Prioritizing disease and trait causal variants at the \textit{TNFAIP3} locus using functional and genomic features

John P. Ray$^{1,12}$, Carl G. de Boer$^{1,2,12}$, Charles P. Fulco$^{1,3}$, Caleb A. Lareau$^{1,4}$, Masahiro Kanai$^{1,5,6}$, Jacob C. Ulirsch$^{1,4}$, Ryan Tewhey$^{1,7}$, Leif S. Ludwig$^{1}$, Steven K. Reilly$^{1,7}$, Drew T. Bergman$^{1}$, Jesse M. Engreitz$^{1,8}$, Robbyn Issner$^{1}$, Hilary K. Finucane$^{1,5}$, Eric S. Lander$^{1,3,9}$, Aviv Regev$^{1,2,9,10,13}$, Nir Hacohen$^{1,11,13}$ & Aviv Regev

Genome-wide association studies have associated thousands of genetic variants with complex traits and diseases, but pinpointing the causal variant(s) among those in tight linkage disequilibrium with each associated variant remains a major challenge. Here, we use seven experimental assays to characterize all common variants at the multiple disease-associated \textit{TNFAIP3} locus in five disease-relevant immune cell lines, based on a set of features related to regulatory potential. Trait/disease-associated variants are enriched among SNPs prioritized based on either: (1) residing within CRISPRi-sensitive regulatory regions, or (2) localizing in a chromatin accessible region while displaying allele-specific reporter activity. Of the 15 trait/disease-associated haplotypes at \textit{TNFAIP3}, 9 have at least one variant meeting one or both of these criteria, 5 of which are further supported by genetic fine-mapping. Our work provides a comprehensive strategy to characterize genetic variation at important disease-associated loci, and aids in the effort to identify trait causal genetic variants.
Genome-wide association studies (GWAS) have revealed >100,000 associations of genetic variants with human traits and diseases (e.g. autoimmune disease), but it remains a challenge to pinpoint the causal variant(s) that account for the association by altering disease risk and determine their functions. This is because they are often in tight linkage disequilibrium (LD) with non-causal variants and, in the vast majority of cases, lie in non-coding regions, where it is more challenging to predict the impact and relevant context of variants.

Most causal variants in the non-coding genome are likely to act through altering transcript abundance in a disease-relevant context. In the relevant context (cell type, tissue source, stimulation, genetic background, and disease status), experimental assays could be used to characterize the relationship between genetic variants and gene regulation. However, there are several challenges in this strategy. First, one or more aspects of the relevant context may be unknown. Second, even in the relevant context, there are many possible impacts of non-coding variants (such as different effects on gene expression or isoform usage), and each would involve a separate experimental assay, highlighting different features. Third, although ideally the relationship would be tested by allelic substitution in the relevant context—for instance, by CRISPR-directed base editing or homologous recombination—this approach is difficult to scale at present. As a result, various assays have been proposed for identifying potentially causal variants, based on the variant’s relation to or impact on different molecular features in a relevant cell type.

These assays can be categorized into four classes, depending on (i) whether they involve observations of natural systems or engineered experimental perturbations and (ii) whether they pertain to a region or an individual variant.

1. Observational assays that characterize the genomic region in which the variant resides. Examples include using ATAC-seq, DNase-I-seq, and H3K27ac ChIP-seq, as well as testing whether the variant lies in spatial proximity to a target gene, based on topological assays such as 4C or HiC.

2. Observational assays that characterize the impact of naturally occurring genetic differences at the variant. Examples include characterizing whether the variant shows allele-specific association with expression of one or more nearby genes or with local chromatin features (that is, an expression quantitative trait locus (eQTL) or a chromatin QTL, respectively), or whether the variant disrupts a transcription factor (TF) motif.

3. Engineered perturbational assays that test the impact of the genomic region containing the variant. Examples include assaying the effect of CRISPR-directed inhibition (e.g., CRISPRi) and activation (e.g., CRISPRa) of the region on the expression of nearby genes or on chromatin organization.

4. Engineered perturbational assays that test the impact of the variant itself. Examples include testing allele-specific enhancer activities in massively parallel reporter assays (MPRAs) and related methods.

These assays have been used in previous studies to suggest particular genetic variants as more likely to impact disease risk. However, we do not know the extent to which each of these assays actually enriches for causal variants.

Here, we reason that assays that usefully prioritize disease-causal variants could be recognized by testing whether they effectively enrich for disease-associated variants among all variants across a region. However, because disease causal variants for most associations are unknown, we use disease-associated variants (which are known and highly enriched for causal variants).

As a proof of concept, we optimize and apply seven assays to characterize all known common genetic variants in the TNFAIP3 locus, a genetic locus associated with multiple autoimmune diseases, and where disease-associated genetic and epigenetic features have been studied extensively. We use cell lines derived from T cells, B cells, and monocytes (U937 or THP-1 monocyte cell lines, GM12878 or BJAB B cell lines, or Jurkat T cell line), representing three major cell lineages that can impact autoimmunity. We find that two criteria are correlated with significant enrichment for the subset of SNPs that show disease/trait-association and, by inference, the subset of SNPs that play a causal role in these associations. These two criteria are: (i) localization within CRISPRi-sensitive regions in one of the cell types, or (ii) localization within open chromatin regions while also showing allele-specific reporter activity by MPRA. We find SNPs that fulfill at least one of these two criteria in 9 of 15 disease/trait-associated TNFAIP3 haplotypes, prioritizing 18 putatively causal SNPs in the locus associated to 15 diseases. By contrast, several other criteria showed no enrichment for disease/trait association. Our results highlight the limitations of using individual assays for implicating a variant as potentially functional, and suggests that a combination of assays, cell types and context will be needed to prioritize variants at disease loci.

Results

The TNFAIP3 locus harbors 15 independent disease associations. As a test case, we investigated the TNFAIP3 locus because it has strong associations to many autoimmune diseases. TNFAIP3 encodes the A20 protein, which is upregulated by NF-kB upon immune stimulation, and dampens pathways that activate NF-kB in a negative feedback loop (Fig. 1a). At least 49 GWASs have identified genome-wide significant SNPs in the TNFAIP3 locus that together are associated with 16 human diseases and phenotypes, including lupus (SLE), rheumatoid arthritis (RA), psoriasis, inflammatory skin disorder (ISD), celiac disease, inflammatory bowel disease (IBD), and multiple sclerosis (MS). Rather than focusing only on disease-associated SNPs (that is, those showing genome-wide-significant associations for one of these diseases as tag SNPs or in tight LD to them), we systematically examined all common SNPs (MAF > 0.01) in the ~300 kb topologically associating domain (TAD) containing TNFAIP3 (based on HiC data from GM12878 B cells and THP-1 monocyte cell lines), and 150 kb on either side of the TAD because it is known that regulatory regions can affect the expression of genes outside of TADs.

We reasoned that studying all common non-coding variants would allow us to derive empirical null distributions for each assay because most variants are not expected to be functional. Accordingly, we selected for analysis all 2776 common variants with minor allele frequency > 0.01 in East Asian or European populations (in 1000 Genomes, see “Methods” section).

