Genetic determinants of co-accessible chromatin regions in T cell activation across humans

Christine S. Cheng[1,15*], 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

[15]Present address: Department of Biology, Boston University, Boston, MA 02215, USA

[*] These authors contributed equally to this work.

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

Abstract

Over 90% 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. In one mechanism, disease-causing variants directly affect the activity of one or more *cis*-regulatory elements in specific cell types leading to dysregulation of gene expression. To identify such cases, we collected Assay for Transposase-Accessible Chromatin (ATAC-seq) profiles from activated primary $CD4^+$ T cells of 105 healthy donors and analyzed them to characterize the inter-individual variability in "ATAC-peaks" of chromatin accessibility and to identify its genetic basis. Interestingly, we found that ATAC-peaks are co-accessible across loci at kilobase and megabase scales, in patterns consistent with 3D chromosome organization as measured by *in situ*-Hi-C in the same cells. Genetic variants associated with ATAC-peaks (ATAC-QTLs) are widespread and those associated with correlated peaks impart the strongest genetic effects. ATAC-QTLs disrupt binding sites for transcription factors important for $CD4^+$ T cell differentiation and activation, overlap and mediate expression QTLs from the same cells, and are enriched for autoimmune disease variants. ATAC-QTLs associated with co-accessible peaks are further enriched in the same chromatin contact domains as the associated peaks and regions of the genome annotated as super enhancers. Accessibility of regulatory elements varies in correlated manner between individuals, and is determined by genetic variation following patterns that reflects the 3D organization of the genome, and mediate genetic effects on gene expression. Our results provide insights into how genetic variants modulate *cis*-regulatory elements, in isolation or in concert, and influence gene expression in primary immune cells that play a key role in many human diseases.

Introduction

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]. Quantitative trait loci (QTL) studies that map genetic variants associated with molecular traits have provided a framework for assessing the gene regulatory effects 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 immune cells [9,10], suggesting that variation in transcriptional regulation in pathogenic cell types are a key driver of human disease. However, because of linkage disequilibrium and the complex context-specific regulation of gene expression by transcription factors (TFs) and the corresponding *cis*-regulatory elements [11,12], it remains difficult to pinpoint the causal genetic variants and to determine the mechanistic basis by which they influence gene expression and disease.

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 mostly 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. A very recent multi-center study associated variability in DNA methylation, histone tail modifications and expression changes with genetic variants in 3 different naïve primary immune cell types (Neutrophils, monocytes and CD4+ CD45RA+T cells)[22]. Although this study provided useful resource on human immune cells, many disease states are associated with activated immune cells rather than the non-proliferative 'quiet' naïve immune cells, highlighting the importance of characterizing cells that are physiologically relevant to disease states.

T cell homeostasis and activation have recently been associated with various disease states including autoimmune diseases [23,24], cancer [25,26] and infectious diseases [27], and genetic variants associated with distinct conditions have been mapped to regulatory regions controlling genes important in different T cell subsets [10]. Here, we performed ATAC-seq on activated $CD4^{\pm}$ T cells from 105 healthy individuals to characterize the extent of natural variability in chromatin organization, identify its genetic basis, and assess its influence on gene expression. Our analysis highlights the co-variability between chromatin features and leverages it to identify co-accessibility relations between multiple *cis*-regulatory elements and to relate those to 3D genome organization. Our work lays the foundation for the critical tasks of mapping the complex gene regulatory relationships between *cis*-regulatory elements in primary human T cells and characterizing how genetic variation contribute to the gene regulatory variability between individual humans.

Results

Changes in T cell chromatin organization in response to activation

We used ATAC-seq [21] to assay $CD4^+$ T cells in two different states: either unstimulated (Th), or stimulated *in vitro* using tetrameric antibodies against CD3 and CD28 for 48 hours (Th_{stim}) (**Fig. 1a**). Aligned reads from six samples (five donors, one pair of replicates) were pooled (**Methods**) for each state yielding a total of 209 million reads for Th_{stim} and 58 million for Th cells. There was a global increase in chromatin accessibility in response to stimulation, with 52,154 chromatin accessibility peaks (ATAC-peaks) in Th_{stim} and 36,486 in Th cells. Downsampling each Th_{stim} sample to the same number of reads as the matching Th sample yielded similar results (24,665 peaks Th_{stim} *vs.* 17,313 Th) suggesting the increased accessibility is not due to differences in depth of sequencing. Of the 63,763 ATAC-peaks identified in at least one state, 27,446 are equally accessible between cell types (shared peaks), 28,017 are more accessible in Th_{stim} cells (FDR, $q < 0.05$), and only 8,298 ATAC-peaks are more accessible in Th cells (FDR, $q < 0.05$) (**Fig. 1b** and **Supplementary Table 1**).

The detected ATAC-peaks were associated with distinctive genomic features and enriched for single nucleotide polymorphisms (SNPs) associated with autoimmune diseases. Specifically, compared to Th-specific ATAC-peaks, Th_{stim}-specific peaks have a higher overlap with enhancers active in conventional T helper cells (T_{conv}, a class that includes Th1 and Th17 cells) [16] and a lower overlap with enhancers active in regulatory (T_{reg}) and naïve Th cells. They also have a higher overlap with enhancers active in Th0 cells (αCD3/αCD28 activated Th cells) and PMA-stimulated T cells, consistent with the polarization-independent activation of our stimulation

protocol (**Fig. 1c**). We identified 9,724 Th_{stim}-specific peaks located in non-coding regions previously unannotated by H3k27Ac [16]. Furthermore, Th_{stim}-specific and shared peaks are more enriched for SNPs associated with autoimmune diseases, most notably inflammatory bowel disease (IBD) and rheumatoid arthritis, than Th-specific peaks (**Fig. 1d**).

