Title

A Compendium of Promoter-Centered Long-Range Chromatin Interactions in the Human Genome

Author

Chan, Marilynn

Publication Date

2019-10-01

Data Availability

The data associated with this publication are available at: https://www.ncbi.nlm.nih.gov/geo/

Peer reviewed

Title:

A Compendium of Promoter-Centered Long-Range Chromatin Interactions in the Human Genome

Authors:

Inkyung Jung[1*†], Anthony Schmitt[2,3*], Yarui Diao[2,4*], Andrew J. Lee[1], Tristin Liu[2], Dongchan Yang[5], Catherine Tan[2], Junghyun Eom[1], Marilynn Chan[6], Sora Chee[2], Zachary Chiang[7], Changyoun Kim[8,9], Eliezer Masliah[8,9,10], Cathy L. Barr[11], Bin Li[1], Samantha Kuan[2], Dongsup Kim[5], Bing Ren[2,12†]

Affiliations:

[1]Department of Biological Sciences, KAIST, Daejeon 34141, Korea

[2]Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA

[3]UCSD Biomedical Sciences Graduate Program, La Jolla, CA 92093, USA

[4]Departments of Cell Biology and Orthopaedic Surgery, Regenerative Next Initiative, Duke University School of Medicine. Durham, NC 27710

[5]Department of Bio and Brain engineering, KAIST, Daejeon 34141, Korea

[6]University of California San Francisco, San Francisco, CA 94158, USA

[7]Department of Bioengineering, UCSD, La Jolla, CA 92093, USA

[8]Molecular Neuropathology Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA

[9]Department Neurosciences, School of Medicine, University of California, San Diego, La Jolla, CA 92093, USA

23[10]Department of Pathology, School of Medicine, University of California,
24San Diego, La Jolla, CA 92093, USA

25[11]Krembil Research Institute, University Health Network, Toronto, and The
26Hospital for Sick Children, Ontario M5T 2S8, Canada

27[12]Department of Cellular and Molecular Medicine, Institute of Genomic
28Medicine, and Moores Cancer Center, University of California at San Diego,
29La Jolla, CA 92093, USA

30

31*These authors contributed equally to this work

32[†]Correspondence to Inkyung Jung (ijung@kaist.ac.kr) and Bing Ren
33(biren@ucsd.edu)

34

35Note:
36All raw and processed data are deposited into GEO database under
37accession number GSE86189. Reviewer's access token is
38mtsxkwgunlipvqb.

39A genome browser session has been set up for visualization of the
40promoter-centered chromatin interactions described in the current study –
41http://epigenomegateway.wustl.edu/browser/?
42genome=hg19&session=IR82F6olpo&statusId=446157315 (Remy the
43chef)

44

45

46

47

48

49

Abstract:

A large number of putative *cis*-regulatory sequences have been annotated in the human genome, but the genes they control remain to be defined. To bridge this gap, we generate maps of long-range chromatin interactions centered on 18,943 well-annotated promoters for protein-coding genes in 27 human cell/tissue types. We use this information to infer the target genes of 70,329 candidate regulatory elements, and suggest potential regulatory function for 27,325 non-coding sequence variants associated with 2,117 physiological traits and diseases. Integrative analysis of these promoter-centered interactome maps reveals widespread enhancer-like promoters involved in gene regulation and common molecular pathways underlying distinct groups of human traits and diseases.

Main Text:

Genome-Wide Association Studies (GWAS) have uncovered thousands of genetic variants associated with human diseases and phenotypic traits[1], but molecular characterization of these genetic variants has been challenging because they are mostly non-coding and lack clear functional annotation. Recent studies have shown that these non-coding variants are frequently marked by chromatin signatures of *cis*-regulatory elements (*c*REs) in cells, leading to the hypothesis that a substantial fraction of variants may act by affecting transcriptional regulation[2,3]. To formally test

this hypothesis, it is critical to define the target genes of *c*REs in the human genome. However, inferring target genes of *c*REs based on linear genomic sequences is not straightforward, since *c*REs can regulate non-adjacent genes over large genomic distances[4-7]. Such long-range regulation can take place because chromatin fibers are folded into a higher-order structure in which distant DNA fragments can be juxtaposed in space[8]. Consequently, mapping spatial contacts between DNA has the potential to uncover target genes of *c*REs. To this end, Chromosome Conformation Capture (3C) techniques such as 4C-seq, ChIA-PET and Hi-C have been developed to determine chromatin interactions in a high throughput manner[9-15]. More recently, Hi-C combined with targeted capture and sequencing (capture Hi-C) has emerged as a cost-effective method to map chromatin interactions for specific regions at high-resolution[16-25].

In order to systematically annotate candidate target genes for the *c*REs in the human genome, we performed capture Hi-C experiments (Fig. 1a; Extended Data Fig. 1) to interrogate chromatin interactions centered at well-annotated human gene promoters for 19,462 protein-coding genes (see Methods). We carried out these experiments with 27 human cell/tissue types including embryonic stem cells, four early embryonic lineages (mesendoderm, mesenchymal stem cell, neural progenitor cells, and trophoblast), two primary cell lines (fibroblast cells and lymphoblastoid cells), and 20 primary tissue types (hippocampus,

dorsolateral prefrontal cortex, esophagus, lung, liver, pancreas, small bowel, sigmoid colon, thymus, bladder, adrenal gland, aorta, gastric tissue, left heart ventricle, right heart ventricle, right heart atrium, ovary, psoas, spleen, and fat) for which reference epigenome maps have previously been produced as part of the Epigenome Roadmap project (Extended Data Fig. 2a; Supplementary Table 1)[26]. We designed and synthesized 12 capture probes for each promoter, six for each of the nearest *Hin*dIII restriction sites upstream and downstream of the transcription start site (TSS). Among 16,720 promoter-containing *Hin*dIII restriction DNA fragments, 14,357 (86%) contain a single promoter, but the 2,363 remaining *Hin*dIII fragments harbor multiple promoters (Extended Data Fig. 2b; see Methods). The robustness and the coverage of capture probe synthesis were validated by sequencing (Extended Data Fig. 2c-f). On average, each capture Hi-C experiment produced 65 million unique, on-target paired-end reads, yielding a total of 1.8 billion valid read pairs, ~30% of which were between DNA fragments >15kb apart (Supplementary Table 2).

To identify the long-range chromatin interactions from the capture Hi-C data, two normalization steps were introduced. First, the biases in capture efficiency of each promoter (Extended Data Fig. 2g, h) were calibrated with the variable “capturability” for each DNA fragment, defined as the fraction of total read counts mapped to the region, using a β-spline regression model (see Methods). Second, significant chromatin

interactions were then identified after normalizing against the distance-dependent background signals (9% and 5% FDR for P-O and P-P interactions, respectively) (see Methods). Focusing on the *Hin*dIII fragments over 15kb away and within 2Mbp of each promoter, we determined a total of 892,013 chromatin interactions (431,141 unique interacting pairs) in one or more of the 27 human cell/tissue types (Fig. 1b; Extended Data Fig. 3a; Supplementary Table 3-5). The median distance between the interacting DNA pairs was 158kb, which is within a similar range of previously reported chromatin loops and eQTL associations (Fig. 1c; Supplementary Table 6)[10,12,13]. The slight discrepancy between pcHi-C interactions and eQTL-associations may be attributed to different experimental approaches, but nevertheless, the two methods give complementary information to each other. Between 13% and 45% pcHi-C interactions detected in a cell or tissue type were unique to that cell/tissue type (Extended Data Fig. 3b). As expected, most of the detected chromatin interactions were within Topologically Associating Domains (TADs) defined in the corresponding tissue/cell type (Extended Data Fig. 3c, d)[27,11].

To demonstrate that pcHi-C could effectively and reproducibly capture long-range chromatin interactions as detected by whole genome *in situ* Hi-C, we compared the pcHi-C data with the *in situ* Hi-C data obtained from four distinct biosamples, including two cell lines (IMR90 lung fibroblast cell line and GM12878 lymphoblastoid cell line[13]) and two

primary tissues - dorsolateral prefrontal cortex and hippocampus (see Methods). Results of pcHi-C experiments accurately recapitulated chromatin loops identified from *in situ* Hi-C assays in all samples, with the area under the receiver operating curve (ROC) ranging between 0.84 and 0.91 (Extended Data Fig. 4a-e) (see Methods). Additionally, we found high reproducibility of pcHi-C chromatin interactions between different donors (average ROC score = 0.85; the average Spearman's rank correlation between replicates = 0.4; Extended Data Fig. 4f-j; Supplementary Table 7; see Methods), and between two independent studies (Extended Data Fig. 4k). The observation that interactions identified in both replicates exhibited the strongest interaction signals, while interactions identified in one replicate were moderately strong in one replicate, but moderately weak in the other replicate (Extended Data Fig. 4l-m), suggests that the interactions that are specific to one replicate may be due to under-sampling of the other replicate.

The chromatin interactome maps allowed us to assign candidate target genes for 70,329 putative *c*REs, defined based on histone H3K27ac signals in each tissue/cell type profiled previously[26], for 17,295 promoters. Each promoter was putatively assigned to 25 *c*REs on average (Extended Data Fig. 5a), while 45% of *c*REs were assigned to one candidate target gene (Extended Data Fig. 5b), similar to the previous observation with DNase I hypersensitivity analysis across diverse human cell types [28]. We took advantage of the existing chromatin datasets collected for the same

tissue/cell types[26], and examined the relationship of the chromatin states between the *c*REs and the target promoters (see Methods). As expected, the fragments that extensively interact with multiple promoters were often found at active chromatin regions, such as TF binding clusters or super enhancer regions (Extended Data Fig. 5c-i; Supplementary Table 8-10; see Methods)[29]. Furthermore, integrative analysis with ChromHMM model revealed that active promoters interact three times more frequently with DNA fragments harboring active enhancers than the bivalent promoters (Fig. 1d). On the other hand, the bivalent promoters interact five times more frequently with genomic regions associated with Polycomb Repressor Complexes than the active promoters (Fig. 1d). Further analysis based on a refined 50-chromatin-state ChromHMM model for 5 cell lines also supports our conclusion (Extended Data Fig. 6).

