Genetic determinants of chromatin accessibility and gene regulation in T cell activation across human individuals

Christine S. Cheng[1*], Rachel E. Gate[2,3*], Aviva P. Aiden[4,5], Atsede Siba[1], Marcin Tabaka[1], Dmytro Lituiev[2], Ido Machol[4], Meena Subramaniam[2,3], Muhammad Shamim[4,10], Kendrick L. Hougen[12],Ivo Wortman[1], Su-Chen Huang[4], Neva C. Durand[4], Ting Feng[6], Philip L. De Jager[1,7,8], Howard Y. Chang[9], Erez Lieberman Aiden[4,10,11], Christophe Benoist[6], Michael A. Beer[12,13], Chun J. Ye[2§], Aviv Regev[1,14§]

[1]Broad Institute of MIT and Harvard, 415 Main Street, Cambridge MA 02142, USA

[2]Institute for Human Genetics, Department of Epidemiology and Biostatistics, Department of Bioengineering and Therapeutic Sciences, UCSF

[3]Biological and Medical Informatics Graduate Program, UCSF

[4]Department of Molecular and Human Genetics, the Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030, USA

[5]Department of Bioengineering, Rice University, Houston, TX 77030, USAs

[6]Division of Immunology, Department of Microbiology and Immunology, Harvard Medical School, Boston, MA 02115

[7]Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences, Department of Neurology and Psychiatry, Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

[8]Harvard Medical School, Boston, MA 02115, USA

[9]Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA

[10]Medical Scientist Training Program, Baylor College of Medicine, Houston, TX 77030, USA

[11]Department of Computer Science, Department of Computational and Applied Mathematics, Center for Theoretical Biological Physics, Rice University, Houston, TX 77030, USAs

[12]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, USA

[13]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, USA

[14]Howard Hughes Medical Institute, Koch Institute of Integrative Cancer Research, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[*] These authors contributed equally to this work.

[§] Corresponding authors. Email: aregev@broadinstitute.org (A.R); jimmie.ye@ucsf.edu (C.J.Y.)

Abstract

The vast majority of genetic variants associated with complex human traits map to non-coding regions, but little is understood about how they modulate gene regulation in health and disease. Here, we analyzed Assay for Transposase-Accessible Chromatin (ATAC-seq) profiles from activated primary $CD4^+$ T cells of 105 healthy donors to identify ATAC-QTLs: genetic variants that affect chromatin accessibility. We found that ATAC-QTLs are widespread, disrupt binding sites for transcription factors known to be important for $CD4^+$ T cell differentiation and activation, overlap and mediate expression QTLs from the same cells and are enriched for SNPs associated with autoimmune diseases. We also identified numerous pairs of ATAC-peaks with highly correlated chromatin accessibility. When we characterize 3D chromosome organization in primary $CD4^+$ T cells by *in situ*-Hi-C, we found that correlated peaks tend to reside in the same chromatin contact domains, span super-enhancers, and are more impacted by ATAC-QTLs than single peaks. Thus, variability in chromatin accessibility in primary $CD4^+$ T cells is heritable, is determined by genetic variation in a manner affected by the 3D organization of the genome, and mediates genetic effects on gene expression. Our results provide insights into how genetic variants modulate chromatin state and gene expression in primary immune cells that play a key role in many human diseases.

The vast majority of disease-associated loci identified through genome-wide association studies (GWAS) (*1-3*) are located in non-coding regions of the genome, often distant from the nearest gene (*4*), suggesting that abnormalities in transcriptional regulation are a key driver of human disease. Quantitative trait loci (QTL) studies that map genetic variants associated with molecular traits provide a framework for interpreting the regulatory and functional role of disease-associated loci. For example, thousands of non-coding variants associated with gene expression (expression QTLs – eQTLs), a significant number of which overlap GWAS loci, have been identified in diverse cell types and tissues (*5*), including in resting (*6-8*) and in stimulated (*9, 10*) immune cells. However, because of linkage disequilibrium and the complex regulation of gene expression, it remains difficult to pinpoint the causal genetic variant and to determine the mechanistic basis for most eQTLs.

Genetic analysis of chromatin organization (*11-15*) provides a powerful complementary approach for identifying genetic variants that affect transcriptional regulation in *cis*-regulatory regions (*16*). In lymphoblastoid cell lines, many genetic variants have been associated with variability in DNase I hypersensitivity (measured by DNase-seq) (*17*) or histone tail modifications (measured by ChIP-seq) (*18-20*). However, both DNase-seq and ChIP-seq are laborious and require large numbers of cells, thus limiting genetic studies using these techniques to cell lines. The recent development of Assay for Transposase-Accessible Chromatin (ATAC-seq), a simple yet efficient two-step protocol (*21*), has opened the way to profiling of chromatin accessibility with small numbers of disease-relevant primary cells isolated from a large human cohort.

Here, we studied the genetic determinants of chromatin accessibility by performing ATAC-seq on primary $CD4^+$ T cells isolated from 105 healthy donors of European descent in the ImmVar Consortium (*10*) (**Fig. 1A**). To this end, we developed an optimized ATAC-seq protocol (**SOM**) that achieved high technical and biological reproducibility (**fig. S2**), highly complex libraries (on average 84% usable nuclear reads, as opposed to 40% prior to optimization) (**fig. S1**) and low mitochrondrial DNA (mtDNA) contamination (on average contamination < 3%, as opposed to 53% prior to optimization).

We first assayed $CD4^+$ T cells (T helper, Th) in two different states: either unstimulated, or activated *in vitro* using anti-CD3 and anti-CD28 antibodies for 48 hours (**Fig. 1A**) to assess global changes in chromatin accessibility after *in vitro* activation. Comparing pooled reads across six samples (five donors, one of which is replicated) in each state (stimulated or unstimulated), we observed a global increase in chromatin accessibility after activation, detecting 36,486 chromatin accessibility peaks (ATAC-peaks) in unstimulated cells and 52,154 ATAC-peaks in activated cells. Of the 63,763 ATAC-peaks identified in at least one of the two states, 27,446 were shared across states, 28,017 were more accessible in activated cells (FDR, $q < 0.05$), but only 8,298 ATAC-peaks were more accessible in unstimulated cells (FDR, $q < 0.05$) (**Fig. 1B** and **table S1**). Activation-specific ATAC-peaks were enriched at enhancers associated with conventional T helper cells (T_{conv}, a class that includes Th1, and Th17 cells) (*16*) and depleted at enhancers associated with regulatory T cells (T_{reg}) and naïve Th cells. They were also enriched in stimulated Th cells and Th0 cells, consistent with our stimulation protocol (**Fig. 1C**).

Similarly, activation-specific ATAC-peaks were enriched for motifs binding transcription factors (TFs) that are important for Th cell activation, including BATF, IRF and AP1 (*22-24*) (**Fig. 1D**). In particular, 1,594 of the 9,724 activation-specific ATAC-peaks located in noncoding regions previously unannotated by H3K27Ac overlapped BATF motifs (16%) (**fig. S3**). TF footprints derived from normalized ATAC-seq reads that span known TF binding sites from ChIP-seq data (*25*) showed increased accessibility of BATF, ISRE and BATF/IRF motifs in stimulated cells, consistent with the known role of these factors in Th cell development and activation (*22-24*) (**Fig. 1E**). In contrast, CTCF and ETS binding motifs were enriched at shared ATAC-peaks showing little change in accessibility following activation (**Fig. 1D** and **1E**). This is consistent with the known role of ETS family transcription factors as pioneer factors that establish the chromatin landscape for T cells (*26-28*). Finally, single nucleotide polymorphisms (SNPs) associated with several autoimmune diseases (most notably inflammatory bowel disease (IBD)) were more enriched in activation-specific and shared ATAC-peaks than in baseline-specific peaks (**Fig. 1F**). These results demonstrate the power of ATAC-seq to identify previously unannotated *cis*-regulatory elements and generate high-resolution TF footprints, providing insights into the regulation of Th cell function and its role in underlying disease mechanisms.

Next, we explored higher order relationships in our ATAC-Seq data using the 105 profiles from activated T cells. We found 1,762 pairs of ATAC-peaks with highly correlated accessibility (linear regression, FDR < 0.05, **table S2, fig. S4**), corresponding to 851 (1.6%) distinct ATAC-peaks of the 52,154 total ATAC-peaks we observed in activated T cells. On average, correlated

peaks were located 313 kb apart, with the closest located only 668 bp apart. The large average distance between correlated peaks suggests that the correlation is unlikely to be the result of local biases in sequencing depth (**fig. S5**).

Next, we characterized the genetic basis of chromatin accessibility in activated T cells by comparing the variability in ATAC-peaks across the 105 individuals with the imputed genotypes of ~10 million SNPs in those individuals (**SOM**). We found 1,790 ATAC-peaks associated with at least one significant local (+/- 20kb) SNP (RASQUAL (*29*), $P < 3.02\times10^{-4}$, permutation FDR < 0.1) (**Fig. 2A**, **table S3**). We term each such associated SNP an ATAC quantitative trait locus (ATAC-QTL) and the corresponding peak an ATAC-QTL-peak. Of the 1,790 ATAC-QTL-peaks, 599 were significantly heritable (GCTA FDR < 0.1) with an average heritability h^2 = 60%. In 580 (97%) of those cases, 36% of the heritability was predicted by the best lead SNP (**Fig. 2B, SOM, table S9**), suggesting that the heritable variability of ATAC-QTL-peaks is largely determined by a single SNP. There were also 6,154 ATAC-QTL-peaks (RASQUAL, $P < 3.02\times10^{-4}$, permutation FDR < 0.1) with distal associations to SNPs located between +/- 20 kb and +/- 500 kb away, but only 2,634 ATAC-QTL-peaks (linear regression, $P < 6.46\times10^{-5}$, permutation FDR < 0.05) with distal associations to SNPs located over 500 kb away. This is consistent with previous observations of limited distal associations to chromatin accessibility traits estimated using DNase I hypersensitivity (*17, 18*).