Analyzing ATAC-peaks in aggregate provide estimates of transcription factor (TF) binding profiles at single nucleotide resolution [21], highlighting key T cell regulators. Th_{stim}-specific peaks are enriched for genomic locations bound by TFs important for Th cell activation or differentiation, including members of the AP-1 super family (*e.g.*, BATF) and interferon response factors (IRFs) [28-30] (**Fig. 1e**). ATAC peaks overlapping both a BATF binding site and the interferon stimulation response element (ISRE) reveal distinct binding footprints in Th_{stim} compared to Th cells (**Fig. 1f**). Conversely, shared peaks are enriched for TFs (*e.g.*, CTCF) known to maintain chromatin state independent of cell type and state [28-30] (**Fig. 1e**), and the footprints estimated from ATAC-peaks overlapping CTCF binding sites do not exhibit condition specific accessibility profiles (**Fig. 1f**). ETS binding sites overlapping shared and condition-specific ATAC-peaks are distinct: shared peaks overlapping ETS binding sites highlight the canonical ETS1 motif (5'-CACTTCCTGT-3'), whereas footprints and *de-novo* TFBS prediction recover a 3' extended motif (5'-CACTTCCTGTCA-3') in Th-specific peaks and a T/G → T (5'-CACTTCC**T**GT-3') at the eighth position in Th_{stim} peaks (**Fig 1g**) [31]. Th-specific ETS1-peaks overlap ETS/RUNX binding sites more than shared or Th_{stim}-specific peaks (OR = 2.7 and 3.9; Fisher's exact test, $P = 2.2x10^{-16}$ and $P = 2.2x10^{-16}$, respectively) (**Fig 1h**), consistent with previous reports that ETS/RUNX binding is specific to Treg enhancers [32,33]. These results

demonstrate the power of ATAC-seq to identify known and novel *cis*-regulatory elements and generate high-resolution TF footprints.

Interindividual variation reveals global and local co-accessibility patterns

Because Th_{stim} ATAC-peaks are more abundant, better overlap with autoimmune disease loci, and are enriched for binding sites for known TFs, we focused on characterizing the interindividual variability in chromatin accessibility in Th_{stim} cells. Specifically, we used an optimized ATAC-seq protocol (**Methods**, **Supplementary Fig. 1**) to profile activated primary $CD4^+$ T cells isolated from 105 healthy donors of European descent in the ImmVar Consortium[10] (**Fig. 2a**). Per sample, we obtained a median of 37 million (MAD +/-13 million) reads from highly complex libraries (on average 84% usable nuclear reads, as opposed to 40% prior to optimization) (**Supplementary Fig. 2**), with low mitochrondrial DNA (mtDNA) contamination (on average contamination < 3%, as opposed to 53% prior to optimization).

Leveraging the variability across 105 individuals, we found strong patterns of co-accessibility at both a global and local level. Globally, we calculated co-accessibility between pairs of 1Mb chromatin accessibility bins (**Fig. 2b**) across the individuals. For every chromosome, we observed significant co-accessibility between regions spanning 42 Mbs on average (FDR < 0.1) (Chr1: **Fig. 2c**, Other chromosomes: **Supplementary Fig. 3**). Locally, we calculated co-accessibility between pairs of ATAC-peaks across the individuals. We found 5,404 pairs of co-accessible ATAC-peaks within 1.5Mb of each other (linear regression, FDR < 0.05,

Supplementary Fig. 4 and 5, Supplementary Table 3 and 4), corresponding to 2,722/52,154 (5.2%) of the distinct ATAC-peaks detected. On average, co-accessible peaks were located 313 kb apart (**Fig. 2d**). Co-accessible peaks are enriched for GWAS SNPs associated with autoimmune diseases (**Fig. 2e**) and more enriched for *BATF*, *ETS1*, and *BATF/IRF* motifs (1.2x, $P < 1x10^{-5}$, hypergeometric test), TSS ($P = 2.3x10^{-195}$, hypergeometric test) and 5' genomic regions ($P < 6.7x10^{-3}$, hypergeometric test) than all Th_{stim} peaks (**Fig. 2f**). They are also enriched in $T_{naïve}$, Th_{stim}, and Th_{17} enhancer subtypes compared to all Th_{stim} peaks ($P < 1.98x10^{-14}$, hypergeometric test) (**Fig. 2g**). The large average distance between co-accessible peaks as well as the enrichment for automimmune associated GWAS SNPs, TFBSs, genomic regulatory regions, and T cell subtype enhancers suggest that co-accessible peaks are unlikely the result of local biases in sequencing and identify important correlated regulated regions.

The observed pattern of co-accessibility is influenced by the 3D conformation of the chromatin, as determined by a loop-resolution *in situ* Hi-C[34] in primary $CD4^+$ T cells activated for 48 hours (**Supplementary Table 2**). Globally, at 1 Mb resolution, the correlation of interaction frequencies across the genome estimated from Hi-C are qualitatively similar and quantitatively correlated (R = 0.36) to co-accessibility patterns estimated from ATAC-seq (**Fig. 2c** and **h**), consistent with previous estimates across single cells[21]. There is also a relationship between locally correlated peaks and the high-resolution 3D structure of the genome. Pairwise correlation of ATAC-peaks is more significant when filtered for regions of Hi-C interaction at 250kb resolution and highly correlated with the Hi-C interaction frequencies (Spearman $\rho = 0.19$) (**Supplementary Fig. 6**).

Local ATAC-QTLs disrupt *cis*-regulatory functions in Th cell enhancers

To define the genetic basis of inter-individual variability in chromatin accessibility and co-accessibility, we compared our ATAC-seq data to genetic variation across the 105 individuals. We found 1,790 ATAC-peaks associated with at least one significant local (+/- 20kb) SNP (RASQUAL[35], $P < 3.02\times10^{-4}$, permutation FDR < 0.1) (**Fig. 3a, Supplementary Fig. 5b**, and **Supplementary Table 4** and **5**). 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, 580 are significantly heritable (average heritability 60%, GCTA FDR < 0.1), with a large proportion of the heritability (average 36%) predicted by the best lead SNP (**Fig. 3b, Methods,** and **Supplementary Table 6**). There are 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].

We found several lines of evidence supporting a model where local ATAC-QTLs ('local' in the genetic sense, where associated SNP is within +/- 20kb of the ATAC-peak) disrupt *cis*-regulatory functions in Th cell enhancers. First, of the 1,790 local ATAC-QTL-peaks, 33% (589) of the lead associated SNPs are located within 2 kb of the ATAC-peak and 18% (327) are located within the