Three lines of evidence demonstrate that the above promoter-centered chromatin interactions contain information on regulatory interactions at each promoter in the corresponding cell/tissue types. First, we compared the chromatin interactions at promoters with regulatory relationships inferred from expression quantitative trait loci (eQTL) in 14 matched tissue-types that were recently reported by the GTEx consortium (see Methods) (Fig. 2a; Extended Data Fig. 7a-c)[30]. For each tissue and cell type, the previously reported eQTLs were highly enriched in the chromatin interactions identified in the corresponding tissue, with enrichment up to five-fold (ovary) (Fig. 2b; Extended Data Fig. 7d and e). A total of 42,627

eQTL associations were detected by pcHi-C chromatin interactions, while only 21,362 were expected by random chance after controlling for linear genomic distances (Supplementary Table 11 and 12). Second, there is significant correlation between activities of *cis*-regulatory sequences and the assigned candidate target gene expression across multiple tissues and cell types, consistent with the purported regulatory relationships. Specifically, the histone modification status of H3K27ac of these *c*REs was significantly correlated with the promoter H3K27ac levels (KS-test P value < 2.2e-16; Extended Data Fig. 7f) and transcription levels of the predicted target genes (KS-test P value < 2.2e-16; Extended Data Fig. 7g) across these tissues/cell types. For example, the *POU3F3* gene expression (second column in Fig. 2c) was highly correlated with H3K27ac signals in the distal *c*RE (first column in Fig. 2c) connected by a tissue-specific chromatin interaction (last column in Fig. 2c). Lastly, cell/tissue-specific *c*RE-promoter pairs connected by pcHi-C interactions are significantly associated with active *c*REs and genes that are specific to the same cell/tissue types. For example, hippocampus specific *c*RE-promoter chromatin interactions are significantly associated with active *c*REs (Fig. 2d) and highly expressed genes, albeit modest, (Fig. 2e) in hippocampus. Significant associations of cell/tissue-specific pcHi-C interactions in active *c*REs and highly expressed genes are found in other cell/tissue types as well (Fig. 2f-h, KS-test P value < 2.2e-16, see Methods). The above results, taken together, strongly suggest that the predicted *c*RE-promoter pairs

could uncover regulatory relationships between the *c*RE and target genes in diverse tissues and cell types.

Widespread promoter-promoter (P-P) interactions have been reported in cultured mammalian cells and a few primary tissues[14,21,31]. The promoter-centered interaction maps obtained from 27 diverse tissues and cell types allowed us to test whether this is a general phenomenon. Indeed, consistent with previous reports, a significant fraction of the chromatin interactions was found between two promoters (9%, n = 79,989, Fisher's Exact test p value < 2.2e-16, Extended Data Fig. 8a). The physical proximity of these promoters is accompanied by a strikingly high correlation in chromatin modification state between the pair of promoters across diverse cell/tissue types (Fig. 3a, b). Previously, several promoter loci have been shown to function as enhancers to regulate distal genes[32-34]. In support of the functional significance of enhancer-like promoters identified in the current study, 6,127 eQTLs match P-P interaction pairs, while only 2,722 eQTLs were expected by random chance (Fig. 3c; Extended Data Fig. 8b-d; Supplementary Table 13 and 14; see Methods). For instance, strong chromatin interactions were found between the *DACT3* and *AP2S1* gene promoter regions, and one significant eQTL (rs78730097) for *DACT3* gene was located in the *AP2S1* promoter in the dorsolateral prefrontal cortex (Fig. 3d). Notably, this eQTL does not show any meaningful genetic association with the adjacent downstream gene (*AP2S1*) or nearby genes, but is exclusively associated with *DACT3*

(Extended Data Fig. 8e), suggesting regulatory potential of the *AP2S1* promoter region in distal *DACT3* gene regulation. To validate the function of enhancer-like promoters, we deleted 2 core promoter regions, where the downstream gene is not expressed but the promoter region shows active chromatin marks, using CRISPR-mediated system (Extended Data Fig. 8f, g; see Methods). Deletion of the *ARIH2OS* core promoter resulted in marked down-regulation of the distal target gene (FDR adjusted p-value = 0.02), *NCKIPSD*, identified by long-range chromatin interactions (Fig. 3e) with no significant or moderate effect on nearby genes (Extended Data Fig. 8h). Importantly, sgRNA-induced mutations in selected eQTLs proximal to transcriptional start sites demonstrated significant down-regulation effect on distal target genes but no significant effect on nearby gene expression in H1-hESC (Fig. 3f; Extended Data Fig. 8i; see Methods). Our results strongly suggest genome-wide presence of enhancer-like promoters in the human genome and provide additional insight into their potential function in distal gene regulation.

The above promoter-centered chromatin interaction maps allowed us to infer the target genes of sequences harboring disease-associated variants and understand the molecular basis of human disease. We focused on 142,633 putative disease/trait-associated genetic variants from a recent public repository of GWAS catalog[1]. Consistent with previous reports[2,35], a significant portion of SNPs (30%, Fisher's Exact test p value < 2.2e-16) were found in putative *c*REs, emphasizing the importance of target gene

identification of *c*REs in functional interpretation of disease associated genetic variants. Since the causal SNPs are unknown in most cases, we also included SNPs that lie outside the previously defined *c*REs for further analysis. In total, we were able to assign target genes for 27,325 SNPs in the list. On average, each SNP was assigned to between 1 and 3 candidate target genes in each cell/tissue type, with the caveat that the precise number of target genes could potentially be affected by the modest resolution of our promoter capture strategy and the heterogeneity of tissue samples (Extended Data Fig. 9a; Supplementary Table 15; see Methods). The above maps therefore provided many more predictions of disease-associated genes than using the nearest neighbor gene predictions alone (one example is provided for the Parkinson disease in Extended Data Fig. 9b, c), with only about 8% of the putative target genes inferred from our promoter-centered chromatin interaction maps were found to be the closest gene to the sequence variant (Extended Data Fig. 9d). To evaluate the validity of target predictions based on the promoter-centered chromatin interaction maps, we focused on 7 GWAS variants that overlap with previously annotated *c*REs and eQTLs in human lymphoblastoid cell line GM12878 cells. We introduced deletions to these elements in GM12878 using CRISPR-Cas9 genome editing tools and examined the expression of predicted target genes using RT-qPCR in the mutant cells and controls. For 5 of the 7 tested *c*REs, genetic perturbation led to down regulation of the predicted distal target genes (Fig. 4a and

Extended Data Fig. 9e-f). This result supports the target gene predictions based on the pcHi-C interactions.

Many diseases and traits could be linked to common molecular pathways, and the identification of these shared molecular pathways can be beneficial in understanding disease pathogenesis and developing treatment. To uncover the common molecular pathways underlying different diseases and physiological traits, we first determined the diseases/traits that share a significant number of common target genes predicted from their respective GWAS-associated SNPs. We grouped 687 traits and diseases into 40 clusters (Fig. 4b; Extended Data Fig. 10a-c; Supplementary Table 16; see Methods). Many physiological traits with known connections are found to be clustered together. For examples, C5 clusters oxygen transport related traits together, C6 groups together traits related to renal functions, and C20 includes vascular function associated traits (Fig. 4b). The above grouping is made possible thanks to the promoter-centered chromatin interactome maps, because the similarities among related traits observed in Fig. 4b were much less evident when we used either GWAS SNPs or nearest genes of the GWAS SNPs to compute the similarities as control experiments (Fig. 4c, d, Extended Data Fig. 10d). Our result suggests the power of target gene identification of GWAS variants to uncover trait-trait associations.

To further understand the common molecular pathways affected in various human diseases, we carried out gene ontology (GO) analysis for the predicted target genes of the GWAS SNPs within each cluster (Supplementary Table 17; see Methods). The enriched GO biological processes suggest potential shared molecular pathways for disease and trait types in each cluster (Fig. 4e, Extended Data Fig. 10e, Supplementary Table 18), including unexpected connections between specific traits. For example, C39 exposes a link between the susceptibility to infectious and autoimmune diseases and the risk of chemotherapeutic toxicity by carboplatin and cisplatin. In support of such link, a putative target gene associated with the response to carboplatin and cisplatin is *ABCF1*, which is involved in inflammatory response[36]. While speculative, the shared molecular pathways uncovered by our analyses may provide new leads for investigation of the molecular basis of complex traits and disease phenotypes.

In summary, we have generated promoter-centered chromatin interactome maps across diverse human cell/tissue types. Our analysis covers a broad range of human tissue types and provides prediction of target genes for over 70,000 putative *cis*-regulatory elements and 27,000 GWAS SNP variants. This resource enables a new approach to understanding the molecular pathways dysregulated in distinct diseases and traits[21]. In future studies, delineation of disease-specific chromatin interactions with clinical samples by comparing our reference chromatin

interaction maps could greatly improve the functional interpretation of many disease and trait associated genetic variants.

It should be noted that the current study only surveys a limited number of human tissues and cell types, and assigned target genes for a small fraction of the putative *cis*-regulatory elements annotated in the human genome. Furthermore, the heterogeneous nature of the tissue samples used in this study prevents us from accessing the cell types in which the identified chromatin interactions occur, except for a few cell lines. Nevertheless, this resource lays the ground for further understanding of human disease pathogenesis and development of new treatment strategies.

Methods

Human tissue samples

Esophagus, lung, liver, pancreas, small bowel, sigmoid colon, thymus, bladder, adrenal gland, aorta, gastric, left heart ventricle, right heart ventricle, right heart atrium, ovary, psoas, spleen, and fat tissues were obtained from deceased donors at the time of organ procurement at Barnes-Jewish Hospital (St. Louis, USA) as described in our previous study[26]. The same tissue types from different donors were combined together during downstream data analysis. Human dorsolateral prefrontal cortex (DLPFC rep1) and hippocampus (HC rep1) 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 31-year-old male donor. Ethics approval was obtained from the University Health Network and The Hospital for Sick Children for the use of these tissues. Another set of human dorsolateral prefrontal cortex (DLPFC rep2) and hippocampus (HC rep2) tissues were obtained from the Shiley-Marcos Alzheimer's Disease Research Center (ADRC). These two samples were from a healthy 80-year-old female donor. Institutional Review Board (IRB) approval was obtained from KAIST for the use of these 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 lymphoblastoid cell lines (GM12878 and GM19240) were cultured and 5 million cells were formaldehyde cross-linked for each Hi-C library. Hi-C was then conducted on the samples as previously described, using *Hin*dIII for Hi-C library preparation[37]. Previously constructed Hi-C libraries[11] were used for human ES cells (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 19,462 well-annotated protein coding gene promoters across 22 autosomes and X chromosome according to GENCODE v19 annotation with confidence level 1 and 2. The annotation confidence level 1 and 2 comprise of genes that are accurately annotated with sufficient validation and manual annotation by combining the manual gene annotation from the Human and Vertebrate Analysis and Annotation (HAVANA) group, automatic gene annotation from Ensembl, and validating by CAGE. Due to the variability of capture efficiency, 19,328 promoter regions (99%) were captured in this study. Among them, 18,943 promoter regions were involved in pcHi-C interactions in one or more cell/tissue types analyzed in this study. For each transcription start site, the two nearest left hand- and right hand-

side *Hin*dIII restriction sites were selected. Six capture oligos were designed to be of 120 nucleotide (nt) length and to have 30nt tiling overhang. Oligos were designed +/- 300bp upstream and downstream of each restriction site. As two restriction sites were chosen for each transcription start site, a total of 12 capture oligos were designed to target each promoter region. Capture sequences that overlap with directly adjacent *Hin*dIII restriction sites were removed. GC contents of 94% capture sequences ranged from 25% to 65%. Some promoters shared the same *Hin*dIII fragment with at least one other, while 14,357 *Hin*dIII fragments (86%) were uniquely assigned to one promoter. The effect of the DNA fragments harboring multiple promoters on the quality of our analytical findings is modest because only 15% of pcHi-C interactions emanated from the promoter sharing DNA fragments, and eliminating these fragments results in no significant changes in our conclusion for both eQTL enrichment test and gene set enrichment analysis. Further, strong correlation of GWAS trait associations remains even after excluding unresolvable promoters. 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[11].