Several lines of evidence support a model where local ATAC-QTLs disrupt *cis*-regulatory functions in Th cell enhancers. First, of the 1,790 local ATAC-QTL-peaks, 33% (589) of the lead

associated SNPs were located within 2 kb of the ATAC-peak and 18% (327) were located within the ATAC-peak width proper (**Fig. 2C**), suggesting that the direct disruption of *cis*-regulatory elements may be an important determinant of the observed variation in accessibility in those cases. Second, local ATAC-QTLs-peaks were more enriched near transcription start sites (TSS) than transcription termination sites (TTS) of the closest gene, supporting a transcriptional role in *cis* (**Fig. 2D**). Third, 77% of local ATAC-QTLs-peaks were in intronic or intergenic regions (**fig. S6A**); of these, 70% lie in regions that were previously identified as enhancers for different Th cell subtypes (**fig. S6B**). Fourth, ATAC-QTL-peaks were enriched for motifs bound by TFs involved in T cell development and activation (*e.g.*, BATF, AP1 and IRF, **Fig. 2E**) as compared to all ATAC-peaks detected in activated cells as background. In fact, 57% of ATAC-QTL-peaks contained either a BATF or an ETS1 motif (a 1.2 fold enrichment compared to all ATAC-peaks in activated cells, $P < 1.79\times10^{-19}$, hypergeometric test; with all peaks as background) and 11% contained both (an 1.3-fold enrichment, $P < 4.02\times10^{-5}$, hypergeometric test). Furthermore, ATAC-QTL-peaks overlapping BATF, ETS1 and CTCF binding sites showed differential accessibility between genotypes at single nucleotide resolution, with the core motif exhibiting the most striking difference in accessibility (**Fig. 2F** and **fig. S7**). This effect extended 1 kb outside the core binding motif, demonstrating that genetic variants may not only cause differential transcription factor binding, but also give rise to long-range effects on chromatin accessibility (**Fig. 2F** and **fig. S8**). Similarly, the effect size of ATAC-QTLs is correlated with SNP motif disruption scores obtained by deltaSVM (*30*), an unbiased analysis to discover *de novo cis*-regulatory elements in ATAC peaks (**Fig. 2G, SOM**). Most (61%) of the ATAC-QTL lead SNPs strongly disrupted one of 11 predicted TF binding sites (TFBSs) (**Fig. 2H**). These 11

predicted TFBS, which arose automatically from this analysis, included many known to act in T cell activation, such as BATF, IRF, NFκB and ATF.

Local ATAC-QTLs were also enriched for GWAS SNPs from autoimmune diseases (**Fig. 2I**), providing a functional context for interpreting these disease associations. For example, rs17293632, an ATAC-QTL SNP, has also been associated with Crohn's disease and IBD in GWAS studies. This SNP is located in the first intron of *SMAD3*, a TF gene involved in the TGF-β signaling pathway that regulates T cell activation and metabolism (*33*). This SNP disrupts a consensus BATF binding site at a conserved position (deltaSVM=-12.72), and results in decreased chromatin accessibility in individuals that possess the alternate allele (**Fig. 2J**). Rs17293632 was also identified as a causal variant by a statistical fine-mapping algorithm that leveraged patterns of linkage disequilibrium but is agnostic to the functional genomic data in this region (*16*). This suggests that rs17293632 may increase susceptibility to Crohn's disease and IBD by disrupting BATF binding at the *SMAD3* locus.

As noted above, we found substantial correlation in accessibility between 1,762 pairs of ATAC-peaks. We next tested the hypothesis that correlated ATAC-peaks could be simultaneously modulated by a single SNP, thereby allowing a single variant to influence multiple *cis*-regulatory regions. Specifically, we found that both local and moderately distal ATAC-QTLs (< 1 Mb) are more strongly associated with correlated peaks than to single peaks (defined as any ATAC-peak that is not part of a correlated peak pair; **Fig. 3A, fig. S8**). Furthermore, correlated peaks that were associated with ATAC-QTLs were more strongly correlated with each other than correlated

peaks that were not associated with an ATAC-QTL (**Fig. 3B** and **table S2**). This suggests that genetic variants could impart stronger co-regulation of *cis*-regulatory regions than epigenetic mechanisms. For example, rs10815868, an ATAC-QTLs associated with a pair of correlated peaks, resides in the 18th intron of *PTPRD*, a tumor suppressor gene, where it disrupts a consensus BATF binding site at a conserved position. This ATAC-QTL is associated with decreased chromatin accessibility in a 4 kb region that contains four highly correlated ATAC-peaks (**Fig. 3C**, yellow box). Notably, an adjacent peak located 5 kb upstream of this region was not affected by this variant (**Fig. 3C,** grey box), suggesting that the genetic control of multiple ATAC-peaks was limited to a defined regulatory region.

To test if the relationship between correlated peaks could be influenced by the 3D conformation of the Th cell chromatin, we performed loop-resolution *in situ* Hi-C (*34*) in primary $CD4^+$ T cells activated for 48 hr. We obtained Hi-C maps at 50 kb resolution and identified 4,614 contact domains (internals of chromatin in which all loci exhibit an enhanced frequency of contact with one another) and 4,419 chromatin loops (pairs of loci that are in closer physical proximity to one another than to neighboring sequences on the chromosome, such as the interaction between a promoter and its distal enhancer) (**fig. S9** and **table S4**). ATAC-peaks overlapping BATF and ETS1 motifs were enriched within Hi-C contact domains (**Fig. 3D**), whereas those overlapping CTCF motifs were enriched at the contact domain boundaries (**Fig. 3D** and **fig. S10A**). This is consistent with previous reports of CTCF enrichment at loop anchors and at contact domain boundaries (*34-37*).

There was a clear relationship between the 3D structure of chromatin, the correlated peaks, and the associated ATAC-QTLs. Specifically, we found that correlated peaks were the most enriched when both peaks were in the same contact domain than when (i) both peaks lie outside of a contact domain, or (ii) only one of the two peaks was in a contact domain, or (iii) the two peaks span a contact domain (**Fig. 3E, fig. S10B,C** and **table S4**; considering all domains, **SOM**). Moreover, correlated ATAC-QTL-peaks exhibited stronger genetic associations if both peaks resided in the same contact domain than if they did not (**Fig. 3F**). This suggests that 3D chromatin structure may explain the ability of a single genetic variant to influence multiple correlated ATAC-peaks.

Furthermore, correlated peaks – whether or not they were associated with an ATAC-QTL – were enriched in super-enhancer regions from $CD4^+$ Th cell (*38*) (Super-enhancers, also called stretch enhancers, are defined as large clusters of contiguous enhancers (*38, 39*).) (**Fig. 3G**, **table S2** and **SOM**). Correlated peaks with an associated ATAC-QTL were more enriched for super-enhancers if they reside in the same contact domain (1.6-fold) (**Fig. 3G**). For example, rs2732588 is an ATAC-QTL associated with three correlated peaks. Individuals who are homozygous alternate at the variant [G→A] tend to exhibit decreased chromatin accessibility in a large 100 kb region containing multiple correlated ATAC-peaks (**Fig. 3H**). The affected region partially overlaps a previously identified super-enhancer in $CD4^+$ T cells (*38*). This super-enhancer overlaps with both coding and intronic regions of *KANSL1*, a chromatin regulator that is part of the nonspecific lethal (NSL) complex controlling expression of constitutively expressed genes (*40, 41*) (**Fig. 3H**). This region containing the super-enhancer and these correlated peaks

was mostly contained inside a Hi-C contact domain, although it also extends immediately outside of the Hi-C contact domain boundary (**Fig. 3H**). Together, these data support a model in which the effects of a genetic variant on the accessibility of nearby loci is influenced by 3D proximity and that these effects are stronger when both the variant and the affected peaks lie in the same contact domain.

To further understand the consequences of ATAC-QTLs on gene regulation, we measured RNA-seq profiles from activated $CD4^+$ T cells from 96 donors (93 with matching ATAC-seq data), and mapped expression QTLs (eQTLs), genetic variants that affect gene expression (**Fig. 4**). We identified 816 genes with at least one significant local eQTL (+/- 500 kb centered around the gene) (RASQUAL, $P < 4.12\times10^{-5}$, permutation FDR < 0.05) (**Fig. 4C**, **fig. S11**, **table S5**), termed eQTL-genes. The 816 eQTL-genes were enriched in HiC contact domains compared to all genes (437 genes intersecting contact domains, 6.6 fold enrichment, $P < 3.44\times10^{-247}$, hypergeometric test), suggesting that eQTL-genes are in regions of 3D chromatin interaction in T cells.

