HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2020 March 09.

Published in final edited form as:

Nat Genet. 2019 October ; 51(10): 1442–1449. doi:10.1038/s41588-019-0494-8.

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

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,†]

[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

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

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

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

[†]Correspondence to Inkyung Jung (ijung@kaist.ac.kr) and Bing Ren (biren@ucsd.edu).
[*]These authors contributed equally to this work

Author Contributions

I.J., A.S., Y.D. and B.R. conceived the study. I.J., A.S., and Y.D. performed experiments with assistance from T.L., C.T., and S.C., I.J., A.J.L., and D.Y. performed data analysis with assistance from J.E., M.C., Z.C., and C.L.B., D.K. supervised data analysis by D.Y., C.K., E.M., and C.L.B. contributed to provide human brain tissue samples. B.L. and S.K. contributed to sequencing and initial data processing. I.J. prepared the manuscript with assistance from A.S., Y.D., A.J.L., J.E., and B.R.. All authors read and commented on the manuscript.

Competing Interests Statement

Bing Ren is a co-founder of Arima Genomics, Inc.. Anthony Schmitt is an employee of Arima Genomics.

Abstract

A large number of putative *cis*-regulatory sequences have been annotated in the human genome, but the genes they control remain poorly 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.

Genome-Wide Association Studies (GWAS) have identified 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), 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; Supplementary 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 (Supplementary 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 (Supplementary Fig. 2b; see Methods). The

robustness and the coverage of capture probe synthesis were validated by sequencing (Supplementary 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 >15 kb 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 (Supplementary 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 promoter-other and promoter-promoter interactions, respectively) (see Methods). Focusing on the *Hin*dIII fragments over 15 kb away and within 2 Mb of each promoter, we determined a total of 892,014 chromatin interactions (431,141 unique interacting pairs) in one or more of the 27 human cell/tissue types (Fig. 1b; Supplementary Fig. 3a; Supplementary Table 3–5). A total of 18,943 promoter regions were involved in at least one significant chromatin interaction in one or more cell/tissue types analyzed in this study. The median distance between the interacting DNA pairs was 158 kb, which is within a similar range of previously reported chromatin loops and eQTL associations (Supplementary Fig. 3b; Supplementary Table 6)[10,12,13,27]. The slight discrepancy between Promoter Capture Hi-C (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 (Supplementary Fig. 3c). As expected, most of the detected chromatin interactions were within Topologically Associating Domains (TADs) defined in the corresponding tissue/cell type (Supplementary Fig. 3d, e)[28,11].

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 (Supplementary Fig. 4a–e) (see Methods). Additionally, we found high reproducibility of pcHi-C chromatin interactions between two biological replicates (average ROC score = 0.85; the average Spearman's rank correlation between replicates = 0.4; Supplementary Fig. 4f–j; Supplementary Table 7; see Methods), and between two independent studies (Supplementary Fig. 4k). The observation that interactions identified in both replicates exhibited the strongest interaction signals, while interactions identified in one replicate were moderately strong but moderately weak in the other replicate (Supplementary 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 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 (Supplementary Fig. 5a), while 45% of *c*REs were assigned to one candidate target gene (Supplementary Fig. 5b), similar to the previous observation with DNase I hypersensitivity analysis across diverse human cell types[29]. 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 (Supplementary Fig. 5c–i; Supplementary Table 8–10; see Methods)[30]. 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. 1c). 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. 1c). Further analysis based on a refined 50-chromatin-state ChromHMM model for 5 cell lines also supports our conclusion (Supplementary Fig. 6).