Promoter Capture Hi-C library construction

Promoter Capture Hi-C library was constructed by performing target enrichment protocol (enriching target promoter-centered proximity ligation fragments from Hi-C library using capture RNA probes). Briefly, we incubated 500ng Hi-C library for 24h at 65°C in a 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. RNA:DNA hybrids were enriched using 50ul T1 streptavidin beads (Invitrogen) through 30min incubation at RT. Next, bead-bound hybrids were washed through a 15min incubation in wash buffer1 (1X SSC and 0.1% SDS) with frequent vortexing, and then washed three times with 500ul of pre-warmed (65 °C) wash buffer2 (0.1X SSC and 0.1% SDS), then finally resuspended in nuclease-free water. The resulting capture Hi-C libraries were amplified while bound to T1 beads, and purified using 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 previously described 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 to human genome hg19 using BWA-mem and manually paired with in house script. Unmapped, non-uniquely mapped, and PCR duplicate reads were removed. Trans-chromosomal read pairs and putative self-ligated products (<15kb read pairs) were also removed. Off-target reads were removed when both read pairs did not match the capture probe sequences. The resulting on-target rates in Promoter Capture Hi-C library ranged from 17% to 44% after removing PCR duplicate reads.

Promoter Capture Hi-C normalization

Interaction frequencies obtained from Promoter Capture Hi-C were normalized in terms of DNA fragment resolution restricted by *Hin*dIII. We defined DNA fragments that spanned each *Hin*dIII restriction site. The start and the end of DNA fragments were defined by taking the midpoint of adjacent upstream and downstream restriction sites, respectively. We merged adjacent DNA fragments if the total length of the DNA fragments was less than 3kb. As a result, 510,045 DNA fragments were defined with a median length of 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 (GC contents, mappability, and effective fragment lengths), RNA probe

synthesis efficiency bias, and RNA probe hybridization efficiency bias. Highly variable RNA probe synthesis efficiency would greatly complicate the control of experimental bias. However, if the efficiency bias was reproducible, the bias can be computationally removed. To prove such bias reproducibility, we performed RNA-seq with two sets of RNA probes that were synthesized independently. The RNA-seq results can quantitatively measure the amount of synthesized RNA probes, which is an indicator of the probe synthesis efficiency. We observed highly reproducible RNA-seq results (Pearson Correlation Coefficient = 0.98), indicating reproducible probe synthesis efficiency. To address the high complexity of different types of experimental biases, we defined a new term named “Capturability”, which refers to the probability of the region being captured. We assumed that “Capturability” represents all combined experimental biases and can be estimated by the total number of capture reads spanning a given DNA fragment divided by the total number of captured reads in *cis*. We found that “Capturability” in each DNA fragment is highly reproducible across samples with 0.95 Pearson correlation coefficient between samples on average. Therefore, we defined universal “Capturability” as the summation of all “Capturability” defined in each sample and normalized raw interaction frequencies by considering “Capturability” of two DNA fragments. During normalization, we processed promoter-promoter interactions and promoter-other interactions independently because promoter regions tend to show very high “Capturability” as our capture probes were designed to target promoter

regions. Also, we only considered promoter-centered long-range interactions over 15kb and within 2Mb from TSS of each gene. We denoted Y_{ij} to represent the raw interaction frequency between DNA fragment i and j and C_i to represent “Capturability” defined in DNA fragment i. We assumed Y_{ij} to follow a negative binomial distribution with mean μ and variance $\mu+\alpha\mu^2$. Here, $\alpha > 0$ is a parameter to measure the magnitude of over-dispersion. We then fitted a negative binomial regression model as follows: $\log u_{ij}=\beta_0+\beta_1 BS(C¿¿i)+\beta_2 BS(C_j)¿$, where u_{ij} is an raw interaction frequency between DNA fragment i and j with coverage C_i and C_j and defined the residual $R_{ij}=Y_{ij}/\exp¿¿$ as a normalized interaction frequency between DNA fragments i and j. BS represents a basis vector obtained from B-spline regression, which applied to a vector of values of input variable, C, during negative binomial regression model fitting for robustness and memory efficient calculation.

Identification of P-P and P-O pcHi-C long-range chromatin interactions

To identify significant pcHi-C chromatin interactions, we removed distance dependent background signals from normalized interaction frequencies. Here, we assumed that normalized interaction frequency R_{ij} follows a negative binomial distribution with mean μ and variance $\mu+\alpha\mu^2$. Similar to the interaction frequency normalization step above, we calculated the expected interaction frequency at a given distance by fitting it to a negative binomial regression model with basis vectors obtained from B-

spline regression of distance between two DNA fragments. We denoted E_d to represent the 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 follows: $(R_{ij} + avg(R)) / (E_d + avg(R))$, where *d* indicates distance between DNA fragment *i* and *j.* We confirm that the average of normalized interaction frequencies against distance dependent background signals are close to one in all distance, indicating the successful elimination of distance dependent background signals using our method. Next, using 'fitDistr' function in propagate R package we found that 3-parmeter Weibull distribution well follows the values of normalized interaction frequencies. Thus, we modeled background distribution of distance normalized interaction frequencies with 3-parmeter Weibull distribution. Based on this, significant long-range chromatin interactions are defined when observed interaction frequencies show lower than 0.01 p-value thresholds by fitting distance background removed interaction frequencies with 3-parameter Weibull distribution. To eliminate false pcHi-C interactions caused by experimental noise, we applied the criteria of minimum raw interaction frequencies (having more than 5 raw interaction frequencies), which is chosen by investigating reproducibility between biological replicates using lymphoblastoid and mesenchymal stem cell. Note that as the interaction frequencies in pcHi-C are mostly zeros or close to zero, the distribution of p-values does not follow the uniform distribution, violating the basic assumption of FDR

calculation, which assumes that the null distribution follows uniform (0,1) distribution. Thus, we simulated normalized interaction frequencies that follow 3-parameter Weibull distribution in a sample specific manner, and computed the estimated FDR through multiple permutations. The estimated FDR through multiple permutation (n=1,000) for P-O and P-P pcHi-C interactions is 9% and 5% on average, respectively

in situ Hi-C experiments and validation of pcHi-C long-range chromatin interactions

The visual inspection of normalized interaction frequencies between IMR90 Promoter Capture Hi-C and high resolution IMR90 Hi-C showed high consistency based on manual inspection despite pcHi-C having only 10% sequencing depth compared to high resolution Hi-C (Extended Data Fig. 4a). Next, we compared the identified pcHi-C interactions with “loops” defined from IMR90, GM12878, dorsolateral prefrontal cortex, and hippocampus tissues using *in situ* Hi-C experiments (Extended Data Fig. 4b-e). Although there is a huge discrepancy between the number of *in situ* Hi-C loops and pcHi-C interactions, we may consider ‘loops’ are a subset of high confident long-range chromatin interactions that involve ‘loop’ domains but cannot cover all promoter-mediated long-range chromatin interactions. Loops of IMR90 and GM12878 *in situ* Hi-C result were obtained from previous publication[13]. Loops of dorsolateral prefrontal cortex and hippocampus were identified using HiCCUPS, distributed with Juicer v1.7.6[13]. The loops were called from Knight-Ruiz normalized 5kb,

10kb, and 25kb resolution data, as these parameters were suggested for a medium resolution Hi-C map by the authors of HiCCUPS. As a result, 7,722 and 8,040 loops were identified from dorsolateral prefrontal cortex and hippocampus, respectively. We compared the identified pcHi-C long-range chromatin interactions to loops of *in situ* Hi-C data and measured the reproducibility in terms of ROC curve (receiver operating characteristic curve), a plot of the true positive rate against the false positive rate at various threshold settings. Here, we set loops as true interactions. We ranked all tested pcHi-C DNA fragment pairs in terms of p-values and then calculated the fraction of true positive and false positive to draw ROC curve. We only considered “loops” emanating from promoter-containing DNA fragments defined in our Promoter Capture Hi-C result. Each point on the ROC curve indicates the true and false positive rate for each 1,000 false positive interactions. The area under the ROC curve is defined as an ROC score and an ROC score of 1 indicates that the rank of DNA fragment pairs matched by loops are always higher than all other tested DNA fragment pairs according to pcHi-C interaction p-values.

Reproducibility of pcHi-C chromatin interactions between biological replicates

The reproducibility of pcHi-C chromatin interactions between biological replicates were measured in terms of ROC curve (Extended Data Fig. 4f). Here, we set pcHi-C interactions identified in one replicate as true interactions. For the other replicate, we ranked all tested DNA fragment

pairs in terms of p-values and then calculated the fraction of true positive and false positive to draw ROC curve. The area under the ROC curve is defined as an ROC score and an ROC score of 1 indicates that the rank of all pcHi-C interactions identified in one replicate is always higher than all other tested DNA fragment pairs in another replicate. Due to different sequencing depths in each replicate, we first defined true interaction sets with one replicate that identified fewer number of pcHi-C interactions than the other replicate, then tested how these true interactions were well detected in the other replicate. Both P-P and P-O interactions were combined together for calculating ROC scores. Each dot in ROC curve indicates the true positive rate at the corresponding false positive rate with increment of 1% of false positive rate. We tested biological replicates in the following 12 tissue/cell types: aorta (AO2/AO3, ROC score=0.79), lung (LG1/LG2, ROC score=0.80), small bowel (SB1/SB2, ROC score=0.83), spleen (SX1/SX3, ROC score=0.80), dorsolateral prefrontal cortex (FC_rep1/FC_rep2, ROC score=0.92), left ventricle (LV1/LV3, ROC score=0.85), mesenchymal stem cell (MSC_rep1/MSC_rep2, ROC Score=0.99), hippocampus (HC_rep1/HC_rep2, ROC score=0.81), gastric (GA2/GA3, ROC score=0.91), lymphoblastoid cell lines (GM12878/GM19240, ROC score=0.98), right ventricle (RV1/RV3, ROC Score=0.83), and pancreas (PA2/PA3, ROC score=0.73). Indeed, we calculated Spearman’s rank correlation of p values between replicates and found that the average Spearman’s rank correlation was around 0.40.

Enrichment of pcHi-C interactions regarding TAD, boundary, and unorganized regions

The TAD annotations for 22 samples by DomainCaller[14] with 2MB windows size were downloaded from the 3DIV database[38]. The regions between TADs were classified as “unorganized” when the gap is longer than 400kb, otherwise, the remaining regions were classified as “boundary”. Then, the types of pcHi-C interactions were classified based on the location of DNA fragment’s centroid.

1. “Within TAD”, if both fragments’ centroids are located in the identical TAD.
2. “Within unorganized region”, if both fragments’ centroids are located in the identical unorganized region.
3. “Between different TADs”, if one fragment’s centroid is located in a TAD while another fragment’s centroid is located in a different TAD.
4. “Between TAD and boundary”, if one fragment’s centroid is located in a TAD while another fragment’s centroid is covered by boundary region.
5. “Between TAD and unorganized region”, if one fragment’s centroid is located in a TAD while another fragment’s centroid is located in an unorganized region.