ATAC-peak proper (**Fig. 3c**), suggesting that the direct disruption of *cis*-regulatory elements is an important determinant of the observed variation in accessibility in those cases. Second, local ATAC-QTLs-peaks are more enriched near transcription start sites (TSS) than transcription termination sites (TTS) of the closest gene, further supporting a transcriptional role in *cis* (**Fig. 3d**). Third, 77% of local ATAC-QTL-peaks are in intronic or intergenic regions ($P < 3.02x10^{-4,}$ hypergeometic test) (**Supplementary Fig. 7a**); of these, 70% lie in regions that are previously identified as enhancers for different Th cell subtypes ($P < 3.12x10^{-50,}$ hypergeometic test) (**Supplementary Fig. 7b**). Fourth, ATAC-QTL-peaks are more enriched for motifs bound by TFs involved in T cell development and activation (*e.g.* BATF, AP1 and IRF, **Fig. 3e**) than all ATAC-peaks detected in activated cells. In fact, 57% of ATAC-QTL-peaks contained either a BATF or an ETS1 motif (1.2x 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 (1.3x enrichment, $P < 4.02\times10^{-5}$, hypergeometric test), suggesting that the perturbation of binding sites for key TFs is a major driver for the observed variation in chromatin accessibility across individuals. Indeed, almost half (48%) of the ATAC-QTL lead SNPs strongly disrupt one of six predicted TF binding sites (TFBSs) (**Fig. 3f)**, including known transcription factors that act in T cell activation, such as BATF, IRF, RUNX1 and ETS. Furthermore, ATAC-QTL-peaks overlapping BATF, ETS1 and CTCF binding sites show differential accessibility between genotypes at single nucleotide resolution, with the core motif exhibiting the most striking difference in accessibility (**Fig. 3g** and **Supplementary Fig. S8**). An extended 1 kb window exhibited weaker but still significant differences, reflecting long range *cis* effects on chromatin accessibility (**Fig. 3g**). Consistent with the footprinting analysis, the effect sizes of ATAC-QTLs

are correlated (ρ=0.648) with SNP motif disruption scores obtained by deltaSVM[36], an unbiased analysis to discover *de novo cis*-regulatory elements in ATAC-peaks (**Fig. 3h, Methods**).

Both ATAC-QTLs and ATAC-peaks exhibit a distinct pattern of accessibility with respect to chromatin 3D structure. ATAC-QTLs and ATAC-peaks overlapping BATF and ETS1 motifs are enriched within Hi-C contact domains, whereas those overlapping CTCF motifs are enriched at the contact domain boundaries (**Fig. 3i**). These results are consistent with previous reports of CTCF enrichment at loop anchors and at contact domain boundaries[34,37-39].

Local ATAC-QTLs are enriched for GWAS SNPs from autoimmune diseases (**Fig. 3j**), providing a functional context for interpreting 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 transcription factor involved in the TGF-β signaling pathway that regulates T cell activation and metabolism[40]. 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. 3k**). This suggests that rs17293632 may increase susceptibility to Crohn's disease and IBD by disrupting BATF binding at the *SMAD3* locus.

Genetic determinants of chromatin co-accessibility

We next tested the hypothesis that the function of multiple *cis*-regulatory elements could be simultaneously modulated by a single SNP. Specifically, we found that both local and moderately distal ATAC-QTLs (< 1 Mb) are more strongly associated with co-accessible peaks than with single peaks (defined as any ATAC-peak that is not part of a co-accessible peak pair; **Fig. 4a**). Co-accessible peaks associated with ATAC-QTLs are more strongly correlated with each other than co-accessible peaks not associated with an ATAC-QTL (**Fig. 4b** and **Supplementary Table 3**). Moreover, co-accessible ATAC-peaks exhibited stronger genetic associations if the peaks and the associated variant reside in the same HiC contact domain than if they did not (**Fig. 4c** and **Supplementary Fig. 9**). For example, rs10815868 is a non-coding variant that resides in the 18th intron of *PTPRD*, a tumor suppressor gene, where it disrupts a consensus BATF binding site at a conserved position. This SNP is associated with decreased chromatin accessibility in a 4 kb region that contains four highly correlated ATAC-peaks (**Fig. 4d**, yellow box). Notably, an adjacent peak located 5 kb upstream of this region was not affected by this variant (**Fig. 4d,** grey box), suggesting that the genetic control of multiple ATAC-peaks was limited to a defined regulatory region. These results suggest that ATAC-QTLs associated with multiple co-accessible *cis*-regulatory elements impart a stronger effect than those associated with a single peak, and are limited by 3D chromatin structure.

Super-enhancers, also called stretch enhancers, are defined as large clusters of contiguous enhancers, and are often bound by master regulators and mediator complexes to drive the transcription of genes involved in cell type specificity[41,42] (**Fig. 4e**). Interestingly, co-accessible peaks – irrespective of their association with an ATAC-QTL – were enriched in super-enhancer

regions previously identified in $CD4^+$ Th cells[41] (**Fig. 4f**, **Supplementary Table 3, Methods**). Co-accessible peaks with an associated ATAC-QTL were 1.2x more enriched in super-enhancers if they reside in the same contact domain (**Fig. 4f**). 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 co-accessible ATAC-peaks (**Fig. 4g**). The affected region partially overlaps a previously identified super-enhancer in $CD4^+$ T cells[41]. 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[43,44] (**Fig. 4g**). The region containing the super-enhancer and these co-accessible peaks was mostly contained within a Hi-C contact domain, although it also extends immediately outside of the Hi-C contact domain boundary (**Fig. 4g**). The high enrichment of super enhancers intersected with co-accessible peaks further supports the hypothesis that we are identifying important regulatory regions which are regulated by both the genetic variation and 3D chromatin structure.

Linking chromatin accessibility to gene expression

We next assessed how variability in chromatin accessibility, including the co-accessibility of multiple *cis*-regulatory elements, could influence gene expression. We measured RNA-seq profiles from Th_{stim} cells from 96 donors (93 with matching ATAC-seq data), and identified 33 genes significantly correlated to at least one ATAC-peak locally (FDR < 0.05). Because of sample size and noise of both ATAC-seq and RNA-seq assays, we further filtered ATAC-peaks

to those with QTLs and found enrichment of significant p-values suggesting that genetic variants impart stronger correlation between peaks and genes (**Fig. 5a**).

In order to assess the sharing of genetic variants between chromatin and expression traits, we mapped genetic variants that affect gene expression (eQTLs). We identified 1,256 genes with at least one significant local eQTL (+/- 500 kb centered around the gene) (RASQUAL, $P < 4.04\times10^{-5}$, permutation FDR < 0.05**, Supplementary Fig. 10, Supplementary Table 4 and 7**), termed eQTL-genes. Among these, 102 lead-eQTLs are also lead-ATAC-QTLs with correlated effect sizes. The majority of these genetic variants (71 out of 102) have effect sizes in the same direction (Spearman rho = 0.73) indicative of activator effects, while 31 have effect sizes in the opposite direction indicative of repressor effects (Spearman $\rho = -0.73$) (**Fig. 5b** and **Supplementary Table 8**). To overcome winner's curse and reduce multiple testing burden, we further filtered genetic variants to 1,790 significant ATAC-QTLs (FDR < 0.05) and found 168 locally-associated genes (+/- 500 kb of the corresponding ATAC-peak, FDR < 0.05, **Fig. 5c,d,** and **Supplementary Table 9**). There was an enrichment for significant associations between gene expression and co-accessibility ATAC-QTLs (Fisher exact test p-value = $2.15\text{x}10^{-8}$, **Fig. 5d**), suggesting that genetic variants that impact accessibility across multiple *cis*-regulatory elements are more likely to impact gene expression.

