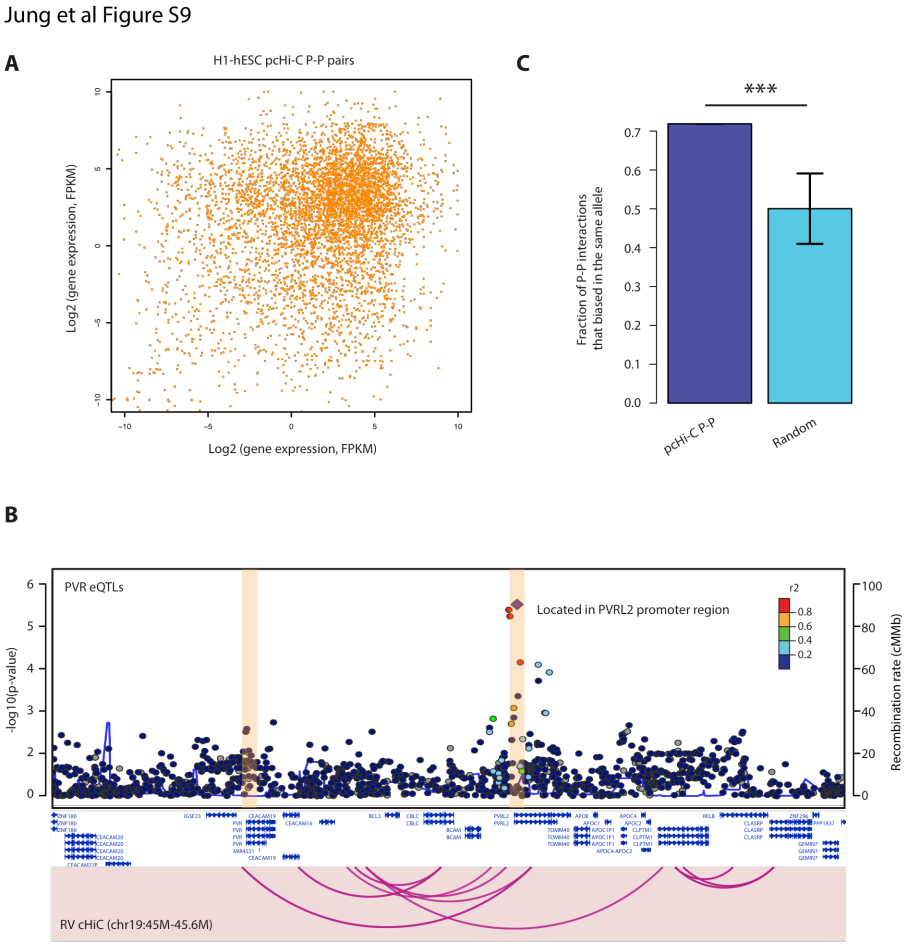


Figure S4.9. Functional promoter-promoter interactions.

(**A**) A scatter plot showing of gene expression FPKM values of significant pcHi-C P-P pairs in H1. Each dot indicates each gene. (**B**) Illustrative locus zoom plot of eQTLs for *PVR* gene expression in right ventricle. Both *PVR* gene promoter region and *PVRL2* promoter that contain significant eQTLs are highlighted in translucent orange. Dots along the locus zoom plot represent SNPs, and their significance of association with *PVR* gene expression is plotted along the left y-axis. Dots are also color-coded based on their LD score with a tag SNP. The blue line traveling across the scatterplot indicates the estimated recombination rate, as plotted along the right y-axis. RefSeq genes position is plotted below the locus zoom plot. Significant Promoter Capture Hi-C interactions in RV were shown as purple in the bottom. (**C**) Bar plots showing the fraction of significant pcHi-C P-P pairs that concordantly biased in the same allele (left, purple), compared to random expectation (right, teal). Standard deviation of error bars was calculated from 1000 times random expectation values.