We next analyzed the locus to estimate the number of SNPs that contribute to disease. Of the 2776 variants, 294 were in tight LD (r² > 0.8) to at least one of 34 ‘tag SNPs’—that is, a SNP reported as having the highest association score in one of the GWASs for the autoimmune and other diseases noted above (Fig. 1c; Supplementary Fig. 2a). Through LD analysis (r² ≥ 0.8) of the tag SNPs, we identified 15 independent haplotypes associated with one or more GWAS traits in Europeans (Fig. 1d; Supplementary Fig. 2b–d); three of these haplotypes also overlapped East Asian disease-associated haplotypes, but with slight differences in the associated SNPs (Fig. 1d; Supplementary Fig. 2d). Notably, fine-mapping of immune-related UK Biobank phenotypes (autoimmune disease (self-reported or diagnosed), self-reported allergy, and eosinophil counts) showed that, despite
limited sample size, all but two of these separately fine-mapped alleles were contained on three of the 15 disease-associated haplotypes from our LD analysis (Supplementary Data 1, 2, see “Methods” section). Collectively, we estimate that at least 15 SNPs in the locus contribute to disease.

While TNFAIP3 is likely to play a role in many disease-relevant cell types, we chose to study T cells, B cells, and monocytes. These important innate and adaptive immune cell types likely play a role in the autoimmune diseases with which the TNFAIP3 locus is associated because their localization in disease-associated tissues, signaling, and function are correlated with disease progression in the clinic and in animal models of disease. T cell-, B cell-, and monocyte-specific accessible chromatin and active histone marks (H3k27ac and H3K4me3 ChIP-seq) are also significantly enriched (compared to other cell types) for GWAS variants (P < 1 × 10⁻⁸) from studies of diseases that had associations in TNFAIP3 according to stratified LD score regression (Fig. 1e; Supplementary Fig. 3a–c). Moreover, deleting TNFAIP3 in these cell types causes systemic autoimmunity in mice.

We studied cell lines derived from these cell types: THP-1 and U937 for monocytes, BJAB and GM12878 for B cells, and Jurkat for T cells. The chromatin accessibility profiles of these cell lines are enriched for autoimmune-associated risk variants similarly to the corresponding primary cells (Supplementary Fig. 3d), and among blood cell types profiled by ATAC-seq, they were most similar to the cell type they represent (Supplementary Fig. 4a), especially at the TNFAIP3 locus (Supplementary Fig. 4b), suggesting that the selected cell lines could serve as models for these cell types.

A panel of assays to annotate genetic variation. We used both observational and perturbational assays to characterize regulatory features in the areas where variants were located, and the variants themselves (Fig. 2).

Using observational assays, we first analyzed regions that contact the TNFAIP3 promoter (primary T cell and GM12878 B cell HiChIP data; ~5 kbp resolution) and regions of accessible chromatin in any of the cell lines (using ATAC-seq in unstimulated and stimulated cells (Supplementary Fig. 5a, b), and publicly available DHS of cell types from the blood). For each variant, we also assessed whether it lies within a region bound by a TF based on ChIP-seq, and whether the variant is predicted to affect TF binding according to its cognate motif (Supplementary Fig. 5c).

Using perturbational assays, we sought to identify regions that can affect TNFAIP3 expression. With CRISPRi (in which KRAB-dCas9 binds to a region targeted by a guide RNA and represses chromatin locally), we identified regions whose inhibition alters TNFAIP3 expression. We targeted all regions with accessible chromatin in either U937, BJAB, or Jurkat cell lines, tiled guides across each element (and up to 100 bp on either side), and identified guides and regions that significantly repress TNFAIP3 expression (see the “Methods” section; Supplementary Fig. 6, Fig. 1 Disease variants in the complex autoimmune-associated TNFAIP3 locus. a TNFAIP3 encodes the A20 protein, which forms part of a negative feedback loop to dampen NF-kB-mediated immune activation. b HIC plots for the lymphoblastoid B cell line GM12878, with color intensity proportional to the interaction frequency between genomic coordinates (x-axis). Boxes indicate the 300 kb high-interaction domain and the 605 kb region used in this study. c, d Genetics of the TNFAIP3 locus. The positions (shared x-axis indicated above c) of variants with respect to the TNFAIP3 gene and a lncRNA (LOC100130476). e GWAS tag SNPs (red) and SNPs in tight LD (greyscale boxes indicating LD to tag SNP) for many immune-related phenotypes (y-axis). d GWAS haplotypes defined by combining all SNPs in tight LD (r² > 0.8) to GWAS tag SNPs for European (CEU; top) and East Asian (CHB/JPT; bottom) populations. Colors are used to help identify shared haplotypes between CEU and CHB/JPT populations. e Autoimmune GWAS signals are enriched in open chromatin of immune cells. Heritability enrichment (color) of disease-associated SNPs in DHS of various tissues (x-axis) for seven autoimmune diseases (y-axis), according to LD-score regression. Also see Supplementary Data 1, 2.
Hits from two strategies enrich for disease-associated SNPs. Ideally, we would assess each assay by directly testing how well it enriches for causal variants among the full set of variants assayed. However, using metrics like ‘precision’ and ‘recall’ would require that the causal variants be known with certainty. Because they are not, we instead tested how well the methods enrich for variants in tight LD with the tag SNP (as these variants are in turn enriched for true causal variants), calculating a ‘pseudo-precision’ and ‘pseudo-recall’. For each assay, we therefore quantified (1) the number of tested SNPs considered ‘hits’ in the assay (n_H), (2) the number of tag SNPs for which at least one SNP in tight LD was tested in the assay (n_T; i.e. recoverable tag SNPs), and (3) the number of tag SNPs for which at least one SNP in tight LD was considered an assay hit (n_T|H; i.e. recovered tag SNPs) (Supplementary Fig. 9a). We next calculated the pseudo-precision and pseudo-recall for GWAS variants for each assay. Here, we define ‘pseudo-precision’ as n_T|H/n_H, representing the fraction of all SNPs considered hits that are recovered tag SNPs, and ‘pseudo-recall’ as n_T|H/n_T, representing the fraction of tag SNPs that are recovered by being in tight LD with one or more hits. These terms are similar to precision and recall except that a single causal SNP might underlie multiple tag SNPs (by being in tight LD to each of them), making a pseudo-precision above 1 possible. By these measures, a highly effective approach would recover all tag SNPs (pseudo-recall = 1) with as few SNP hits as possible (high pseudo-precision). In the calculation of pseudo-precision and pseudo-recall, we did not consider GWAS tag SNPs that had no assayed variants in tight LD with that tag SNP (including the tag SNP itself) in order not to falsely penalize the assays for technical failures (e.g., lack of PAM site for CRISPR or poor coverage in MPRA). We conducted these analyses for all variants and for the subset of variants that lie in accessible chromatin in one of the three blood cell types studied (because GWAS variants are enriched in accessible chromatin) and accessibility data is readily available for many cell types (Fig. 3a, b).

To determine whether the pseudo-precision/pseudo-recall performance of each method is better than expected by chance, we created an empirical null distribution by randomly permuting the hit status among the assayed SNPs (1000 permutations) or by shifting the hit status of each SNP to the next adjacent assayed SNP (Supplementary Fig. 9b, c). The shift approach preserves positional clustering of hits inherent to LD and to some of the assays (e.g. CRISPRi, open chromatin). This reduces inflation of positive hits within the null that may occur by permutation, where the permuted hits may be in LD with many more tag SNPs than are possible given the clustered nature of the assay (thus increasing pseudo-precision and pseudo-recalls) (see the “Methods” section). Both shifting and permutation yielded similar results for SNPs in tight LD with GWAS tag SNPs (Fig. 3c, Supplementary Fig. 9d). For each method, we compared the pseudo-precision and pseudo-recall of actual data to the null distribution. We did this both for all variants (Fig. 3c) and the variants located in accessible chromatin in the three blood cell types (Fig. 3d).