We next assessed the causal relationship between genetic variants, chromatin accessibility and expression variability. The observed p-values of association between gene expression and genetic variants in *cis* are similarly distributed independent of conditioning on each most

correlated ATAC-peak (**Fig. 5e** and **Supplementary Table 10**). On the other hand, the observed p-values of correlation between gene expression and chromatin accessibility are less significant after conditioning on each best-associated eQTL (**Fig. 5e**). For example, before conditioning on rs174556, *FADS2* expression and chromatin accessibility at chr11:61595257-61595730 are correlated (P-value = 9.9×10^{-11}) and after conditioning the correlation is far less significant (P-value = 0.094) (**Fig. 5f**). These results suggest that the effect of genetic variants on gene expression are mediated by ATAC-peaks. Notably, rs174556 is an ATAC-QTL associated with a pair of co-accessible peaks, resides in a 25 kb region between two Hi-C contact domains, where the alternative allele disrupts a CTCF binding site (**Fig. 5g**). 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*, and rs174556 is also identified as an eQTL for both *FADS1* and *FADS2* in our T cell dataset (**Fig. 5g**). *FADS1* and *FADS2*, two fatty acid desaturases (FADS), regulate inflammation, promote cancer development, 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 thereby increase disease risk.

Discussion

Here, we integrated genetic variation and ATAC-seq data from primary activated $CD4^+$ T cells from 105 healthy donors to identify multiple *cis*-regulatory elements, which we characterized through variability in chromatin accessibility (co-accessible peaks), genetic variation, and genomic structure. We found that ATAC-seq profiles identify important regulatory regions and are reflective of the 3D structure of the genome. ATAC-peaks and correlated peaks are enriched for TF binding motifs, which allowed for the identification of context specific differential ETS1 motifs. ATAC-QTLs are heritable, fall close to their peak proper, and impact TF binding. We found widespread genetic control of co-accessible peaks, in a manner consistent with the 3D organization of the genome. Thus, there are regions of the chromatin that are co-accessible, which means that accessibility at one enhancer element is affected by genetic variation at another. From a molecular standpoint, one might hypothesize cooperative synergy between interacting enhancers, positive feedback reinforcing the activity of an enhancer's partners. Alternatively, genetic variant could potentially affect the composition of different subpopulations of CD4+ T cells such as effector Th cells, regulatory Treg cells and natural killer T cells, which is also a biologically interesting phenomenon. With the recent advancement of single cell resolution epigenomic[50] and transcriptomic[51,52] analysis, it will enable us to detect ATAC-QTLs and eQTLs in each subpoulation of cells from a heterogeneous 'cell cloud'. 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 dynamic, cooperative multiple *cis*-regulatory elements and 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.

Materials and Methods

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 **Supplementary Table 4**. 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 (**Supplementary Table 4**). 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 (**Supplementary Table 4**). 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 [53], 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) [54]. For each sample, mitochondrial reads and multiply-mapped reads were filtered out using BEDtools (function intersectBed) [55]. After filtering, we had an median of 37 million (MAD +/- 13 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) [56], 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 [57] and Epigenomics Roadmap [12] 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. 3e,** 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.

Transcription factor footprinting

Using the Homer suite tool annotatePeaks –m and –mbed options, we found all instances of BATF. ISRE, BATF/IRF, ETS1, and CTCF motifs in shared, differentially accessible stimulated and unstimlated ATAC-peaks. Next, we determined the per-basepair coverage +/- 1 kb around the center of the motif, only using cut-site reads and splitting the reads into those that are on the motif strand or on the opposite strand. Final TF footprints were derived from median normalized reads[57].

Outlier analysis and sample mix-up analysis

We kept samples there were highly correlated for downstream analyses (Pearson r > 0.68, SOM), where we developed an optimized ATAC-seq protocol (**SOM**) that achieved high technical and biological reproducibility (**fig. S1**), highly complex libraries (on average 84% usable nuclear reads, as opposed to 40% prior to optimization) (**fig. S2**), and low mitochrondrial DNA (mtDNA) contamination (on average contamination < 3%, as opposed to 53% prior to optimization). Quality of all ATAC-seq samples were assessed again, only keeping samples that contain a minimum of 8 million QC-passed reads (median of 37 million, MAD +/-13 million) and high inter-sample correlation (Pearson r > 0.68, SOM). ATAC-Seq profiles from the 105 individuals were further filtered for samples who had that were also predicted with < 0.93 identity by descent to identify multiple *cis*-regulatory elements [58]. To identify sample mix-ups, we used the software VerifyBamID [58], where 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 [35] 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 included 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 [59] 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 [54] 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.

Co-accessible peak analysis

To identify co-accessible 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[60]. 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 adjusted for covariates as described above and three principal components. Then, we identified 851/1,762 co-accessible peaks (387 unique ATAC-peaks) with ATAC-QTLs (FDR < 0.05), of which 159 co-accessible peaks (93 unique ATAC-peaks) with ATAC-QTLs reside in contact domains. To ensure that the co-accessible peaks were enriched in contact domains, we permuted the position of the contact

domains while keeping the length of the contact domains constant and performed the same analysis (**Supplementary Fig. 11**).

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[61]. 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 expected counts for filtered genes into RASQUAL[35], 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[35] 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 included as covariates. The RASQUAL "-r" option was used for permutations to determine the empirical null distribution and compute a FDR, ultimately retaining a total of 816 eQTLs at FDR<0.05.

GWAS enrichment

The GREGOR suite [62] 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 [63] 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.[41]). 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[64,65] 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[36]. 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. [66]. Restricted maximum likelihood heritability (h^2) estimates were calculated using GCTA software [67] 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[66]. 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.

Relating gene expression and chromatin accessibility

To analyze the relationship between gene expression, chromatin accessibility, and genetic variation, we performed conditioning analysis in a following way. Gene expression residualized for its 16 first principal components (PCs) and biometric data was correlated to local genetic variations (SNPs within 500 kb from TSS); statistics of the highest association and its residual were recorded, and the residual was further correlated to the local peaks residualized for their first four PCs and biometric data to account for population structure and other factors (if not indicated otherwise); highest association of the residual to the peak intensities was reported. Similarly, gene expression residualized for its first 16 PCs and biometric data was correlated to local peak intensities and the statistics of highest association and its residuals were recorded; the residual was further correlated to local genotypes and the highest association was reported in the Q-Q plot.

