natureresearch

https://doi.org/10.1038/s41588-020-00745-3

Supplementary information

Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants

In the format provided by the authors and unedited

SUPPLEMENTARY METHODS

PBMC processing

Peripheral blood mononuclear cells (PBMC) were obtained from leukapheresis samples by density gradient centrifugation and cryopreserved in liquid nitrogen. For the isolation of immune cell types of interest, cryopreserved PBMCs were thawed, washed, stained directly with cocktails of fluorescently conjugated antibodies or pre-enriched for total B cells using the 'Human B Cell Isolation Kit II' (Miltenyi Biotec), following the manufacturer's instructions before staining with antibodies and sorted on a BD FACSAria II (Becton Dickinson) using the gating strategies as described [1]. The FACS-sorted cells were washed and fixed using 1% formaldehyde, as described [2], for ChIP-seq and HiChIP assays.

Semi-automated micro-scaled ChIP-seq for H3K27ac

Cell types from 6 donors (as described above) were utilized for this analysis. A total of 30 ChIP-Seq assays were performed as described previously [3]. Briefly, sheared chromatin from each cell type was immunoprecipitated with a polyclonal anti-H3K27ac antibody (1.5μg/sample) (C15410196; Diagenode) by use of an automated ancillary liquid handler SX-8G IP-Star from Diagenode. Immunoprecipitated chromatin was captured, washed, Illumina library adaptors integrated by transposase-based method [3] and library prepared by PCR amplification. Libraries were sequenced on an Illumina HiSeq 2500 sequencer to obtain 50-bp single-end reads.

Luciferase assay

500 ng of luciferase reporter plasmid for the indicated cis-regulatory sequence was transfected into 2.0 × 10^5 cells using the Neon Transfection System (Thermo Fisher Scientific) according to the manufacturer's protocol (settings: 1,600 V, 10 ms, 3 pulses). Fresh media (as described above) was then added and cells were maintained for 72 hours. After 72 hours cells were washed with PBS, and harvested. Preparation of cytosolic extracts with equal number of cells and luciferase assays were carried out using a Nano-Glo® Luciferase Assay System (Promega, Madison, WI), according to the supplier's protocol, and results were quantitated on an envision multiplate reader. All luciferase reporter plasmid was cloned by Twist biosciences.

qRT-PCR

Total RNA was extracted using the miRNeasy Micro Kit (Qiagen); cDNA was reverse-transcribed with the SuperScript III First-Strand Synthesis System (Life Technologies). Real-time PCR was performed using the Fast Start Universal SYBR Green Master Mix (Roche); see Supplementary Table 1F for primer sequences. Data were acquired on the QuantStudio 6 Flex (Applied Biosystems); all results are presented relative to expression in control crRNA and tracrRNA duplex transfected cells. Transcript levels were normalized using housekeeping gene *YWHAZ* as control in each sample.

Lentivirus production and generation of dCas9-KRAB expressing cell lines

For virus production, 5 X 10^6 HEK293T cells were plated in a 10 cm plate. The following day, plasmid encoding lentivirus (pHR-SFFV-KRAB-dCas9-P2A-mCherry) was co-transfected with MISSION® Lentiviral Packaging Mix into the cells using JetPRIME® transfection reagent (Polyplus transfection) according to the manufacturer's instructions. Supernatant containing viral particles were collected 48 hours after transfection and filtered. Virus was concentrated using Lenti-X concentrator (Clontech) and stored at -80°C. For virus infection, 1 ml of virus, and polybrene (final 4 mg/ml, Sigma Aldrich) was added to 1 million cells (Jurkat cell or GM12878 cells) and centrifuged for 90 minutes at 1400 X g at 30°C. After centrifugation, cells were incubated at 37°C for two hours, and then viral supernatant was replaced with fresh medium (as described above). Fresh media was then added every other day and cells were maintained for 45 days. After 45 days KRAB-dCas9 expressing cells were selected by sorting mCherry expressing cells.

Alignment and peak calling for ChIP-seq data

ChIP-seq reads were aligned using bowtie2 [4] with respect to hg19 reference genome, using the parameters *-k 4 --mm --threads 8 -X 2000*. Uniquely mapped reads with MAPQ >= 30 were retained, and duplicate reads were discarded using Picard (http://broadinstitute.github.io/picard). For each cell type, we then merged the resulting de-duplicated aligned reads (in .bam format) using samtools (http://samtools.sourceforge.net/) for all six donors, to produce aggregate ChIP-seq reads for each cell type. These merged alignment files were then applied to MACS2 - version 2.1.1 [5] for peak calling, using the parameters *-f BAM -g 'hs' --nomodel --extsize 147 --*

keep-dup 1 -q 0.01. In addition, we also merged alignments for all cell types, and then called MACS2 (with the same settings) to obtain ChIP-seq peaks for all cell types.