Three lines of evidence support 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; Supplementary Fig. 7a–c)[27]. 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) (Supplementary Fig. 7d,e). A total of 42,627 eQTL associations were detected by P-O 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 levels of H3K27ac in these *c*REs were significantly correlated with both the promoter H3K27ac levels (Supplementary Fig. 8a) and transcription levels of the predicted target genes (Supplementary Fig. 8b) across these tissues/cell types. For example, *POU3F3* gene expression (second column in Fig. 2b) was highly correlated with H3K27ac signals in the distal *c*RE (first column in Fig. 2b) connected by a tissue-specific chromatin interaction (last column in Fig. 2b). 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. 2c) and highly expressed genes, albeit modestly, (Supplementary Fig. 8c) 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. 2d–f; 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[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 < 2.2×10^{-16}, Supplementary Fig. 9a). 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[32–35]. 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; Supplementary Fig. 9b–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 (NC_000019) for *DACT3* gene was located in the *AP2S1* promoter in the dorsolateral prefrontal cortex (Supplementary Fig. 10a). 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* (Supplementary Fig. 10b), 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 (Supplementary Fig. 10c, d; Supplementary Table 15; 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. 3d) with no significant or moderate effect on nearby genes (Supplementary Fig. 10e). 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. 3e; Supplementary Fig. 10f; Supplementary Table 16; 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 42,633 putative disease/trait-associated genetic variants from a recent public repository of GWAS catalog[1]. Consistent with previous reports[2,36], a significant portion of SNPs (30%, Fisher's Exact test *P* value < 2.2×10^{-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 (Supplementary Fig. 11a; Supplementary Table 17; 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's

disease in Supplementary Fig. 11b, 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 (Supplementary Fig. 11d). 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 the human lymphoblastoid cell line GM12878. We introduced deletions to these elements in GM12878 cells 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; Supplementary Fig. 11e–f; Supplementary Table 18; see Methods). 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; Supplementary Fig. 12a–c; Supplementary Table 19; 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, Supplementary Fig. 12d). 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 20; see Methods). The enriched GO biological processes suggest potential shared molecular pathways for disease and trait types in each cluster (Fig. 4e, Supplementary Fig. 12e, Supplementary Table 21), 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[37]. 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 systematic 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[38]. 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 oligonucleotide sequences were designed to be of 120 nucleotide (nt) length and to have 30 nt tiling overhang. Oligonucleotides were designed ± 300 bp upstream and downstream of each restriction site. As two restriction sites were chosen for each transcription start site, a total of 12 capture oligonucleotides 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 oligonucleotide design generated 280,445 unique probe sequences including randomly selected capture regions (i.e. gene deserts). Single-stranded DNA oligonucleotides were then synthesized by CustomArray Inc. Single-stranded DNA oligonucleotides contained universal forward and reverse primer sequences (total length 31 nt), whereby the forward priming sequence contained a truncated SP6 recognition sequence that was completed by the overhanging forward primer during PCR amplification of the oligonucleotides. 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.

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 500 ng Hi-C library for 24 h at 65 °C in a humidified hybridization chamber with 2.5 μg human Cot-1 DNA (Life Technologies), 2.5 μg salmon sperm DNA (Life Technologies), and p5/p7 blocking oligonucleotides with hybridization buffer mix (10× SSPE, 10 mM EDTA, 10× Denhardts solution, and 0.26% SDS) and 500 ng RNA probes. RNA:DNA hybrids were enriched using 50 μl T1 streptavidin beads (Invitrogen) through 30 min incubation at room temperature (RT). Next, bead-bound hybrids were washed through a 15 min incubation in wash buffer1 (1× SSC and 0.1% SDS) with frequent vortexing, and then washed three times with 500 μl of pre-warmed (65 °C) wash buffer2 (0.1× 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 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 (< 15 kb 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 3 kb. As a result, 510,045 DNA fragments were defined with a median length of 4.8 kb. 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 15 kb and within 2 Mb 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 expected interaction frequency between DNA fragment i and j with coverage C_i and C_j and defined the residual $R_{ij} = Y_{ij}/\exp\left(\hat{\beta}_0 + \hat{\beta}_1 BS\left(C_i\right) + \hat{\beta}_2 BS\left(C_j\right)\right)$ 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 two independently prepared 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 (Supplementary 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 (Supplementary 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 5 kb, 10 kb, and 25 kb 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 (two different donors for tissues and two independently cultured cells for cell lines) was measured in terms of ROC curve (Supplementary 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 2 Mb windows size were downloaded from the 3DIV database[39]. The regions between TADs were classified as "unorganized" when the gap is longer than 400 kb, 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 cells, mesendoderm, mesenchymal stem cells, trophoblast, and IMR90, using a 50-state ChromHMM model produced by the Roadmap Epigenomics Project[36]. 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 50-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 cells (GM12878) and human embryonic stem cells (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[40]. 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 number of TFs within cluster (5 or 10), minimum length of cluster from 100 bp to 1,600 bp, and maximum gap length within cluster from 100 bp to 51.2 kb. Final TF clusters were defined when the region was detected as TF clusters more than 50 times from 100 different parameter sets.