Annotation of ChromHMM 18-chromatin state to DNA fragments

The pre-calculated chromatin state annotations were downloaded from the 18-state ChromHMM model established by Roadmap Epigenomics Project. As the genomic proportion of promoter and enhancer regions are relatively low, we assigned the chromatin states to DNA fragments based on the following priority order (TssA-EnhA1-EnhA2-TssFlnk-TssFlnkU-

TssflnkD-EnhG1/G2-EnhWk-TssBiv-Enhbiv). For instance, the chromatin state of a fragment was assigned as TssFlnkU, if the fragment contained two annotations TssFlnkU and EnhWk. EnhG1 and EnhG2 annotations were merged because of their low occurrence percentage. We considered two promoter types (TssA and TssBiv) according to ChromHMM annotations and investigated the preference of their interacting partners. For each promoter type, the occurrence of each chromatin status at interacting DNA fragments was divided by the total number of interacting DNA fragments. This fraction value of each chromatin status was normalized against the genomic fraction of each chromatin status. KS-test was performed to measure the statistical significance of each chromatin status at interacting DNA fragments between TssA and TssBiv promoters.

Analysis with a 50-chromatin-state ChromHMM model

To supplement our analysis with the ChromHMM 18-chromatin state model, we conducted in-depth investigations with 5 samples, including H1 embryonic stem cell, mesendoderm, mesenchymal stem cell, trophoblast, and IMR90, using a 50-state ChromHMM model produced by the Roadmap Epigenomics Project[35]. The ChromHMM model utilized combination of 29 chromatin marks to generate a 50-state ChromHMM model. To be consistent with the 18-state ChromHMM model, we used the same definition for TssA and TssBiv promoter containing fragments, but chromatin state of their interacting partners was further refined using the

150-state ChromHMM model. The statistical test was performed as described in the analysis with the 18-chromatin-state ChromHMM model.

Identification of extensively interacting DNA fragments

In order to identify DNA fragments that showed extensive long-range interactions with multiple promoters, we systematically defined these promiscuously interacting DNA fragments from P-P pcHi-C interaction maps and P-O pcHi-C interaction maps, respectively. For each cell or tissue-type, we selected frequently interacting DNA fragments with multiple promoters in terms of 0.01 Poisson p value cutoff.

Identification of TF clusters in H1-hESC and GM12878

Transcription factor ChIP-seq datasets on human lymphoblastoid cell lines (GM12878) and human embryonic stem cell (H1-hESC) were collected from ENCODE. These ChIP-seq reads were aligned against human genome hg19 using BWA-mem with default parameters. Non-uniquely mapped, low quality (MAPQ<10), and PCR duplicate reads were removed. Peak calling of individual ChIP-seq experiments was performed with MACS2 callpeak with default parameters [39]. 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 remove parameter dependent bias, we retrieved TF clusters 40 times with various parameter sets as following; minimum # of TFs within cluster (5 or 10), minimum length of cluster from 100bp to

6851600bp, and maximum gap length within cluster from 100bp to 51.2kb.
686Final TF clusters were defined when the region was detected as TF
687clusters more than 50 times from 100 different parameter sets.
688
689**Enrichment analysis of TF clusters and super-enhancers**
690In order to calculate the enrichment of TF clusters or super-enhancers at
691extensively interacting DNA fragments (EIF), we counted the number of
692matched TF clusters and super-enhancers. The list of super-enhancers
693was obtained from the 3DIV database[38]. Permutation test was performed
694to calculate the expected values. Using Bedtools shuffleBed, we
695generated random genomic locations that resemble actual TF clusters
696with the same size but different genomic coordinates. Bedtools
697intersectBed identified any overlap between EIF and TF clusters or random
698genomic locations. Standard deviations of error bars in the random
699genomic locations were calculated from 10,000 random data sets. In order
700to test the enrichment of TF clusters compared to typical TF peaks, we
701generated random genomic locations that resemble actual TF clusters
702with the same size but different genomic coordinates matched to typical
703TF peaks. Standard deviations of error bars in the typical TF peaks were
704calculated from 10,000 random data sets. Similarly, enrichment analysis
705of super-enhancers was conducted by generating random genomic
706locations of the same size as super-enhancers but at different genomic
707coordinates. We also conducted the enrichment test with typical
708enhancers. We revealed that P-O EIFs highly co-exist with super-enhancer

regions, rather than typical enhancers and genomic background for most of the samples, except two samples, lymphoblastoid cell lines and gastric tissue. Note that half of lymphoblastoid P-O EIFs are co-occupied with typical enhancers that are classified as super-enhancers in other cell/tissue types.

Comparison between eQTL associations and P-O interactions

In order to test the enrichment for P-O pcHi-C chromatin interactions in significant eQTL associations, we compared P-O pcHi-C interactions to significant eQTL associations in the matching tissue types. The eQTL associations were downloaded directly from GTEx Portal (downloaded on Nov. 10th, 2017) for all matching tissue types (n=14, adrenal gland, aorta, dorsolateral prefrontal cortex, brain hippocampus, sigmoid colon, esophagus, left heart ventricle, liver, lung, ovary, pancreas, small intestine terminal ileum for small bowel, spleen, and stomach for gastric). First, the significant eQTLs defined by GTEx (q-value ≤ 0.05) were filtered so that only the eQTL variants within the fragments that involve P-O pcHi-C interactions remain for comparison. Then, we removed pcHi-C interactions beyond 1Mb in distance to match the range of eQTL association, and discarded eQTL associations with distance below 15kb to match the valid interaction cutoff. The filtered, significant eQTL associations were compared with pcHi-C and randomized interactions in the same condition. Here, we only considered P-O pcHi-C interactions with DNA fragments that do not harbor multiple promoters. For the random

expectation, we generated a simulated pcHi-C interaction pool by creating all possible combinations of DNA fragments with no TSS and the protein coding genes that exist within the distance range. The pcHi-C interactions that exist in any of the tissue/cell type were removed from the control interaction pool for the enrichment analysis. To avoid variation caused by the difference in distance between pcHi-C interactions and eQTL associations, we created distance matched control, in which the number of pcHi-C interactions was stored at the interval of 40kb, and the same number of interactions was drawn randomly from the control interaction pool. The number of randomized interactions drawn from each chromosome was matched to the pcHi-C interactions. The standard deviation was obtained by permuting the random expectation with 1,000 iterations and was used to calculate the statistical confidence

To illustrate the filtering process of the eQTL data, for example, the 549,763 significant eQTLs in adrenal gland were reduced to 237,181 after collecting eQTLs located in the DNA fragments without TSS and discarding eQTL association with the distance below 15kb and with a pseudogene target. This filtered set of significant eQTL associations was used for enrichment test for both pcHi-C and randomized interactions. The number of total tested significant eQTL association, 19,996 in case of adrenal gland, in Supplementary Table 11, indicates the number of significant eQTLs located in the DNA fragments that are associated with the pcHi-C interactions in the corresponding cell/tissue type.

Variations in H3K27ac signals at promoters and *c*REs connected by P-O interactions

We conducted correlation analysis of H3K27ac signals across all available cell/tissue types for each promoter-*c*RE pair connected by P-O interactions in at least one cell/tissue type analyzed. First, we defined putative distal *cis*-regulatory elements (*c*REs) marked by H3K27ac peaks across all 27 cell/tissue types. We merged these elements if the peaks are within 3kb of each other, then we defined *c*RE-containing DNA fragment when the DNA fragment harbors at least one *cis*-regulatory element. When a DNA fragment contained both TSS and *c*RE, we defined the fragment as a promoter-containing DNA fragment instead of a *c*RE-containing DNA fragment because our experiment is designed to target promoter regions. We used input normalized H3K27ac RPKM values by taking log2 transformation as H3K27ac signals at promoters and *c*REs. Pearson correlation coefficient values were calculated for each promoter-*c*RE pair connected by pcHi-C interactions after excluding *c*REs spanning adjacent DNA fragments and visualized as a box plot. Random expectation values were calculated after randomization of H3k27ac signals at promoters and *c*REs. Distance matching random expectation values were calculated after random selection of *c*RE-promoter pairs by controlling distance information as same as identified *c*RE-promoter pairs.

Analysis of H3K27ac signals at *c*REs and expression of target genes connected by cell/tissue specific *c*RE-promoter pairs

In order to investigate cell/tissue-specific *c*RE-promoter pairs, for each cell/tissue-type unique *c*RE-promoter pairs were collected and then distance normalized pcHi-C interaction frequencies of corresponding P-O pcHi-C interactions were obtained across all cell/tissue types. We only considered a unique P-O interaction pair when multiple *c*REs are located in a same DNA fragment and target a same promoter. The cell/tissue-specific *c*RE-promoter pairs exhibit strong enrichment of pcHi-C interaction frequencies in the corresponding cell/tissue type but depleted in other cell/tissue types, validating the cell/tissue-specificity of these *c*RE-promoter pairs. Statistical significance of pcHi-C interaction frequencies was tested by conducting KS-test between mean of pcHi-C interaction frequencies in the matched cell/tissue types (values in diagonal in Fig. 2f) and those in other cell/tissue types (values in off diagonal in Fig. 2f).

In order to investigate cell/tissue-specific activity of *c*REs connected by cell/tissue-specific *c*RE-promoter pairs, we identified group of *c*REs that are connected by unique *c*RE-promoter pairs for each cell/tissue type. After that, H3K27ac signals were calculated for each *c*RE across all cell/tissue types and these values were normalized by taking *z*-score transformation to obtain relative H3K27ac enrichment signals. The mean values of normalized H3K27ac signals were calculated for each group of *c*REs in each cell/tissue type. KS-test was performed between the mean

values in the corresponding cell/tissue types (values in diagonal in Fig. 2g) and those in other cell/tissue types (values in off diagonal in Fig. 2g).

In order to investigate expression levels of target genes connected by cell/tissue-specific cRE-promoter pairs, we first defined a group of genes that are connected by unique *c*RE-promoter pairs more than twice for each cell/tissue-type. After that, gene expression levels (FPKM) were calculated for each gene across all cell/tissue types. Relative gene expression levels were obtained by taking z-score transformation for each gene across all cell/tissue types. The mean values of z-score transformed FPKM values were calculated for each group of genes in each cell/tissue type. KS-test was performed between the mean values in the corresponding cell/tissue types (values in diagonal in Fig. 2h) and those in other cell/tissue types (values in off diagonal in Fig. 2h).

Comparison between eQTL associations and P-P chromatin interactions

In order to assess the enrichment for promoter-promoter pcHi-C interactions in the significant eQTL associations, we computed the number of P-P pcHi-C interactions matched to the significant eQTL associations (downloaded on Nov. 10th, 2017). For the tested tissue types (n=13, adrenal gland, aorta, dorsolateral prefrontal cortex BA9, hippocampus, sigmoid colon, left ventricle, liver, lung, ovary, pancreas, small intestine terminal ileum for small bowel, spleen, and stomach for gastric), we

considered only the eQTLs that are located within 2.5kb from a TSS of a protein coding gene. For accurate comparison, we removed P-P chromatin interactions beyond 1Mb in distance to match the range of eQTL association, and discarded eQTL associations with distance below 15kb to match the valid interaction cutoff. Finally, the significant eQTLs were filtered by collecting only the eQTLs within the fragments that involve P-P pcHi-C interactions in the corresponding cell/tissue and by removing eQTLs that target pseudogenes. Then, the number of filtered significant eQTLs that match P-P pcHi-C interactions was counted. The DNA fragments that harbor multiple promoters were removed from the analysis. For the random expectation, we created a control pool of all P-P pairs within the range of 15kb to 1Mb, selected the same number of random P-P pairs as used in significant eQTL comparison, and counted the matched number of random P-P pairs with P-P pcHi-C interactions. The P-P pcHi-C interactions that exist in any of the tissue/cell type were removed from the control interaction pool for the enrichment analysis. In addition, to avoid variation caused by the difference in distance between pcHi-C interactions and eQTL associations, we created distance matched control, in which the number of pcHi-C interactions was stored at the interval of 40kb, and the same number of interactions was drawn randomly from the randomized interaction pool. In addition, the number of randomized interactions drawn from each chromosome was matched to the pcHi-C interactions. The standard deviation was obtained by permuting the

random expectation with 1,000 iterations and was used to calculate the statistical significance.

