Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk

Yakir A. Reshef, Hilary K. Finucane, David R. Kelley, Alexander Gusev, Dylan Kotliar, Jacob C. Ulirsch, Farhad Hormozdiari, Joseph Nasser, Luke O’Connor, Bryce van de Geijn, Po-Ru Loh, Sharon R. Grossman, Gaurav Bhatia, Steven Gazal, Pier Francesco Palamara, Lu...
further characterize the transcriptional programs underlying our complex trait associations via gene-set enrichment analyses using gene sets from the Molecular Signatures Database (MSigDB)\textsuperscript{20,21}.

**Results**

**Overview of methods.** Our method for quantifying directional effects of signed functional annotations on disease risk, SLDP regression, relies on the fact that the signed marginal association of a SNP to disease includes signed contributions from all SNPs tagged by that SNP. Given a signed functional annotation with a directional effect on disease risk, the vector of marginal SNP effects on disease risk will therefore be proportional (in expectation) to a vector quantifying each SNP’s aggregate tagging of the signed annotation, which we call the signed LD profile of the annotation. Thus, our method detects directional effects by assessing whether the vector of marginal SNP effects and the signed LD profile are systematically correlated genome wide.

More precisely, under a polygenic model\textsuperscript{22} in which true causal SNPs are correlated with a signed functional annotation, we show that

$$E(\hat{a} \mid v) = r_f \sqrt{h^2} R v$$  \hspace{1cm} (1)$$

where $\hat{a}$ is the vector of marginal correlations between SNP alleles and a trait, $v$ is the signed functional annotation (re-scaled to norm 1) reflecting, for example, the signed effect of a SNP on transcription factor binding, $R$ is the LD matrix, $h^2$ is the SNP heritability of the trait, and $r_f$ is the correlation between the vector $v$ and the vector of true causal effects of each SNP, which we call the *functional correlation*. Equation (1), together with an estimate of $h^2$, allows us to estimate $r_f$ by regressing $\hat{a}$ on the signed LD profile $R v$ of $v$. To improve power, we use generalized least-squares regression to account for redundancy among linked SNPs. We assess statistical significance by randomly flipping the signs of entries of $v$, with consecutive SNPs being flipped together in large blocks (~300 blocks total), to obtain a null distribution and corresponding $P$ values and false discovery rates (FDRs). We perform a multiple regression that explicitly conditions on a ‘signed background model’ corresponding to directional effects of minor alleles in five equally sized minor allele frequency (MAF) bins, which could reflect confounding due to genome-wide negative selection or population stratification. We note that SLDP regression requires signed effect size estimates $\hat{a}$ and quantifies directional effects, in contrast to stratified LD score regression\textsuperscript{1}, which analyzes unsigned $\chi^2$ statistics and quantifies unsigned heritability enrichment. Details of and intuition for the method are described in the Methods section and the Supplementary Note; we have released open-source software implementing the method (see URLs).

We applied SLDP regression using a set of 382 signed annotations $v$, constructed using the Basset software\textsuperscript{23}, each quantifying the predicted effects of SNP alleles on binding of a particular transcription factor in a particular cell line. The resulting annotations were sparse, with only 0.2% of SNPs having non-zero entries on average (Methods and Supplementary Table 1).

**Simulations.** We performed simulations with real genotypes, simulated phenotypes and our 382 signed transcription factor binding annotations to assess null calibration, robustness to confounding and power (Methods).

We first performed null simulations involving a heritable trait with no unsigned enrichment or directional association to any of our 382 annotations. The resulting $P$ values were well calibrated (Fig. 1a, Supplementary Table 2, and Supplementary Fig. 1a).

We next performed null simulations involving a trait with unsigned enrichment but no directional effects; these simulations were designed to mimic unsigned genomic confounding as might arise from the co-localization of transcription factor binding sites with other enriched regulatory elements\textsuperscript{24} (Methods). We again observed well-calibrated $P$ values (Fig. 1b). It is notable that our method is well calibrated even though it has no knowledge of the unsigned genomic confounder; this contrasts with unsigned enrichment approaches, in which unsigned genomic confounders must be carefully accounted for and modeled\textsuperscript{25}.