Enrichment analysis of TF clusters and super-enhancers

In order to calculate the enrichment of TF clusters or super-enhancers at extensively interacting DNA fragments (EIF), we counted the number of matched TF clusters and super-enhancers. The list of super-enhancers was obtained from the 3DIV database[39]. Permutation test was performed to calculate the expected values. Using Bedtools shuffleBed, we generated random genomic locations that resemble actual TF clusters with the same size but different genomic coordinates. Bedtools intersectBed identified any overlap between EIF and TF clusters or random genomic locations. Expected values in the random genomic locations were calculated from 10,000 random data sets. In order to test the enrichment of TF clusters compared to typical TF peaks, we generated random genomic locations that resemble actual TF clusters with the same size but different genomic coordinates matched to typical TF peaks. Standard deviations of error bars in the typical TF peaks were calculated from 10,000 random data sets. Similarly, enrichment analysis of super-enhancers was conducted by generating random genomic locations of the same size as super-enhancers but at different genomic coordinates. We also conducted the enrichment test with typical enhancers. 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 1 Mb in distance to match the range of eQTL association, and discarded eQTL associations with distance below 15 kb 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 40 kb, 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 15 kb 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.

Statistics

We used the Kolmogorov-Smirnov (KS) test to compare distributions between two groups as a nonparametric test without assumptions of normality. We also used permutation test to calculate empirical *P* values, which does not make any assumptions on the underlying distribution of the data.

Supplementary Note

Further information on Methods is available in the Supplementary Note.

Reporting Summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data Availability

All raw and processed data have been deposited in the GEO database under accession number GSE86189. Visualization of processed Promoter Capture Hi-C data is available at http://www.3div.kr/capture_hic.

Code Availability

Code for pcHi-C interaction detection can be made available on request. For other data analysis, we used publicly available software.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank members of the Ren laboratory for critical suggestions in the course of this work. We are thankful to N. Nariai (UCSD) for sharing LD information. This work was funded in part by the Ludwig Institute for Cancer Research (to B. Ren), NIH (1R01ES024984, to B. Ren), the Ministry of Science, ICT, and Future Planning through the National Research Foundation in Republic of Korea (2017R1C1B2008838 to I. Jung), Korean Ministry of Health and Welfare (HI17C0328 to I. Jung), and SUHF Fellowship (to I. Jung).

References

1. Welter D et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42, D1001–6 (2014). [PubMed: 24316577]
2. Maurano MT et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 337, 1190–5 (2012). [PubMed: 22955828]
3. Hindorff LA 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). [PubMed: 19474294]
4. Lettice LA 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). [PubMed: 12837695]
5. Uslu VV et al. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. Nat Genet 46, 753–8 (2014). [PubMed: 24859337]
6. Claussnitzer M et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. N Engl J Med 373, 895–907 (2015). [PubMed: 26287746]
7. Smemo S et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 507, 371–5 (2014). [PubMed: 24646999]
8. Yu M & Ren B The Three-Dimensional Organization of Mammalian Genomes. Annu Rev Cell Dev Biol 33, 265–289 (2017). [PubMed: 28783961]
9. de Wit E et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature 501, 227–31 (2013). [PubMed: 23883933]
10. Sanyal A, Lajoie BR, Jain G & Dekker J The long-range interaction landscape of gene promoters. Nature 489, 109–13 (2012). [PubMed: 22955621]
11. Dixon JR et al. Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331–6 (2015). [PubMed: 25693564]