Visualization of eQTL-supported P-P and P-O chromatin interactions

The pcHi-C interactions that matched significant eQTLs were visualized by LocusZoom[40]. We collected and merged significant and all tested eQTLs for each tissue type and extracted the relevant p-values and SNP IDs for a queried gene. Then, LocusZoom was run with default parameters to show the pcHi-C interaction and its eQTL associations surrounding the region.

Experimental validation of enhancer-like function of promoters

H1-hESC was cultured in mTeSR1 medium on Matrigel coated plates[33]. To knockout the core promoter regions of *ZNF891* (chr12:133706994) and *ARIH2OS* (chr3:48956862) in H1-hESC, we utilized CRISPR/Cas9 RNP method as previously described by Diao, et al.[33]. Briefly, we used *in vitro* synthesized CRISPR crRNA and CRISPR tracrRNA (IDT) with the sequences specified below.

ZNF891 sgRNAs 5p-1: GCGTCCGTGACGCACAGACC
ZNF891 sgRNAs 5p-2: GACCAGGCCCCTCTGCGGGG
ZNF891 sgRNAs 3p-1: AGGCTGGGGCGCGTGCGTAA
ZNF891 sgRNAs 3p-2: GTGCGTAACGGTGTGTGTTG
ZNF891 genotyping primer 5p: GTCCTCAGTGCCTGCCTC
ZNF891 genotyping primer 3p: CAGCAACAGCAAAACAGAGAAC
ARIH2OS sgRNAs 5p-1: GCTCCCAAAGATGACTCGAG
ARIH2OS sgRNAs 5p-2: GACTCGAGTGGTGAGCCCCG
ARIH2OS sgRNAs 3p-1: GGAGAAGTCATCCAAGAACG

ARIH2OS sgRNAs 3p-2: CGCTATGACAGAAAGTTCTA
ARIH2OS genotyping primer 5p: CATCTAGGCCCTCTCTCCCT
ARIH2OS genotyping primer 3p: TCAGCAATTTCGTTTCAAAATC

Each of the core promoter was knocked-out by two sets of sgRNA pairs to avoid the potential off-target effect caused by CRISPR/Cas9 genome editing. The Cas9 recombinant protein was purchased from NEB (Cat M0386M) and the Cas9/crRNA/tracRNA was assembled *in vitro* by following the previously described protocol[33]. The RNP complex was electro-transfected into POU5F1-eGFP hESC reporter line with Neon Transfection System 10µl kit (ThermoFisher Scientific, Cat#: MPK1096) with the default electrotransfection protocol. Seven days after post-transfection, individual colonies were picked and expanded, followed by genotyping and in-depth analysis. After genotype validation, we performed RNA-seq using Ovation® RNA-Seq System V2 (NuGEN) as previously described[41].

RNA-seq data analysis between WT and mutants upon promoter deletion

Raw RNA-seq fastq files were aligned to the reference genome (hg19) using BWA-mem. Duplicate reads were discarded with Picard to avoid any artifact caused by the amplification step originated from Ovation® RNA-Seq System V2 (NuGEN). Then, FPKM values were calculated using Cufflinks with GENCODE v19 annotation. Reproducibility between biological replicates were measured (PCC of FPKM for WT = 0.98, *ZNF891*

promoter deletion clone #1 =0.99, *ZNF891* promoter deletion clone #2 = 0.99, *ZNF891* promoter deletion clone #3=0.99 , and *ARIH2OS* promoter deletion clone #1 =0.98). FPKM values of *ZNF84* and *NCKIPSD* were investigated as distal target genes of *ZNF891* and *ARIH2OS*, respectively, between mutant and WT to test the effect of deletion of core promoters on distal target genes.

Experimental validation of promoter-proximal eQTL distal target genes

In order to validate the distal target genes of promoter-proximal eQTLs identified by pcHi-C results, we designed sgRNA sequences targeting +/- 5bp of the eQTLs in H1-hESC and cloned the sgRNAs into lentiCRISPRv2 backbone, followed by lentiviral preparation, infection and Puromycin selection as previously described[33]. Two weeks after the infection, single clones were selected and genotyped to confirm the mutations on the targeted eQTL sites (genotyping PCR results are listed in the oligo file). Total RNA was purified from each single clone and subjected to RT-qPCR analysis as previously described (Genotyping PCR primers are listed in the oligo file)[33]. To conduct statistical analysis, two separate sgRNAs were generated, which target the same eQTL. Then, three clones were isolated and cultured for a single sgRNA in order to induce the knockout, and each of these clones was considered as a biological replicate. Each clone was consisted of technical triplicates for the stable measurement of the expression during RT-qPCR experiment.

chr9:139305041_1 sgRNA in H1-hESC: GCCTTGGGCCGTCGGCGAGGGGG
chr9:139305041_2 sgRNA in H1-hESC: TGGGCCGTCGGCGAGGGGGAGGG
chr17:18128865_1 sgRNA in H1-hESC: GCGGGGCCGGGCCTGCACGGGGG
chr17:18128865_2 sgRNA in H1-hESC: CGCGCGGGGCCGGGCCTGCACGG
chr14:104029246_1 sgRNA in H1-hESC: CGAAGCCCGAGGAAGCGCGGCGG
chr14:104029246_2 sgRNA in H1-hESC: CGGCAGGGTCGCGAAGCCCGAGG
chr3:184032262_1 sgRNA in H1-hESC: GGCAAATCCCATGTGCTCGGCGG
chr3:184032262_2 sgRNA in H1-hESC: GGGGGCAAATCCCATGTGCTCGG

chr9:139305041_F genotyping primer: CGCTGGTAGCCCGACATC
chr9:139305041_R genotyping primer: CCCCGCTTCAGTCGTCAC
chr17:18128865_F genotyping primer: CCCAGTTCACCATTGTCTGG
chr17:18128865_R genotyping primer: AACCGAAACTTCATCATCTTGC
chr14:104029246_F genotyping primer: GAGGCAGCCTGGAGTGAC
chr14:104029246_R genotyping primer: GAGAAAGGTCTTCTTCCCCG
chr3:184032262_F genotyping primer: AATGAACTAAAGAATCGCGGAA
chr3:184032262_R genotyping primer: CACAGACGTAGTCCACAACCAT

Experimental validation of distal target genes for disease-associated genetic variants

In order to validate the distal target genes of disease-associated genetic variants (GWAS-SNPs) identified by pcHi-C results, we designed sgRNA sequences targeting +/- 5bp of the GWAS-SNPs in lymphoblastoid cells, and cloned the sgRNAs into lentiCRISPRv2 backbone as described above. Two weeks after the infection, single clones were selected and genotyped to confirm the mutations on the targeted GWAS-SNP sites (genotyping PCR results are listed in the oligo file). Total RNA was purified from each single clone and subjected to RT-qPCR analysis as previously described (Genotyping PCR primers are listed in the oligo file)[33]. To conduct

statistical analysis, two separate sgRNAs were generated, which target the same GWAS SNP. Then, two clones were isolated and cultured for a single sgRNA in order to induce the knockout, and each of these clones was considered as a biological replicate. Each clone was consisted of technical triplicates for the stable measurement of the expression during RT-qPCR experiment.

chr5.96297527 sgRNA in GM12878: TGCCATTCAGTCTATAGATCTGG
chr17.38032460 sgRNA in GM12878: TGGGCTTTGGCTGGGCGCAGTGG
chr17.38023745 sgRNA in GM12878: GGGCTCCATCCCTACAGAAAAGG
chr3.52707026 sgRNA in GM12878: GAGTTTTGCTCTTATTGTCCAGG
chr3.52703615 sgRNA in GM12878: AGTTATTACAAATAACATCATGG
chr3.52728804 sgRNA in GM12878: TCCTGGAAGATAGCATGCGTGGG
chr3.52706724 sgRNA in GM12878: GGTCTCGAACTCCTGCACTCAGG

chr5: 96297527_F genotyping primer: ACCAGTTTACACGAATCATCCC
chr17:38032460_F genotyping primer: TAGAGACAGAGTTTCGCCCTGT
chr17:38023745_F genotyping primer: TGGGCTCTCTCTACTAACCAGC
chr3:52707026_F genotyping primer: TGACAGCAAGAGAGGAAAGATG
chr3:52703615_F genotyping primer: TCAAATGAAGTTCCAGGAGACA
chr3:52728804_F genotyping primer: ACTTGGTAAGGCAGATGGAGAC
chr3:52706724_F genotyping primer: GTTCAAGTGATTCTCCTGCCTC
chr5: 96297527_R genotyping primer: ACTTCATCATGGGCAGTAAACC
chr17:38032460_R genotyping primer: AGGACCATTCTGTTTTCCTTCA
chr17:38023745_R genotyping primer: GTGACCTTGCTTTAAAAATGGG
chr3:52707026_R genotyping primer: AGGTGGGAGAATTGCTTGAAC
chr3:52703615_R genotyping primer: AACCTGTCAGCTAAGGTTCCAA
chr3:52728804_R genotyping primer: GCAAATTCAACCTAATCCGAAG
chr3:52706724_R genotyping primer: ATGCCTGTAATCCCAACACTTT

Extended GWAS-SNPs with high LD structure

GWAS-SNPs were obtained from GWAS catalogue database (version1.0.1, downloaded on February 2018) and selected with p-value cutoff of 10^{-5} with minor manual curations. As GWAS-SNPs obtained from GWAS catalog 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 that had tight LD scores (>0.8) across all five populations (AFR, AMR, EAS, EUR, and SAS). With the p-value cutoff of 10^{-5}, we collected 42,674 significant GWAS-SNPs across 2,310 GWAS mapped traits and expanded this list to 180,893 by including LD information. Then, putative target genes of GWAS-SNPs were identified by aggregating all unique pcHi-C interactions. We noted that the cutoff value of high LD association is arbitrarily determined by considering a stringent cutoff value presented in a set of previous studies to minimize additional noise in the data analysis.

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

The list of putative disease associated genes was downloaded from GeneCard database, obtaining 9,989 disease-associated genes. Then, we defined putative target genes of GWAS-SNPs associated with Parkinson disease by using pcHi-C interactions or the nearest gene information, respectively. Then, we counted the number of matched disease-associated genes in each set of putative GWAS-SNP target genes.