We next performed null simulations to assess whether our method remains well calibrated in the presence of confounding due to genome-wide directional effects of minor alleles on both disease risk and transcription factor binding, which could arise due to genome-wide negative selection or population stratification (Methods). $P$ values were well calibrated for the default version of the method, which conditions on the 5-MAF-bin signed background model, but were not well calibrated without conditioning on this model (Fig. 1c). The incorrect calibration that we observe when we do not include our signed background model could potentially be explained by genome-wide negative selection against decreased transcription factor binding\textsuperscript{26} resulting in a bias in the sign of the entries of our annotations (Supplementary Fig. 2). We condition on the signed background model in all analyses in this paper unless stated otherwise.

Finally, we performed causal simulations with true directional effects to assess the power and establish the unbiasedness of SLDP regression (Methods). The method is well powered to detect directional effects corresponding to a functional correlation of 2–6% (Fig. 2a, Supplementary Table 3, and Supplementary Figs. 3–5),
Analysis of gene expression across 48 GTEx tissues. We next applied SLDP regression to GTEx eQTL across 48 tissues (average n = 214). We first tested each of our 382 transcription factor binding annotations for a directional effect on expression in each of the 48 tissues in turn, analogous to our previous analysis of molecular traits in blood (Supplementary Table 6). For each significant association, we then assessed for tissue specificity by checking whether the association remained at least as significant when conditioning on average eQTL effects across tissues (Methods).

Our analysis yielded 2,330 annotation-tissue expression associations at per-trait FDR < 5%, representing 651 distinct transcription factor-tissue expression associations, of which 30 were robustly tissue-specific in our conditional analysis (Fig. 4 and Supplementary Table 7). We detected both known and novel associations. For example, our results recapitulate known activating roles for FOXA1 and FOXA2 in pancreas and other gastrointestinal tissues, early B-cell factor 1 (EBF-1) in lymphocytes, hepatocyte nuclear factors 4y and 4x (HNF4G and HNF4A) in liver, PU.1 in spleen, and FOs in fibroblasts and nerve tissue. We also detected ubiquitous activating signatures for the transcription pre-initiation complex members POL2, TAF1, and TBP (90% of the 28 tissues with a sample size above 150). Our results were concordant with transcription factor-tissue associations identified via a purely gene expression-based analysis (Methods and Supplementary Fig. 7).

Our analysis also uncovered many previously unknown associations. For example, our most significant association in aorta is a previously unreported activating role for GABPA, one of several transcription factors whose binding sites are enriched near aortic aneurysm-specific genes. In addition, our top, and only, association in the brain tissue substantia nigra is TAF1. Neurodegeneration in the substantia nigra is a hallmark of Parkinson’s disease, and TAF1 was recently shown to be the causal gene in a rare form of Parkinsonism. Our analysis links these two facts, potentially shedding light on the mechanism of TAF1’s role in Parkinsonism.

Our tissue-specific analysis (Methods) also suggests new master-regulatory relationships for further exploration (Fig. 4). For example, we detected a robust tissue-specific activating role for CEBPB in pancreas, where it was our top result. Although CEBPB is not a classic pancreatic transcription factor, it is expressed in pancreatic beta cells specifically under metabolic stress. We also identified a robust tissue-specific activating role in skeletal muscle for MAFF, a transcription factor whose expression is increased by an order of magnitude in muscle tissue after exercise (Methods). MAFF is typically considered a repressor, and we identified it as such in our blood histone quantitative trait loci analysis; the positive association here suggests a tissue-specific function in muscle, perhaps via recruitment of an as-yet uncharacterized activator. Finally, we identified robust tissue-specific roles for CTCF as a repressor in tibial artery and an activator in the brain tissue putamen. While CTCF is known to be capable of both repression and activation, these results suggest that its repressive/activating role varies meaningfully across tissues.

Our results also demonstrate how our method can offer insights into non-tissue-specific aspects of transcriptional regulation. For example, YY1, a pioneer transcription factor that has recently attracted considerable interest, has been theorized via detailed experimental work to mediate enhancer–promoter interaction. However, YY1

Analysis of molecular traits in blood. Transcription factor binding is known to affect gene expression and other molecular traits, and regulatory relationships in blood are particularly well characterized. We therefore applied SLDP regression to 12 molecular traits in blood with an average sample size of n = 149 to further validate the method.

We first analyzed cis-eQTL data based on RNA-sequencing experiments in three blood cell types from the BLUEPRINT consortium (Supplementary Table 4). We tested each of our 382 transcription factor binding annotations for a directional effect on aggregate expression in each of the three blood cell types (Methods).