Relative to all variants, most of the methods (ATAC-seq on our cell lines, Blood DHS + ATAC-seq on our cell lines, TF ChIP + motif, L- and T-MPRA, and CRISPRa) did not show a significant enrichment for GWAS variants (Fig. 3c). However, CRISPRi showed 7.5-fold enrichment for GWAS variants (95% C.I., [0.9375; ∞]), albeit not significant (P = 0.087, empirical P-value with genome-shifts null) (Fig. 3a, c, Supplementary Fig. 9a).
After restricting our analysis to variants located in accessible chromatin in the three blood cell types, several of the methods (CRISPRa and TF ChIP + motif) again showed no significant enrichment for GWAS variants. However, T-MPRA showed significant enrichment ($P = 0.011$, empirical $P$-value with genomic-shifts null; 1.44-fold enrichment for GWAS, 95% CI [1.04; 5.2]; Fig. 3d, Supplementary Fig. 9e).

Both L-MPRA and T-MPRA showed greatly increased pseudo-precision with only marginally reduced pseudo-recall when restricting attention only to variants in accessible chromatin (Fig. 3d, Supplementary Figs. 9e and 10). This may be because many variants have the capacity to alter expression when tested in an enhancer assay (such as MPRA), but do not reside in a region of accessible chromatin in the relevant cell types and thus do not alter disease risk. Although L-MPRA performed well for variants in accessible chromatin, having the highest pseudo-precision of any assay, there was limited power to evaluate L-MPRA because only four variants (in tight LD to 15 tag SNPs) out of the 19 L-MPRA hits were in accessible chromatin ($P = 0.128$, empirical $P$-value with genomic-shifts null; Fig. 3d).

For CRISPRi, pseudo-precision and pseudo-recalls changed little when focusing only on variants in accessible chromatin (Fig. 3a–d, Supplementary Fig. 10), but pseudo-precision was less significant ($P = 0.215$, empirical $P$-value with genomic-shifts null) because some of the SNPs tested lay just outside (within 100 bp) regions of accessible chromatin (Fig. 3c, d, Supplementary Fig. 9d, e).

We also considered another alternative proxy for causal variants, using credible sets from fine-mapping studies (Supplementary Data 10), determining, in this case, the number of credible sets ($n_T$) that were recovered ($n_T/TH$) by containing one or more assay hits ($n_H$). Although the SNPs in a credible set are more likely to be causal than when doing LD expansion, the limited availability of fine-mapping data restricted this analysis and reduced our statistical power. We calculated the pseudo-precision and pseudo-recall for GWAS variants for each assay in an analogous way (Supplementary Fig. 9f–k). The rates from the credible set-based analysis generally showed similar trends to the tag SNP approach, but were less significant due to the reduced sample size (Fig. 3c, d vs. Supplementary Fig. 9d–k); in addition, pseudo-precision was necessarily reduced for fine mapping due to reduced number of association signals, but with no change in assays hits.

**Prioritization of variants in disease-associated haplotypes.** Finally, we used our analysis of genomic assays to prioritize SNPs on each disease-associated haplotype (Fig. 4, Supplementary Data 3). We annotated as high-priority those variants that were hits in at least one of the two assays with the best performance (CRISPRi for all variants and T-MPRA variants in accessible chromatin), finding a total of 18 such high-priority variants (Fig. 4, asterisks). Of the 15 disease-associated haplotypes, nine included one or more of these 18 SNPs. These included five SNPs that had been fine-mapped in the UK Biobank, lying in 95% credible sets representing associations with allergy, all autoimmune diseases combined, and eosinophil counts (Fig. 4, Table 1).

Several of these high-priority variants had other evidence supporting a role in disease. For example, rs6927172 is the only variant in the relevant cell types and thus do not alter disease risk. Although L-MPRA performed well for variants in accessible chromatin, having the highest pseudo-precision of any assay, there was limited power to evaluate L-MPRA because only four variants (in tight LD to 15 tag SNPs) out of the 19 L-MPRA hits were in accessible chromatin ($P = 0.128$, empirical $P$-value with genomic-shifts null; Fig. 3d).

For CRISPRi, pseudo-precision and pseudo-recalls changed little when focusing only on variants in accessible chromatin (Fig. 3a–d, Supplementary Fig. 10), but pseudo-precision was less significant ($P = 0.215$, empirical $P$-value with genomic-shifts null) because some of the SNPs tested lay just outside (within 100 bp) regions of accessible chromatin (Fig. 3c, d, Supplementary Fig. 9d, e).

We also considered another alternative proxy for causal variants, using credible sets from fine-mapping studies (Supplementary Data 10), determining, in this case, the number of credible sets ($n_T$) that were recovered ($n_T/TH$) by containing one or more assay hits ($n_H$). Although the SNPs in a credible set are more likely to be causal than when doing LD expansion, the limited availability of fine-mapping data restricted this analysis and reduced our statistical power. We calculated the pseudo-precision and pseudo-recall for GWAS variants for each assay in an analogous way (Supplementary Fig. 9f–k). The rates from the credible set-based analysis generally showed similar trends to the tag SNP approach, but were less significant due to the reduced sample size (Fig. 3c, d vs. Supplementary Fig. 9d–k); in addition, pseudo-precision was necessarily reduced for fine mapping due to reduced number of association signals, but with no change in assays hits.

**Prioritization of variants in disease-associated haplotypes.** Finally, we used our analysis of genomic assays to prioritize SNPs on each disease-associated haplotype (Fig. 4, Supplementary Data 3). We annotated as high-priority those variants that were hits in at least one of the two assays with the best performance (CRISPRi for all variants and T-MPRA variants in accessible chromatin), finding a total of 18 such high-priority variants (Fig. 4, asterisks). Of the 15 disease-associated haplotypes, nine included one or more of these 18 SNPs. These included five SNPs that had been fine-mapped in the UK Biobank, lying in 95% credible sets representing associations with allergy, all autoimmune diseases combined, and eosinophil counts (Fig. 4, Table 1).

Several of these high-priority variants had other evidence supporting a role in disease. For example, rs6927172 is the only variant in the relevant cell types and thus do not alter disease risk. Although L-MPRA performed well for variants in accessible chromatin, having the highest pseudo-precision of any assay, there was limited power to evaluate L-MPRA because only four variants (in tight LD to 15 tag SNPs) out of the 19 L-MPRA hits were in accessible chromatin ($P = 0.128$, empirical $P$-value with genomic-shifts null; Fig. 3d).

For CRISPRi, pseudo-precision and pseudo-recalls changed little when focusing only on variants in accessible chromatin (Fig. 3a–d, Supplementary Fig. 10), but pseudo-precision was less significant ($P = 0.215$, empirical $P$-value with genomic-shifts null) because some of the SNPs tested lay just outside (within 100 bp) regions of accessible chromatin (Fig. 3c, d, Supplementary Fig. 9d, e).

We also considered another alternative proxy for causal variants, using credible sets from fine-mapping studies (Supplementary Data 10), determining, in this case, the number of credible sets ($n_T$) that were recovered ($n_T/TH$) by containing one or more assay hits ($n_H$). Although the SNPs in a credible set are more likely to be causal than when doing LD expansion, the limited availability of fine-mapping data restricted this analysis and reduced our statistical power. We calculated the pseudo-precision and pseudo-recall for GWAS variants for each assay in an analogous way (Supplementary Fig. 9f–k). The rates from the credible set-based analysis generally showed similar trends to the tag SNP approach, but were less significant due to the reduced sample size (Fig. 3c, d vs. Supplementary Fig. 9d–k); in addition, pseudo-precision was necessarily reduced for fine mapping due to reduced number of association signals, but with no change in assays hits.