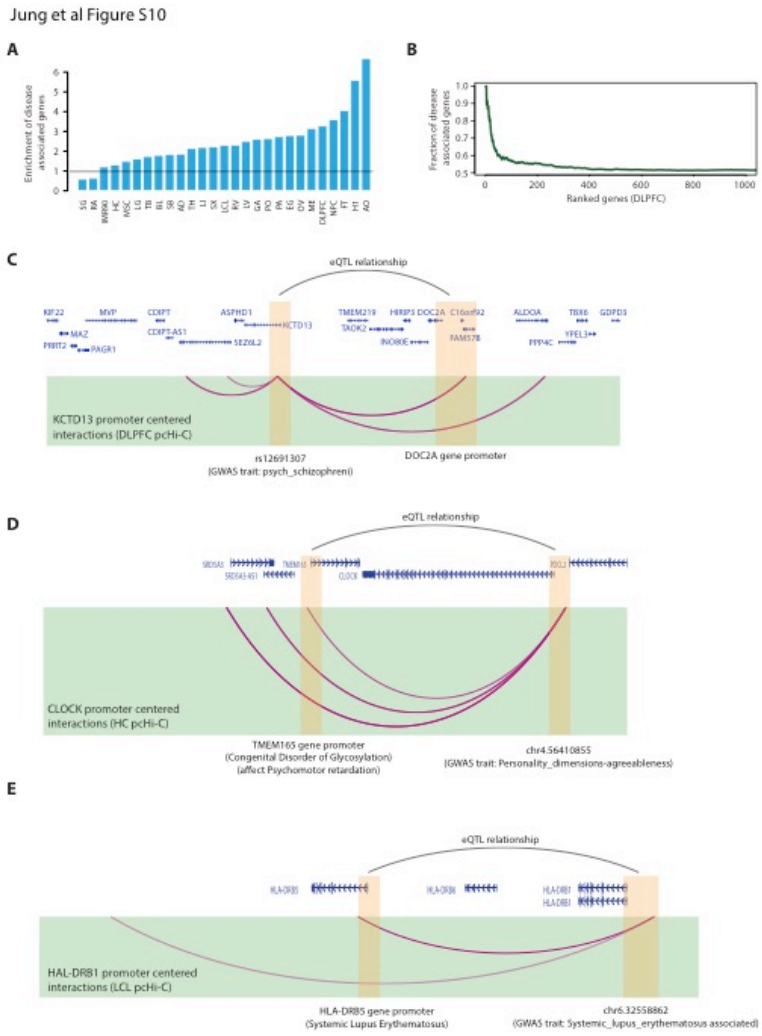


Figure S4.10. Promoter located GWAS-SNPs and their putative distal target genes.

(**A**) Barplot showing the enrichment of disease-associated genes (y-axis) in top 20 high confident putative target genes of GWAS-SNPs identified by pcHi-C maps across cell and tissue types (x-axis). Black line indicates no enrichment of disease-associated genes compared to other genes. (**B**) Line plot showing the fraction of disease associated genes in GWAS-SNP putative target genes identified by DLPFC Promoter Capture Hi-C data. The genes are ranked (x-axis) in terms of frequency targeted by GWAS-SNPs. (**C-E**) Illustrative several examples showing putative target genes of GWAS-SNPs that resided in promoter proximal regions and target distal genes by both eQTL relationships and Promoter Capture Hi-C result.

Table S4.1. List of cell/tissue types analyzed in this study

Samples	Tissue/cell type
AD_STL002	Adrenal Gland
AO_STL002	Aorta
AO_STL003	Aorta
BL_STL001	Bladder
DLPFC	Dorsolateral Prefrontal Cortex
EG_STL002	Esophagus
FT_STL002	Fat
IMR90	Fibroblasts
GA_STL002	Gastric
GA_STL003	Gastric
HC	Hippocampus
H1	Human ES cells
LV_STL001	Left Ventricle
LV_STL003	Left Ventricle
LI_STL011	Liver
LG_STL001	Lung
LG_STL002	Lung
GM12878.rep1	Lymphoblasts
GM12878.rep2	Lymphoblasts
GM19240	Lymphoblasts
MSC.rep1	Mesenchymal Stem Cell
MSC.rep2	Mesenchymal Stem Cell
ME	Mesendoderm
NPC	Neural Progenitor Cell
OV_STL002	Ovary
PA_STL002	Pancreas
PA_STL003	Pancreas
PO_STL001	Psoas
PO_STL003	Psoas
RA_STL003	Right Artrium
RV_STL001	Right Ventricle
RV_STL003	Right Ventricle
SG_STL001	Sigmoid Colon
SB_STL002	Small Bowel
SB_STL001	Small Bowel
SX_STL001	Spleen
SX_STL003	Spleen
TH_STL001	Thymus
TB	Trophoblast