12. Jin F et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290–4 (2013). [PubMed: 24141950]
13. Rao SS et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–80 (2014). [PubMed: 25497547]
14. Dixon JR et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–80 (2012). [PubMed: 22495300]
15. Tang Z et al. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell 163, 1611–27 (2015). [PubMed: 26686651]
16. Sahlen P et al. Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution. Genome Biol 16, 156 (2015). [PubMed: 26313521]
17. Jager R et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat Commun 6, 6178 (2015). [PubMed: 25695508]
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). [PubMed: 25938943]
19. Dryden NH et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res 24, 1854–68 (2014). [PubMed: 25122612]
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). [PubMed: 26616563]
21. Javierre BM et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369–1384 e19 (2016). [PubMed: 27863249]
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). [PubMed: 28475875]
24. Rubin AJ et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. Nat Genet 49, 1522–1528 (2017). [PubMed: 28805829]
25. Orlando G et al. Promoter capture Hi-C-based identification of recurrent noncoding mutations in colorectal cancer. Nat Genet (2018).
26. Leung D et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature 518, 350–354 (2015). [PubMed: 25693566]
27. Consortium GT Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–60 (2015). [PubMed: 25954001]
28. Schmitt AD et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep 17, 2042–2059 (2016). [PubMed: 27851967]
29. Thurman RE et al. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012). [PubMed: 22955617]
30. Whyte WA et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–19 (2013). [PubMed: 23582322]
31. Zhang Y et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature 504, 306–310 (2013). [PubMed: 24213634]
32. Rajagopal N et al. High-throughput mapping of regulatory DNA. Nat Biotechnol 34, 167–74 (2016). [PubMed: 26807528]
33. Diao Y et al. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. Nat Methods 14, 629–635 (2017). [PubMed: 28417999]
34. Engreitz JM et al. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature 539, 452–455 (2016). [PubMed: 27783602]
35. Dao LTM et al. Genome-wide characterization of mammalian promoters with distal enhancer functions. Nat Genet 49, 1073–1081 (2017). [PubMed: 28581502]
36. Roadmap Epigenomics C et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–30 (2015). [PubMed: 25693563]

37. Richard M, Drouin R & Beaulieu AD ABC50, a novel human ATP-binding cassette protein found in tumor necrosis factor-alpha-stimulated synoviocytes. Genomics 53, 137–45 (1998). [PubMed: 9790762]

38. Lieberman-Aiden E et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289–93 (2009). [PubMed: 19815776]

39. Yang D et al. 3DIV: A 3D-genome Interaction Viewer and database. Nucleic Acids Res 46, D52–D57 (2018). [PubMed: 29106613]