**Prioritization of variants in disease-associated haplotypes.** Finally, we used our analysis of genomic assays to prioritize SNPs on each disease-associated haplotype (Fig. 4, Supplementary Data 3). We annotated as high-priority those variants that were hits in at least one of the two assays with the best performance (CRISPRi for all variants and T-MPRA variants in accessible chromatin), finding a total of 18 such high-priority variants (Fig. 4, asterisks). Of the 15 disease-associated haplotypes, nine included one or more of these 18 SNPs. These included five SNPs that had been fine-mapped in the UK Biobank, lying in 95% credible sets representing associations with allergy, all autoimmune diseases combined, and eosinophil counts (Fig. 4, Table 1).

Several of these high-priority variants had other evidence supporting a role in disease. For example, rs6927172 is the only
**Fig. 4 Prioritizing variants on disease-associated haplotypes.** A summary of assay results and genetics data for all SNPs on each disease-associated haplotype. Each subpanel represents a different haplotype, with associated traits and the haplotype number are indicated on the top. For each SNP (x axes), the total number of assay hits is shown in the bar graph (top) with SNPs that are hits in CRISPRi or T-MPRA hits in accessible chromatin marked with an asterisk. Results from each assay are shown in the middle, with hits in red, and SNPs that are assayed but were not hits in gray for each of the seven assays (y-axis). The vertical black bars above accessible chromatin SNP status indicate SNPs that were in accessible chromatin in our tested cell lines. Fine-mapped immune-related traits from UK Biobank (UKBB), including SNPs in the 95% credible set (CS—blue) and those that have a posterior inclusion probability > 10% (PIP10—green) are second from the bottom. The population-specific SNPs contained within each disease-associated haplotype are indicated (bottom) with orange for European (CEU) and purple for East Asian (CHB/JPT). Also see Supplementary Data 3.

High-priority variant on haplotype 6 (which lay in accessible chromatin and scored in the T-MPRA assay, but not in the CRISPRi assay); this variant is associated with many diseases, including RA, SLE, celiac, T1D, and asthma, and it is a fine-mapped SNP in our analysis of combined autoimmune disease in the UK Biobank and in previously reported studies of ulcerative colitis, RA, and celiacia1 (posterior inclusion probability (PIP) = 0.1343; Table 1, Supplementary Data 2). This variant also has evidence of allele-specific ATAC-seq and allele-specific ChIP-seq for the TFs NF-kB and JunD in lymphoblastoid cell lines44,45 and allele-specific ATAC-seq and allele-specific ChIP-seq for the NF-kB1 p50 subunit in primary CD4 T cells46. It appears to interact with the TNFAIP3 promoter by 3C, has allele-specific reporter activity according to a luciferase assay, and lays in a region that affected TNFAIP3 expression based on 11–12 bp CRISPR-induced deletions46,47. Only two of the other 10 variants on the haplotype had evidence of impact (with rs111710107 only being in accessible chromatin, and rs111231590 having allele-specific reporter activity according to both T- and L-MPRA assays).

Similarly, rs643177 is one of two high-priority variants on haplotype 9 (laying in accessible chromatin and a hit in T-MPRA assay, but not tested in CRISPRi due to the lack of a suitable guide-RNAs). This variant also had evidence of interaction with the TNFAIP3 promoter according to HiChIP, and had allele-specific reporter expression in L-MPRAs. rs643177 is a fine-mapped psoriasis SNP1 and has evidence of allele-specific binding of the TF Pou2f1 (Table 1). The other high-priority variant on haplotype 9 is rs559766217, which was a hit in the CRISPRi assay, is in accessible chromatin and contacts the TNFAIP3 promoter according to HiChIP. Four of the 17 other SNPs on the haplotype have some evidence of impact (including rs538522 and rs598493, which interact with the TNFAIP3 promoter according to HiChIP;
Table 1 Disease-associated variants positive for CRISPRi or chromatin accessibility with T-MPRA.

| SNP ID         | Associated trait                                    | Tehranchi asATAC | Tehranchi asChIP | Fine mapped UKBB 95% CS (SuSiE) | Other evidence | Haploype | Hit in assays                   |
|---------------|-----------------------------------------------------|------------------|------------------|---------------------------------|----------------|----------|---------------------------------|
| rs200820567   | Allergy, ISD, RA, SLE, eosinophil counts, IgA deficiency, Sjögren’s | x                |                  | Eosinophil counts (PIP = 0.03); Allergy (PIP = 0.04) | Fine mapped in Adrianto et al. (SLE) | 2        | T-MPRA + accessible chromatin   |
| rs14831465    | Allergy, ISD, RA, SLE, eosinophil counts, IgA deficiency, Sjögren’s |                  |                  |                                 | Fine mapped in Adrianto et al. (SLE) | 2        | T-MPRA + accessible chromatin   |
| rs112497003   | Allergy, ISD, RA, SLE, eosinophil counts, IgA deficiency, Sjögren’s |                  |                  | Eosinophil counts (PIP = 0.01)   |                                 | 2        | T-MPRA + accessible chromatin   |
| rs111883038   | Allergy, ISD, RA, SLE, eosinophil counts, IgA deficiency, Sjögren’s |                  |                  | Eosinophil counts (PIP = 0.01)   |                                 | 2        | T-MPRA + accessible chromatin   |
| rs69277172    | Celiac, IDB, RA, Asthma, IgA deficiency, Sjögren’s, ISD, T1D, primary biliary cirrhosis | x                |                  | Combined Autoimmune (PIP = 0.13) |                                 | 6        | T-MPRA + accessible chromatin   |
| rs643177      | ISD, psoriasis                                      |                  |                  |                                 | Fine mapped in Farh et al. (Psoriasis, PIP = 0.15) | 9        | L-MPRA + accessible chromatin; T-MPRA + accessible chromatin |
| rs59086769    | Urine metabolites                                   |                  |                  |                                 |                                 | 5        | T-MPRA + accessible chromatin   |
| rs1002658     | Celiac                                              |                  |                  |                                 |                                 | 14       | T-MPRA + accessible chromatin   |
| rs11758213    | MS                                                  | x                |                  |                                 |                                 | 15       | T-MPRA + accessible chromatin   |
| rs939527      | MS                                                  |                  |                  |                                 |                                 | 4        | T-MPRA + accessible chromatin   |
| rs12201430    | Blood metabolites                                   |                  |                  |                                 |                                 | 15       | T-MPRA + accessible chromatin   |
| rs12192746    | Blood metabolites                                   |                  |                  |                                 |                                 | 4        | T-MPRA + accessible chromatin   |
| rs34654849    | MS, IgA deficiency, RA                              |                  |                  |                                 |                                 | 1        | T-MPRA + accessible chromatin   |
| rs7358137     | MS, IgA deficiency, RA                              |                  |                  |                                 |                                 | 1        | T-MPRA + accessible chromatin   |
| rs5029924     | Allergy, ISD, RA, SLE, eosinophil counts, IgA deficiency, Sjögren’s |                  |                  |                                 | BIAB asATAC and fine mapped in Farh et al. (SLE, PIP = 0.09) | 2        | T-MPRA + accessible chromatin; CRISPRi |
| rs5029926     | Allergy, ISD, RA, SLE, eosinophil counts, IgA deficiency, Sjögren’s |                  |                  |                                 |                                 | 2, 3     | CRISPRi                           |
| rs10499197    | Allergy, ISD, RA, SLE, eosinophil counts, IgA deficiency, Sjögren’s |                  |                  |                                 |                                 | 2        | T-MPRA + accessible chromatin; CRISPRi; L-MPRA + accessible chromatin |
| rs58905141    | Allergy, ISD, RA, SLE, eosinophil counts, IgA deficiency, Sjögren’s |                  |                  | Eosinophil counts (PIP = 0.02)   |                                 | 2        | CRISPRi                           |
| rs559766217   | ISD, Psoriasis                                      |                  |                  |                                 |                                 | 9        | CRISPRi                           |