Table S4.2. Number of processed reads

Tissue	Trans_target	Cis_target	Self_target	Total_read
AD	21189399	16562245	15701347	258605662
AO	21377741	18073853	19497645	258079376
BL	19127398	16708591	20757180	293419345
DLPFC	14693048	15456589	11210022	190395167
EG	3979711	2548027	3846385	57973772
FT	3418500	3155609	4042855	63282029
GA	33563029	17767829	31333830	283205529
LCL	52538294	61292977	33491869	424713922
H1	16280869	18490970	16652197	207432338
HC	19096377	18868512	16072830	164780406
IMR90	16534785	22902355	16810314	168302264
LG	37427610	17759143	26821237	389985523
LI	16169255	13920351	20183454	253040895
LV	43954536	20527913	26589758	350765301
ME	10351330	16852133	15223522	96531837
MSC	8477823	31160058	16619223	144584470
NPC	27684792	8439809	15900447	128845471
OV	27058343	12047340	17769330	241662445
PA	22450998	13975650	18959814	244227059
PO	9000450	5916505	6930829	87699425
RA	3601248	2012229	3486904	52853723
RV	34167423	17182471	20265717	281406073
SB	35145481	20051901	28740253	323151259
SG	2016586	1740077	2486683	46896920
SX	18189701	10225803	16381391	192495180
TB	9259870	9422392	6745688	68223281
TH	35147513	14994797	21521930	246537737

Table S4.3. Number of significant long-range promoter-centered interactions from pcHi-C

Tissue/cell-type	Number of interactions
AD	22964
AO	28876
BL	29856
DLPFC	26285
EG	991
FT	1705
GA	11564
LCL	37344
H1	45397
HC	17346
IMR90	49332
LG	12708
LI	27346
LV	20141
ME	45480
MSC	73242
NPC	8669
OV	5749
PA	8446
PO	6026
RA	923
RV	16172
SB	10891
SG	567
SX	11777
TB	26046
TH	15731

Table S4.4. Total number of interaction hotspots (Poisson P value < 0.01)

Tissue/cell type	Number of interaction hotspots
AD	743
AO	952
BL	952
DLPFC	743
GA	475
GM	1200
H1	1319
HC	676
IMR90	1420
LG	475
LI	952
LV	587
ME	1420
PA	377
RV	587
SB	402
SX	439
TB	952
TH	512
MSC	1319
NPC	402

Table S4.5. List of TF ChIP-seq data to define GM12878 TF clusters

TF	Number of peaks	SRA accession ID
ATF2	25922	SRX190236
BATF	56935	SRX100583
BCL3	27935	SRX100387
BCLAF1	12138	SRX100554
BHLHE40	30235	SRX150509
CHD2	20388	SRX150458
CREB1	32576	SRX190216
CTCF	34460	SRX150690
EBF1	30572	SRX150455
EGR1	19053	SRX190186
ELF1	34951	SRX100541
EP300	18036	SRX150641
MAX	14357	SRX150597
MAZ	18017	SRX150363
MEF2A	25785	SRX100556
MTA3	17290	SRX190185
MXI1	16048	SRX150510
NFATC1	22246	SRX190235
NFYB	15000	SRX150586
PAX5	48119	SRX100436
PBX3	31282	SRX100577
PML	14882	SRX190227
RAD21	48659	SRX150412
RUNX3	90498	SRX190349
SIN3A	10673	SRX150411
SMC3	24493	SRX150456
SP1	49101	SRX100408
SPI1	64814	SRX100576
SRF	12766	SRX100395
STAT5A	10225	SRX190177
TBL1XR1	11263	SRX150732
TBP	16668	SRX150732
TCF12	49332	SRX100434
ZNF143	34116	SRX150692
ZNF384	10930	SRX186607

Table S4.6. List of TF ChIP-seq data to define H1-hESC TF clusters

TF	Number of peaks	SRA accession ID
ATF2	23883	SRX190198
BACH1	15407	SRX150659
CEBPB	24803	SRX150375
CHD1	7427	SRX186640
CHD2	12886	SRX150377
CREB1	36040	SRX190352
CTBP2	20048	SRX150542
CTCF	58646	SRX067423
E2F6	46692	SRX190355
EGR1	8405	SRX100475
EP300	14776	SRX100587
GABPA	20837	SRX100469
HDAC2	18928	SRX186668
JUND	19605	SRX100574
KDM4A	29948	SRX186675
MAFK	13639	SRX150372
MAX	82405	SRX190354
MXI1	8604	SRX150373
NANOG	18796	SRX100482
PHF8	20069	SRX100482
RAD21	68538	SRX150459
RBBP5	18346	SRX186780
REST	20314	SRX100410
SAP30	22515	SRX186768
SIN3A	37272	SRX150369
SP1	29549	SRX100422
SP4	16194	SRX190199
TAF1	31733	SRX100495
TAF7	9255	SRX100546
TBP	26635	SRX150383
TCF12	26383	SRX100472
TEAD4	53037	SRX190301
USF1	43955	SRX100471
YY1	46580	SRX100558
ZNF143	39522	SRX150593