Clustering of GWAS mapped traits based on putative target gene similarities

The “mapped traits” were obtained from GWAS catalog database (version1.0.1, downloaded on February 2018), and paired with their corresponding GWAS SNPs. Then, putative target genes for each GWAS SNP were obtained by the unique and aggregated pcHi-C interactions. After defining putative target genes and their target frequency for each trait, we constructed a 1442 by 1442 correlation matrix where each entry indicates a similarity score between the mapped traits in terms of the Pearson correlation coefficient (PCC), for which only the traits with a total gene count greater than 5 were considered. The correlation matrix was subjected to K-means clustering (n=30) using Euclidean distance, and the cluster containing ungrouped terms was excluded in further analysis to eliminate miscellaneous terms. To avoid having a predetermined number of clusters, the remaining 687 traits were rearranged in a correlation matrix in terms of their hierarchical relationship (Pearson uncentered and complete linkage). The final hierarchically clustered correlation matrix showed a clear organization of 40 clusters with a threshold of dendrogram height, 0.9. Fig. 4c was drawn by using the nearest gene of GWAS SNPs. After defining the list of nearest genes for each mapped trait, we again measured the similarity between the mapped traits by calculating the Pearson correlation coefficient. We presented similarity values between the mapped traits as in the same order of mapped traits in Fig. 4b.

Similarly, Fig. 4d was drawn by using the GWAS SNPs alone. We measured the similarity of the mapped traits by calculating the Pearson correlation coefficient between GWAS SNPs of each trait, and presented the values as in the same order of mapped traits in Fig. 4b.

Analysis of functional enrichment using DAVID

To identify the enriched biological pathways in the GWAS mapped traits for the clusters, we extracted putative target genes associated with each cluster. Then, we performed Gene Ontology (GO) analysis using DAVID (6.8 version) to obtain the list of enriched biological pathways for each cluster with a cutoff p-value of 10^{-3} by using the GO_BP annotation selection. After that we constructed 40 (number of clusters) by 126 (number of GO_BP annotations) matrix where each entry indicates -log10(p-value) of corresponding GO_BP annotation. Next, we performed hierarchical clustering in respect to the enriched biological pathways by Pearson correlation matric and average linkage parameter. In Supplementary Table 17, we presented GO_BP annotation information.

To see the effect of multiple TSS co-existing in a DNA fragment during gene set enrichment analysis, we calculated the number of genes that are located in the defined DNA fragments for all genes and the genes in cluster 38. To see the effect of fragment-sharing TSS of genes on the enriched biological pathways, we submitted the genes in cluster 38 for enriched pathway analysis using three different queries; 1) total genes in

the cluster, 2) random selection of genes in case of fragment-sharing, and 3) after removal of the fragment-sharing genes, as illustrated in Supplementary Table 18. We did not observe any significant effect on gene set enrichment analysis caused by promoters shared by the same *Hin*dIII fragment with at least one other promoter.

References

1. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001-6 (2014).
2. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
3. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
4. Lettice, L.A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725-35 (2003).
5. Uslu, V.V. *et al.* Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nat Genet* **46**, 753-8 (2014).
6. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373**, 895-907 (2015).
7. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371-5 (2014).
8. Yu, M. & Ren, B. The Three-Dimensional Organization of Mammalian Genomes. *Annu Rev Cell Dev Biol* **33**, 265-289 (2017).
9. de Wit, E. *et al.* The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227-31 (2013).
10. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109-13 (2012).
11. Dixon, J.R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-6 (2015).
12. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-4 (2013).
13. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
14. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).
15. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611-27 (2015).
16. Sahlen, P. *et al.* Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. *Genome Biol* **16**, 156 (2015).
17. Jager, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6**, 6178 (2015).
18. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* **47**, 598-606 (2015).
19. Dryden, N.H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res* **24**, 1854-68 (2014).
20. 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).
21. Javierre, B.M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e19 (2016).
22. Freire-Pritchett, P. *et al.* Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *Elife* **6**(2017).
23. Siersbaek, R. *et al.* Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation. *Mol Cell* **66**, 420-435 e5 (2017).
24. Rubin, A.J. *et al.* Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat Genet* **49**, 1522-1528 (2017).
25. Orlando, G. *et al.* Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer. *Nat Genet* (2018).

112526. Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across
1126 human tissues. *Nature* **518**, 350-354 (2015).
112727. Schmitt, A.D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially
1128 Active Regions in the Human Genome. *Cell Rep* **17**, 2042-2059 (2016).
112928. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome.
1130 *Nature* **489**, 75-82 (2012).
113129. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-
1132 enhancers at key cell identity genes. *Cell* **153**, 307-19 (2013).
113330. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot
1134 analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
113531. Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer
1136 long-range associations. *Nature* **504**, 306-10 (2013).
113732. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat Biotechnol*
1138 **34**, 167-74 (2016).
113933. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element
1140 identification in mammalian cells. *Nat Methods* **14**, 629-635 (2017).
114134. Engreitz, J.M. *et al.* Local regulation of gene expression by lncRNA promoters,
1142 transcription and splicing. *Nature* **539**, 452-455 (2016).
114335. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human
1144 epigenomes. *Nature* **518**, 317-30 (2015).
114536. Richard, M., Drouin, R. & Beaulieu, A.D. ABC50, a novel human ATP-binding
1146 cassette protein found in tumor necrosis factor-alpha-stimulated synoviocytes.
1147 *Genomics* **53**, 137-45 (1998).
114837. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions
1149 reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
115038. Yang, D. *et al.* 3DIV: A 3D-genome Interaction Viewer and database. *Nucleic Acids*
1151 *Res* **46**, D52-D57 (2018).
115239. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137
1153 (2008).
115440. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association
1155 scan results. *Bioinformatics* **26**, 2336-7 (2010).
115641. Dahl, J.A. *et al.* Broad histone H3K4me3 domains in mouse oocytes modulate
1157 maternal-to-zygotic transition. *Nature* **537**, 548-552 (2016).
1158

1159

1160**Supplementary Information** is linked to the online version of the paper

1161at www.nature.com/ng.

1162

1163**Acknowledgement**

1164We thank members of the Ren laboratory for support and critical

1165suggestions throughout the course of this work. We are thankful to Dr.

1166Naoki Nariai (UCSD) for sharing LD information. This work was funded by

1167in part by the Ludwig Institute for Cancer Research (to B.R.), NIH

(U54HG006997 and 1R01ES024984, to B.R.), the Ministry of Science, ICT, and Future Planning through the National Research Foundation in Republic of Korea (2017R1C1B2008838 to I. J.), Korean Ministry of Health and Welfare (HI17C0328 to I. J.), and SUHF Fellowship (to I.J.).

Author Contributions

IJ, AS, YD and BR conceived the study. IJ, AS, and YD performed experiments with assistance from TL, CT, and SC. IJ, AJL, and DY performed data analysis with assistance from JE, MC, ZC, and CLB. DK supervised data analysis by DY. CK, EM, and CLB contributed to provide human brain tissue samples. BL and SK contributed to sequencing and initial data processing. IJ prepared the manuscript with assistance from AS, YD, AJL, JE, and BR. All authors read and commented on the manuscript.

Author Information

 The authors declare no competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.R. (biren@ucsd.edu) or I.J (ijung@kaist.ac.kr). All raw and processed data have been deposited in the GEO database under accession number GSE86189.

1192

1193

9950
100

Figure Legends

Figure 1. Genome-wide mapping of promoter-centered chromatin interactions in diverse human tissues and cell types.

a, A schematic of the pcHi-C procedure. **b,** Barplots of normalized promoter-centered chromatin interaction frequencies (y-axis) emanating from the *ADAMTS1* promoter (translucent gray). The identified chromatin interactions are shown below the axis (purple loops). Highlighted in translucent yellow are cell/tissue type specific interactions. **c,** Barplots of the number of chromatin interactions that span a given genomic distance are shown. Orange line indicates the accumulated fraction of chromatin interactions from all 27 tissues/cell types. **d,** Boxplots showing the fold enrichment of the interaction frequencies between the active (colored dots) or bivalent promoters (gray dots) and each chromatin state. The 17 chromatin states shown were obtained by processing 18-state ChromHMM model after merging genic enhancer 1 and 2 annotations. KS-tests were performed between interactions originating from active promoter regions (colored dots) and those from bivalent promoters (gray dots) (** p value < 0.01 and *** p value < 0.001). The chromatin states that interact more frequently with active promoters than bivalent promoters were highlighted in translucent yellow. The chromatin states that interact more frequently with bivalent promoters than active promoters were highlighted in translucent blue. Whiskers correspond to the highest and lowest points within 1.5× the interquartile range.

Figure 2. Inference of target genes of *cis*-regulatory sequences from pcHi-C data.

a, Illustrative LocusZoom plot of eQTLs for *VLDLR* (top) and pcHi-C interactions originating from the *VLDLR* promoter region in aorta tissue (bottom). Dots along the LocusZoom plot represent the P-values of SNPs' association with *VLDLR* gene expression levels in the aorta (data obtained from GTEx). Dots are also color-coded based on their Linkage Disequilibrium (LD) scores with a tagging SNP. The blue bars indicate the recombination rate. **b,** Barplots showing fold enrichment between the number of eQTL-associations matched to P-O pcHi-C interactions and that of distance matched random P-O pcHi-C interactions for 12 corresponding tissue types. P-O interactions in all 12 tissues were significantly enriched for eQTL associations (empirical p value < 0.01). The dotted line indicates the expected fold-enrichment (i.e. 1). Error bars indicate standard deviation obtained by 1,000 random trials. **c,** An illustrative example of tissue specifically expressed gene, showing positive correlation between the chromatin state (H3K27ac) at a distal *c*RE and expression levels (RNA-seq) of the promoter connected by long-range chromatin interactions. The significant chromatin interaction between the *POU3F3* promoter and a distal *c*RE marked by H3K27ac ~350kb upstream in hippocampus (HC) tissue is shown at the top. Shown below are H3K27ac signals and locations of genes. The bar plots at the lower half show the H3K27ac signals at the distal *c*RE (left), the transcript levels of the *POU3F3* (middle), and the normalized pcHi-C interaction frequencies between the

POU3F3 promoter and the distal *c*RE (right). **d,** Boxplots illustrating the H3K27ac signals after quantile normalization at the *c*REs exhibiting hippocampus specific pcHi-C interactions with putative target promoters. These *c*REs are marked by higher levels of H3K27ac in hippocampus than in other cell/tissues types (KS-test p value < 0.005). Whiskers correspond to the highest and lowest points within 1.5× the interquartile range. **e,** Boxplots showing transcript levels of the putative target genes predicted by hippocampus specific pcHi-C interactions. Genes are significantly expressed in hippocampus compared to other cell/tissues types (KS-test p value < 0.005) except dorsolateral prefrontal cortex (KS-test p value 0.27) and mesenchymal stem cell (KS-test p value 0.02). Whiskers correspond to the highest and lowest points within 1.5× the interquartile range. **f-h**, Heatmaps demonstrate the enrichment of pcHi-C interactions for cell/tissue-specific *c*RE-promoter pairs (column) in the corresponding cell/tissue type (row) (f), z-score transformed H3K27ac signals (column) at the promoter associated *c*REs (row) (g), and z-score transformed FPKM values (column) of RNA-seq at the *c*REs' putative target genes (row) (h). Color indicates mean values of distance normalized pcHi-C interaction frequencies for H1 (n=5,096), MES (n=3,380), MSC (n=5,188), NPC (n=1,295), TB (n=5,830), HC (7,100), FC (n=15,733), IMR90 (n=5,313), LG (n=1,101), LI (n=2,656), PA (n=2,751), SB (n=1,072), TH (n=2,233), GA (n=1,511), LV (n=1,501), PO (n=865), RV (n=1,049), SX (n=9,228), AD (n=1,998), AO (n=4,407), and LCL (n=10,283) (f), z-score transformed H3K27ac signals for H1 (n=5,813), MES (n=3,951), MSC (n=5,790), NPC