Variants that are positive for either chromatin accessibility with T-MPRA or CRISPRi are listed with their associated trait, and whether they were also positive in Tehranchi et al. as having allele-specific ATAC (asATAC) or asChIP-seq for TFs in LCLs. Our fine-mapping data using UKBB traits for the 95% credible set variants are included, and other fine-mapping data or evidence for SNP functionality is listed in Other Evidence. The haplotype for the SNP is listed in Haplotype.

Discussion

GWASs effectively narrow down the search for causal variants to a small set of candidates, but determining which of the candidates contributes to disease risk remains a challenge. Because disease-causal variants are likely to be correlated with functionally relevant genomic features in the cell types in which they act, it should be possible to use genomic features to help inform the search for disease-associated variants—provided that the relevant cell types are known and can be studied (which remains a serious limitation).

To study the potential utility of various genomic features for prioritizing non-coding variants, we studied seven genomic assays in three disease-relevant cell types to assess to the extent to which they enrich for disease-associated variants within a set of 2776 common non-coding SNPs in the TNAIP3 locus. We found significant enrichment among high-scoring SNPs for two methods: (1) variants present in CRISPRi-responsive regulatory regions and (2) variants present in accessible chromatin that also showed allele-specific reporter activity by T-MPRA. These two criteria identified 18 TNAIP3 variants associated with 15 diseases on 9 haplotypes; potential functional roles for these variants in immunity were supported by additional published data (such as allele-specific ATAC-seq, ChIP-seq, and genetic fine-mapping). By contrast, the other genomic features did not provide significant enrichment.

Our data support two prioritization schemes (CRISPRi and accessible chromatin with T-MPRA) as viable methods for enriching for causal variants in the TNAIP3 locus. However, since perturbational methods (e.g. CRISPRi, MPRA) cannot
Methods

variants will require the development and integrated analysis of a strategy to help guide future variant characterization studies at cell types and conditions. Interestingly, many haplotypes contained features and functions in three main immune cell types. While our region with many genetic associations through analysis of variant associated loci, variants, and cell types.

Cell culture and stimulation of immune cells

BJAB (DSMZ, cat. no. ACC 757), Jurkat (ATCC, cat. no. TIB-152), U937 (ATCC, cat. no. CRL-19592), THP-1 (ATCC, cat. no. TIB-202), and GM12878 (Coriell, cat. no. GM12878 LCL from B-Lymphocyte) cell lines were cultured using RPMI 1640 (Thermo Fisher, 2180002) containing 10% fetal bovine serum (FBS, VWR, 97068-091), 1% Pen/strep (VWR, 45000-652), 1% γ-glutamine (Thermo Fisher, 25030081), and 1% HEPS (Sigma, H9887-100ML). Cells were maintained at a culture density between 100K and 1M cells/mL. Jurkat T cells were stimulated with 2.5 μg/mL of anti-CD3 (Biologend, 3170304) and 10ng/mL of PMA (Sigma, P1358-1MG) for 1h prior to harvesting for CRISPRi and MPRA, and 1 and 4 h for ATAC-seq experiments. BJAB and GM12878 B cells were stimulated with 2.5 μg/mL of anti-lgM (Sigma-Alrich, 86620270) and 2 μg/mL of anti-cD40 (Thermo Fisher, 14-0409-02) for 2 h for CRISPRi and MPRA, and 1 and 4 h for ATAC-seq and 4C (BJAB) experiments. THP-1 and U937 monocytes were stimulated with 100 ng/mL LPS (Invivogen, thl-pkps) for 2 h for CRISPRi and MPRA, and 1 and 4 h for ATAC-seq and 4C (U937) monocytes.

Lentivirus preparation. HEK293T cells were grown using DMEM (VWR, 45000-316) with 10% FBS, 1% Pen/strep, 1% γ-glutamine, 1% HEPS (10DMEM). Cells were passed at 80% confluence for each passage. To make lentivirus, media was aspired from the adherent cells and Trypsin EDTA (#63897) was added to a final concentration of 8%. The cells were pipetted up and down until they were in a single cell suspension. They were then counted and plated in a six-well plate at 500K cells/well in 2 mL 10DMEM. The next day, when the cells were ~70% confluent, they were transfected. pSV2-G (0.1 μg; Addgene, 8454), pPA2X (1 μg; Addgene, 12260), and the donor plasmid (1 μg), were added to 125 μl of OPTI-MEM and mixed. 6 μl of the TransIT-LT1 (Mirus Bio, MIR2000) transfection reagent was added to a separate tube of 125 μl OPTI-MEM (Thermo Fisher, 31985062) and mixed. The OPTI-MEM LT1 mixture was then added to the OPTI-MEM plasmid mixture, mixed, and incubated at RT for 15 min. The mixture was then added dropwise to the well. The plate was then swirled to ensure distribution of the mixture and effective transfection. The cells were put at 37°C to incubate overnight and the media was changed at 24h post-transfection, this time using 10DMEM with 1% BSA (Sigma, A9779). The cells were then incubated at 37°C for 16 h, and the supernatant was harvested. The viral supernatant was spun at 500 x g for 5 min to separate cellular debris, and stored at 4°C for up to 3 months.

GWAS immune cell enrichments. Heritability enrichment traits of interest (Supplementary Fig. 3) in cell lines and cell types were estimated using a stratified LD-score regression (s-LDSC) model as previously described15 by interpreting the cell type-specific enrichment. For broad tissue enrichments, DNase Hypersensitivity peaks and H3K27ac and H3K4me1 ChIP-seq peaks were overlapped with common variants to compute heritability. For tissue broad tissue enrichments, DNase Hypersensitivity peaks and H3K27ac and H3K4me1 ChIP-seq peaks were overlapped with common variants to compute heritability. The −log10P-values for the s-LDSC regression terms for each specific annotation were shown as a measure of enrichment.

HiChIP data and analysis. H3K27ac HiChIP data previously generated41 were downloaded in.fastq format from GEO accession GSE120166. Biological and technical replicates of Th17, Naïve T-cell, and GM12878 H3K27ac fluorescence were pooled and aligned with Hi-C Pro63, Virtual 4C plots (Supplementary Fig. 5) using a resolution of 2.5 kb and a rolling mean of 2.5 windows41. Per-fragment estimates of interaction strength to the TFAP3 promoter were generated using hicchip44 and normalizing to the total number of unique fragments in each library. We used a normalized interaction score of 20 to annotate regions as TFAP3 interacting.