Acknowledgements

Chapter 4, in full, has been submitted for publication of the material as it may appear in *Science*. Inkyung Jung, Anthony Schmitt, Yarui Diao, Dongchan Yang, Zachary Chiang, Marilynn Chan, Catherine Tan, Cathy Barr, Bin Li, Samantha Kuan, Dongsup Kim, Bing Ren. The dissertation author was the co-primary investigator and the co-primary author of this material.

Chapter 5

Conclusion

Summary

We have conducted several studies to either develop new technologies for mapping 3D genome architecture, or to investigate genome architecture across human tissues using genome-wide chromatin architecture mapping technologies. First, we have developed the Capture-HiC technique, which we show can be used to obtain ultra-high resolution maps of interaction profiles for user-defined loci throughout the genome[1]. When combined with the HaploSeq algorithm[2], we also show the capability of Capture-HiC (termed Targeted HaploSeq) to obtain high-resolution, accurate, and complete haplotype phasing information for the MHC and KIR loci. In ongoing work from our lab and in collaboration with Jerry Morris (UCSD), we are extending this work into patient samples in a proof-of-concept study to determine if targeted HaploSeq can match donor and recipients in transplant clinics by way of improved MHC locus phasing. Second, we have performed Hi-C analysis in twenty-one human cell lines and primary tissues, and have discovered a novel 3D structure that we have termed frequently interacting regions (FIREs). FIREs are the most highly locally interactive sequences in the genome, and through integration with other epigenomic datasets[3, 4], we find that FIREs are sample-specific, positioned near cell identity genes and towards the center of TADs, mediated by Cohesin, and enriched for active enhancers and disease-associated genetic variation. FIREs are also promiscuously interactive loci with several significant local interaction partners, of which many are also other FIREs.

This analysis has highlighted several important points about the interaction landscape of enhancer-bearing loci in human tissues. First, although it is known that enhancers impart their function through long-range chromatin interactions, it is surprising that the most highly locally interactive sequences in the genome are enriched for enhancers, rather than other regulatory sequences such as promoters. Second, it reinforces a model whereby enhancers are highly interactive with their local neighborhood, which may include additional enhancer(s) as well as promoter(s). This brings to light an important distinction, which is

that enhancers may only impart a gene-regulatory function on a single gene, but this function is mediated by a more complex local interaction network potentially involving simultaneous enhancer-enhancer and enhancer-promoter looping. An alternative hypothesis is that enhancers are highly interactive regions "searching" for their correct interaction partner. This searching may be somewhat stochastic, until a correct protein-protein interaction is established between the enhancer and its suitable regulatory target. In this way, enhancers interact highly with its neighboring loci, but only impart a function on a single interaction target. In the context of these two models, it is intriguing to think of how deleterious genetic variants abrogate the function of enhancers, but in order to more finely address these questions, one must analyze interaction profiles at the resolution of individual *cis*-regulatory elements.

To better understand the gene-regulatory function of DNA looping between *cis*-regulatory elements, as well as the potential impact of disease-associated variants, we have implemented promoter Capture-HiC[5] to map the interaction profiles of nearly 20,000 well-annotated gene promoters across twenty seven human cell lines and primary adult tissue types. We have analyzed this invaluable resource of interaction maps to identify tissue-specific promoter-enhancer interactions and interactions hotspots that may be involved in complex gene regulation networks. We also suggest a widespread role for promoters to regulate distal gene expression through interaction with other promoters, an event termed enhancer-like promoter elements (EPLs). Lastly, we utilize rich annotations of disease-associated variants from GWAS studies[6] to systematically pinpoint the target genes of thousands of genetic variant loci. Notably, this study provides a wealth of critical information linking disease-associated risk loci to target genes in the disease-relevant tissue types, a significant advance in the post-GWAS era, and ultimately helps link 3D DNA looping, to both gene regulation mechanisms and candidates for disease pathogenesis.