We detected 409 significant associations at per-trait FDR < 5%, representing 107 distinct transcription factor-blood cell type expression associations (Fig. 3a and Supplementary Table 5a). All of the detected associations were positive, implying that greater binding of these transcription factors leads to greater expression (in aggregate across genes); 78% of the associations involved transcription factors annotated as activating but not repressing in UnitProt (Fig. 3a and Methods). As expected, many of the detected associations recapitulate known aspects of transcriptional regulation, including the pro-transcriptional roles of RNA polymerase II and other members of the transcription pre-initiation complex (PIC) as well as roles of transcription factors unrelated to the PIC but known to have activating activity. We obtained similar results in an independent set of whole-blood eQTLs based on expression array experiments from the Netherlands Twin Registry (NTR) (Fig. 3b,c and Supplementary Table 5b,c).

We next conducted a similar analysis using histone QTLs (H3K27me1 and H3K27ac) and methylation QTLs for the three cell types in the BLUEPRINT data set. We detected 645 significant associations at per-trait FDR < 5%, four of which were negative (Fig. 3d,e and Supplementary Table 5d,e). Again, the majority of the positive associations (82%) involved unambiguously activating transcription factors. The four negative associations involved MAFK and MAFF, both of which lack a transactivation domain, and CTCF, which is known to act as an insulator. Many of the detected associations recover known aspects of histone mark biology and match a prior analysis of allelic imbalance in chromatin immunoprecipitation sequencing (ChIP-seq) data.

similar to values observed in analyses of real traits (see “Analysis” sections below). Notably, the power of the method is improved dramatically by its use of generalized least-squares regression to account for redundant information (Fig. 2a) as well as by its modeling of untapped causal SNPs via the signed LD profile (Supplementary Fig. 3). In all instances, our method produced either unbiased or nearly unbiased estimates of functional correlation and related quantities (Fig. 2b and Supplementary Fig. 6).

**Fig. 2 | Simulations assessing power, bias, and variance.** a, Power curves comparing SLDP regression using generalized least-squares (that is, weighting) to an ordinary (that is, unweighted) regression of the summary statistics on the signed LD profile. Error bars indicate standard errors of power estimates. b, Assessment of bias and variance of the SLDP regression estimate of r, across a range of values of the true r. Blue box and whisker plots depict the sampling distribution of the statistic, while the red dots indicate the estimated sample mean and the red error bars indicate the standard error around this estimate. Both a and b are conducted at realistic sample size (47,360) and heritability (0.5). Numerical results are reported in Supplementary Table 3.
knockdown experiments have shown a mix of upregulation and downregulation of many genes, presumably due to downstream regulatory cascades. In contrast, our analysis, which due to its use of eQTLs is able to focus exclusively on cis-regulatory effects, shows a robust, predominantly activating role for YY1 across 25 tissues.

Analysis of 46 diseases and complex traits. We applied SLDP regression to 46 diseases and complex traits with an average sample size of 289,617 (URLs and Supplementary Table 8). We first tested each of our 382 transcription factor binding annotations against each of the 46 traits in turn (Table 1 and Supplementary Table 9). For each significant association, we then characterized the implicated transcriptional programs by evaluating 10,325 gene sets from MSigDB22,24 (URLs) for enrichment among the genomic regions driving the association (controlling for LD and co-localizing genes; Methods) (Table 1 and Supplementary Table 10).

Our analysis yielded 77 significant annotation-trait associations at per-trait FDR <5%, spanning six diseases and complex traits (Fig. 5 and Supplementary Table 9a) and representing 12 independent transcription factor–trait associations (after pruning correlated annotations; Table 1 and Supplementary Note). Our results were 4.3× enriched for autoimmune disease associations (Supplementary Note). We verified empirically that our results are not driven by directional effects of minor alleles (Supplementary Table 9b and Supplementary Note), and we computed a lower bound on the number of independent transcription factor binding sites contributing to each association (74 on average; Table 1, Supplementary Fig. 8, and Methods).