Figure legends

Figure 1. Chromatin dynamics in human T cell activation. (**a**) ATAC-seq experimental overview. (**b**) Differential ATAC-peaks. Shown are ATAC-peaks (columns) in six individuals (rows) before (top, Th specific) and 48hr after (bottom, Th_{stim} specific) 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[12] in $CD4^+$ T cells pooled samples from either Th specific cells, Th_{stim} specific cells or shared. (**d**) 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 present in only Th specific (left), only Th_{stim} specific cells (middle), or are shared (right). (**e**) 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 present in only Th specific cells (left), only Th_{stim} specific cells (middle), or are shared (right). (**f-h**) TF footprinting. Shown are for each TF motif (as identified in ENCODE[62], indicated on top), aggregated plots of mean chromatin accessibility (y axis) in Th specific (purple) or Th_{stim} specific (red) along TF binding site (x axis; log(bp from center of the TF motif). (f) BATF, ISRE, and BATF/IRF motifs in stimulated-specific peaks (three left panels) and CTCF in shared peaks (right panel). (**g**) ETS1 binding sites in Th specific (left) and Th_{stim} specific ATAC-peaks (right). (**h**) ETS1/RUNX combination TF binding sites in Th specific peaks. (**i**) Fraction of the number of Th specific, Th_{stim} specific, and shared peaks with an ETS1 binding sites also containing a BATF TF binding sites (blue) or an ETS1/RUNX binding site (green).

Figure 2. Interindividual chromatin co-accessibility is constrained by chromosome architecture. (**a**) Overview. (**b**) Schematic of co-accessible regions across individuals. (**c**) Inter-individual co-variation of chromatin accessibility. Heat map shows the pair-wise Pearson correlation coeffcient (colorbar) in chromatin accessibility across 105 ATAC-seq profiles in Th_{stim} specific peaks binned into 1 Mb windows for Chr 1 (rows, columns). (**d**) Correlation of Hi-C interactions at 1 Mb resolution for Chr 1. (**e**) Histogram shows the distribution (density, Y axis) of the distances (X axis) between co-accessible peaks (black) compared to co-accessible computed on permuted ATAC-peaks (grey). (**f**) Co-accessible peaks are enriched for GWAS variants. Shown are the enrichments (X axis) in co-accessible peaks and their associated significance (Y axis) for loci associated with the indicated disease or phenotype. (**g**) Enrichment (Y axis) of genomic annotations (X axis) overlapping co-accessible peaks compared to all Th_{stim} peaks. (**h**) Enrichment (Y axis) of T cell enhancer annotations (Y axis) overlapping correlated peaks compared to all Th_{stim} peaks.

Figure 3. 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**) Unsupervised analysis associates ATAC-QTLs with key TF binding sites show key motifs. Shown are the motifs for 6 TFs associated with most of the large gkmSVM weights, and the proportion of the overall disruption (%, bottom) explained by ATAC-QTLs. (**g**) ATAC-QTLs affect binding motifs in an allele specific manner. Shown are for each of three indicated TF binding site (as identified in ENCODE[62]), 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. (h) The effect sizes of lead ATAC-QTL SNPs (X axis) are well correlated with deltaSVM scores (Y axis) for

these variants. (**i**) Relation between contact domains, and stimulated ATAC-peaks, and ATAC-QTLs associated with TF binding motifs. Shown are the distributions (Y axis, density) of position (X axis) of ATAC-peaks (density histogram, grey) and ATAC-QTLs (blue) that overlap either an ETS1, CTCF, or BATF binding sites. Positions (X axis) are relative to Hi-C chromatin contact domain boundaries (dotted red lines). (**j**) 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. (**k**) 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 4. Genetic determinants of co-accessible 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**) Co-accessible peak correlations. Q-Q plots of the *P* value for the correlation of genetically controlled correlated peaks (blue) and all correlated peaks (red). (**c**) Contact domains impact genetically controlled co-accessible peaks. Q-Q plots of the *P* value for the association of SNPs to either peak, when the co-accessible peaks are both inside a contact domain (blue) or are outside a Hi-C contact domain (red). (**d**) Example of genetically controlled co-accessible 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). (**e**) Schematic of co-accessible peaks in super enhancers. (**f**) 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). (**g**) 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 5. Association of chromatin accessibility and gene expression. (**a**) QQ-plot of RNA-seq and ATAC-peak associations distinguishing cases where the ATAC-peak is an ATAC-QTL-peak (red) or not (blue). (**b**) Correlation of effect sizes between ATAC-QTLs (X axis) and eQTLs (Y axis). (**c**) Manhattan plot of shared eQTL and ATAC-QTLs; negative log10-pvalue of eQTL association to SNPs that are ATAC-QTLs (FDR<0.05) is shown on y-axis; significant associations of gene expression with single ATAC-QTLs are highlighted in red and to multipeak ATAC-QTLs are highlighted in blue. (**d**) QQ-plot of gene expression associations to the SNPs that are lead associations to ATAC-peaks; multipeak ATAC-QTLs at FDR < 0.05 (blue), single ATAC-QTL-peaks at FDR < 0.05 (red), and insignificant lead associations at FDR > 0.05 (grey). (**e**) Conditioning analysis of association between gene expression, chromatin state and genotypic variation. QQ-plot representing p-values of the best per-gene association of gene expression to genotypes (i.e. eQTLs, RNA ~ GT, yellow) and the p-values of the best associations of their residuals to the ATAC peaks (RNA~CA□GT, green). Similarly, p-values of the best per-gene associations of gene expression to the ATAC peaks (RNA~CA, red) and p-values of association of their residual to genotypes (RNA ~ GT□CA, blue). (**f**) Scatterplot of FADS1 expression (Y axis) and chromatin accessibility at chr11:61,595,257-61,595,730 colored by genotype before (left) and after (right) conditioning on rs174556. (**g**) 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. McCarthy, M.I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-69 (2008).
2. Visscher, P.M., Brown, M.A., McCarthy, M.I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7-24 (2012).
3. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**, 95-108 (2005).
4. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
5. Stranger, B.E. *et al.* Population genomics of human gene expression. *Nat Genet* **39**, 1217-24 (2007).
6. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-11 (2013).
7. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* **24**, 14-24 (2014).
8. Raj, T. *et al.* Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519-23 (2014).
9. Lee, M.N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).
10. Ye, C.J. *et al.* Intersection of population variation and autoimmunity genetics in human T cell activation. *Science* **345**, 1254665 (2014).
11. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
12. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
13. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
14. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).
15. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).
16. Farh, K.K. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337-43 (2015).
17. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390-4 (2012).
18. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750-2 (2013).
19. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science* **342**, 747-9 (2013).
20. Kilpinen, H. *et al.* Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science* **342**, 744-7 (2013).

21. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-8 (2013).
22. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414 e24 (2016).
23. Ohashi, P.S. T-cell signalling and autoimmunity: molecular mechanisms of disease. *Nat Rev Immunol* **2**, 427-38 (2002).
24. Kronenberg, M. & Rudensky, A. Regulation of immunity by self-reactive T cells. *Nature* **435**, 598-604 (2005).
25. Speiser, D.E., Ho, P.C. & Verdeil, G. Regulatory circuits of T cell function in cancer. *Nat Rev Immunol* **16**, 599-611 (2016).
26. Restifo, N.P., Dudley, M.E. & Rosenberg, S.A. Adoptive immunotherapy for cancer: harnessing the T cell response. *Nat Rev Immunol* **12**, 269-81 (2012).
27. Belkaid, Y. & Rouse, B.T. Natural regulatory T cells in infectious disease. *Nat Immunol* **6**, 353-60 (2005).
28. Kurachi, M. *et al.* The transcription factor BATF operates as an essential differentiation checkpoint in early effector CD8+ T cells. *Nat Immunol* **15**, 373-83 (2014).
29. Li, P. *et al.* BATF-JUN is critical for IRF4-mediated transcription in T cells. *Nature* **490**, 543-6 (2012).
30. Murphy, T.L., Tussiwand, R. & Murphy, K.M. Specificity through cooperation: BATF-IRF interactions control immune-regulatory networks. *Nat Rev Immunol* **13**, 499-509 (2013).
31. http://broadinstitute.github.io/picard.
32. Samstein, R.M. *et al.* Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. *Cell* **151**, 153-66 (2012).
33. Hollenhorst, P.C. *et al.* DNA specificity determinants associate with distinct transcription factor functions. *PLoS Genet* **5**, e1000778 (2009).
34. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
35. Kumasaka, N., Knights, A.J. & Gaffney, D.J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat Genet* **48**, 206-13 (2016).
36. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**, 955-61 (2015).
37. Hou, C., Zhao, H., Tanimoto, K. & Dean, A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc Natl Acad Sci U S A* **105**, 20398-403 (2008).
38. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* **137**, 1194-211 (2009).
39. Splinter, E. *et al.* CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev* **20**, 2349-54 (2006).
40. Delisle, J.S. *et al.* The TGF-beta-Smad3 pathway inhibits CD28-dependent cell growth and proliferation of CD4 T cells. *Genes Immun* **14**, 115-26 (2013).

41. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
42. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-19 (2013).
43. Dias, J. *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 Dev* **28**, 929-42 (2014).
44. Lam, K.C. *et al.* The NSL complex regulates housekeeping genes in Drosophila. *PLoS Genet* **8**, e1002736 (2012).
45. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).
46. Fan, Y.Y. *et al.* Characterization of an arachidonic acid-deficient (Fads1 knockout) mouse model. *J Lipid Res* **53**, 1287-95 (2012).
47. Barrie, A. *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**, 268-73 (2011).
48. Sakata, D., Yao, C. & Narumiya, S. Prostaglandin E2, an immunoactivator. *J Pharmacol Sci* **112**, 1-5 (2010).
49. Stroud, C.K. *et al.* Disruption of FADS2 gene in mice impairs male reproduction and causes dermal and intestinal ulceration. *J Lipid Res* **50**, 1870-80 (2009).
50. Buenrostro, J.D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-90 (2015).
51. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-201 (2015).
52. Macosko, E.Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-14 (2015).
53. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
54. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-95 (2010).
55. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-2 (2010).
56. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
57. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976-87 (2014).
58. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-48 (2012).
59. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* **100**, 9440-5 (2003).
60. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).
61. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

62. Schmidt, E.M. *et al.* GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics* **31**, 2601-6 (2015).
63. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-89 (2010).
64. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M.A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**, e1003711 (2014).
65. Ghandi M, M.-N.M., Ghareghani N, Lee D, Garraway L, Beer MA. gkmSVM, an R package for gapped-kmer SVM. *Bioinformatics.* **Apr 19**(2016).
66. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-8 (2015).
67. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).

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. Chromatin dynamics in human T cell activation.