Technical Challenges, Implications and Future Perspectives

In recent years, the applications of chromosome conformation capture data have broadened beyond 3D genome architecture mapping, to now include haplotype phasing[1, 7, 8], genome assembly[9, 10] and

deconvoluting mixtures of microorganisms[11, 12]. With respect to my work developing targeted HaploSeq, certain technical challenges remain to realize the ultimate goal of bringing this technology to prospective clinical use.

In terms of assay performance, an ideal targeted HaploSeq platform would be able to call SNPs *de novo* from the Hi-C data with high accuracy and sensitivity, and then accurately and completely phase those same SNPs. In order to call SNPs *de novo* with high sensitivity, one must be have adequate sequence coverage across the entire locus (for targeted HaploSeq) or genome (for HaploSeq). In Hi-C, this is problematic since sequence coverage has been conventionally limited by the choice of restriction enzyme (RE) used to prepare the Hi-C library. For example, when preparing libraries with HindIII, a theoretical maximum of 23% of the genome can be covered due to the relative paucity of HindIII cut sites throughout the genome[13], which has limited the amount of SNPs covered by sequence reads to 22-27%[1, 8]. Hi-C has also been performed using more frequently cutting RE in flies[14], and recently methodological advancements in the Hi-C protocol have brought 4-cutters to use in human samples[15]. However, even a single 4-cutter only has an 83% theoretical genome coverage maximum, indicating that nearly 20% of the genome will be "blind" to *de novo* SNP detection from Hi-C data alone. Going forward, I envision SNP detection from Hi-C will be greatly aided by the use of multiple buffer-compatible RE, such as combinations of 4-cutter and 6-cutter RE. At a glance, combining a single 4-cutter and 6-cutter increases the genome coverage to 90%, and additional enzymes could theoretically be added. Alternatively, other methods for chromatin fragmentation during Hi-C have the potential to improve genome coverage. For example, DNase has been used to prepare Hi-C libraries, and has been shown to yield 62% genome coverage with shallow sequencing (40M reads)[13]. If this were increased to a typical 30-35X genome, or used in targeted HaploSeq, one may achieve nearly complete genome coverage, though the upper bounds are currently unknown. Another option is chromatin fragmentation using micrococcal nuclease, which has been shown to prepare Hi-C libraries in yeast cells[16]. Theoretically, this assay would obtain sequence coverage where any nucleosomes are positioned in a given cell, and since nucleosome positioning may be relatively dynamic in a cell population, Hi-C from MNase fragmentation could be the superior enzymatic approach. Lastly, non-enzymatic approaches, such as mechanical shearing, have used to fragment

chromatin for 4C (http://www.nature.com/protocolexchange/protocols/1979) and suggested to work well for Hi-C, though no published data exists to examine genome coverage[13]. Finally, one can partially circumvent the problem of reads mapping adjacent to cut sites by performing Hi-C, but instead preparing such libraries for long-read sequencing platforms such as PacBio or Oxford Nanopore. Though experimentally possible, the high cost and relatively low coverage from the current PacBio instrument precludes use on large genomes, such as humans.

In addition to the sequence coverage limitation, the Hi-C assay itself, and therefore Haploseq, is too laborious and expensive for clinical adoption. The most rapid Hi-C protocol published to date still takes 3-4 hands-on working days, while industry-standard NGS workflows are typically single-day automated procedures. Additionally, a single Hi-C experiment in its current form can cost >$300 in reagents alone, which also precludes clinical adoption. The combination of speed, automation-compatibility, and cost efficiency must be dramatically improved to enable use of Haploseq or targeted Haploseq in clinical settings.

Despite the robustness and efficiency of the most improved Hi-C protocol[15], several shortcomings persist that prevent further study of chromatin organization in important biological contexts. First, Hi-C has traditionally been used to study genome organization from cell populations, requiring >2M cells for a single experiment. Recent modifications to the Hi-C protocol have enabled Hi-C analysis of single-cells[17, 18], but close examination of these methods still reveal significant shortcomings, such as the low number of single-cells analyzed in a single experiment (throughput), or the low number of detectable interactions per cell (sensitivity), respectively. I suspect that implementation of microfluidics technology, such as the 10X Genomics instruments, will facilitate drastically increased throughput and sensitivity for single-cell chromosome conformation capture analysis.