We assessed the mediation of genetic effects on gene expression through chromatin accessibility. 71 genetic variants, corresponding to 103 unique genes, were simultaneously the lead SNP for an eQTL and an ATAC-QTL with correlated effect sizes (Pearson rho = 0.52, **Fig. 4A, table S6**). Because in-sample overlapping of eQTL and ATAC-QTLs has limited power to detect shared genetic variants, we used a feature selection prediction framework (*43, 44*) to detect associations between imputed ATAC-peaks and gene expression. For 1,790 significant ATAC-QTL-peaks,

we identified 104 genes significantly associated with predicted local (+/- 500 kb) ATAC-peak accessibility in 96 individuals (FDR < 0.05, 10 permutations, **Fig. 4B,C, table S7**), of which 67 were eQTL-genes. These results demonstrate how the inherent hierarchy in how genetic variants modulate chromatin state and gene expression enables the use of imputed chromatin accessibility to simultaneously identify genetically controlled genes and to annotate *cis* regulatory elements mediating the genetic effect.

For example, rs174556, an ATAC-QTL associated with a pair of correlated peaks, resides in a 25 kb region between two Hi-C contact domains, where the alternative allele disrupts a CTCF binding site (**Fig. 4D**). Rs174556 is linked (D'=1, R^2=0.79) with rs102275, a variant previously associated with Crohn's disease (*45*). The associated correlated peaks span the promoters of *FADS1* and *FADS2*, two fatty acid desaturase (FADS) genes. Rs174556 is also identified as an eQTL for both *FADS1* and *FADS2* in our T cell dataset (**Fig. 4C**). FADS1 and FADS2 regulate inflammation, promote cancer development by regulating the metabolism of arachiodonic acid and its derivatives prostaglandin E2, and *FADS2* knockout mice develop dermal and intestinal ulcerations (*46-49*). Given the well-known role of CTCF in maintaining the integrity of chromatin domain boundaries and insulation of transcriptional activities, abolishing CTCF binding may abolish the insulation, opening chromatin and causing increased expression of both target genes. These results suggest that variability in chromatin accessibility may underlie variability in gene expression and may increase disease risk.

In conclusion, we integrated genetic variation and ATAC-seq data from primary activated $CD4^+$ T cells from 105 healthy donors to identify and characterize genetic variants that contribute to variability in chromatin accessibility. We found widespread genetic control of chromatin accessibility, some of which affected multiple ATAC-peaks in a manner consistent with the 3D organization of the genome. Integrating genotyping, ATAC-seq and RNA-seq data provided causal anchors for predicting and explaining the variability in molecular traits in a manner consistent with known modes of transcriptional regulation. We did not find significant distal effects, consistent with reports that measured chromatin state by DNAse-I-seq (*17*), but unlike studies that measured chromatin state using ChIP-seq (*18-20*). Predicting variability in gene expression between individuals based on chromatin state is significantly impacted by technical and biological variability in the assays, but is helped by leveraging genetic variation as causal anchors. It is possible that our ability to detect weaker interactions and predict gene expression could significantly improve with increased sample sizes and deeper sequencing.

Our findings, derived from large scale mapping of epigenetic quantitative traits in primary human cells implicated in many diseases, provide a molecular framework for the interpretation of disease-causing variants, focused on modeling how genetic variants could alter local chromatin structure to modulate gene expression. Future studies that use other disease-relevant primary cells and tissues will help pinpoint causal disease variants and understand the regulatory mechanism underlying common disease.

Supplementary Online Materials

Study subjects and genotyping

Healthy subjects between the ages of 18 to 56 (avg. 29.9) enrolled in the PhenoGenetic study (*8*) were recruited from the Greater Boston Area and gave written informed consent for the studies. Individuals were excluded if they had a history of inflammatory disease, autoimmune disease, chronic metabolic disorders or chronic infectious disorders. Genotyping demographics of the donors are listed in **table S8**. Genotyping using the Illumina Infinium Human OmniExpress Exome BeadChips (704,808 SNPs are common variants [MAF > 0.01] and 246,229 are part of the exomes, respectively; Illumina Inc., San Diego, CA) has been previously described (*16*). The genotype success rate was at least 97%. We applied rigorous subject and SNP quality control (QC) that includes: (1) gender misidentification; (2) subject relatedness; (3) Hardy-Weinberg Equilibrium testing; (4) use concordance to infer SNP quality; (5) genotype call rate; (6) heterozygosity outlier; and (7) subject mismatches. We excluded 1,987 SNPs with a call rate < 95%, 459 SNPs with Hardy-Weinberg equilibrium p-value < 10^{-6}, and 63,781 SNPs with MAF < 1% from the 704,808 common SNPs (a total of 66,461 SNPs excluded).

We used the IMPUTE2 software (version: 2.3.2) to impute the post-QC genotyped markers from the entire Immvar cohort (N = 688) using reference haplotype panels from the 1000 Genomes Project (The 1000 Genomes Project Consortium Phase III) that contain a total of 37.9 Million SNPs in 2,504 individuals with ancestry from West Africa, East Asia, and Europe. After genotype imputation, we extracted the genotypes for 108 individuals assayed for chromatin

accessibility and gene expression. Additional filtering for SNPs with MAF < 0. 1 in our cohort, resulted in 4,558,693 and 4,421,936 common variants tested for chromatin accessibility and gene expression assays, respectively.

Preparation and activation of primary human $CD4^+$ T cells

$CD4^+$ T cells were isolated and stimulated as previously described (*10*). Briefly, $CD4^+$ T cells were isolated from whole blood by negative selection using RosetteSep human $CD4^+$ T cell enrichment cocktail (STEMCELL Technologies Inc., Vancouver, BC) and RosetteSep density medium gradient centrifugation. Isolated $CD4^+$ T cells were placed in freezing container at -80°C for overnight, and then moved into a liquid nitrogen tank for long-term storage. On the day of activation, $CD4^+$ T cells were thawed in a 37°C water bath, counted and resuspended in RPMI-1640 supplemented with 10% FCS, and plated at 50,000 cells per well in a 96 well round-bottom plate. Cells were either left untreated or stimulated for 48 hours with beads conjugated with anti-CD3 and anti-CD28 antibodies (Dynabeads, Invitrogen #11131D, Life Technologies) at a cell:bead ratio of 1:1. At each time point, cells were further purified by a second step positive selection with $CD4^+$ Dynabeads (Invitrogen #11145D, Life Technologies).

ATAC-seq profiling

ATAC-seq profiles were collected for 139 individuals (**table S8**). We performed ATAC-seq as previously described (*21*), with a modification in the lysis buffer to reduce mitochondrial DNA contamination, while maintaining high complexity of nuclear reads. 200,000 purified $CD4^+$ T cells were lysed with cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$

and 0.03% tween20). Immediately after lysis, nuclei were spun at 500g for 8 minutes at 4°C. After pelleting the nuclei, we carefully removed the supernatant and resuspended the nuclei with Tn5 transposase reaction mix (25 ul 2X TD buffer, 2.5 ul Tn5 transposase, and 22.5 ul nuclease-free water) (Illumina Inc). The transposition reaction was performed at 37°C for 30 minutes. Immediately after the transposition reaction, DNA was purified using a Qiagen MinElute kit. Libraries were sequenced on an Illumina HiSeq 2500 sequencer to an average read depth of 42 million (+/- 38 million) per sample (fig. S2), with low mtDNA contamination (0.30%-5.39%, 1.96% on average), low rates of multiply mapped reads (6.7%-56%, 19% on average) and a relatively high proportion of usable nuclear reads (60%-92%, 79% on average).

RNA-seq profiling

RNA-seq profiles were collected for 95 individuals, of which 93 have matching ATAC-seq profiles (**table S8**). RNA was isolated using Qiagen RNeasy Plus Mini Kit and RNA integrity was quantified by Agilent RNA 6000 Nano Kit using the Agilent Bioanalyzer. Purified RNA were converted to RNA-seq libraries using a previously published protocol (*50*), where reverse transcription was carried out based on the SMART template switching method and the resulting cDNA was further tagmented and PCR amplified using Nextera XT DNA Sample kit (Illumina) to add the Illumina sequencing adaptors. Samples were sequenced on Illumina HiSeq 2500 to an average depth of 16.9 million reads per sample (+/- 8.7 million).

***In situ*-Hi-C**

$CD4^+$ T cells were isolated from commercially available fresh blood of healthy individuals (Research Blood Components). $CD4^+$ T cells were stimulated for 48 hours with beads conjugated with anti-CD3 and anti-CD28 antibodies and then crosslinked with 1% formaldehyde for 10 min at room temperature. *In situ*-Hi-C was performed as previously described (*34*). Libraries were sequenced using Illumina HiSeq and NextSeq, to produce ~1 billion 100bp paired-end reads.

Alignment of ATAC-seq reads

25bp ATAC-seq reads were aligned to the human genome assembly (hg19) with the Burrows Wheeler Aligner-MEM (version: 0.7.12) (*51*). For each sample, mitochondrial reads and multiply-mapped reads were filtered out using BEDtools (function intersectBed) (*52*). After filtering, we had an average of 42 million (+/- 38 million) reads per sample.