Of our 12 independent transcription factor–trait associations, five refine emerging theories of disease while seven are previously unknown. Due to space restrictions, we highlight two relationships from each category (Fig. 6 and Supplementary Table 11; Fig. 7 and Supplementary Table 12), providing discussion of additional relationships in the Supplementary Note. We begin with the transcription factor–trait associations that build on previous knowledge (Fig. 6). First, we detected a positive association of the GWAS locus (Supplementary Table 13). Our result suggests that BCL11A causes intellectual disability not via regulation of a few key disease genes but rather via binding throughout the genome causing modulation (in cis) of genes comprising a broad transcriptional program relevant to brain function or development (see Discussion). Furthermore, our MSigDB gene-set enrichment analysis allows us to characterize this putative transcriptional program as being significantly enriched for genes involved in mTOR signaling and in cholesterol metabolism (Fig. 6a and Supplementary Table 10). mTOR is an intellectual disability gene with links to cholesterol defects in brain cholesterol metabolism have been linked to central nervous system disease, and BCL11A has also been linked to lipid levels. These observations raise the possibility that mTOR causes intellectual disability by interacting with BCL11A to regulate cholesterol metabolism in the developing brain (Supplementary Note).
Fig. 4 | Analysis of GTEx eQTL using signed LD profile regression. We plot polarized $-\log_{10}(P)$ values for all significant associations as a heatmap. Columns denote the 36 GTEx tissues (of 48 GTEx tissues tested) with significant associations. Rows denote the 67 transcription factors (of 75 transcription factors tested) with significant associations, collapsing all annotations corresponding to a single transcription factor into one row and displaying in each cell the most significant result. Cells with dots indicate associations that show robust evidence for tissue-specificity in our conditional analysis (see main text and Methods). Cells indicated in outline correspond to associations described in the main text, with dashed outline indicating known associations and solid outline indicating previously unknown associations or associations supporting emerging theories. GWAS data are described in Supplementary Table 6, and the statistical method and multiple comparisons adjustments are described in the Methods. Numerical results are reported in Supplementary Table 7.
Table 1 | Independent transcription factor–trait associations from analysis of diseases and complex traits using signed LD profile regression

| Trait                 | Top transcription factor (No.) | $r_f$  | $P$          | $q$       | Minimum no. of sites | Top 2 significant MSigDB enrichments                                                                 |
|-----------------------|--------------------------------|--------|--------------|-----------|----------------------|-------------------------------------------------------------------------------------------------------|
| Years of education    | BCL11A (1)                     | 2.4%   | $3.9 \times 10^{-5}$ | $1.5 \times 10^{-2}$ | 104                  | • Cholesterol homeostasis  
  • ↑ on mTOR inhibition  
  • ↓ on immunosuppression  
  • regulation of reproductive process  
  • ↑ on mTOR inhibition  
  • Androgen response  
  • Regulated by NF-κB in response to TNF |
| Crohn’s               | POL2* (20)                     | 5.3%   | $4.8 \times 10^{-5}$ | $1.5 \times 10^{-2}$ | 74                   | • ↑ BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Anorexia              | SP1 (1)                        | −8.9%  | $1.1 \times 10^{-4}$ | $4.0 \times 10^{-2}$ | 30                   | • ↑ on BCL6 knockout  
  • ↓ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| HDL                   | FOS (1)                        | 4.8%   | $1.2 \times 10^{-4}$ | $4.6 \times 10^{-2}$ | 19                   | • ↑ BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Eczea                 | CTCF (12)                      | 2.7%   | $1.4 \times 10^{-4}$ | $3.4 \times 10^{-2}$ | 106                  | • ↑ on BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Crohn’s               | ELF1 (1)                       | 4.9%   | $1.6 \times 10^{-4}$ | $1.5 \times 10^{-2}$ | 58                   | • ↑ on BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Crohn’s               | POL2 (1)                       | 4.4%   | $2.6 \times 10^{-4}$ | $1.5 \times 10^{-2}$ | 50                   | • ↑ on BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Lupus                 | CTCP* (36)                     | −5.0%  | $3.6 \times 10^{-4}$ | $4.4 \times 10^{-2}$ | 100                  | • ↑ on BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Crohn’s               | TBP (1)                        | 5.4%   | $4.9 \times 10^{-4}$ | $1.5 \times 10^{-2}$ | 54                   | • ↑ on BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Crohn’s               | E2F1 (1)                       | 4.3%   | $6.4 \times 10^{-4}$ | $2.7 \times 10^{-2}$ | 90                   | • ↑ on BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Crohn’s               | IRF1 (1)                       | 4.7%   | $9.8 \times 10^{-4}$ | $1.5 \times 10^{-2}$ | 90                   | • ↑ on BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of NF-κB |
| Crohn’s               | ETS1 (1)                       | 6.1%   | $1.4 \times 10^{-3}$ | $1.5 \times 10^{-2}$ | 114                  | • ↑ on BCL6 knockout  
  • ↑ on IL21 stimulation  
  • ↓ on PPARγ activation  
  • Transcription co-repressor activity  
  • ↓ in fibroblast early serum response  
  • ↓ on ALK knockdown  
  • Targets of MYC |