(n=1,631), TB (n=6,616), HC (7,712), FC (n=15,389), IMR90 (n=6,146), LG (n=1,345), LI (n=3,224), PA (n=3,211), SB (n=1,310), TH (n=2,717), GA (n=1,903), LV (n=1,741), PO (n=1,087), RV (n=1,296), SX (n=10,077), AD (n=2,342), AO (n=5,179), and LCL (n=10,945) (g), and *z*-score transformed FPKM values for H1 (n=1,589), MES (n=1,024), MSC (n=1,587), NPC (n=450), TB (n=1,920), HC (2,339), FC (n=4,830), IMR90 (n=1,743), LG (n=310), LI (n=870), PA (n=845), SB (n=293), TH (n=747), GA (n=460), LV (n=368), PO (n=281), RV (n=295), SX (n=3,054), AD (n=550), AO (n=1,381), and LCL (n=3167) (h). KS-test was performed between pcHi-C interaction frequencies, *z*-score transformed H3K27ac signals, and *z*-score transformed FPKM values in the matched cell/tissue types (values in diagonal in each heatmap) and those in other cell/tissue types (values in off diagonal in each heatmap), demonstrating significant association of *c*RE-promoter pairs with cell/tissue-specific *c*RE H3K27ac signals and gene expression (KS-test p value < 2.2e-16).

Figure 3. Enhancer-like promoters involved in regulation of distal target genes.

a, Browser snapshots of the *TMED4* locus showing H3K27ac signals and promoter-centered chromatin interactions. Shown at the RefSeq genes (top), H3K27ac histone modification signals as measured by ChIP-seq (middle) and promoter-centered chromatin interactions detected from pcHi-C experiments (bottom). Highlighted in translucent blue are promoter-promoter pairs showing highly correlated H3K27ac signal and

significant pcHi-C interactions. Highlighted in gray is an adjacent promoter of the *TMED4*. Shown below are Pearson correlation coefficient (PCC) values based on H3K27ac signals and links based on pcHi-C interactions, with MSC as the acronym for mesenchymal stem cell. **b,** Density plots showing distributions of PCC values of H3K27ac (blue, median of PCC=0.45, n=48,893), H3K4me1 (orange, median of PCC=0.67, n=48,893), and H3K4me3 (green, median of PCC=0.64, n=48,893) signals for P-P pcHi-C interactions. As a control, a density plot of PCC distributions of H3K27ac signals for randomly selected promoter-promoter pairs is shown (gray, median of PCC=0.02, n=48,142). X-axis indicates PCC of histone modification signals between promoter-promoter pairs across 27 cell/tissue types. **c,** A pie chart showing the fraction of unique P-P interactions matched by eQTL associations, of which 5.7% are P-P interactions (n=1,976) in 12 matched tissue types (n=34,880). **d,** An illustrative LocusZoom plot of eQTLs for *DACT3* gene expression in dorsolateral prefrontal cortex. Both the *DACT3* gene promoter region and the *AP2S1* gene promoter that contains significant eQTLs are highlighted in translucent orange, dots along the LocusZoom plot represent SNPs, and their significance of association with the *DACT3* gene expression is plotted along the left y-axis. Dots are also color-coded based on their LD score with a tag SNP (rs78730097). The blue line indicates the estimated recombination rate, as plotted along the right y-axis. Gene expression levels detected by RNA-seq and RefSeq genes are plotted below the LocusZoom plot. **e,** Illustrative genome browser snapshot of RNA-seq

results between control and mutant clones with deletion of the core promoter regions of the *ARIH2OS*. In both control and mutant cells, the *ARIH2OS* gene was not expressed, but the expression of the *NCKIPSD* gene, which displays chromatin interactions with the *ARIH2OS* gene promoter, was significantly down-regulated in the mutant clones (FDR adjusted p value from cuffdiff = 0.02). **f,** Genome browser snapshot showing the promoter containing an eQTL targeted by sgRNAs and its distal target gene, *ABCF3*, together with H3K27ac and chromatin accessibility (DNaseI). The relative mRNA expression levels of the *ABCF3* quantified by RT-qPCR are shown below, which were significantly down-regulated in both mutants (*** one-tailed KS-test p value < 0.001). Error bars indicate standard deviation of three mutant clones with technical triplicates.

Figure 4. Analysis of human diseases and physiological traits based on the putative target genes of GWAS-SNPs.

a, Genome browser snapshot showing multiple *c*REs harboring GWAS-SNPs and their common target gene, *NT5DC2,* together with signals of H3K27ac (ChIP-seq) and chromatin accessibility (DNaseI). The DNA fragments containing all these *c*REs interact with the *NT5DC2* gene promoter region as evidenced by pcHi-C analysis (arcs). The relative mRNA expression levels of the *NT5DC2* upon induced mutation of GWAS-SNPs with sgRNA were quantified by RT-qPCR as shown below. Error bars indicate standard deviation of two mutant clones with technical triplicates

(KS-test, ** p value < 0.01, *** p value < 0.001). **b,** Hierarchical clustering of human diseases and traits based on similarities of the putative target genes of trait-associated SNPs and SNPs in LD. The color intensity of each dot indicates Pearson correlation coefficient (PCC) of the putative target genes between two diseases or traits. Color bars on the left and top demarcate the clusters. **c, d,** Shown are similarities, as measured by Pearson correlation coefficient (PCC), between traits in the same order as Fig. 4b, based on either the nearest genes of the GWAS SNPs (c) or the GWAS SNPs alone (d). The color intensity of each dot indicates PCC of target gene similarities between two traits. **e,** Hierarchical clustering of GO biological processes (each column, n=126) for the trait clusters defined in Fig. 4b (each row, n=40). Each entry indicates –log10(p-value) value of GO biological processes in the corresponding cluster. Several representative biological processes are highlighted.

Extended Data Figure Legends

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

a, Schematic of probe design for Promoter Capture Hi-C experiments. For each promoter (black rectangle), two flanking *Hin*dIII cut sites were identified. A 15bp buffer was then added to each side of the *Hin*dIII cut site, followed by allocation of three 120-mer capture probes to the same sites, with a 30bp shift between the adjacent probes. In total, 12 capture probes were designed for each promoter and all probes were targeted to the Watson Strand. **b,** Schematic workflow of custom RNA probe synthesis. Single stranded DNA (ssDNA) probe synthesis by CustomArray, Inc., is shown from top to bottom; PCR amplification with SP6 recognition sequence completion and purification, BsrDI digestion 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, followed by combining the mixes and overnight incubation to bind probes to Hi-C template. Then, preparation of streptavidin beads and wash buffers, followed by binding of 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.

Extended Data Figure 2. Overview of samples and capture probe quality control.

a, Schematic overview of 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,** Bar plots showing the fraction of number of TSS in a DNA fragment. **c,** Scatter plot showing the reproducibility of probe density from RNA-seq data between two probe synthesis experiments. Each dot on the scatter plot represents a single promoter and the value is the aggregated probe density from all probes assigned to that given promoter. **d,** Venn diagram showing the number of targeted regions that contain detectable probe density based on RNA-seq of the capture probes from each replicate of probe synthesis. **e,** 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. **f,** 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 *Hin*dIII cut sites and GENCODE gene annotations, illustrating that the vast majority of probe density is only found around *Hin*dIII restriction sites flanking promoters. **g, h,** Histograms of the probe densities measured by RNA-seq (x-axis) in each promoter from replicate 1 (g) and replicate 2 (h) of probe synthesis.

Extended Data Figure 3. General characterization of promoter-centered long-range interactions.

a, Identified pcHi-C chromatin interactions across multiple cell/tissue types are plotted in Genome Browser, with the darkness of blue corresponding to the strength of interactions. RefSeq genes are presented below the snapshot. **b,** Fraction of pcHi-C interactions uniquely detected in one cell/tissue type (green) or also detected in other cell/tissue types (orange). The average fraction of cell/tissue-specific interactions is not over-estimated due to the number of tested samples (at 22 samples the fraction of cell/tissue-specific interactions reach plateau) and tissue-heterogeneity (similar trend was observed when we only considered pcHi-C interactions obtained from cell lines). **c,** Snapshot of a locus showing promoter-centered long-range interactions from pcHi-C data in H1-hESC (bottom, purple loops) in the context of TAD annotations (blue rectangles) identified from Hi-C data (top, red) in H1-hESC. RefSeq genes are shown at the bottom. **d,** Fraction of P-O pcHi-C chromatin interactions in the context of TAD annotations with the respective cell/tissue types.

Extended Data Figure 4. Validation of Promoter Capture Hi-C Interactions.

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. The top two tracks show histone modification signals for H3K4me3 and H3K27ac, followed by RefSeq

genes. Below are pcHi-C chromatin in IMR90 (blue loops) and promoter-centered chromatin interactions from high-resolution Hi-C data in IMR90 (reddish brown loops). **b-e,** ROC plots illustrating the prediction performance of Promoter Capture Hi-C result for *in situ* Hi-C loops anchored at promoters in lymphoblastoid (b), IMR90 (c), hippocampus (d), and dorsolateral prefrontal cortex (e). Promoter centered interactions for *in situ* Hi-C loops were considered as true interactions, and ROC plots were drawn for the corresponding pcHi-C result. ROC scores are shown in the ROC plot. **f,** ROC plots showing the reproducibility of pcHi-C chromatin interactions between biological replicates. pcHi-C interactions from one replicate were used as true interactions, and ROC plots were drawn for the other replicate. **g-k**. Venn diagrams presenting the number of commonly identified pcHi-C interactions between biological replicates for lymphoblastoid (g), dorsolateral prefrontal cortex (h), mesenchymal stem cell (i), lymphoblastoid processed by CHICAGO (j), and GM12878 with previously published pcHi-C data[18] (k). Hypergeometric p-values are shown together. **l-m**, Illustration of interaction intensity in the replicates of lymphoblastoid (l) and mesenchymal stem cells (m), depending on the replicate consistency. Whiskers correspond to the highest and lowest points within 1.5× the interquartile range.