ATAC-seq peak identification

Filtered ATAC-seq reads from matched unstimulated and activated $CD4^+$ T cell ATAC-seq reads from 6 individuals were merged (separately for unstimulated and for activated cells) using the Samtools function "merge". Peaks were called on the unstimulated $CD4^+$ T cell merged bam file and the activated $CD4^+$ T cell merged bam file using MACS2 –callpeak (with parameters --nomodel, --extsize 200, and --shift 100), such that there were 36,486 unstimulated peaks with an average width of 520 bp (+/- 319 bp) and 52,154 activated $CD4^+$ T cell peaks with an average width of 483 bp (+/- 344 bp) (Benjamini-Hochberg FDR < 0.05). The activated and unstimulated $CD4^+$ T cell peaks were further merged (using the BEDtools "merge" function), to a total of 63,763 jointly called peaks. A matrix of the coverage for each of the 63,763 peaks in each of the

12 samples was used as input for calling differential peaks. Differential peaks between activated and unstimulated conditions were identified using the DESeq2 R package (version 3.2) (*53*), with 8,298 regions more accessible in unstimulated $CD4^+$ T cells, and 28,017 regions more accessible in activated $CD4^+$ T cells (FDR < 0.05).

Enrichment of transcription factor binding motifs

We used the Homer suite (*52*), which uses ChIP-seq data from the ENCODE (*25*) and Epigenomics Roadmap (*15*) projects, to determine transcription factor enrichment within our ATAC-peaks, using the findMotifsGenome.pl (with parameters hg19 and –size given). For the analysis in **Fig. 2E,** we used an additional parameter –bg, using the activated $CD4^+$ T cell peaks as background, instead of the local background generated by HOMER. This allowed us to determine which transcription factors were more enriched in our ATAC-QTLs than all background ATAC-peaks.

Outlier and sample mix-up analysis

Using the software VerifyBamID (*54*), we matched each ATAC-seq and RNA-seq sample with each genotyping array. Samples were identified as those with the highest fIBD, and those with designated labels not matching the VerifyBamID predicted labels were flagged as sample mix-ups. We switched the designated label to the predicted label for cases where the fIBD > 90%. 15 out of the 139 total ATAC-seq samples were re-labeled and 4 out of the 110 total RNA-seq samples were re-labeled. For the ATAC-seq samples: 18 do not have genotypes, 3 are outliers, 1 did not match anyone. For the 110 RNA-seq samples: 8 samples do not have genotypes, 5 are

outliers, 1 did not match anyone. 111 ATAC-seq samples and 96 RNA-seq samples were used in the final analysis after filtering for outliers (average mean correlation to others samples < 0.7). In the pilot study, there were 5 people total, 1 person was repeated for a total of 6 samples, none were genotyped.

Mapping of ATAC-QTLs

We mapped local ATAC-QTLs by running RASQUAL (*29*) on the 52,154 peaks identified in activated $CD4^+$ T cells and 4,558,693 imputed genetic variants, testing variants within a 40-kb window of each ATAC-peak, and filtering for a minor allele frequency of 10% to remove rare variants. The input to RASQUAL is the number of reads in each peak quantified using BEDtools "coverage" with using uniquely mapped nuclear reads for each individual. Duplicated fragments were kept for quantification. Date and time of visit, sex, age, race, ethnicity, height, weight, BMI, blood pressure, sequencing batch, and preparation batch were removed as covariates, along with four principal components to minimize confounding factors. Using the RASQUAL "-r" option, 10 random permutations for each feature were generated. Then, the empirical null distributions and P-values were compared using the R qvalue (*55*) package, for a total of 1,790 local ATAC-QTLs at a FDR of 0.1. Distal ATAC-QTLs were similarly mapped at a window of > 40-kb but < 1 Mb to attain a total of 7,301 distal ATAC-QTLs.

Hi-C data analysis

Data were processed using a custom pipeline that uses BWA (*51*) to map each read separately and Hi-C contact domains and chromatin loops were identified as previously described (*34*).

Determination of distance from ATAC-seq peak to contact domains

We determined the distance from each feature of interest to the middle of the closest contact domain. We analyzed the following features: (**1**) ATAC-peaks; (**2**) ATAC-peaks containing a significant genetic association ("ATAC-QTL-peaks"); ATAC-peaks containing (**3**) BATF, (**4**) ETS1, or (**5**) CTCF motifs; and ATAC-QTL-peaks containing (**6**) BATF, (**7**) ETS1, or (**8**) CTCF motifs. We normalized the distances from each feature to the closest domain by the length of the domain. In order to determine that the distribution of the distance between a given feature and a contact domain is different than the null distribution, we kept the length of each contact domain constant and shuffled the positions of the contact domain. The distances from the feature to the contact domain were binned into 30 bins and divided by the binned distances between a given feature and the shuffled contact domains to determine enrichment at each position.

Correlated peak analysis

To identify correlated peaks, we tested for correlation between every pair of the 52,154 ATAC-peaks within 1.5 Mb of each other using a linear regression in Matrix eQTL (*56*). To correlate peaks, we first normalized the ATAC-peaks by (**1**) removing sequencing depth bias using a median normalization, (**2**) standardizing the matrix by subtracting out the mean and dividing by the standard deviation; and (**3**) quantile normalizion of the matrix (Bolstad BM (2016). *preprocessCore: A collection of pre-processing functions*. R package version 1.34.0, https://github.com/bmbolstad/preprocessCore). Next, we removed covariates (as described for

mapping of ATAC-QTLs) and three principal components. Then, we identified 851/1,762 correlated peaks (387 unique ATAC-peaks) with ATAC-QTLs (FDR < 0.05), of which 159 correlated peaks (93 unique ATAC-peaks) with ATAC-QTLs reside in contact domains. To ensure that the correlated peaks were enriched in contact domains, we again permuted the position of the contact domains while keeping the length of the contact domains constant and performed the same analysis (**fig. S10**).

RNA-seq analysis

25bp paired end RNA-seq reads were aligned to the hg19 using UCSC transcriptome annotations. Expression levels (expected counts) were determined using RSEM (*57*). We applied TMM normalization to the expected counts using the edgeR package and filtered for genes that had TMM count > 1 in at least 75% of the samples. For the mapping of eQTLs, we inputted raw counts for filtered genes into RASQUAL which performs internal normalization. For the repeatability, heritability and predictability analyses, we used log transformed TMM counts of filtered genes in order to fit generalized linear models.

eQTL fine-mapping

RASQUAL (*29*) was used to map eQTLs within a 1 Mb window of a gene, using gene expression levels (TPM) from RSEM and genotypes filtered as above for a minor allele frequency of 0.1. To minimize confounding factors, 16 principal components, date and time of visit, sex, age, race, ethnicity, height, weight, BMI, blood pressure, sequencing batch, and preparation batch were removed as covariates. The RASQUAL "-r" option was used for

permutations to determine the empirical null distribution and compute a FDR, retaining 816 eQTLs at FDR<0.05.

GWAS enrichment

The GREGOR suite was used for calculating the enrichment of loci from GWAS in features of interest (*56*): (**1**) peaks differentially accessible in activated CD4+ T cells; (**2**) peaks differentially accessible in unstimulated CD4+ T cell peaks, and (**3**) peaks shared in both conditions. The 95% confidence interval was calculated from the $\log_{10}$(odds ratio), where the odds ratio was the number of GWAS overlaps / expected overlaps.

Overlap of traits in genomic regions

The Homer suite tool annotatePeaks.pl was used to determine the number of quantitative traits intersecting each genomic feature of interest. For enhancers, we specified the –ann parameter for the T cell H3K27ac enhancer annotation (*16*).

Enrichment in Super-enhancer regions

Using the BEDtools intersect function, we calculated how many of the correlated peaks, genetically-associated correlated peaks, and genetically-associated correlated peaks that fall in contact domains are also in stimulated T helper super-enhancers (as reported in Hinsz et al (*38*)). To calculate an enrichment score, for each pair of correlated peaks, we fixed one peak and mirrored the second peak by the peak distance to preserve the genomic properties of the correlated peak, while breaking any correlation to annotated enhancers. To calculate a

confidence interval for the enrichment score, we shuffled the position of super-enhancers, while maintaining the length of the super-enhancer 10 times and calculated the enrichment score of intersected features to the mirrored features with the shuffled super-enhancers.

Gkm-SVM and deltaSVM

We ran gkm-SVM (*31, 32*) on 24,745 300bp ATAC-peaks centered on MACS summits using default parameters and an equal size GC matched negative set, excluding from training any region containing a SNP to be scored by deltaSVM, and repeated with 5 independent negative sets, and averaged the deltaSVM predictions, as previously described (*30*). We then calculated deltaSVM for each SNP within 200bp of the peak signal of an ATAC-QTL with peak p-value < 10^{-5}, scoring 663 SNPs in 500 loci. We find a Pearson correlation of C=0.611 between ATAC-QTL beta and the largest deltaSVM SNP. 442 of the peak p-value SNPs had the largest deltaSVM, but 58 flanking SNPs scored more highly than the peak p-value SNP and disrupt immune associated TF binding sites. While the gkm-SVM weights fully specify the deltaSVM score, for interpretation we associated the large gkm-SVM weights with the most similar TF PWM from a catalog of JASPAR, Transfac, Uniprobe, and Homer motifs.