For each of 12 independent associations at per-trait FDR < 5% after pruning correlated annotations ($R^2 > 0.25$), we report the associated trait, the transcription factor of the most significant annotation and the number of correlated annotations with significant associations; the estimated functional correlation $r_f$, $P$ value, $q$ value, and minimum number of transcription factor binding sites contributing to the association; and the top two significant MSigDB gene-set enrichments among loci driving the association. Linked transcription factors producing significant associations: *TAFI, TBP; *RAD21. See Supplementary Table 10 for full gene set names and enrichment $q$ values (all $< 5 \times 10^{-5}$). GWAS data are described in Supplementary Table 8, gene set data are described in the Methods, and the statistical method and multiple comparisons adjustments are described in the Methods. LMPF, lymphoid-primedpluripotent progenitor; GMP, granulocyte-monocyte monopoecuss.

Fig. 5 | Analysis of diseases and complex traits using signed LD profile regression. For each disease or complex trait with at least one significant result, we plot $-\log_{10}(P)$ against estimated effect size for each of the 382 annotations analyzed. Points are colored by transcription factor identity, with transcription factors with no significant associations for each trait colored in gray. Larger points denote significant results. The number of significant results for each trait is: Crohn’s disease, 26; lupus, 36; years of education, 1; eczea, 12; HDL, 1; anorexia, 1. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the Methods. Numerical results are reported in Supplementary Table 9a.
Fig. 6 | Highlighted transcription factor binding-complex trait associations that refine emerging theories of disease. a, b, For each of BCL11A-Years of education (a) and IRF1-Crohn’s disease (b), we display plots of the marginal correlation $\alpha$ of SNP to trait versus the signed LD profile $R_v$ of the annotation in question, with SNPs averaged in bins of 4,000 SNPs with similar $R_v$ values and a larger bin around $R_v = 0$; Manhattan plots of the trait GWAS signal near the associated transcription factor; and the top two significant MSigDB gene-set enrichments among the loci driving the association, with error bars indicating standard errors. GWAS data are described in Supplementary Table 8, gene set data are described in the Methods, and the statistical method and multiple comparisons adjustments are described in the Methods. Numerical results are reported in Supplementary Table 11.

Second, we detected a positive association between genome-wide binding of interferon regulatory factor 1 (IRF1) and Crohn’s disease (Fig. 6b), a case in which existing GWAS evidence has been suggestive but not conclusive. Although IRF1 lies in a locus associated with Crohn’s disease in multiple GWAS54–76 (one of the earliest Crohn’s disease associations74), this locus remains mysterious. Strong LD makes it challenging to determine which variant(s) is causal, and high gene density (23 protein-coding genes within 500 kb of IRF1) complicates the task of determining which gene is affected by any putative causal variant, resulting in several genes56,78,79 being previously nominated as potentially causal. For example, a recent large-scale fine-mapping study76 narrowed the causal signal to eight SNPs including rs2188962, an eQTL for SLC22A5 in immune and gut epithelial cells52,80 but also for IRF1 in blood33. Transcriptome-wide association studies have also been inconclusive51–83. Our result provides genome-wide evidence for a causal link between IRF1 and Crohn’s disease that, unlike single-locus approaches, is not fundamentally limited by LD and pleiotropy near the IRF1 gene (see Discussion). The top results in our MSigDB gene-set enrichment analysis strengthen our finding: the regions driving this association are most significantly enriched for genes involved in production of type I interferon and regulation of nuclear division (Fig. 6b and Supplementary Table 10), matching well-known roles of IRF184,85. We note that several other transcription factor–trait associations from our analysis implicate causal genes at established GWAS loci, including ELF1-Crohn’s disease and ETS1-Crohn’s disease, with...
gene-set enrichments suggesting connections to existing Crohn’s disease drugs and to the role of autophagy in Crohn’s disease pathogenesis, respectively (Table 1 and Supplementary Note).