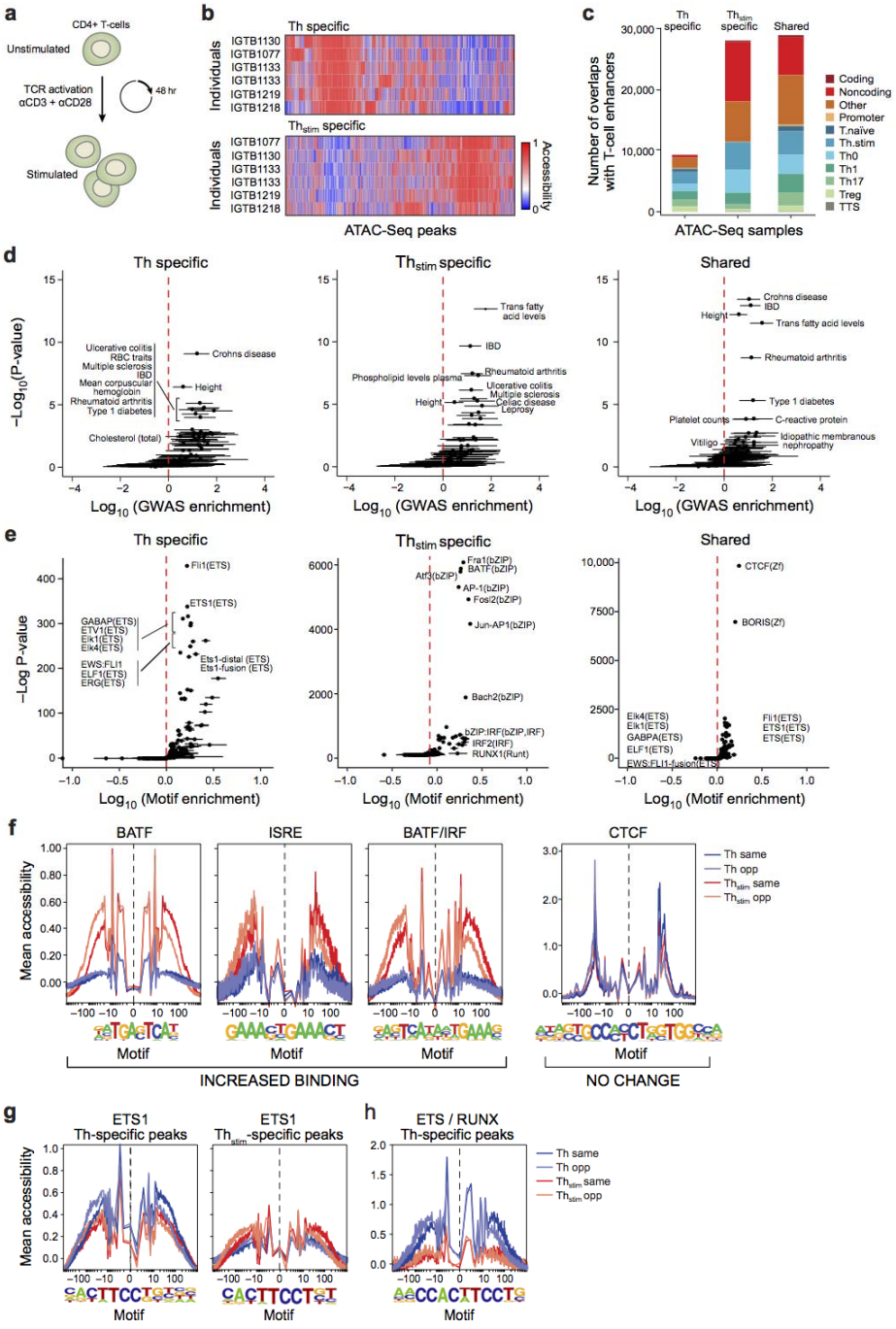

Figure 2. Interindividual chromatin co-accessibility is constrained by chromosome architecture.

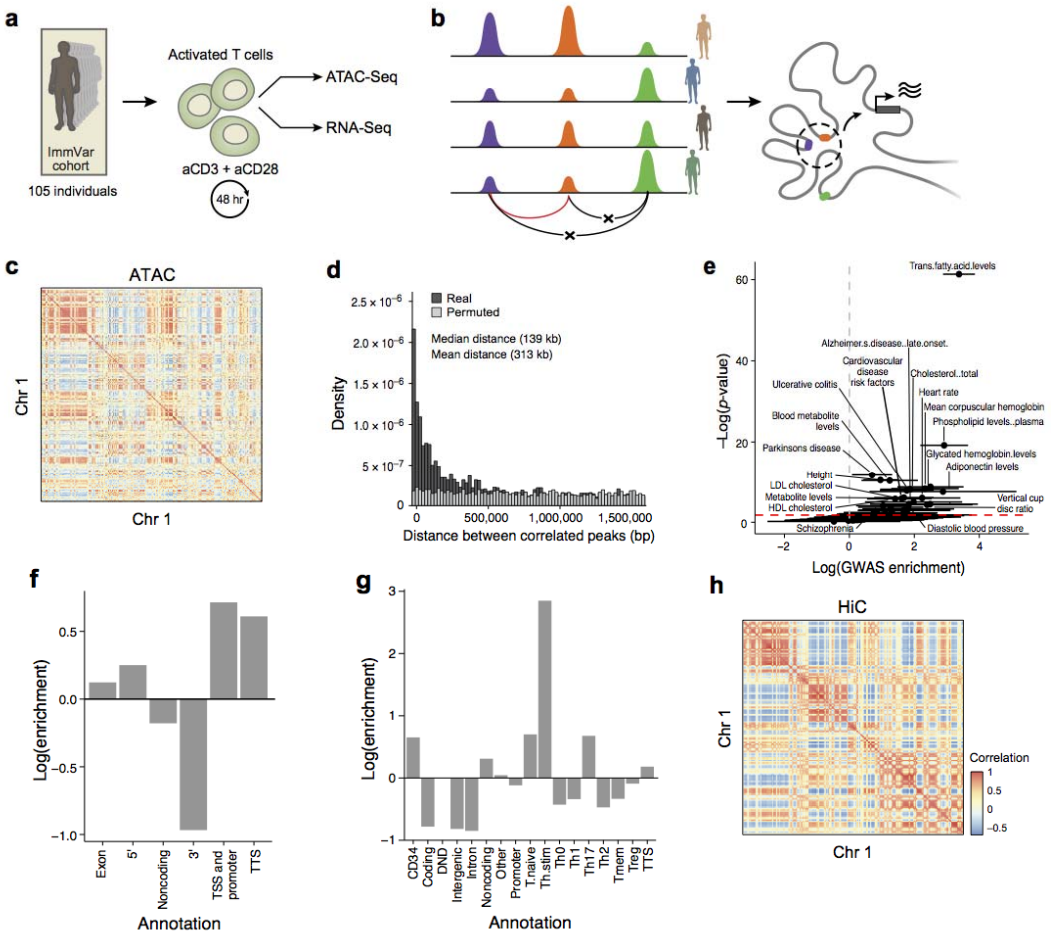

Figure 3. Genetic variants that affect chromatin states in human T cell activation.