Hi-C has been invaluable for mapping *pairwise* interactions genome-wide in many contexts, such as in response to exogenous stimuli[19, 20], manipulation of architectural proteins[21-23], or for charting 3D genomes across several cell types[15, 24, 25], such as the work described in this dissertation. However, the technical limitations of Hi-C to detect more complex *multi-way* interactions arise from at least two flaws; 1) Hi-C depends on chromatin digestion and re-ligation, so only spatially proximal DNA that can

efficiently undergo these molecular biology steps are detectable by Hi-C. This means that many spatially proximal sequences are likely to go undetected, due to natural inefficiencies in the Hi-C molecular biology steps. 2) Hi-C libraries are currently only prepared for short-read sequencing (SRS) on Illumina instruments. Illumina SRS can only sequence 500bp DNA fragments, and therefore are likely to only detect pairwise ligation events on a single sequenced DNA fragment. Recent analysis of Hi-C libraries prepared using 4-cutters estimates that only 0.05% of sequenced fragments actually contain more than 2 Hi-C restriction fragments (i.e. a multi-way interaction), thereby precluding the analysis of multi-way interactions from Hi-C data[26]. To circumvent this problem, Darrow and colleagues have proposed the use of a modified 3C protocol using a pseudo-2-cutter RE, which fragments the genome into much smaller fragments compared to 4-cutters and increases the likelihood of a multi-way interaction to be present in an Illumina SRS sequenced DNA fragment. This has been shown to increase the frequency of muti-way interactions from 0.05% to 0.6% (13-fold increase), however this is still too infrequent for high resolution multi-way interaction mapping in large human genomes, but potentially suitable for smaller genomes such as fly or yeast. Going forward, one promising unpublished technology, namely Genome Architecture Mapping (GAM) from the Pombo Lab, may be able to gain deeper insights into the complex multi-way interaction hubs occurring in nuclei as seen in microscopy studies. GAM is a promising technique whereby nuclei are fixed and sectioned onto a microscopy slide, and then individual nuclei "slices" are laser-captured and DNA from each nuclei slice is prepared for NGS. Given that each cell has a unique planar slice through its nuclei, the spatial proximity of any set of DNA sequences can be inferred from how frequently they are detected in the same planar slice. Though this method does not depend on the numerous sequential molecular biology steps in Hi-C, it does depend on efficient library preparation from the scarce about of DNA collected from individual nuclei slices and the resolution is somewhat limited to the thickness of the planar slice.

One final technical hurdle towards better understanding chromatin organization through 3C technologies is mapping chromatin dynamics over time. Currently, the obvious brute force approach is to crosslink cells at many points across a time interval and perform high-resolution comparative Hi-C, however, this is unlikely to give much insight into chromatin dynamics since Hi-C is a cell population

assay, meaning too much noise from random chromatin motion will be present in the data to draw clear conclusions. Instead, I believe advances in high-resolution microscopy techniques and sophisticated computational algorithms will be better suited to detect chromatin motion and looping in individual cells across the dimension of time.

In addition to technical obstacles related to *mapping* 3D genomes, significant technical challenges also remain with respect to the faithful *analysis* of 3D genomes. First, as high-resolution Hi-C data is becoming more readily available[15, 20], so are the number of approaches for identifying a statistically significant DNA looping interaction. Thus far, the field is divided in how to computationally identify a "significant interaction". Two lines of thought divide the 3D genomics field on this matter, depending on one's belief of which background model should be used to identify a significant pairwise interaction. Choice of a global background model has resulted in the calling of ~1,000,000 significant pairwise contacts[20], while a local background model has resulted in identifying only 10,000 significant contacts[15]. Going forward, I believe that the determination of which model is more "accurate" will be greatly aided by the fields continuing ability to decipher between a *statistically significant* interaction compared to a *functional* interaction. In other words, a complete set of significant interactions and functional interactions may indeed be overlapping, however, not entirely synonymous. For example, if loci A-B are significant interacting, but loss of the A-B interaction (via genetic manipulation) has no detectable quantitative effects (via chromatin state, expression, etc), then it's likely that the interaction has no direct and discernable function. Therefore, the highest performing computational algorithm for *functional* 3D genomics will have the highest sensitivity and specificity to detect functional interactions, even if they are not statistically significant. In order to train these algorithms to detect functional interactions, significant work remains to identify and characterize a reference set of functional interactions. Currently, projects in our lab as part of the 4D Nucleome Project are working to characterize the interactions of 100 enhancers using CRISPR-Cas9 technology, in order to identify and describe the biochemical and DNA looping characteristics of functional and non-functional interactions. Additional work from our lab, and others, are developing high-throughput functional screening tools for a similar purpose. At the end of the day, the best algorithms will

likely incorporate several data types (ChIP-seq, Hi-C, RNA-seq, etc) to predict function DNA looping interactions.