Reproducibility analysis for HiChIP

We performed reproducibility analysis between two replicates of individual donors by first constructing the union of FitHiChIP-L loops significant in at least one replicate. Contact counts of those loops in both replicates (missing values are replaced by zero) are used to plot the scatterplots. R function *cor()* is used to find the correlation of contact counts.

Principal component analysis (PCA) for HiChIP samples

We first applied FitHiChIP-L (peak-to-all foreground; loose or peak-to-all background) on donor-specific HiChIP samples from the five immune cell types. Then, for each cell type (consisting of either six or nine samples, including duplicates), we extracted 20,000 most significant (FDR < 0.01) loops occurring in all the samples. In total for the five cell types, we extracted 100K loops, of which 52,648 were unique loops. For each of these loops, we listed their FitHiChIP-L significance (FDR values) for all samples (replacing missing entries with 1) to form a feature matrix M of 52648 x 30 dimension. The matrix M was applied on the R function *prcomp* for PCA. Individual samples in the PCA were plotted using the R package *factoextra*.

Finding interacting ChIP-seq peaks from HiChIP

Peak were called by MACS2 on the merged H3K27ac ChIP-seq tracks (across 6 donors) for each individual cell types. These H3K27ac peaks were inputted as reference peaks to FitHiChIP to process the 70M read merged HiChIP dataset for the corresponding cell type. The resulting significant interaction calls were used to determine *interacting* H3K27ac peaks by checking whether the 5 Kb bin(s) that overlaps a given ChIP-seq peak (i.e., peak bin) is involved in a significant HiChIP interaction. Similar overlap computations and classifications were carried out for PCHiC data using CHiCAGO loop calls. We computed all overlaps using *bedtools intersect* routine (minimum 1 bp overlap).

Annotation of interacting bins as promoter, enhancer or other

We defined an interacting bin as a promoter (P) bin if it lies within 5 Kb of a reference TSS site. An interacting bin was labeled as enhancer (E) if it overlaps (minimum 1 bp) with reference H3K27ac ChIP-seq peaks from merged tracks and is not a promoter bin. A bin not in promoter or enhancer category was labeled as other (O).

Enrichment of different histone marks in promoters and enhancers

ChIP-seq bigwig tracks of corresponding cell types for the histone marks H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3 were downloaded for all available replicates from the IHEC data portal (https://epigenomesportal.ca/ihec/grid.html?build=2017-10&assembly=1&cellTypeCategories=1). For each histone mark, we merged the multiple bigwig files from replicates using utility functions *bigWigMerge* and *bedGraphToBigWig* [6].

These merged bigwig tracks along with the H3K27ac ChIP-seq tracks for these cell types were used to find the enrichment of promoters and enhancers participating in HiChIP loops for the different cell types. The *computeMatrix* and *plotHeatmap* of the package Deeptools were used to plot the heat map and distribution plot.

Reanalysis of HLA transcript levels using HLApers

We applied HLApers pipeline (https://github.com/genevol-usp/HLApers) for *in silico* HLA mapping and obtaining transcript expression with some modifications. This pipeline generates personalized HLA index files for individual samples, which are used to estimate sample specific HLA genotype. We applied HLApers with its default settings, aside from customizing it to support our single-end reads. We executed HLApers by using STAR as the aligner and Salmon for quantifying the transcripts. Reference HLA transcript annotations and fasta sequences were obtained from the latest release of IMGTHLA database (https://github.com/ANHIG/IMGTHLA.git) for HLA. We downloaded Gencode version 30 reference fasta and gene annotation (GTF) files corresponding to the hg19 human reference genome (https://www.gencodegenes.org). The R libraries *TxDb.Hsapiens.UCSC.hg19.knownGene, tximport* and *tximportData* were used to generate gene expression from transcript quantification, for both non-HLA and HLA transcripts. We then used matrixeQTL to obtain eQTLs from the HLApers-computed gene expression values for HLA genes. For all downstream analysis, we replaced the HLA eQTLs from the initial DICE release with this revised analysis.