ATAC-seq. We used the FAST-ATAC protocol62. 10,000–20,000 cells were sorted into RPMI 1640 containing 10% fetal bovine serum. The cells were centrifuged at 500 x g for 5 min at 4°C. All of the supernatant was aspirated, ensuring that the pellet was not disturbed. The pellet was then resuspended in the ligation mixture. The ligation mixture was normalized to 25 μl 2X TD Buffer (Illumina, 15027866), 2.5 μl TD Enzyme Buffer (Illumina, 150238061), 0.5 μl TD Enzyme (Illumina, 150238051) and mixed at 300 RPMs at 37°C for 30 min on an Eppendorf Thermomixer. Immediately after the incubation, samples were purified using a minElute kit (Qiagen,
Individual off-targets are aggregated into a single guide using:

\[
M = \left[ \begin{array}{c} 0, 0, 0.014, 0, 0.395, 0.317, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.815, 0.604, 0.685, 0.583 \end{array} \right]
\]

\[
S_{\text{guide}} = \frac{100}{1 + \frac{1}{\sum_{i=1}^{n} h_i(h_i)} + \frac{1}{n_{\text{mm}}}}
\]

CRISPRI screens. The guide libraries targeting the TNFAIP3 locus for CRISPRi and CRISPRa are available in Supplementary Data 4 and 7. To design the guide library, we aligned to the genome using BWA-MEM and then filtered for unique guide sequences with a minimum percent identity of 95% in the region surrounding TNFAIP3 (chr6:13784700–13784850, hg19) with a maximum 10,000 mismatches with up to three mismatches (parameter -n 3 -i 15 -e 10000 -y -a -s). Using this set of potential target locations in the genome, off-target score was calculated using the method of Hsu et al.68. Briefly, single off-targets were calculated as \(e \) moves over positional mismatches between guide and off-target, where the \( m \) is as below and \( d \) is mean pairwise distance between mismatches:

\[
\Pi_{i=1}^{n} (1 - W(x) + \frac{1}{(x + d)^{4 + 1}}) \frac{1}{n_{\text{mm}}}
\]
the library. A pseudocount was added to each guide count consisting of one read for every 100,000 total reads sequenced in that bin, corresponding to a prior that there is no expression difference for cells containing the guide. For each guide, the mean expression for that guide was estimated by maximizing the likelihood of the observed guide counts for each bin under this model, given that guide’s overall abundance. A z-score was estimated for each guide corresponding to how much the mean TNCAP3 expression of cells containing that guide differed from those containing non-targeting guides by subtracting the mean of the non-targeting guides.

In order to get element-level statistics, the z-scores for each guide were combined in two ways: a significance z-score (proportional to a signed P-value), and an effect-size z-score (the average z-score of guides targeting the element). Significance z-scores were calculated by applying Stouffer’s method to the individual guide’s z-scores. In order to correct these significance z-scores for the noise of the assay, they were scaled by the standard deviation of Stouffer z-scores calculated from the non-targeting guides. These scaling factors were calculated independently for every number of guides per targeted element n (since the noise in the Stouffer z-score depends on the number of guides used to calculate it). For example, Stouffer z-scores for elements targeted with n = 5 guides were normalized by the standard deviation of non-targeting Stouffer z-scores, each calculated from randomly sampled groups of five non-targeting guides. Here, non-targeting Stouffer z-scores were calculated by sampling the non-targeting guides into groups of size n, including each non-targeting guide 10 times total, and calculating a set of Stouffer z-scores from each sampling, and using the standard deviation of these z-scores to scale the significance z-scores for each element for that n. P-values were then calculated from these z-scores, considering only one-tailed tests (downregulation for CRISPRi and upregulation for CRISPRa). For an element to be considered significantly regulating (downregulation for CRISPRi and upregulation for CRISPRa), the number of guides per targeted element was at least 5, the significance z-score for an element was at least 2 (this was the empirical 95% confidence bound of non-targeting guides), and the effect-size z-score for an element was at least 2 (this was the empirical 95% confidence bound of non-targeting guides). The significance z-score and effect-size z-score for each element were calculated separately for each donor.

MPRA oligosynthesis and cloning was adapted from refs.16,71, tagging each element with an average of ~250 DNA barcodes. Oligos were synthesized by Agilent Technologies containing 150 bp of genomic context and 15 bp of adapter sequences (GTTCGGCCTCCCAGCAGCGGTCGTCG; Supplementary Data 3; Supplementary Data 8; 180 bp total). 20 bp barcodes and additional adapter sequences were added by performing 28 emulsion PCR reactions each 50 μL reactions, by the standard deviation of non-targeting Stouffer z-scores calculated from each sampling, and using the standard deviation of these z-scores to scale the significance z-scores for each element for that n. P-values were then calculated from these z-scores, considering only one-tailed tests (downregulation for CRISPRi and upregulation for CRISPRa). For an element to be considered significantly regulating (downregulation for CRISPRi and upregulation for CRISPRa), the number of guides per targeted element was at least 5, the significance z-score for an element was at least 2 (this was the empirical 95% confidence bound of non-targeting guides), and the effect-size z-score for an element was at least 2 (this was the empirical 95% confidence bound of non-targeting guides). The significance z-score and effect-size z-score for each element were calculated separately for each donor.

The mean expression for that guide was estimated by maximizing the likelihood of the observed guide counts for each bin under this model, given that guide’s overall abundance. A z-score was estimated for each guide corresponding to how much the mean TNCAP3 expression of cells containing that guide differed from those containing non-targeting guides by subtracting the mean of the non-targeting guides. For each guide, the mean expression for that guide was estimated by maximizing the likelihood of the observed guide counts for each bin under this model, given that guide’s overall abundance. A z-score was estimated for each guide corresponding to how much the mean TNCAP3 expression of cells containing that guide differed from those containing non-targeting guides by subtracting the mean of the non-targeting guides.

To insert the TNCAP3 promoter and GFP ORF, 20 μg of mpRar:TNCAP3:gfp plasmid was linearized by XbaI (NEB, R0145S) and KpnI-HF (NEB, R3142S) and 1x cutsmart buffer (NEB, R7204S) in a 50 μL volume for 3.5 h at 37 °C, followed by SPRI cleaning. An amplicon containing 165 bp of the TNCAP3 ORF, GFP open reading frame and a 15 bp 5’ flanking sequence was amplified by Gibson assembly (NEB, E2611) using 1.1 μg of oligos and 1 μg of digested vector in a 40 μL reaction incubated for 60 min at 50 °C followed by AMPure XP SPRI purification and elution in 20 μL of EB. Half of the ligated vector was then transformed into 100 μL of 10-beta e.coli (NEB, C3020K) by electroporation (2 kV, 200 μF) and 10 μL of the transformed bacteria was plated in duplicate on LB plates containing 100 μg/mL of carbenicillin and grown for 9 h at 37 °C prior to plasmid purification (Qiagen, 12963). For each of the aliquots we plated serial dilution plates on SOC (Supplementary Data 1) and estimated a library size of ~3.2 × 10^5 CFUs representing ~250 barcodes per allele.

For an element to be calculated from these z-scores, considering only one-tailed tests (downregulation for CRISPRi and upregulation for CRISPRa). For an element to be considered significantly regulating (downregulation for CRISPRi and upregulation for CRISPRa), the number of guides per targeted element was at least 5, the significance z-score for an element was at least 2 (this was the empirical 95% confidence bound of non-targeting guides), and the effect-size z-score for an element was at least 2 (this was the empirical 95% confidence bound of non-targeting guides). The significance z-score and effect-size z-score for each element were calculated separately for each donor.