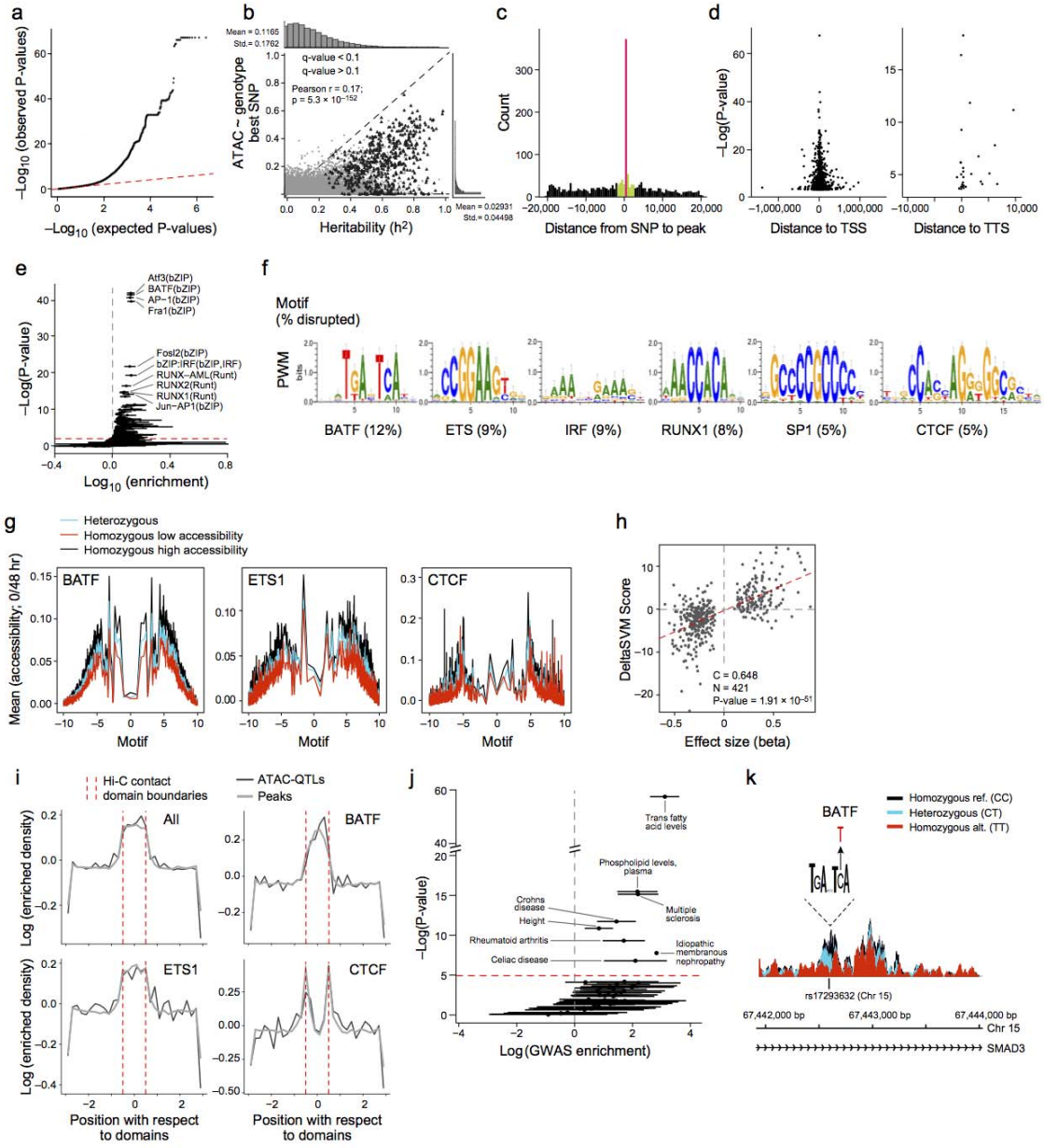

Figure 4. Genetic determinants of co-accessible peaks.

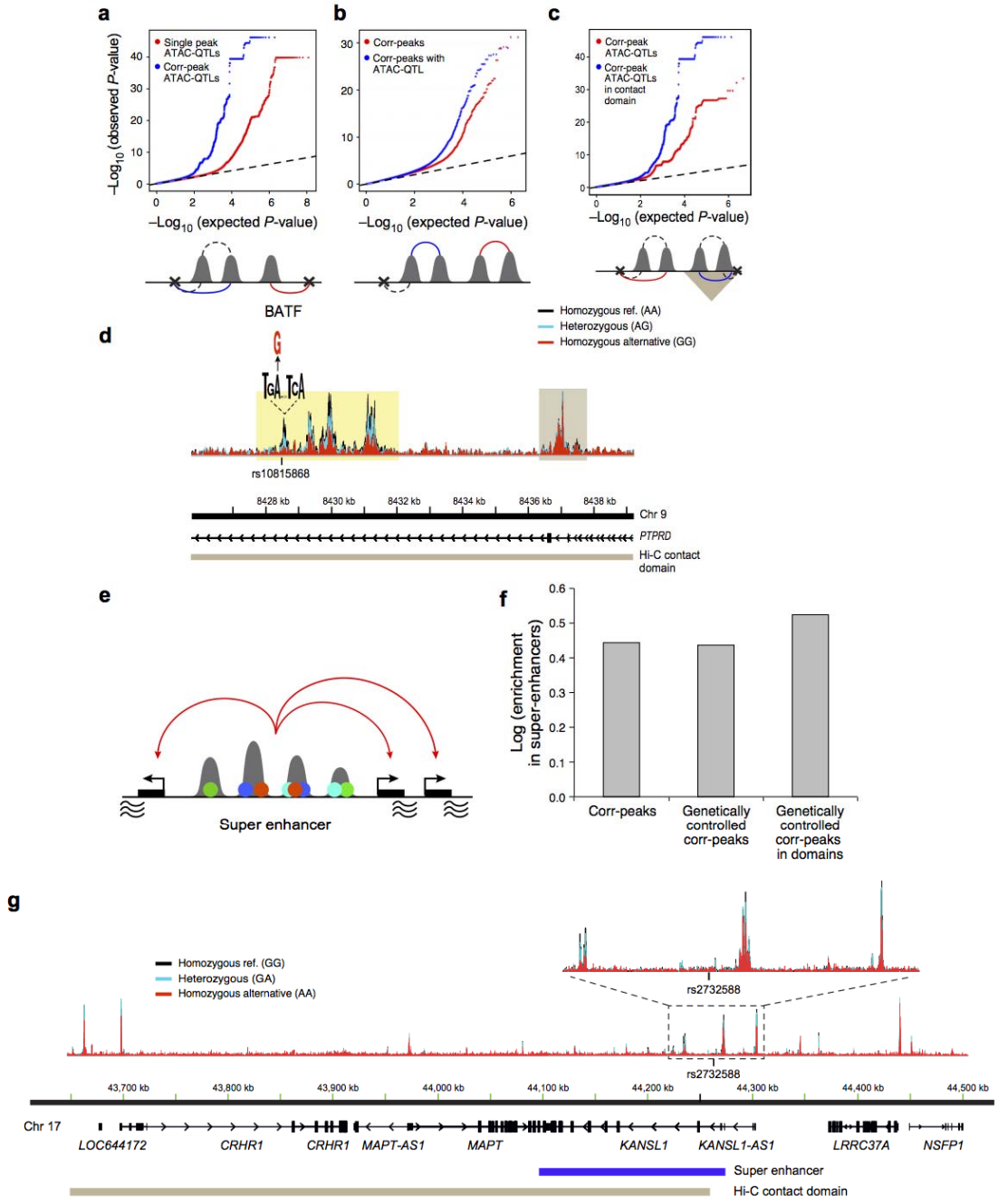

Figure 5. Association of chromatin accessibility and gene expression.