Heritability and prediction of gene expression and ATAC-peaks

Data for predictability and heritability analysis of gene expression and ATAC-peaks was prepared in following way. Each ATAC-peak was residualized against its 4 principal components and patient data covariates (date and time of visit, sex, age, height, weight, systolic and diastolic blood pressure) for cross-validated prediction studies, or were included as fixed

effects for the heritability analysis. Analogously, we used 16 principal components in the analyses of heritability and predictability of gene expression.

Repeatability was calculated by leveraging repeated measures of gene expression in 25 individuals whose cells were sampled on two different dates two years apart. Elastic net model prediction analysis was performed using glmnet R package with the L1 ratio set to 0.5 with 5-fold cross-validation. For joint prediction of gene expression, we chose distinct weights for ATAC-peak and genotype features using a grid search approach (parameter grid 2^-3 to 2^3) to maximize the mean cross-validation R^2. Reported R^2 estimates are calculated as in Gamazon et al. (*43*). Restricted maximum likelihood heritability (h^2) estimates were calculated using GCTA software (*58*) with algorithm 0 and no constraints on heritability (i.e. h^2 can be less than 0). For the gene expression predictability analysis, we used genotype and ATAC-peak features +/- 500 kb from the transcription start site of the gene. For the heritability and predictability analysis of ATAC-peaks, we used genotypes +/- 500 kb and +/- 20 kb from the center of each ATAC-peak respectively.

Association of gene expression to genetically imputed ATAC-peaks

In order to assess the ability to predict gene expression from genotypes mediated by ATAC-peaks, we associated gene expression to imputed intensities of ATAC-peaks, similar to the approach proposed by Gamazon et al (*43*). First, we split the available dataset into a training set and a test set of equal size (2 x 46). We used the training set to estimate the effect of genetic variants on ATAC-peaks using ordinary linear regression. Next, we applied the estimated effects

to the test set to predict the intensity of ATAC-peaks. We correlated the imputed ATAC-peaks with gene expression and report the mean R^2 statistic for each gene from the analysis of three random partitions of training and test sets. We derived empirical *p*-value and FDR estimates on the R^2 statistic by shuffling the gene expression matrix 10 times. These results are compared to those obtained from associating genotype data with gene expression in the full cohort of 92.

Figure legends

Figure 1. Chromatin dynamics in human T cell activation. (A) Study overview. (B) Differential ATAC-peaks. Shown are ATAC-peaks (columns) in six individuals (rows) before (top) and 48hr after (bottom) activation of primary T cells with anti-CD3/CD28 antibodies. (C) ATAC-peaks in Th cell enhancers. Bar chart shows the number of ATAC-peaks that overlap previously identified enhancers (blue and green) and other genomic features (red and orange) in different Th cell subtypes (*15*) in $CD4^+$ T cells pooled samples from either unstimulated cells, cells activated for 48hr or both. (D) Transcription factor motif enriched in ATAC-peaks. Shown are the enrichments (X axis) and significance (Y axis) for transcription factor motifs in ATAC-peaks that are only present in unstimulated $CD4^+$ T cells (left), 48hr activated $CD4^+$T cells (middle), or in both (right). (E) TF footprinting. Shown are for each TF motif (as identified in ENCODE (*59*)), aggregated plots of mean chromatin accessibility (Y axis, mean ATAC-seq signal) along the TF binding site (X axis) in unstimulated (bottom) or 48hr activated (top) $CD4^+$ T cells. (F) GWAS variant enrichment in ATAC-peaks. Shown are the enrichments (X axis) and significance (Y axis) for loci associated with the indicated disease or phenotype in ATAC-peaks

that are only present in unstimulated $CD4^+$ T cells (left), 48hr activated $CD4^+$ T cells (middle), or in both (right).

Figure 2. Genetic variants that affect chromatin states in human T cell activation. (A) ATAC-QTLs. Q-Q plot for all tests of association between activated ATAC-peaks and variants within 40 kb regions centered on the target ATAC-seq peak (red dashed line – expected). (B) Heritability of chromatin state. Shown are out-of-sample R^2 predictability of ATAC-peaks based on genotypes +/- 20 kb of each peak (y-axis) as a function of heritability h^2 estimated based on genotypes within +/- 500 kb of each peak (x-axis). Solid triangles: significantly heritable peaks (q-value < 0.1). (C) ATAC-QTLs are close to associated ATAC-peaks. Shown is a distribution of the distances of ATAC-QTLs and their associated ATAC-peaks (X axis, bp). Red: SNP contained in associated ATAC-peak. Green: SNP within 2 kb of associated ATAC-peak. (D) ATAC-QTL-peaks are closer to TSS than to TTS. Shown are the distributions of the distances of ATAC-QTL-peaks to the closest TSS (left) or TTS (right). (E) Transcription factor motifs that are enriched in ATAC-QTLs. Shown are the enrichments (X axis) and significance (Y axis) for transcription factor binding sites in ATAC-QTL-peaks. (F) ATAC-QTLs affect binding motifs in an allele specific manner. Shown are for each of three indicated TF binding site (as identified in ENCODE (*59*)), aggregated plots of mean chromatin accessibility of ATAC-QTL-peaks overlapping each TFBS (Y axis, mean ATAC-seq signal) along the TF binding site (X axis, log2 distance) for heterozygote (light blue), homozygous with high ATAC-seq signal (red) and homozygous with low ATAC-seq signal (black) genotypes. (G, H) Unsupervised analysis associates ATAC-QTLs with key TF binding sites. (G) The effect sizes of lead ATAC-QTL

SNPs (X axis) are well correlated with deltaSVM scores (Y axis) for these variants. (H) Key motifs. Shown are the motifs for 11 TFs associated with most of the large gkmSVM weights, and the proportion of the overall disruption (%, bottom) explained by ATAC-QTLs. (I) GWAS variants enrichment for ATAC-QTL-peaks. Shown are the enrichments (X axis) and significance (Y axis) for ATAC-QTL-peaks overlapping variants associated with indicated disease or phenotype. (J) ATAC-QTL and GWAS variant disrupting TF binding site. Shown is ATAC-QTL rs17293632 on chromosome 15 and the overlapping binding site for BATF. ATAC-seq profiles were combined between individuals with homozygous reference genotype (black), heterozygous genotype (light blue) and homozygous alternative genotype (red).

Figure 3. Chromatin architecture constraints ATAC-QTLs and their ability to impact correlated peaks. (A) ATAC-QTLs. Q-Q plots of the *P* value for the association of ATAC-QTLs for correlated ATAC-QTL-peaks (blue) and single ATAC-QTL-peaks (red). (B) Correlated peak correlations. Q-Q plots of the *P* value for the correlation of genetically controlled correlated peaks (blue) and all correlated peaks (red). (C) Example of genetically controlled correlated peaks. Shown is ATAC-QTL rs10815868 on chromosome 9 associated with four ATAC-peaks, overlapping a BATF binding site, where the alternative allele (A→G) is predicted to disrupt the site. ATAC-seq profiles were combined between individuals with homozygous reference genotype (black), heterozygous genotype (light blue) and homozygous alternative genotype (red). The correlated peaks are within a Hi-C contact domain (grey bar). (D) Relation between contact domains and ATAC-QTLs associated with TF binding motifs. Shown are the distributions (Y axis, density) of position (X axis) of ATAC-QTLs (density histogram,

black) and ATAC-peaks (density histogram, grey) that contain either an ETS1, CTCF, or BATF binding sites relative to Hi-C chromatin contact domain boundaries (dotted red lines). (E) Contact domains impact the extent of correlation between correlated peaks. Bar plot of the enrichments (y axis) of the position of pair of peaks relative to a contact domain for correlated peaks where both peaks (blue) or only one peak (purple) reside in the same contact domain, both cross a contact domain boundary (green) or both are outside a contact domain (red), compared to their respective backgrounds, obtained by fixing one peak and calculating the proportion of overlap by flipping the location of second peak (**SOM**). (F) Contact domains impact genetically controlled correlated peaks. Q-Q plots of the *P* value for the association of SNPs to either peak, when the correlated peaks are both inside a contact domain (blue) or are outside a Hi-C contact domain (red). (G) ATAC-QTLs in super-enhancers. Shown are the enrichment of proportion of features overlapping super-enhancers (Y axis) for each of correlated peaks, genetically controlled correlated peaks, and genetically controlled correlated peaks in a domain compared to randomly shuffled domains (X axis). Error bars are obtained by permuting the position of the super-enhancers, while keeping the length of the super-enhancer constant. (H) An example of an ATAC-QTL associated with a correlated peak (rs2732588) residing in a CD4+ T cell super-enhancer. The super-enhancer and a HiC contact domain are marked by purple and grey bars, respectively. ATAC-seq profiles were combined between individuals with homozygous reference genotype (black), heterozygous genotype (light blue) and homozygous alternative genotype (red).

Figure 4. ATAC-QTLs and the effects on gene expression. (A) Correlation of effect sizes between ATAC-QTLs (X axis) and eQTLs (Y axis). (B) Overlap of number of genes associated with imputed ATAC-peaks on the test set (purple) or with genotypes on the test set (light blue) . (C) Manhattan plot for the tagged loci for the imputed ATAC-peak mediating eQTLs, with the associations to rs968568 highlighted in green. (D) An example ATAC-QTL (rs174556) on chromosome 11 that is also an eQTL for *FADS1* and *FADS2*. The alternative allele (C→T) impacts the binding site for CTCF. ATAC-seq (top) and RNA-seq (bottom) profiles were combined between individuals with homozygous reference genotype (black), heterozygous genotype (light blue) and homozygous alternative genotype (red).