To insert the TNCAP3 promoter and GFP ORF, 20 μg of mpRar:TNCAP3:gfp plasmid was linearized by XbaI (NEB, R0145S) and KpnI-HF (NEB, R3142S) and 1x cutsmart buffer (NEB, R7204S) in a 50 μL volume for 3.5 h at 37 °C, followed by SPRI cleaning. An amplicon containing 165 bp of the TNCAP3 ORF, GFP open reading frame and a 15 bp 5’ flanking sequence was amplified by Gibson assembly (NEB, E2611) using 1.1 μg of oligos and 1 μg of digested vector in a 40 μL reaction incubated for 60 min at 50 °C followed by AMPure XP SPRI purification and elution in 20 μL of EB. Half of the ligated vector was then transformed into 100 μL of 10-beta e.coli (NEB, C3020K) by electroporation (2 kV, 200 μF). Electroporation was immediately followed by 8 μL aliquots of the plasmid. Replicates’ Benjamini–Hochberg FDRs were less than q(0.1) (i.e. combined FDR < 0.1, and both replicates close to significant independently) and for which the direction of expression change was identical. In cases where there were more than two replicates, we included only the two replicates for which the TNCAP3 promoter positive control guides showed the strongest effect. Element- and guide-level data are available in Supplementary Data 5, 6.

MPRA oligosynthesis and cloning was adapted from refs.16,71, tagging each element with an average of ~250 DNA barcodes. Oligos were synthesized by Agilent Technologies containing 150 bp of genomic context and 15 bp of adapter sequences (GTTCGGCCTCCCAGCAGCGGTCGTCG; Supplementary Data 3; Supplementary Data 8; 180 bp total). 20 bp barcodes and additional adapter sequences were added by performing 28 emulsion PCR reactions each 50 μL reactions, by the standard deviation of non-targeting Stouffer z-scores calculated from each sampling, and using the standard deviation of these z-scores to scale the significance z-scores for each element for that n. P-values were then calculated from these z-scores, considering only one-tailed tests (downregulation for CRISPRi and upregulation for CRISPRa). For an element to be considered significantly regulating (downregulation for CRISPRi and upregulation for CRISPRa), the number of guides per targeted element was at least 5, the significance z-score for an element was at least 2 (this was the empirical 95% confidence bound of non-targeting guides), and the effect-size z-score for an element was at least 2 (this was the empirical 95% confidence bound of non-targeting guides). The significance z-score and effect-size z-score for each element were calculated separately for each donor.

To insert the TNCAP3 promoter and GFP ORF, 20 μg of mpRar:TNCAP3:gfp plasmid was linearized with XbaI (NEB, R0145S) and KpnI-HF (NEB, R3142S) and 1x cutsmart buffer (NEB, R7204S) in a 50 μL volume for 3.5 h at 37 °C, followed by SPRI cleaning. An amplicon containing 165 bp of the TNCAP3 ORF, GFP open reading frame and a 15 bp 5’ flanking sequence was amplified by Gibson assembly (NEB, E2611) using 1.1 μg of oligos and 1 μg of digested vector in a 40 μL reaction incubated for 60 min at 50 °C followed by AMPure XP SPRI purification and elution in 20 μL of EB. Half of the ligated vector was then transformed into 100 μL of 10-beta e.coli (NEB, C3020K) by electroporation (2 kV, 200 μF). Electroporation was immediately followed by 8 μL aliquots of the plasmid. Replicates’ Benjamini–Hochberg FDRs were less than q(0.1) (i.e. combined FDR < 0.1, and both replicates close to significant independently) and for which the direction of expression change was identical. In cases where there were more than two replicates, we included only the two replicates for which the TNCAP3 promoter positive control guides showed the strongest effect. Element- and guide-level data are available in Supplementary Data 5, 6.
following conditions: 95 °C for 20 s, 6 cycles (95 °C for 20 s, 62 °C for 15 s, 72 °C for 30 s), 72 °C for 2 min. Amplified material was SPRI purified using a 0.6x bead/sample ratio and then diluted to 30 µl of ER. Sequencing indexes were then attached using 20 µl of the eluted product and the same reaction conditions as for the tag-seq protocol, except the number of enrichment cycles was lowered to 5. Samples were pooled and sequenced using 2 × 150 bp chemistry on Illumina HiSeq 2500 at NextSeq.

MPRA RNA output and DNA input sequencing reads were mapped to the known tag sequences using a custom python script (quantifyRNATags.py; available from https://github.com/Carldeboer/MPRAs, allowing for up to four mismatches within the constant region (the common sequence before the tag sequence) and no mismatches within the tag sequence. The barcode counts were filtered and tags having fewer than 30 reads in the input (DNA) or 4 reads in the output (RNA) were excluded from subsequent analysis. The log(DNA/RNA) ratio (expression) was calculated using raw counts, scaled so that the median expression is 0, and the expression levels G+C-content normalized such that the mean expression for every %G+C = 0. Finaly, to eliminate instances where the tag sequence modifies the apparent expression level, any tags containing any one of eight blackballed 5-mer DNA sequences were excluded. Blackballed 5-mers were defined as those for which the absolute value of the average expression level of all tags containing that 5-mer was >0.15.

SNPs were tested for allele-specific reporter activity by a two-sided Student’s t-test, comparing the normalized log(RNA/DNA) expression values for the tags for allele A compared to the tags for allele B. Only SNPs for which we had at least 80 good tags between the two alleles were tested. P-values were corrected for multiple hypothesis testing by Benjamin–Hochberg FDR correction. Only SNPs that had an FDR < 0.1 for at least two of the replicates and where the direction of allele-specific reporter activity was consistent between all replicates were considered to be significant.

Predicted TF-binding perturbation. In order to find TFs whose motifs were disrupted, both alleles were scanned for each SNP with human and selected mouse motifs from CIS-BP32 using VEP33 and a custom VEP module implementing the GOMER approach34 for motif scanning (https://github.com/Carldeboer/VEP_GOMER). In order to be considered a motif disruption, the region surrounding the SNP must be either perfectly conserved or at least 95% conserved between the two alleles. Since some assays could not assay every variant, only assays containing >1% SNPs being perturbed per motif were included. Both code and motifs for this analysis are available here: https://github.com/Carldeboer/VEP_GOMER.

Data integration and analysis. In order to gauge how much each assay enriched for GWAS signal (as in Fig. 3), we considered all GWAS tag SNPs. Since the set of causal SNPs remains unknown, we must use the set of potentially causal SNPs as an enriched gold standard (e.g. fine mapped variants, or SNPs in tight LD (r² > 0.8), as used here). However, a single causal variant could underlie multiple GWAS tag SNPs, for instance, if the causal SNP is in tight LD with both tag SNPs in the GWAS population. Although a single GWAS tag SNP could represent multiple underlying causal SNPs, we expect this to be uncommon, and a potential explanation featuring fewer causal SNPs should be favored. With these considerations in mind, we evaluated each assay for its ability to identify GWAS tag SNPs by being in tight LD to hits. Since some assays could not assay every variant, only assays containing variants included. Similarly, if a tag SNP had no assayed SNPs in tight LD, that tag SNP was not included in evaluation of the assay since it could not have been recovered by the assay. The pseudo-precision and pseudo-recall were calculated for each assay as described in the “Results” section. The enrichment analysis using credible sets instead of tag SNPs was performed identically, but instead of evaluating the recovery of tag SNPs by being in tight LD to hits, we evaluated the recovery of credible sets by having one or more hits within each credible set.