Conditional eQTL analysis

We used MatrixEQTL with linear regression model to perform forward stepwise conditional eQTL analysis[7]. We used the same genotype and gene expression data in the conditional eQTL analysis as employed in the eQTL study, i.e. DICE data for non-HLA genes and HLApers output for HLA genes. In the default conditional analysis, FDR was computed in the very first step using the Benjamini-Hochberg method across all cis-eQTL tests within each chromosome. Subsequent iterations did not re-compute FDR, but rather used a fixed p-value threshold corresponding to the 5% FDR of the first step, as suggested[7]. At each step, for each gene with at least one *cis*-eQTL (+/- 1Mb) below the FDR or p-value threshold, its most significant SNP (lead SNP) was added as a covariate in order to identify additional independent eQTLs. We repeated this process for a maximum of 20 iterations to identify conditionally independent eQTLs for each gene. Consistent with GTEX estimates[8], only 10-18% of eGenes reported additional independent eQTLs (E2, E3, …) beyond the first iteration of the conditional mapping (E1) across all cell types (**Extended Data Fig 3A**). For 19-26% of the eGenes the identified conditionally independent eQTLs coming from any iteration directly matched the pieQTLs compared to 12-15% matching the promoter eQTLs (**Extended Data Fig 3B**).

As many pieQTLs or promoter eQTLs may not come as the top SNP but may indeed be in high LD with it (**Extended Data Fig 3C**), we expanded the set of SNPs coming from each iteration of conditional analysis to those that are in high LD (R^2>0.8) with the lead SNP (E1) or the top conditionally independent SNP (E2, E3, …).We used PLINK v1.90b3w [9] to compute the linkage disequilibrium (LD) between the lead SNPs generated at each step of the conditional analysis, and the promoter proximal eQTLs (located within

2.5Kb of respective TSS) or pieQTLs. If a gene does not have any promoter proximal eQTL or pieQTL, corresponding LD is set as zero. The *ggplot2* package in R was used to plot the LD score density plots for promoter eQTLs and pieQTLs. For 68-73% of the eGenes, we found a pieQTL in the expanded set, whereas these percentages were 44-51% for the promoter eQTLs across the five cell types (**Extended Data Fig 3D**).

Allele-specific mapping of HiChIP data

For individual samples in different cell types, we again applied HiC-Pro pipeline [10] with previously described parameters and settings but this time in the allele-specific mapping mode. Briefly, HiC-Pro in this mode first performs both a global and local alignment of HiChIP reads to a masked hg19 genome reference (*bedtools maskfasta*) using SNPs from the genotyping data generated in the initial DICE database release [1]. Then using all imputed SNP data from donors, HiC-Pro assigns reads to either of the parental genomes (G1 or G2) for heterozygous loci. Downstream analysis including allele-specific screening of reads overlapping pieQTLs was performed using the valid read pairs output files of HiC-Pro. We determined the number of allele-specific interactions/contacts by counting the number of valid read pairs overlapping a given pieQTL in samples that are heterozygous for that pieQTL (*e.g.*, *GAB2* pieQTL). For each such read, the base pair at the pieQTL location was further confirmed by looking up the sequence at that specific coordinate.

Cell specific eQTL, pieQTL, eGene and pieGene counts

For each cell type, we extracted the following sets of eQTLs or eGenes: A) all eQTLs, B) subset of A within 2.5Kb of TSS of respective genes, C) subset of A falling within 10Kb of TSS of respective genes, D) eQTLs overlapping with ChIP-seq peaks, E) subset of D within 2.5Kb of TSS of respective genes, F) subset of D falling within 10Kb of TSS of respective genes, G) Set of direct and indirect pieQTLs, H) eGenes corresponding to the set of eQTLs in set A, I) eGenes corresponding to the set of eQTLs in set D, J) list of promoters interacting with the pieQTLs in set G through HiChIP interactions. For each of these sets, cell specificity and common elements in multiple cell types were listed and plotted in pie-charts.

Inspecting LD between pieQTLs and eQTLs proximal to the TSS

We used PLINK v1.90b3w [9] to compute linkage disequilibrium (LD) between the eQTLs. The parameter *--ld-window-kb* was fixed at 10000 (i.e. 10 Mb) since we considered pieQTLs up to 10 Mb distance. The parameter *--ld-window* (maximum number of intermediate variants between two SNPs) was fixed at 1000,000. As mentioned before, PheGenI database [11] having information of continental 'super populations' (AFR, AMR, EAS, EUR, SAS) based on data from the phase 3 of the 1,000 Genomes Project [12] was used for generating LD statistic. The parameter *--r2* was set as 0.8, indicating that an LD > 0.8 in any of the five super-populations would indicate tight genetic linkage between a pair of SNPs. For each cell type c, suppose G_c denotes the set of genes having at least one promoter-proximal eQTL (distance from TSS < 2.5 Kb or 10 Kb depending on the proximality threshold used) and simultaneously at least one pieQTL having significant HiChIP loops with that gene. Using the LD analysis, we listed for each gene $g \in G_c$, LD values between all pairs of its promoter-proximal eQTLs and pieQTLs interacting with *g*. For each cell type *c*, we then plotted the following statistics:

1. Count and fraction of genes *g* out of total genes G_c such that at least one pair of promoter-proximal eQTL and pieQTL of *g* has LD > 0.8 (i.e. significant tight linkage).
2. Per gene fraction of pieQTLs which are in tight linkage with one or more promoter proximal eQTLs of that gene.
3. For each chromosome, fraction of all pieQTLs which are in tight linkage with one or more promoter proximal eQTLs of the corresponding interacting promoters.

Inspecting LD between ultra-long pieQTLs and eQTLs (all or proximal to the TSS)

Here we considered ultra-long pieQTLs (i.e. pieQTLs having distance > 1 Mb from the interacting promoters). We computed LD between all pairs of SNPs in the catalog of SNP-trait associations (as described before) with PLINK using the settings *--ld-window-kb* 10000 *--ld-window* 1000000 *--r2* 0.8. For each cell type *c*, here we considered the set of eGenes G_c having at least one ultra-long interacting pieQTL. For each gene $g \in G_c$, we inspected LD between its ultra-long pieQTLs and either all eQTLs or only promoter proximal eQTLs (distance from TSS < 10 Kb), and reported the fraction of genes or pieQTLs showing tight linkage with one or more eQTLs / promoter proximal eQTLs for each cell type.

Fine mapping of eQTLs

We used FINEMAP (version 1.3.1) [13] to perform statistical fine-mapping of the DICE eQTLs. Statistics for individual eQTLs such as reference and alternate alleles, MAF, FDR, beta, z-scores were used from our earlier work [1]. Fine-mapping was performed with stepwise conditioning *(--cond)* option in the FINEMAP package. For all other options, we used default settings in FINEMAP.

Enrichment of ultra-long pieQTLs and testable SNPs compared to the distance matched random simulated SNPs

We considered P-E loops > 1 Mb from FitHiChIP significant loops and extracted SNPs from the interacting bins and also from their +/- 1 bins. These are the set of testable SNPs *T* for ultra-long pieQTL mapping within distance 1 – 10 Mb from the interacting promoters. To construct the set of distance matched random SNPs (denoted by the set *R*), we checked individual bins *b* interacting with a promoter by FitHiChIP significant loops > 1 Mb, and randomly sampled 50 SNPs from the bins located within 5 Kb to 105 Kb (i.e. a span of 100 Kb) from the bin *b*. We have also simulated 50 SNPs from the bins within 5 Kb to 105 Kb of the bin that is of equal distance to the same promoter but in opposite direction. By definition, the distance distribution of SNPs in *R* is matched to that of *T*, the set of actually tested SNPs for ultra-long pieQTL discovery. We also selected another set of distance-matched random SNPs that overlap active *cis*-regulatory regions as defined by H3K27ac peaks, namely *A*. For all SNPs within the sets *T*, *R* and *A*, we computed the p-values of their association with the expression of their “target” genes using MatrixEQTL [14]. We used default settings of the MatrixEQTL with a linear regression model except setting the *cis* distance to 10000000 (i.e. 10 Mb). Compared to both sets of SNPs, overlapping randomly selected regions or H3K27ac peaks, the promoter interacting SNPs used for association testing showed a very strong enrichment in identifying SNPs with statistically significant associations to the expression of their target gene for each cell type (**Extended Data Fig. 6**).

REFERENCES

1. Schmiedel, B.J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715 e16 (2018).
2. Seumois, G. *et al.* Epigenomic analysis of primary human T cells reveals enhancers associated with TH2 memory cell differentiation and asthma susceptibility. *Nat Immunol* **15**, 777-88 (2014).
3. Youhanna Jankeel, D., Cayford, J., Schmiedel, B.J., Vijayanand, P. & Seumois, G. An Integrated and Semiautomated Microscaled Approach to Profile Cis-Regulatory Elements by Histone Modification ChIP-Seq for Large-Scale Epigenetic Studies. *Methods Mol Biol* **1799**, 303-326 (2018).
4. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-9 (2012).
5. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
6. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
7. Dobbyn, A. *et al.* Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. *Am J Hum Genet* **102**, 1169-1184 (2018).
8. Consortium, G.T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).

9. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).

10. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**, 259 (2015).

11. Ramos, E.M. *et al.* Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet* **22**, 144-7 (2014).

12. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).

13. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-501 (2016).

14. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).