References

1. M. I. McCarthy *et al.*, Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-369 (2008).
2. P. M. Visscher, M. A. Brown, M. I. McCarthy, J. Yang, Five years of GWAS discovery. *Am J Hum Genet* **90**, 7-24 (2012).
3. J. N. Hirschhorn, M. J. Daly, Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108 (2005).
4. M. T. Maurano *et al.*, Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).
5. B. E. Stranger *et al.*, Population genomics of human gene expression. *Nature genetics* **39**, 1217-1224 (2007).
6. T. Lappalainen *et al.*, Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511 (2013).
7. A. Battle *et al.*, Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* **24**, 14-24 (2014).
8. T. Raj *et al.*, Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519-523 (2014).
9. M. N. Lee *et al.*, Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
10. C. J. Ye *et al.*, Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
11. E. P. Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
12. M. B. Gerstein *et al.*, Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
13. S. Neph *et al.*, An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).
14. R. E. Thurman *et al.*, The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
15. C. Roadmap Epigenomics *et al.*, Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).
16. K. K. Farh *et al.*, Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-343 (2015).
17. J. F. Degner *et al.*, DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-394 (2012).
18. M. Kasowski *et al.*, Extensive variation in chromatin states across humans. *Science* **342**, 750-752 (2013).
19. G. McVicker *et al.*, Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747-749 (2013).
20. H. Kilpinen *et al.*, Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744-747 (2013).

21. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).
22. M. Kurachi *et al.*, The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8+ T cells. *Nat Immunol* **15**, 373-383 (2014).
23. P. Li *et al.*, BATF-JUN is critical for IRF4-mediated transcription in T cells. *Nature* **490**, 543-546 (2012).
24. T. L. Murphy, R. Tussiwand, K. M. Murphy, Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. *Nat Rev Immunol* **13**, 499-509 (2013).
25. P. Kheradpour, M. Kellis, Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic acids research* **42**, 2976-2987 (2014).
26. C. M. Evans, R. G. Jenner, Transcription factor interplay in T helper cell differentiation. *Brief Funct Genomics* **12**, 499-511 (2013).
27. R. M. Samstein *et al.*, Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151**, 153-166 (2012).
28. E. V. Rothenberg, The chromatin landscape and transcription factors in T cell programming. *Trends Immunol* **35**, 195-204 (2014).
29. N. Kumasaka, A. J. Knights, D. J. Gaffney, Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature genetics* **48**, 206-213 (2016).
30. D. Lee *et al.*, A method to predict the impact of regulatory variants from DNA sequence. *Nature genetics* **47**, 955-961 (2015).
31. M. Ghandi, D. Lee, M. Mohammad-Noori, M. A. Beer, Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology* **10**, e1003711 (2014).
32. M.-N. M. Ghandi M, Ghareghani N, Lee D, Garraway L, Beer MA., gkmSVM, an R package for gapped-kmer SVM. *Bioinformatics.* **Apr 19**, (2016).
33. J. S. Delisle *et al.*, The TGF-beta-Smad3 pathway inhibits CD28-dependent cell growth and proliferation of CD4 T cells. *Genes Immun* **14**, 115-126 (2013).
34. S. S. Rao *et al.*, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
35. C. Hou, H. Zhao, K. Tanimoto, A. Dean, CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 20398-20403 (2008).
36. J. E. Phillips, V. G. Corces, CTCF: master weaver of the genome. *Cell* **137**, 1194-1211 (2009).
37. E. Splinter *et al.*, CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development* **20**, 2349-2354 (2006).
38. D. Hnisz *et al.*, Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013).
39. W. A. Whyte *et al.*, Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).

40. J. Dias *et al.*, Structural analysis of the KANSL1/WDR5/KANSL2 complex reveals that WDR5 is required for efficient assembly and chromatin targeting of the NSL complex. *Genes & development* **28**, 929-942 (2014).
41. K. C. Lam *et al.*, The NSL complex regulates housekeeping genes in Drosophila. *PLoS Genet* **8**, e1002736 (2012).
42. A. E. Taylor *et al.*, Mendelian randomization in health research: using appropriate genetic variants and avoiding biased estimates. *Econ Hum Biol* **13**, 99-106 (2014).
43. E. R. Gamazon *et al.*, A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics* **47**, 1091-1098 (2015).
44. A. Gusev *et al.*, Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**, 245-252 (2016).
45. A. Franke *et al.*, Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature genetics* **42**, 1118-1125 (2010).
46. Y. Y. Fan *et al.*, Characterization of an arachidonic acid-deficient (Fads1 knockout) mouse model. *J Lipid Res* **53**, 1287-1295 (2012).
47. A. Barrie *et al.*, Prostaglandin E2 and IL-23 plus IL-1beta differentially regulate the Th1/Th17 immune response of human CD161(+) CD4(+) memory T cells. *Clin Transl Sci* **4**, 268-273 (2011).
48. D. Sakata, C. Yao, S. Narumiya, Prostaglandin E2, an immunoactivator. *J Pharmacol Sci* **112**, 1-5 (2010).
49. C. K. Stroud *et al.*, Disruption of FADS2 gene in mice impairs male reproduction and causes dermal and intestinal ulceration. *J Lipid Res* **50**, 1870-1880 (2009).
50. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
51. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).
52. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
53. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
54. G. Jun *et al.*, Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-848 (2012).
55. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445 (2003).
56. A. A. Shabalin, Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-1358 (2012).
57. B. Li, C. N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).
58. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
59. E. M. Schmidt *et al.*, GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601-2606 (2015).

Acknowledgements: We thank the ImmVar participants. We would like to thank Jason Buenrostro for critical reading of the manuscript and advice on ATAC-seq analysis, Jenna Pfiffner and Charles Fulco for initial experimental help with ATAC-seq, Alicia Schep for ATAC-seq nucleosome free caller, Natasha Asinovski and Ho-keun Kwon for help setting up primary T cell cultures and members of the Regev laboratory for discussions. M.B. and K.L.H. are supported by NIH HG007348 to M.B., H.Y.C. is supported by NIH grant P50-HG007735, C.S.C is supported by the NIH through a Ruth L. Kirschstein National Research Service Award (F32-DK096822). This work was supported by the Klarman Cell Observatory at the Broad Institute. A.R. is a Howard Hughes Medical Institute Investigator.