We next discuss two selected transcription factor–trait associations that were previously unknown (Fig. 7). First, we detected a positive association between genome-wide binding of CTCF and eczema (Fig. 7a). We do not observe a GWAS signal for eczema at the CTCF locus. This could be because the CTCF gene is under strong selective constraint (probability of loss-of-function intolerance =1.00, greater than 99.9% of genes) and highlights the potential of our method to uncover causal roles for genes that harbor relatively little variation. The top two significant MSigDB gene-set enrichments for CTCF-eczema are convergent: genes upregulated in Treg cells on knockout of the inflammatory regulator BCL6, and genes upregulated in response to stimulation by the immune signaling molecule IL21, which is a known regulator of BCL6 activity (Fig. 7a and Supplementary Table 10). These enrichments, because they pertain to genes putatively regulated in cis by CTCF to cause eczema, suggest a detailed cascade that we hypothesize to modulate eczema risk. IL21 signaling regulates BCL6, which in turn regulates CTCF to activate a broad transcriptional program that increases eczema risk. This hypothesis makes three predictions: it predicts that BCL6 modulates CTCF activity, and it predicts that IL21 and BCL6 each affect eczema risk. Indeed, we determined that BCL6 has many binding sites near the CTCF promoter in publicly available ChIP-seq data (Supplementary Table 14), and the IL21 and BCL6 genes each fall in eczema GWAS loci (in each case along with seven other protein-coding genes within 500 kb). Thus, the association between CTCF binding and eczema that we detected nominate causal genes at two different existing eczema GWAS loci and provides a parsimonious mechanism that explains their effect on eczema via a regulatory cascade that activates a CTCF-mediated transcriptional program.

Second, we detected a negative association between genome-wide binding of SP1 and anorexia (Fig. 7b), a heritable trait for which no loci reach genome-wide significance in our GWAS data. SP1 levels observationally correlate negatively with psychiatric conditions such as schizophrenia (which is positively genetically correlated with anorexia), but this association has not been shown to be causal and has not previously been observed in GWAS of psychiatric traits. Our MSigDB gene-set enrichment results for this association yielded significant enrichments for an androgen response gene set and an mTOR signaling gene set (Fig. 7b and Supplementary Table 10). (Years of education, for which an mTOR signaling gene set was also among the top two MSigDB enrichments, is also significantly positively genetically correlated with anorexia, the median rank of the top-scoring mTOR gene set across the 10 other independent transcription factor–complex trait associations was 1,123, of 10,325.) The androgen response result is intriguing given the sex-imbalanced nature of this phenotype. The mTOR signaling result is noteworthy given the well-established connections between mTOR, caloric restriction, and growth; it also suggests that a link between SP1 and mTOR could explain prior observations tying SP1 to insulin, appetite, and energy metabolism. mTOR has also been shown to play an important role in androgen signaling, suggesting a potential unification of these two signals.

We provide additional discussion of other transcription factor–trait associations in the Supplementary Note (Supplementary Fig. 9 and Supplementary Tables 15 and 16).

Discussion
We have introduced a method, signed LD profile regression, for identifying genome-wide directional effects of signed functional annotations on diseases and complex traits. Our approach allows us to draw fine-grained biological conclusions that are not confounded by simple genomic co-localization of functional elements. The directional relationships we identify concretely implicate broad disease-relevant transcriptional programs. Our characterization of these programs via gene-set enrichment analyses yields detailed hypotheses about disease mechanisms that in several cases mechanistically link existing GWAS loci and disparate molecular evidence into a parsimonious mechanism mediated by the associated transcription factor.

Our method differs from unsigned GWAS enrichment methods by assessing whether a systematic genome-wide correlation exists between a signed functional annotation and the (signed) true causal effects of SNPs on disease, rather than assessing whether a set of SNPs have large effects on a disease without regard to the directions of those effects. Our method also differs from single-locus GWAS methods in that a consistent genome-wide directional effect across a large set of transcription factor binding sites (Table 1) is less susceptible to pleiotropy, LD, and allelic heterogeneity. Finally, our method differs from genetic correlation and Mendelian randomization analyses, which can be confounded by reverse causality and pleiotropic effects; in contrast, the sequence-based nature of our annotations makes them ideal instrumental variables for the effect of transcription factor binding on the trait of interest (Supplementary Note).