In practice, identifying significant *changes* in DNA looping in response to experimental manipulation or disease context is also of critical importance. For example, many studies to date have challenged the 3D genome "system" by adding a stimulus or by genetically manipulating the cell line[19-23, 26]. Here, a key question simply is how does a given experimental manipulation change the 3D genome, and naturally, what changes are significant? One potential approach may be to identify functional interactions in each condition independently, and then compare between conditions, while another approach would be to identifying the statistically significant *differential* interactions between conditions, analogous to identifying differentially expressed genes in RNA-seq data. Thus far, the recent diffHiC analysis package[27] has made significant advances towards this goal, but ultimately it is the differential *functional* interactions that have biological consequences. In other words, detecting a significantly differential interaction frequency is a computational task, while detecting differential functional interactions requires deeper insight into the function of a given interaction.

Lastly, as the 3D genomics field moves forward, there is dire need to unify on best practices for data generation and analysis, as well as terminology use to describe 3D genome structural features. For example, several labs around the world each have their own adaptation to Hi-C and Hi-C analysis pipelines. When experimental and computational methods differ, it makes the interpretation of results across studies very difficult, and may lead to the discovery of new structural features that may simply be a result of inconsistent experimental or computational methods. Further, terminologies in the field have become infiltrated with a plethora of new terms to define essential the same structural features. For example, there seems to be increasing confusion about the terminologies 'TADs', 'sub-TADs', 'contact domains', 'loop domains', and 'insulated neighborhoods'[15, 24, 28, 29]. The trouble with the latter 4 terminologies is they refer to highly overlapping regions of the genome and essentially annotate the same sequences. Although they were defined using different 3C-derived technologies (5C, Hi-C, and ChIA-PET), they all essentially describe chromatin interaction domains, mediated by a single outer-most loop structure. It seems clear that genome folding is organized hierarchically[30, 31], but much concerted efforts going forward needs to be spent

reviewing the 3D structures proposed in the literature, and unifying a set of distinct chromatin architectural features. Hi-C has now been around for seven years, and with lower sequencing costs and improved Hi-C protocols, Hi-C data is becoming increasing more common. Going forward, I envision a unified effort to bring the 3D genomics community to a common ground in terms of protocols, computational methods, and structural feature definitions. I also envision the next few years to include the continued generation of Hi-C maps across >100 cell and tissue types (analogous to ChIP-seq and ENCODE), as well as the explosion of high-throughout functional screening and the 3D genomics community focusing on *functional* interaction mapping and detection. Lastly I imagine a steady increase in the number of studies linking alternations in the 3D genome to disease pathogenesis, which has already been demonstrated in a number of seminal studies[32, 33].

In summary, I feel the work presented in this dissertation has made considerable contribution to the 3D genomics field through development of chromatin architecture mapping technologies and through analysis of chromatin organization across dozens of human tissues and cell types. I hope that many of the core insights gained through these studies will have a broader impact on clinical haplotyping, gene regulation, and interpreting the function of disease-associated genetic variation.

References

1. Selvaraj, S., Schmitt, A.D., Dixon, J.R. & Ren, B. Complete haplotype phasing of the MHC and KIR loci with targeted HaploSeq. *BMC Genomics* **16**, 900 (2015).

2. Selvaraj, S., R Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature biotechnology* **31**, 1111-8 (2013).

3. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

4. Consortium, R.E., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Wang, J., Ward, L.D., Sarkar, A., Quon, G., Kheradpour, P., Coarfa, C., Harris, R.A., Ziller, M.J., Schultz, M., Eaton, M.L., Pfenning, A., Wang, X., Polak, P., Karlic, R., Amin, V., Wu, Y.-c., Sandstrom, R.S., Ray, P., Wu, J., Kulkarni, A., Lister, R., Hong, C., Gascard, P., Carles, A., Mungall, A.J., Moore, R., Chau, E., Tam, A., Zhou, X., Li, D., Zhang, B., Mercer, T.R., Neph, S.J., Siebenthall, K.T., Thurman, R.E., Canfield, T., Hansen, R.S., Kaul, R., Sabo, P.J., Beaudet, A.E., Boyer, L., Jager, P.D. & Peggy, J. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

5. Schoenfelder, S., Furlan-magaril, M., Mifsud, B., Tavares-cadete, F., Sugar, R., Javierre, B.-m., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., Dimitrova, E., Dimond, A., Edelman, L.B., Elderkin, S., Tabbada, K., Darbo, E., Andrews, S., Herman, B., Higgs, A., Leproust, E., Osborne, C.S., Mitchell, J.a., Luscombe, N.M. & Fraser, P. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, 1-16 (2015).

6. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).

7. de Vree, P.J., de Wit, E., Yilmaz, M., van de Heijning, M., Klous, P., Verstegen, M.J., Wan, Y., Teunissen, H., Krijger, P.H., Geeven, G., Eijk, P.P., Sie, D., Ylstra, B., Hulsman, L.O., van Dooren, M.F., van Zutven, L.J., van den Ouweland, A., Verbeek, S., van Dijk, K.W., Cornelissen, M., Das, A.T., Berkhout, B., Sikkema-Raddatz, B., van den Berg, E., van der Vlies, P., Weening, D., den Dunnen, J.T., Matusiak, M., Lamkanfi, M., Ligtenberg, M.J., Ter Brugge, P., Jonkers, J., Foekens, J.A., Martens, J.W., van der Luijt, R., van Amstel, H.K., van Min, M., Splinter, E. & de Laat, W. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat Biotechnol* **32**, 1019-25 (2014).

8. Selvaraj, S., Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* **In Press** (2013).

9. Kaplan, N. & Dekker, J. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol* **31**, 1143-7 (2013).

10. Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. & Shendure, J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* **31**, 1119-25 (2013).

11. Burton, J.N., Liachko, I., Dunham, M.J. & Shendure, J. Species-Level Deconvolution of Metagenome Assemblies with Hi-C-Based Contact Probability Maps. *G3 (Bethesda)* (2014).

12. Beitel, C.W., Froenicke, L., Lang, J.M., Korf, I.F., Michelmore, R.W., Eisen, J.A. & Darling, A.E. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* (2014).

13. Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C.B., Krumm, A., Shendure, J., Blau, C.A., Disteche, C.M., Noble, W.S. & Duan, Z. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat Methods* (2014).

14. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. & Cavalli, G. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-72 (2012).

15. 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. & Aiden, E.L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).

16. Hsieh, T.H., Weiner, A., Lajoie, B., Dekker, J., Friedman, N. & Rando, O.J. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108-19 (2015).

17. Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A. & Fraser, P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).

18. Ramani, V., Deng, X., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z. & Shendure, J. Massively multiplex single-cell Hi-C. *bioRxiv* (2016).

19. Le Dily, F., Bau, D., Pohl, A., Vicent, G.P., Serra, F., Soronellas, D., Castellano, G., Wright, R.H., Ballare, C., Filion, G., Marti-Renom, M.A. & Beato, M. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. *Genes Dev* **28**, 2151-62 (2014).

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

21. Sofueva, S., Yaffe, E., Chan, W.C., Georgopoulou, D., Vietri Rudan, M., Mira-Bontenbal, H., Pollard, S.M., Schroth, G.P., Tanay, A. & Hadjur, S. Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J* **32**, 3119-29 (2013).

22. 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., Flicek, P., Dekker, J. & Merkenschlager, M. Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome Res* **23**, 2066-77 (2013).

23. Zuin, J., Dixon, J.R., van der Reigden, 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., Grosveld, F.G., Ren, B. & Wendt, K.S. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci* **111**, 996-1001 (2014).

24. Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).

25. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V.V., Ecker, J.R., Thomson, J.A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).

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

27. Lun, A.T. & Smyth, G.K. diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16**, 258 (2015).

28. 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. & Young, R.A. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-87 (2014).

29. Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., Bland, M.J., Wagstaff, W., Dalton, S., McDevitt, T.C., Sen, R., Dekker, J., Taylor, J. & Corces, V.G. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-95 (2013).

30. Gibcus, J.H. & Dekker, J. The hierarchy of the 3D genome. *Mol Cell* **49**, 773-82 (2013).

31. Gorkin, D.U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell* **14**, 762-775 (2014).

32. Lupianez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A. & Mundlos, S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-25 (2015).

33. Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., Reddy, J., Borges-Rivera, D., Lee, T.I., Jaenisch, R., Porteus, M.H., Dekker, J. & Young, R.A. Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-8 (2016).