40. Zhang Y et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137 (2008). [PubMed: 18798982]

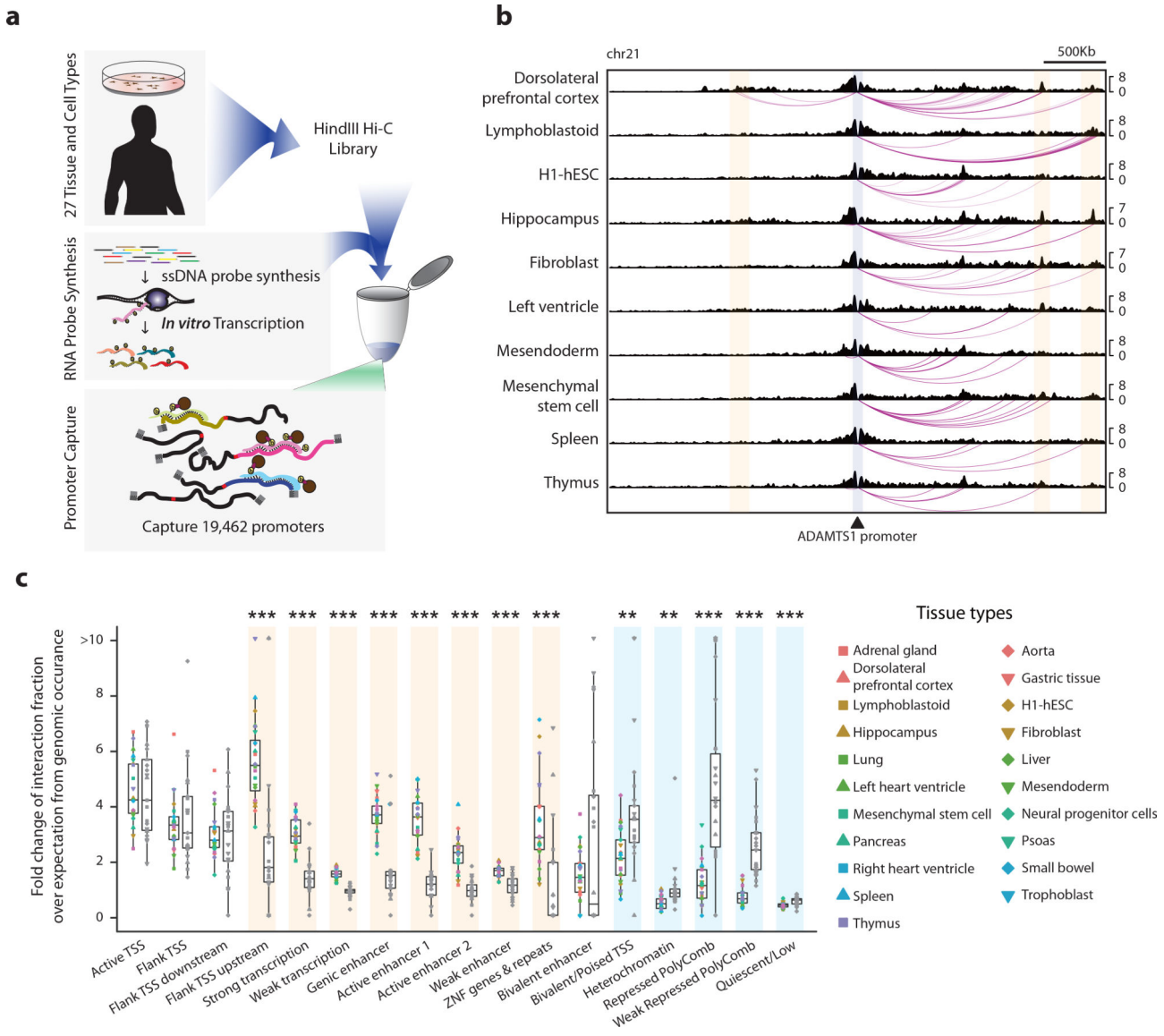

Figure 1. Genome-wide mapping of promoter-centered chromatin interactions in diverse human tissues and cell types.

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,** 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. Two-sided KS tests were performed between interactions originating from active promoter regions (colored dots) and those from bivalent promoters (gray dots) for the samples listed on the right (n = 21) (** *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. For the boxplots, the box represents the interquartile range (IQR), and the whiskers correspond to the highest and lowest points within 1.5 × IQR.