The genome-wide nature of our method means that our results constitute instances in which transcription factors affect traits via coordinated regulation of gene expression throughout the genome (‘a genome-wide model’) rather than via regulation of one or a small number of key disease genes (‘a local model’). This distinction has potential implications for drug development as well as attempts to elucidate disease mechanisms (Supplementary Note). For example, as we have shown, the genome-wide nature of the putative transcriptional programs identified by our method allows us to characterize and interpret these programs by aligning them with existing gene sets, leading in some cases to detailed mechanistic hypotheses.

There exist many potentially effective methods for constructing signed transcription factor binding annotations and many potential data sets on which to train them. We present an initial exploration of alternative annotations generated using some of these, along with a discussion of potential signed annotations besides transcription factor binding annotations, in the Supplementary Note (Supplementary Figs. 10–14 and Supplementary Tables 1–20).

We note several limitations of signed LD profile regression. First, although our results are less susceptible to confounding due to their signed nature, they are not immune to it: in particular, our method cannot distinguish between two transcription factors that are close binding partners and thus share sequence motifs, and it likewise cannot distinguish between binding of the same transcription factor in different cell types, as the resulting annotations could be highly correlated. Second, we used annotations constructed using data from cell lines, which is non-ideal because chromatin dynamics in cell lines do not necessarily match those in real tissue; we note, however, that although this reduces our power and the effect sizes we see, it does not introduce false positives into our results. Third, the interpretability of our MSigDB gene-set enrichment analysis is limited by the potential for distinct gene sets to have overlapping membership and for co-expressed genes to be included in the same gene sets; however, we believe this is somewhat ameliorated by the fact that we treat blocks of genes together in our empirical null (Methods). Due to space restrictions, additional limitations are discussed in the Supplementary Note.

Despite these limitations, signed LD profile regression is a powerful new way to leverage functional genomics data to draw mechanistic conclusions from GWAS about both diseases and underlying cellular processes.
115. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

116. Venkataraman, A. et al. A toolbox of immunoprecipitation-grade monoclonal antibodies to human transcription factors. *Nat. Methods* (2018).

**Acknowledgements**

We thank C. de Boer, L. Dicker, J. Engreitz, T. Finucane, N. Friedman, R. Gumpert, M. Kanai, S. Kim, X. Liu, M. Mitzenmacher, J. Perry, S. Reilly, D. Reshef, S. Raychaudhuri, A. Schoech, P. Saberi, R. Tewhey, O. Troyanskaya, P. Turley, O. Weissbrod, J. Zhou, and the CGTA discussion group for helpful discussions. This research was conducted using the UK Biobank Resource under Application No. 16549 and was supported by US National Institutes of Health grants No. U01 HG009379, R01 MH101244, R01 MH109978, and R01 MH107649. Y.A.R. was supported by award No. T32GM007753 from the National Institute of General Medical Sciences, the National Defense Science and Engineering Graduate Fellowship, and the Paul and Daisy Soros Foundation. H.K.F. was supported by the Fannie and John Hertz Foundation and by Eric and Wendy Schmidt. F.H. is supported by National Institute of Health award No. T32 DK110919. L.P. is supported by National Institutes of Health award No. R00HG008399. R.P.A. is supported by NSF grant No. IIS-1421780. Computational analyses were performed on the Orchestra High Performance Compute Cluster at Harvard Medical School, which is partially supported by grant No. NCRR 1S10RR028832-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

**Author contributions**

Y.A.R. and A.L.P. designed the study. Y.A.R., H.K.F., D.R.K., A.G., F.H., J.N., and P.-R.L. analyzed data. Y.A.R. and A.L.P. wrote the manuscript with assistance from H.K.F., D.R.K., A.G., D.K., J.C.U., F.H., J.N., L.O., B.v.d.G., P.-R.L., S.R.G., G.B., S.G., P.F.P., L.P., N.P., and R.P.A.

**Competing interests**

D.R.K. is employed by the Calico Life Sciences LLC. The rest of the authors declare no competing interests.

**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0196-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to Y.A.R. or A.L.P.