Figure 1

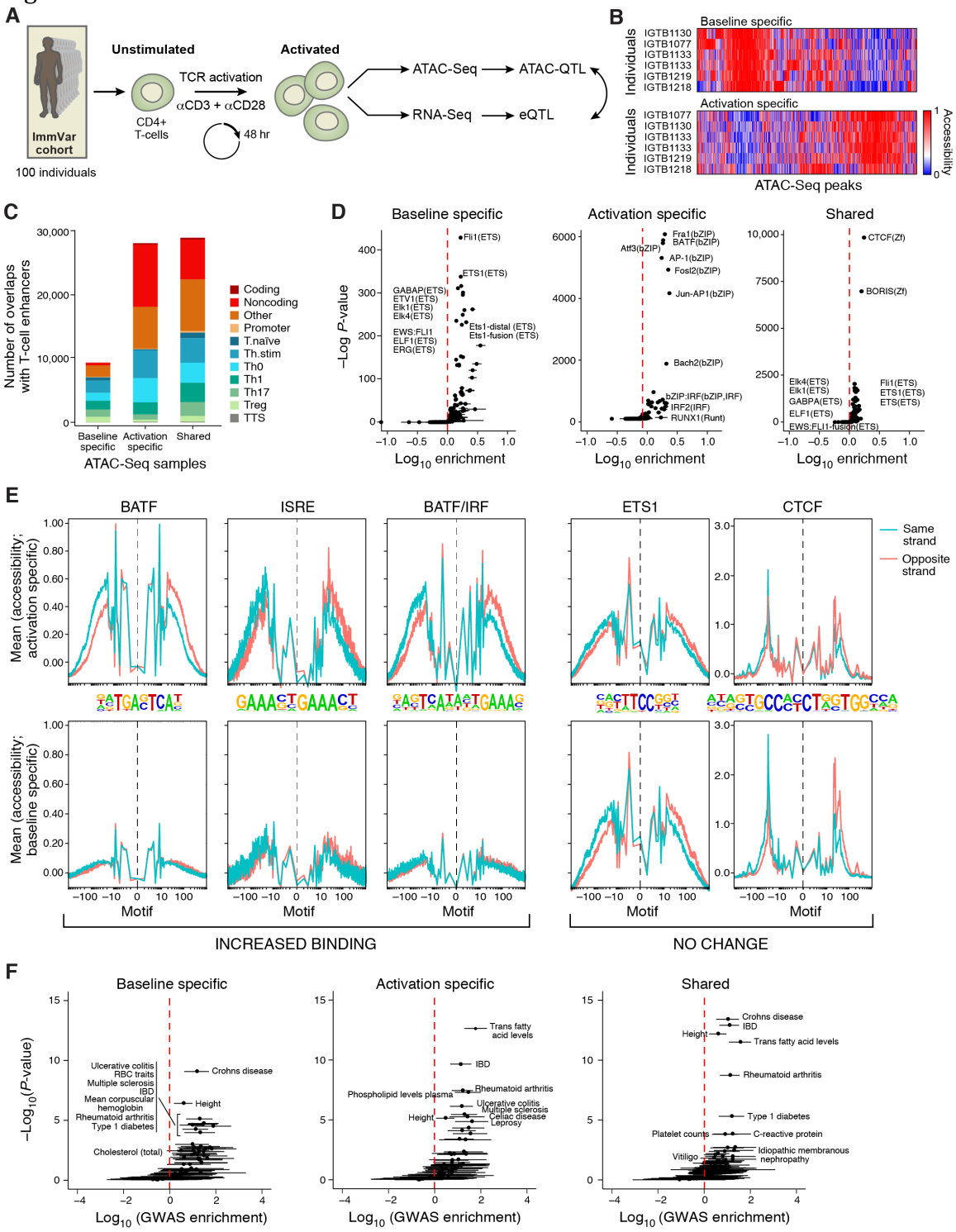

A
ImmVar cohort
100 individuals
Unstimulated
CD4+ T-cells
TCR activation
αCD3 + αCD28
48 hr
Activated
ATAC-Seq
ATAC-QTL
RNA-Seq
eQTL
B
Baseline specific
Activation specific
Individuals
IGTB1130
IGTB1077
IGTB1133
IGTB1219
IGTB1218
Accessibility
ATAC-Seq peaks
C
Number of overlaps with T-cell enhancers
Coding
Noncoding
Other
Promoter
T.naïve
Th.stim
Th0
Th1
Th17
Treg
TTS
Baseline specific
Activation specific
Shared
ATAC-Seq samples
D
Baseline specific
Activation specific
Shared
−Log P-value
Log10 enrichment
Fli1(ETS)
ETS1(ETS)
GABAP(ETS)
ETV1(ETS)
Elk1(ETS)
Elk4(ETS)
EWS:FLI1
ELF1(ETS)
ERG(ETS)
Ets1-distal (ETS)
Ets1-fusion (ETS)
Fra1(bZIP)
BATF(bZIP)
Atf3(bZIP)
AP-1(bZIP)
Fosl2(bZIP)
Jun-AP1(bZIP)
Bach2(bZIP)
bZIP:IRF(bZIP,IRF)
IRF2(IRF)
RUNX1(Runt)
CTCF(Zf)
BORIS(Zf)
Elk4(ETS)
Elk1(ETS)
GABPA(ETS)
ELF1(ETS)
EWS:FLI1-fusion(ETS)
Fli1(ETS)
ETS1(ETS)
ETS(ETS)
E
BATF
ISRE
BATF/IRF
ETS1
CTCF
Same strand
Opposite strand
Mean (accessibility; activation specific)
Mean (accessibility; baseline specific)
Motif
INCREASED BINDING
NO CHANGE
F
Baseline specific
Activation specific
Shared
−Log10(P-value)
Log10 (GWAS enrichment)
Crohns disease
Ulcerative colitis
RBC traits
Multiple sclerosis
IBD
Mean corpuscular hemoglobin
Height
Rheumatoid arthritis
Type 1 diabetes
Cholesterol (total)
Trans fatty acid levels
IBD
Rheumatoid arthritis
Phospholipid levels plasma
Ulcerative colitis
Multiple sclerosis
Celiac disease
Leprosy
Height
Crohns disease
IBD
Height
Trans fatty acid levels
Rheumatoid arthritis
Type 1 diabetes
Platelet counts
C-reactive protein
Vitiligo
Idiopathic membranous nephropathy

Figure 2

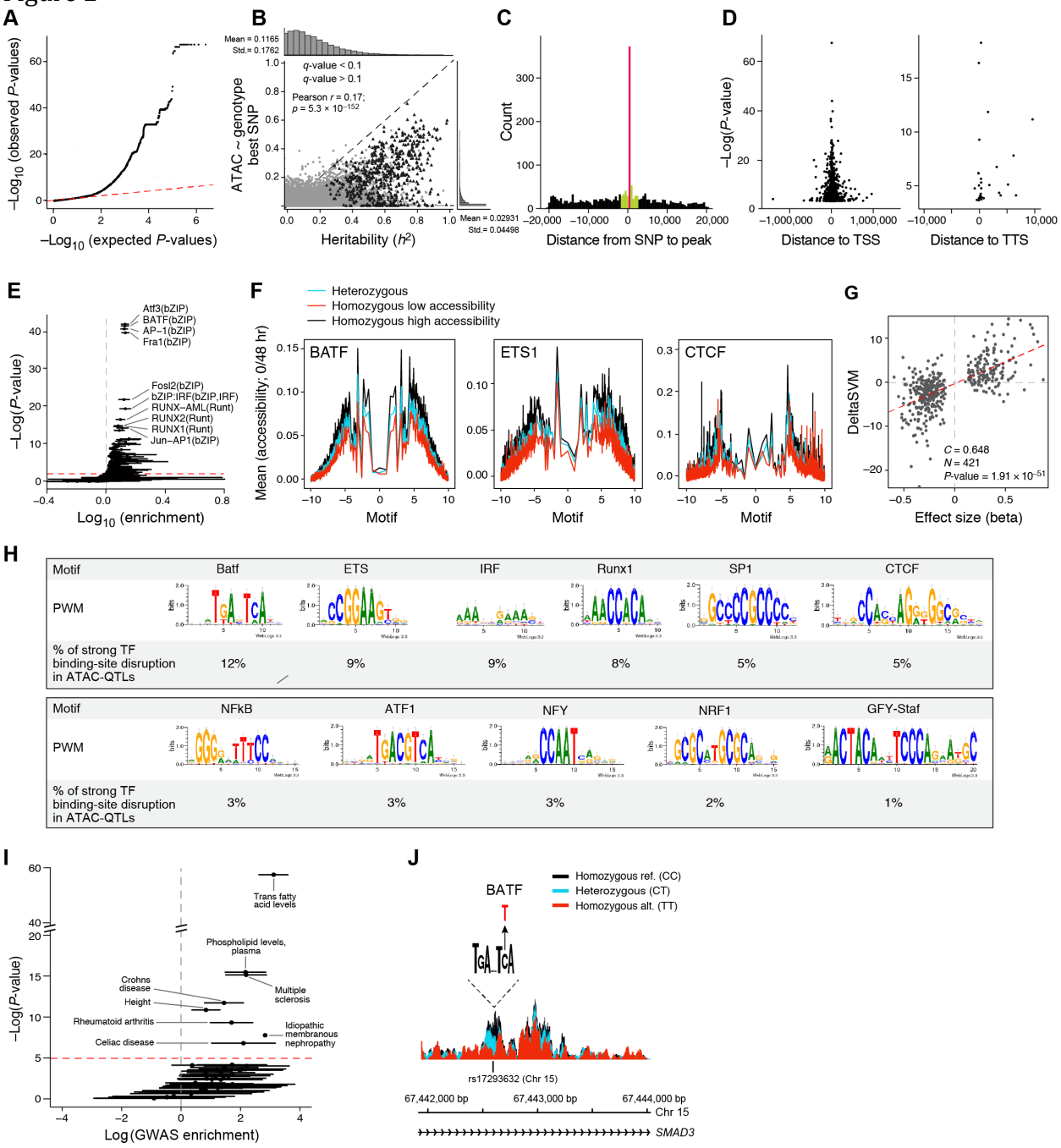

A
−Log10 (observed P-values)
−Log10 (expected P-values)
B
Mean = 0.1165
Std.= 0.1762
q-value < 0.1
q-value > 0.1
Pearson r = 0.17;
p = 5.3 × 10−152
ATAC ~ genotype best SNP
Heritability (h2)
Mean = 0.02931
Std.= 0.04498
C
Count
Distance from SNP to peak
D
−Log(P-value)
Distance to TSS
Distance to TTS
E
Atf3(bZIP)
BATF(bZIP)
AP−1(bZIP)
Fra1(bZIP)
Fosl2(bZIP)
bZIP:IRF(bZIP,IRF)
RUNX−AML(Runt)
RUNX2(Runt)
RUNX1(Runt)
Jun−AP1(bZIP)
−Log(P-value)
Log10 (enrichment)
F
Heterozygous
Homozygous low accessibility
Homozygous high accessibility
Mean (accessibility; 0/48 hr)
BATF
ETS1
CTCF
Motif
G
DeltaSVM
C = 0.648
N = 421
P-value = 1.91 × 10−51
Effect size (beta)
H
Motif
Batf
ETS
IRF
Runx1
SP1
CTCF
PWM
% of strong TF binding-site disruption in ATAC-QTLs
12%
9%
9%
8%
5%
5%
Motif
NFkB
ATF1
NFY
NRF1
GFY-Staf
PWM
% of strong TF binding-site disruption in ATAC-QTLs
3%
3%
3%
2%
1%
I
Trans fatty acid levels
Phospholipid levels, plasma
Crohns disease
Height
Multiple sclerosis
Rheumatoid arthritis
Idiopathic membranous nephropathy
Celiac disease
−Log(P-value)
Log(GWAS enrichment)
J
BATF
Homozygous ref. (CC)
Heterozygous (CT)
Homozygous alt. (TT)
rs17293632 (Chr 15)
67,442,000 bp
67,443,000 bp
67,444,000 bp
Chr 15
SMAD3