In order to gauge the significance of enrichment for each assay with limited tag SNPs, we created an empirical null distribution by randomizing the data. Since some of the assays (e.g. DHS, CRISPRi/a) have an inherent clustering of their hits (i.e. SNPs within the same enhancer will share the same hit status), our null aimed to preserve this clustering. Specifically, the null was derived by ordering the assayed SNPs by genomic position and reassigning hit status H₉ₙ = H₉ₙ(mod n) for every possible i (0 ≤ i < n; where n is the number of assayed SNPs and mod is the modulo operation), and, each time (i.e. for each value of i), calculating pseudo-precision and pseudo-recall. P-values represent the fraction of this empirical null with at least as high a pseudo-precision and pseudo-recall as that observed from the actual data. We also tested random permutation of the SNP hit status as an alternative to shifting. In this approach, we used 100 independent random permutations of the SNP hits to create the null model. Although both approaches yielded similar results (Fig. 3 genomic shifts and Supplementary Fig. 9d, permutation), we opted to focus on the random null created by shifting hit status; the random permutations fail to capture the clustering of hits that results from genomic proximity and shared hit origins (e.g. adjacent SNPs in the same open chromatin region). For example, if there was only a single functional enhancer with CRISPRi which contained 10 SNPs, and the clustering of these SNPs prioritizes only one GWAS signal in one region, the distribution of these SNPs randomly could result in as many as 10 GWAS positive results in a null permutation test. These null distributions form a straight line in Fig. 3 because the numerator for both is the number of GWAS tag SNPs recovered (n₉₀) and the denominators for both pseudo-precision and pseudo-recall are invariant across the randomization (nₑₚ for pseudo-precision and nₑₚ for pseudo-recall).

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

Data availability
Raw and processed sequencing data for this study are available on NCBI GEO, under accession “GSE16793”. Other sources for data that support our findings are available from: 1000 Genomes, ENCODE, CHIP-Atlas, ImmuneBase, and GWAS Catalog.

Code availability
CRISPR analysis software is available at the following link: https://github.com/Carldeboer/MAUDE. Both code and motifs for TF binding motif analysis are available at the following link: https://github.com/Carldeboer/VEP_GOMER. Code for processing MPRA data is available at the following link: https://github.com/Carldeboer/MPRAs.

References
1. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).
2. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am. J. Hum. Genet. 95, 535–552 (2014).
3. Ullrich, J. C. et al. Identification of functional regulatory variants in mouse and human. Science 337, 1190–1195 (2012).
4. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med. 373, 905–907 (2015).
5. Gupta, R. M. et al. A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. Cell 170, 522–533.e515 (2017).
6. Mohanam, V. et al. Clorft06 is a colitis risk gene that regulates stability of epithelial adherens junctions. Science 359, 1161–1166 (2018).
7. Sankaran, V. G. et al. Human fetal hemoglobin expression is regulated by the development-stage-specific repressor BCL11A. Science 322, 1839–1842 (2008).
8. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Assessing the three-dimensional organization of the mammalian genome. Cell 159, 1665–1680 (2014).
9. Hsu, J. Y. et al. CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. Nat. Methods 15, 992–993 (2018).
10. Simoncovic, D. R. et al. Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature 549, 111–115 (2017).
11. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic interaction deconvolution of CRISPR tiling screen data. Nat. Methods 10, 1213–1218 (2013).
12. Degner, J. F. et al. DNase I sensitivity QTLs are a major determinant of human hematopoiesis at kilobase resolution. Nat. Genet. 43, 1341–1347 (2011).
13. Xiao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. 38, 1341–1347 (2006).
14. Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680 (2014).
15. Hsu, J. Y. et al. CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. Nat. Methods 15, 992–993 (2018).
16. Tewhey, R. et al. Direct identification of hundreds of expression—modulating variants using a multiplexed reporter assay. Cell 165, 1519–1529 (2016).
17. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell 165, 1530–1545 (2016).
18. Vockley, C. M. et al. Massively parallel quantification of chromatin looping. Genome Res. 25, 1206–1214 (2015).
19. Ma, A. & Malynn, B. A. A20: linking a complex regulator of ubiquitylation to immunity and human disease. Nat. Rev. Immunol. 12, 774–785 (2012).
20. Calderon, D. et al. Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat. Genet. 51, 1494–1505 (2019).
21. Sokhi, U. K. et al. Dissection and function of autoimmunity-associated TNFAIP3 (A20) gene enhancers in humanized mouse models. Nat. Commun. 9, 658 (2018).
22. Adrianto, I. et al. Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nat. Genet. 43, 253–258 (2011).
23. Wang, S. et al. TNFAIP3 gene expression and mimics a molecular phenotype associated with systemic lupus erythematosus. Genes Immun. 17, 165–170 (2016).
24. Wang, S. et al. Enhancement of TNFAIP3 promoter to influence A20 expression. PLoS Genet. 9, e1003750–1003710 (2013).
25. Coenraets, R. et al. T cell antigen receptor stimulation induces MALT1 paracaspase-mediated cleavage of the NF-κB inhibitor A20. Nat. Immunol. 9, 263–271 (2008).
26. Housley, W. J. et al. Genetic variants associated with autoimmunity drive NFκB signaling and responses to inflammatory stimuli. Sci. Transl. Med. 7, 291ra293–291ra293 (2015).
27. Phanestir, D. H. et al. Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development. Mol. Cell 67, 1037–1048. e1036 (2017).
28. Luppino, J. M. et al. Cohesin promotes interactions across domain boundaries to regulate gene expression. bioRxiv 1663, 231–265 (2019).
29. Weber, M. et al. Type II monocytes modulate T cell-mediated homeostasis and prevents colitis and spondyloarthritis. J. Clin. Investig. 129, 357–367 (2019).
30. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single nucleotide resolution. Nature 547, 173–178 (2017).
31. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27, 718–719 (2011).
32. Housley, W. J. et al. Genetic variants associated with autoimmunity drive NFκB signaling and responses to inflammatory stimuli. Sci. Transl. Med. 7, 291ra293–291ra293 (2015).
33. Phanestir, D. H. et al. Static and dynamic DNA loops form AP-1-bound activation hubs during macrophage development. Mol. Cell 67, 1037–1048. e1036 (2017).
34. Luppino, J. M. et al. Cohesin promotes interactions across domain boundaries to regulate gene expression. bioRxiv 1663, 231–265 (2019).
35. Weber, M. et al. Type II monocytes modulate T cell-mediated homeostasis and prevents colitis and spondyloarthritis. J. Clin. Investig. 129, 357–367 (2019).
36. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single nucleotide resolution. Nature 547, 173–178 (2017).
37. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics 27, 718–719 (2011).
Competing interests
AR is a co-founder and equity holder of Celsius Therapeutics, a founder of Immunitas, and SAB member of ThermoFisher Scientific, Asimov Neogene Therapeutics, and Syros Pharmaceuticals. NH is a co-founder and equity holder of Neon Therapeutics. ESL serves on the Board of Directors for Codiak BioSciences and Neon Therapeutics, and serves on the Scientific Advisory Board of F-Prime Capital Partners and Third Rock Ventures; he is also affiliated with several non-profit organizations including serving on the Board of Directors of the Innocence Project, Count Me In, and Biden Cancer Initiative, and the Board of Trustees for the Parker Institute for Cancer Immunotherapy. ESL has served and continues to serve on various federal advisory committees.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-15022-4.

Correspondence and requests for materials should be addressed to A.R. or N.H.

Peer review information Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.