Extended Data Figure 5. Integrative analysis of long-range chromatin interactions with epigenome.

a, b, Shown are histograms of number of interacting *c*REs per promoter (a) and number of interacting promoters per *c*RE (b). Y-axis indicates frequency of the corresponding value in x-axis. **c,** Depiction of identified long-range promoter-centered interactions across a 0.84Mb locus in lymphoblastoid (top). Shown below are histone modification signals obtained from ChIP-seq analyses[35], as well as accessible chromatin regions measured from DNaseI hypersensitivity assay. **d,** Depiction of extensively interacting DNA fragments (EIF) from P-P and P-O interactions, and transcription factor (TF) binding clusters identified in GM12878 cells for the same region shown in Extended Data Fig. 5c. Below are 67 TF binding profiles obtained from TF ChIP-seq results performed in GM12878 cells. Highlighted in translucent blue are overlapping EIF and TF binding clusters. EIF was defined in each cell/tissue type by selecting frequently interacting DNA fragments with multiple promoters in terms of 0.01 Poisson p value cutoff. **e, f,** Bar plots showing the number of P-O EIF overlapping with TF clusters compared to random expectation in lymphoblastoid (e) and H1-hESC (f). Error bars indicate standard deviation of expectation values calculated by using typical TF peaked regions (blue) and generating random genomic regions (green). Empirical p-value shows statistical significance (*** p value < 0.001). **g, h,** Bar plots showing the number of P-P EIF overlapping with TF clusters compared to random expectation in lymphoblastoid (g) and H1-hESC (h). Error bars indicate standard deviation of expectation values calculated by using typical TF peaked regions (blue) and generating random genomic regions (green).

Empirical p-value shows statistical significance (*** p value < 0.001). **i,** An array of bar plots showing the number of P-O EIF overlapping with super-enhancers (first bar plot, orange), compared to typical enhancers (middle bar plot, blue) and random genomic regions (last bar plot, purple). Error bars indicate standard deviation of expectation values obtained by 10,000 permutations. Empirical p-value showed statistical significance for all tested cell/tissue types compared to random genomic regions (p value < 0.0001).

Extended Data Figure 6. Enrichment of long-range chromatin interactions at various chromatin states generated by a 50-state ChromHMM model.

a, Boxplots showing the fold change of interaction frequencies between active/bivalent promoters and each chromatin state over expected values. The 50 chromatin states (E01-E50) were obtained from the 50-state ChromHMM model. KS-tests were performed between active promoters and bivalent promoters (two adjacent boxplots) (** p value < 0.01 and *** p value < 0.001). The chromatin states that interact more frequently with active promoters than bivalent promoters were highlighted in pink asterisk. The chromatin states that interact more frequently with bivalent promoters than active promoters were highlighted in blue asterisk. Whiskers correspond to the highest and lowest points within 1.5× the interquartile range. **b,** A heatmap showing an emission parameter matrix of each chromatin state in which each row corresponds to a different

chromatin state and each column corresponds to an emission probability of a chromatin mark shown at the top. The pre-calculated emission parameter heatmap was downloaded from the 50-state ChromHMM model established by Roadmap Epigenomics Project.

Extended Data Figure 7. Validation of P-O interactions with eQTL associations.

a-c, Illustrative LocusZoom plots of eQTLs for the *HS3ST1* (a), the *METTL25* (b), and the *DAAM1* (c) gene expression in left ventricle, dorsolateral prefrontal cortex, and aorta, respectively. RefSeq genes are plotted below the LocusZoom plot. Identified pcHi-C interactions are shown as loops (purple) in the bottom. **d,** Array of bar plots showing the number of matched eQTL associations between P-O pcHi-C chromatin interactions after exclusion of DNA fragment shared promoters and random expectation across 14 matched tissue types from GTEx database. All P-O pcHi-C interactions are significantly enriched by eQTL associations compared to random P-O pcHi-C interactions with or without distance match (* empirical p-value <0.05, ** empirical p-value <0.01, *** empirical p-value <0.001). Error bars indicate standard deviation of random expectation values. **e**, Density plots showing the number of unique eQTLs per P-O pcHi-C interaction fragment and randomized interactions. No significant difference between pcHi-C interactions and randomized interactions (KS-test p value > 0.05) except pancreas (p value = 0.02), gastric (p value = 0.009), and lung (p value = 0.03). **f,** Shown are

boxplots of the distribution of PCC between H3K27ac signals in *c*RE-promoter pairs connected by pcHi-C interactions after exclusion of multiple fragment spanning *c*REs (Orange, n=154,055), compared to the distribution of random expectation with matched distance (dark gray, n=154,055) and without matched distance (gray, n=154,055). We only considered P-O pairs where other DNA fragments are marked by H3K27ac peaks in at least one cell/tissue type analyzed. We also excluded two fragments spanning *c*REs. KS-test was performed between P-O pairs and random control, demonstrating that P-O pairs showed significant positive correlation (*** Welch's t-test p value < 2.2e-16). Whiskers correspond to the highest and lowest points within 1.5× the interquartile range. **g,** Similar to Extended Data Fig. 7e, but the distribution of PCC between H3K27ac signals at a *c*RE and target gene expressions of the *c*RE connected by pcHi-C interactions. KS-test was performed between P-O pairs (orange, n=154,055), distance matched random control (dark gray, n=154,055), and random control (gray, n=154,055), revealing that P-O pairs showed significant positive correlation (*** Welch's t-test p value < 2.2e-16). Whiskers correspond to the highest and lowest points within 1.5× the interquartile range.

Extended Data Figure 8. Functional analysis of promoter-promoter interactions.

a, Pie chart showing the fraction of promoter-promoter interactions (P-P) among all pcHi-C interactions. The fraction of P-P pcHi-C interactions

modestly decrease to 6.5% after excluding fragment that harbor multiple promoters. **b,** An array of bar plots showing the number of eQTL associations matched to P-P pcHi-C interactions (left, purple), compared to random expectation with matched distance (middle, blue) and without matched distance (right, light blue). Each bar plot represents analysis of a different tissue. Error bars indicate standard deviation of random expectation values. Empirical p values are shown at the top (* < 0.05, ** < 0.01, *** < 0.001). **c, d,** Illustrative LocusZoom plots of *FHOD1* eQTLs (c) and *POFUT2* eQTLs (d) in left ventricle and aorta, respectively. Promoters that contain significant eQTLs and target promoters are highlighted in translucent orange. Dots along the LocusZoom plot represent SNPs, and their significance of association with *FHOD1* and *POFUT2* gene expression is plotted along the left y-axis, respectively. The blue line traveling across the scatterplot indicates the estimated recombination rate, as plotted along the right y-axis. RefSeq genes and RNA-seq are plotted below the LocusZoom plot. pcHi-C interactions are shown as purple in the bottom. **e,** Bar plot showing the eQTL associations between the SNP rs78730097 and surrounding genes, showing the most significant association with the distal gene *DACT3*. Y-axis indicates –log10(eQTL association p value). **f, g,** Bar plots showing FPKM values of distal target gene expressions upon deletion of core promoter regions of the *ARIH2OS* (f) and the *ZNF891* (g). Two biological replicates of one mutant clone for the *NCKIPSD* and two biological replicates of three mutant clones for the *ZNF84* were measured using RNA-seq, respectively.

FDR-adjusted p value obtained from cuffdiff is shown together. N.S indicates statistically non-significant. **h**, Bar plots showing FPKM values of two nearby genes of the *ARIH2OS* and one nearest gene of the *NCKIPSD* (y-axis) upon deletion of core promoter regions of the *ARIH2OS*. The *ARIH2*, a DNA fragment sharing gene with the *ARIH2OS*, is excluded. FDR-adjusted p value obtained from cuffdiff is shown together. Corresponding gene name is shown on the top of bar plots. **i**, The relative mRNA expression levels of distal target genes (orange) and nearby genes (gray) of promoter-proximal eQTLs quantified by RT-qPCR are shown. Error bars indicate standard deviation from total six mutant clones for two separate sgRNAs with technical triplicates. One-sided KS-test p values are shown together on the top of each bar plot (*** p value < 0.001).

Extended Data Figure 9. Identification of target genes of disease-associated genetic variants.

a, Illustration of the strategy to identify target genes of each GWAS trait. An example result is shown for Alzheimer disease. Both known and novel target genes were identified according to literature search. **b,** Venn diagram showing number of target genes by pcHi-C interactions and by nearby gene information for the GWAS-SNPs associated with Parkinson disease. **c,** Number of matched disease-associated genes in each group of target genes identified in Parkinson disease. **d,** Fraction of distal genes (blue) and nearby genes (gray) among the identified target genes of GWAS-SNPs based on pcHi-C interactions (left). Expected fraction is shown

by calculating the fraction of nearby genes when we consider a nearest gene over 15kb as a GWAS-SNP target gene (right). **e,** Barplot showing the relative mRNA expression levels of *GNL3* upon induced mutation of GWAS-SNPs with sgRNA as quantified by RT-qPCR as a control. Error bars indicate standard deviation of two mutant clones with technical triplicates. **f,** Barplot showing RT-qPCR results of relative target gene expression (y-axis) between mutant and control. Error bars indicate standard deviation of two mutant clones with technical triplicates. The mutants showing significant down regulation of target genes are shown in orange (KS-test, ** p value < 0.01, *** p value < 0.001). sgRNA target GWAS-SNP genomic coordinate, rsID, associated disease, distal target gene information, high LD SNP on coding region, and related publication PMID information are shown together.

Extended Data Figure 10. Analysis of disease-disease associations.

a, Illustration of the strategy to calculate the similarity between GWAS mapped traits using target gene similarity information. **b, c,** Shown are similarities, as measured by Pearson correlation coefficient (PCC), between traits in the same order as Fig. 4b based on similarities of the putative GWAS-SNP target genes without shared promoters (b) and without genes located in HLA and HIST locus (c). The color intensity of each dot indicates Pearson correlation coefficient (PCC) of the putative target genes between two diseases or traits. **d,** Shown are similarities, as

measured by Pearson correlation coefficient (PCC), between traits based on the 5 nearest genes of the GWAS SNPs. The color intensity of each dot indicates PCC of target gene similarities between GWAS mapped traits. **e,** Bar plots showing the fraction of number of TSS in a DNA fragment between all TSS and TSS corresponding genes in cluster 38 of Fig. 4b.

Supplementary Tables

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

Supplementary Table 2. Number of processed pcHi-C reads

Supplementary Table 3. List of P-O interactions

Supplementary Table 4. List of P-P interactions

Supplementary Table 5. Number of significant pcHi-C promoter-centered interactions

Supplementary Table 6. The list of mean and median distance of pcHi-C and eQTL associations in each cell/tissue type

Supplementary Table 7. The numbers and fractions of overlapped interactions between replicates

Supplementary Table 8. Total number of extensively interacting DNA fragments (Poisson P value < 0.01)

Supplementary Table 9. List of TF ChIP-seq data used to define GM12878 TF clusters

Supplementary Table 10. List of TF ChIP-seq data used to define H1-hESC TF clusters

Supplementary Table 11. Summary of matched eQTL-associations with P-O pcHi-C interactions

Supplementary Table 12. List of P-O pcHi-C interactions and matched eQTL relationships

Supplementary Table 13. Summary of matched eQTL-associations with P-P pcHi-C interactions

Supplementary Table 14. List of P-P pcHi-C interactions and matched eQTL relationships

Supplementary Table 15. Summary of average number of target genes of GWAS-SNPs

Supplementary Table 16. List of putative target genes of GWAS-SNPs

Supplementary Table 17. List of GWAS mapped traits and enriched GO biological processes in Fig. 4b

Supplementary Table 18. Enriched pathway analysis of Cluster 38 in Fig. 4b