Figure 3

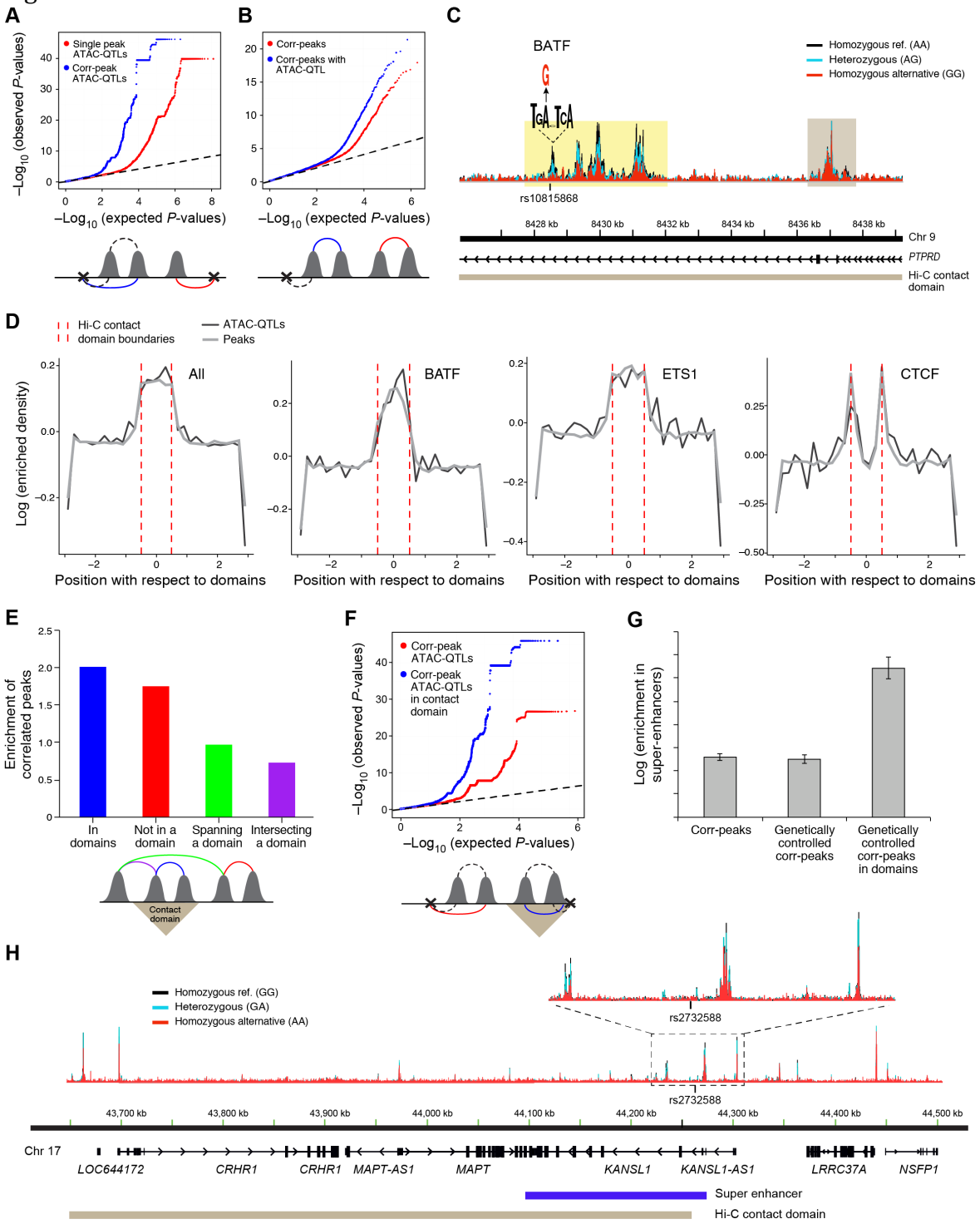
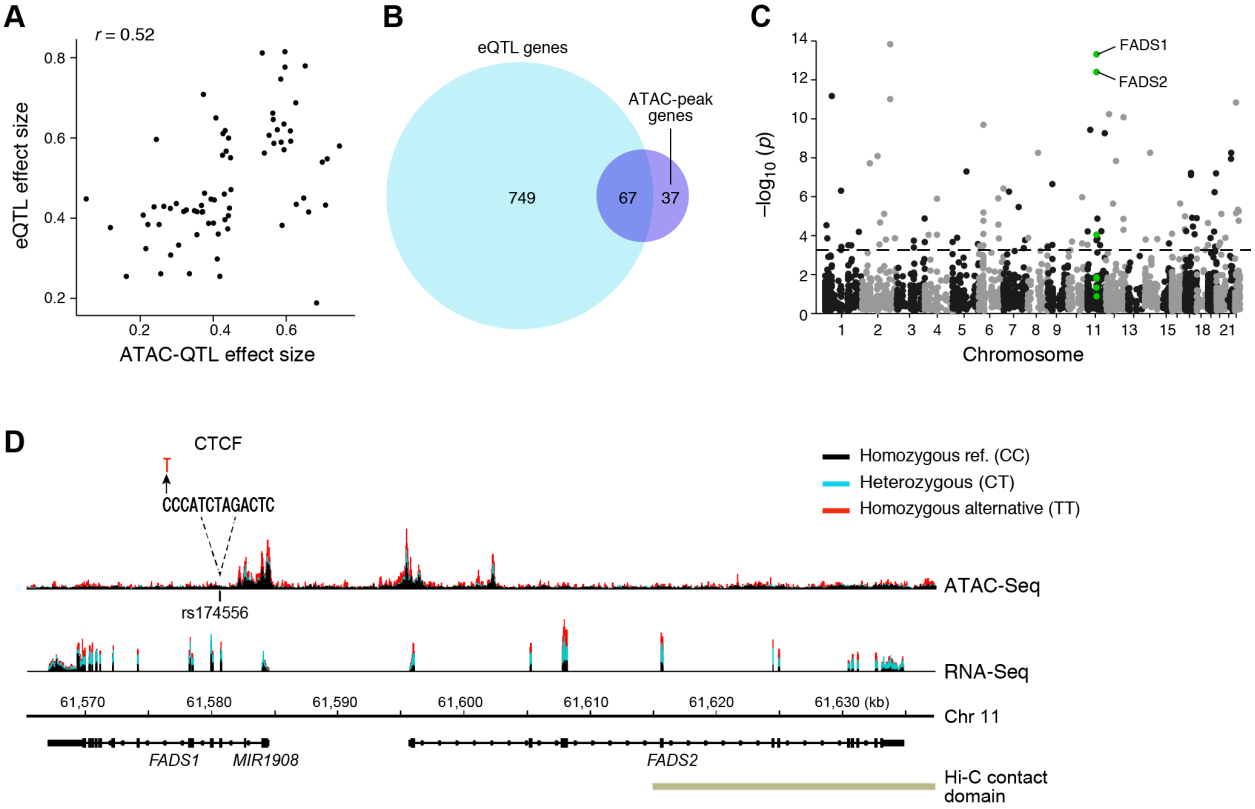

A
Single peak ATAC-QTLs
Corr-peak ATAC-QTLs
–Log10 (observed P-values)
–Log10 (expected P-values)
B
Corr-peaks
Corr-peaks with ATAC-QTL
–Log10 (expected P-values)
C
BATF
Homozygous ref. (AA)
Heterozygous (AG)
Homozygous alternative (GG)
rs10815868
8428 kb
8430 kb
8432 kb
8434 kb
8436 kb
8438 kb
Chr 9
PTPRD
Hi-C contact domain
D
Hi-C contact domain boundaries
ATAC-QTLs
Peaks
All
BATF
ETS1
CTCF
Log (enriched density)
Position with respect to domains
E
Enrichment of correlated peaks
In domains
Not in a domain
Spanning a domain
Intersecting a domain
Contact domain
F
Corr-peak ATAC-QTLs
Corr-peak ATAC-QTLs in contact domain
–Log10 (observed P-values)
–Log10 (expected P-values)
G
Log (enrichment in super-enhancers)
Corr-peaks
Genetically controlled corr-peaks
Genetically controlled corr-peaks in domains
H
Homozygous ref. (GG)
Heterozygous (GA)
Homozygous alternative (AA)
rs2732588
43,700 kb
43,800 kb
43,900 kb
44,000 kb
44,100 kb
44,200 kb
44,300 kb
44,400 kb
44,500 kb
Chr 17
LOC644172
CRHR1
CRHR1
MAPT-AS1
MAPT
KANSL1
KANSL1-AS1
LRRC37A
NSFP1
Super enhancer
Hi-C contact domain

Figure 4


A
r = 0.52
eQTL effect size
ATAC-QTL effect size
B
eQTL genes
ATAC-peak genes
749
67
37
C
FADS1
FADS2
−log10 (p)
Chromosome
D
CTCF
CCCATCTAGACTC
rs174556
Homozygous ref. (CC)
Heterozygous (CT)
Homozygous alternative (TT)
ATAC-Seq
RNA-Seq
61,570
61,580
61,590
61,600
61,610
61,620
61,630 (kb)
Chr 11
FADS1
MIR1908
FADS2
Hi-C contact domain