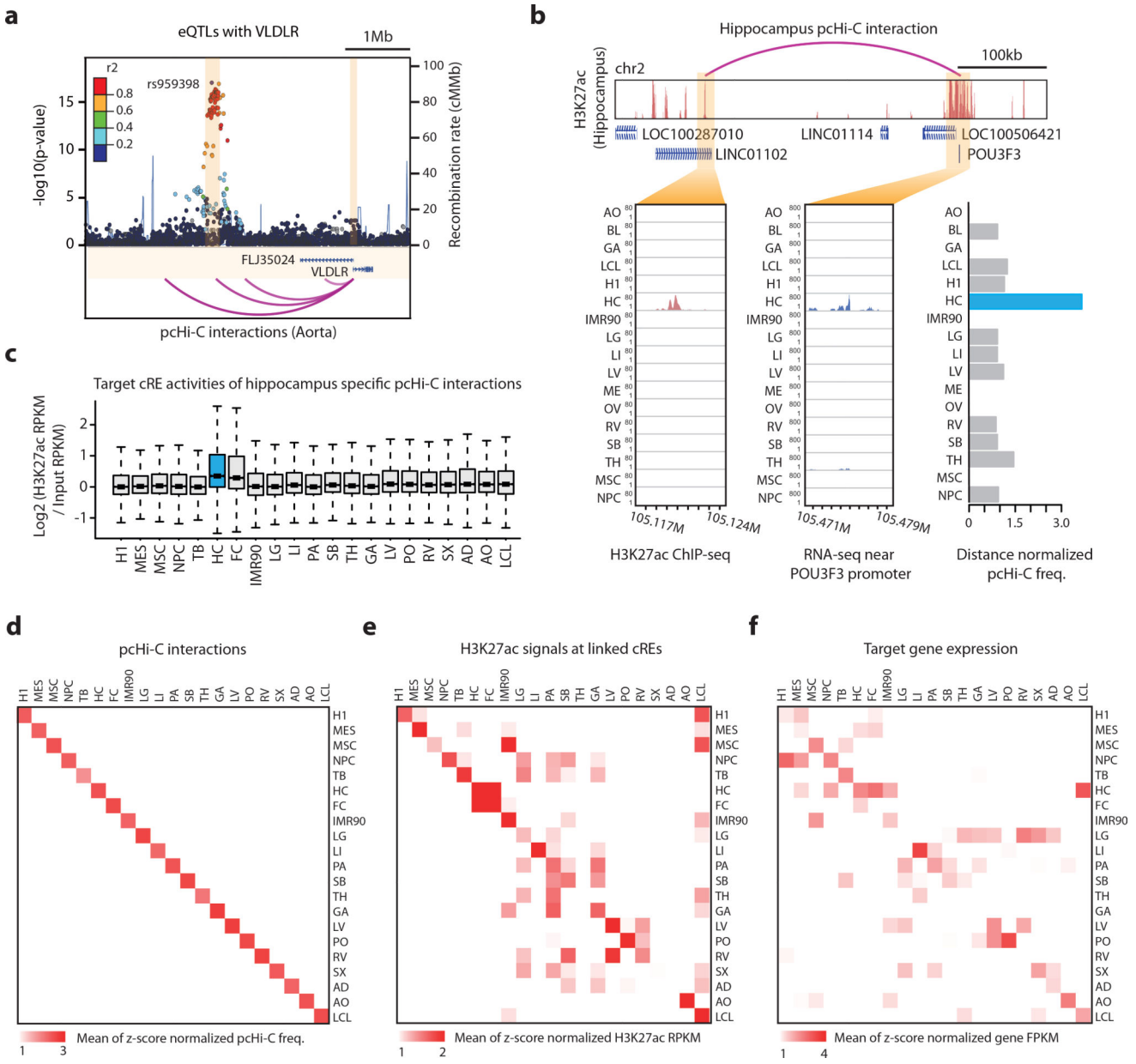

Figure 2. Inference of target genes of *cis*-regulatory sequences from pcHi-C data.

a, Illustrative LocusZoom plot of eQTLs for *VLDLR* (top) and pcHi-C interactions in aorta tissue (bottom). Highlighted in translucent yellow are the *VLDLR* promoter and an eQTL connected by a pcHi-C interaction. Dots 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 scores with a tagging SNP. The blue bars indicate the recombination rate. **b,** Browser snapshots of the *POU3F3* locus, showing positive correlation between the H3K27ac signals at a distal *c*RE (bottom left) and expression levels (bottom middle) of the promoter connected by long-range chromatin interactions (bottom right). The significant chromatin interaction between the *POU3F3* promoter and a distal *c*RE is shown at the top (translucent yellow). **c,** Boxplots illustrating

the H3K27ac signals at the *c*REs (n = 7,712) connected by hippocampus (HC, colored by blue) specific pcHi-C interactions. These *c*REs are marked by higher levels of H3K27ac in hippocampus than in other cell/tissues types (one-sided KS test *P* value < 0.005). For the boxplots, the box represents the interquartile range (IQR), and the whiskers correspond to the highest and lowest points within 1.5 × IQR. **d-f**, Heatmaps demonstrate the enrichment of pcHi-C interactions (column in Fig. 2d), z-score transformed H3K27ac RPKM values at *c*REs (column in Fig. 2e), and z-score transformed RNA-seq FPKM values at the *c*REs' putative target genes (column in Fig. 2f) for given cell/tissue-specific *c*RE-promoter pairs in the corresponding cell/tissue type (rows in Fig. 2d-f). KS test was performed between pcHi-C interaction frequencies, z-score transformed H3K27ac RPKM values, and z-score transformed RNA-seq 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 (two-sided KS test *P* value < 2.2×10^{-16}).

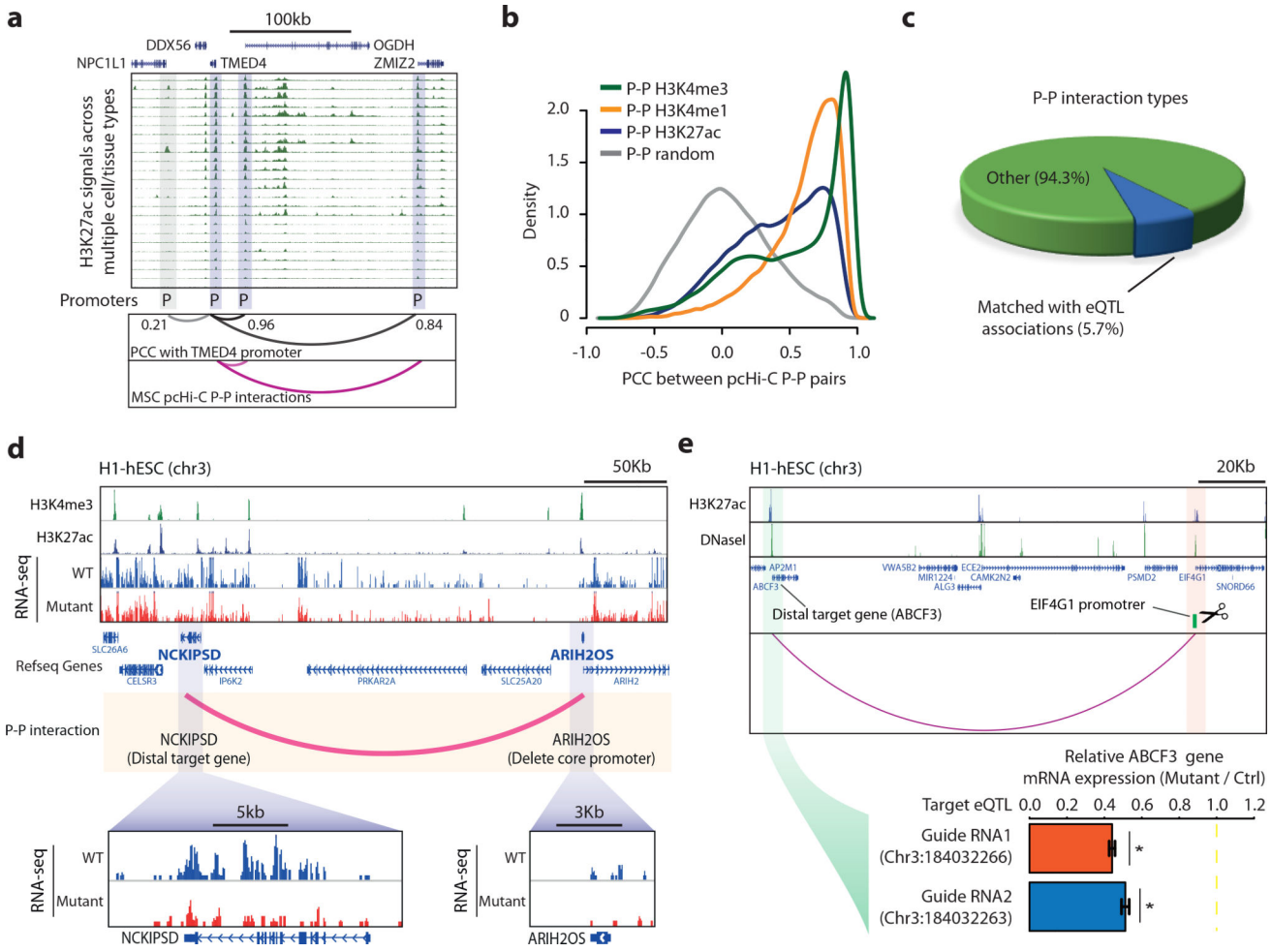

Figure 3. Enhancer-like promoters involved in regulation of distal target genes.

a, Browser snapshots of the *TMED4* locus showing the RefSeq genes (top), H3K27ac signals (middle, n = 24), and pcHi-C chromatin interactions (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 (x-axis) 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 together with a random expectation (gray, median of PCC = 0.02, n = 48,142). **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,** Browser snapshots of RNA-seq results between control (n = 2) and mutant (n = 2) clones with deletion of the core promoter regions of the *ARIH2OS*. 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 obtained from cuffdiff with two mutant clones = 0.02). **e,** Browser snapshots showing the promoter

containing eQTL (translucent yellow with a scissors symbol) targeted by sgRNAs and its distal target gene, *ABCF3* (translucent green), together with H3K27ac and chromatin accessibility (DNase I). The relative mRNA expression levels of the *ABCF3* quantified by RT-qPCR are shown below (* one-sided KS test *P* value < 0.05 derived from three mutant clones). Error bars indicate standard deviation of three mutant clones and y-axis indicates mean values.

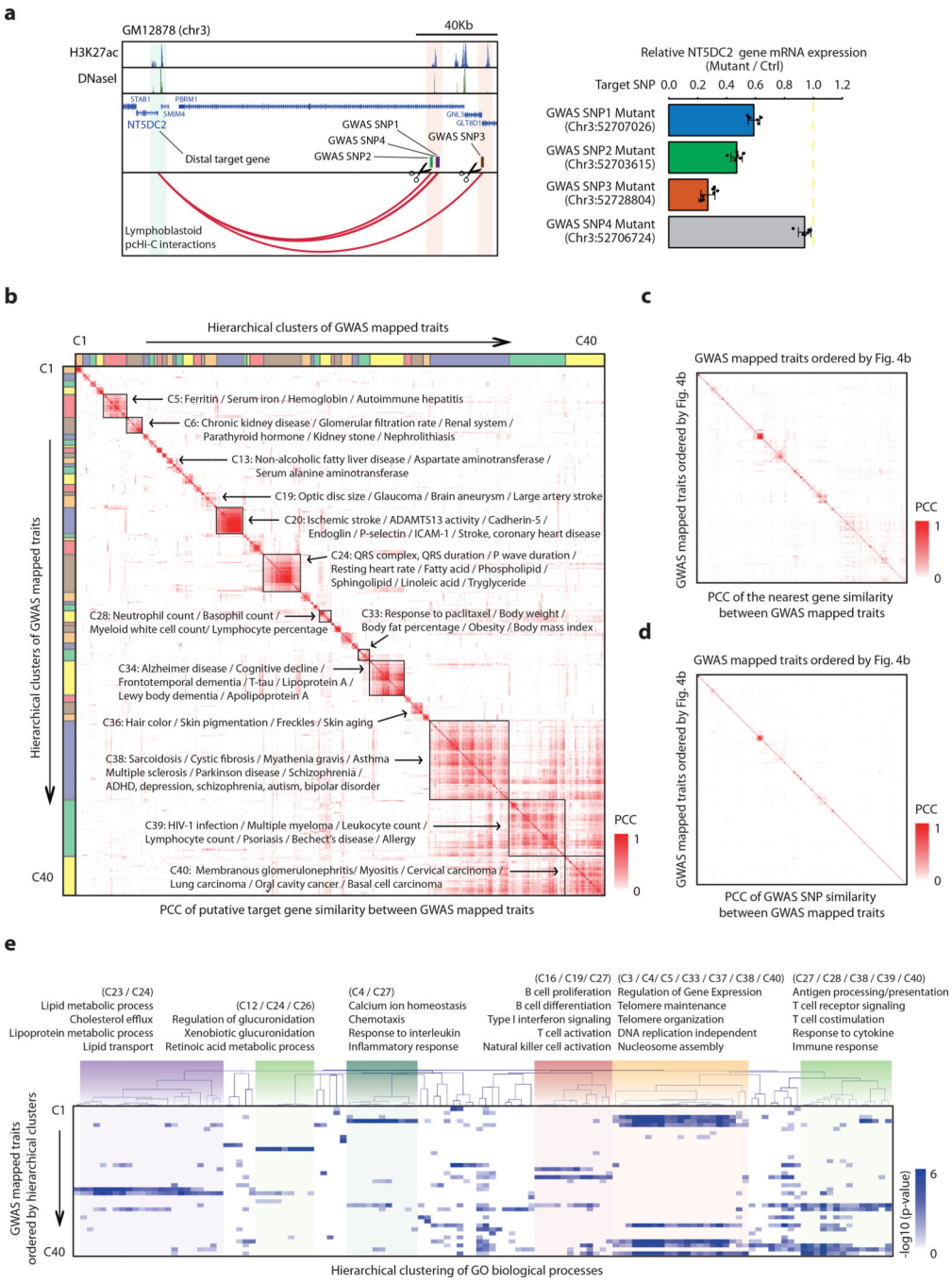

Figure 4. Analysis of human diseases and physiological traits based on the putative target genes of GWAS SNPs.

a, Browser snapshots showing multiple *c*REs harboring GWAS-SNPs (translucent yellow with a scissors symbol) and their common target gene, *NT5DC2* (translucent green), together with signals of H3K27ac (ChIP-seq) and chromatin accessibility (DNase I) (left). The DNA fragments containing 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 mutations of GWAS SNPs with sgRNAs were quantified by RT-

qPCR (right). Error bars indicate standard deviation of two mutant clones with technical triplicates and y-axis indicates mean values. **b,** Hierarchical clustering of human diseases and traits (n = 687) 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 PCC, between traits (n = 687) 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 $-\log_{10}(P \text{ value})$ of GO biological processes in the corresponding cluster obtained from DAVID. Several representative biological processes are highlighted.