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

Signed LD profile regression. Intuition. Our method for quantifying directional effects of signed functional annotations on disease risk, signed LD profile regression, relies on the following intuition. Suppose there are M SNPs and we are given a signed functional annotation, specified by a length-M vector \( v \), with a directional linear effect on disease risk. For example, \( v \) might be a vector whose \( m \)th entry is the effect of SNP \( m \) on binding of some transcription factor. If we knew the length-M vector \( \beta \) of the true causal effects of the same SNPs on a trait, we could simply regress \( \beta \) on \( v \) to evaluate whether there is a non-trivial signed association across SNPs \( m \) between \( v_m \) and \( \beta_m \). In reality, we cannot do this because we do not observe \( \beta \); instead we observe a vector, denoted \( \delta \), of GWAS summary statistics describing the marginal correlation of every SNP to our trait of interest. This vector differs from \( \beta \) because it includes both causal and tagging effects, plus statistical noise. Specifically, it can be shown mathematically that, in expectation, \( \delta \) will equal the matrix-vector product \( R\beta \) where \( R \) is the M × M LD matrix. Therefore, just as \( \beta \) would be proportional to \( v \) in the presence of a signed effect, \( \delta(\hat{R}\beta) \) would likewise be proportional to \( Rv \), which is a vector capturing each SNP's aggregate tagging of the signed annotation. This means that instead of regressing \( \beta \) on \( v \) (which is impossible since we do not observe \( \beta \)), we can regress \( \delta \) on \( Rv \). We call the vector \( Rv \) the signed LD profile of \( v \), and thus our method is called signed LD profile regression. The remainder of our technical material is oriented toward: (i) weighting this regression to achieve optimal power; (ii) being able to efficiently perform the required computations; (iii) determining the proper way to test the null hypothesis of no signed effect; and (iv) controlling for potential confounding due to directional effects of minor alleles.

Model and estimates. Let \( M \) be the number of SNPs in the genome. We assume a linear model:

\[ y | \beta, \sigma^2 \sim N(\beta^T x, \sigma^2 I_n) \]

where \( y \in \mathbb{R}^n \) and \( x \in \mathbb{R}^n \) are the standardized genotype vector and phenotype, respectively, of a randomly chosen individual from some population, \( \beta \in \mathbb{R}^n \) is a vector of true causal effects of each SNP on phenotype, and \( \sigma^2 \) represents environmental noise. Given a signed functional annotation \( v \in \mathbb{R}^n \), we then model

\[ \beta | v \sim [\mu_v, \sigma_v^2] \]

with the scalar \( \mu \) represents the genome-wide directional effect of \( v \) on \( \beta \), \( \sigma^2 \) represents other sources of heritability unrelated to \( v \), and the notation \([\cdot, \cdot]\) is used to specify the mean and covariance of the distribution without specifying any higher moments.

Although we can estimate \( \mu \), its value depends on the units of the annotation and the heritability of the trait. Because of this, we focus instead on the functional correlation \( r_v \), which re-scales \( \mu \) to be dimensionless and is defined as

\[ r_v := \text{corr}(x^T \beta, x^T \nu) = \frac{\nu^T Rv}{k_v} \]

where \( k_v \) is \( \text{var}(x^T \beta) \) the SNP heritability of the phenotype and \( R = E(xx^T) \in \mathbb{R}^{n\times n} \) is the (signed) population LD matrix of the genotypes. The quantity \( r_v \) can be interpreted as a form of genetic correlation; the value of \( r_v \) cannot exceed the proportion of SNP heritability explained by SNPs with non-zero values of \( v \). (Note that \( r_v \) can also be defined as a correlation between \( \beta \) and \( v \); this definition is approximately equivalent in expectation under our random effects model, provided \( \nu^T v \approx |v|^2 \).) We additionally estimate \( k_v = \sigma^2 / \hat{\sigma}_v^2 \), the total phenotypic variance explained by the signed contribution of \( v \) to \( \beta \), as well as \( \hat{k}_v = k_v / k_v^2 = r_v \).

For annotations with small support, these quantities are expected to be small. For more detail on derivations and computational considerations, see the Supplementary Note.

Null hypothesis testing. To test the null hypothesis \( H_0: \mu = 0 \) (or, equivalently, \( H_0: \hat{R} = 0 \)), we permute the genome to approximately 300 blocks of approximately the same size with the block boundaries constrained to fall on estimated recombination hotspots\(^{121} \). We then define the null distribution of our statistic as the distribution arising from independently multiplying \( v \) by one independent random sign per block. We perform this empirical sign-flipping many times to obtain an approximation of the null distribution and corresponding \( P \) values.

Controlling for covariates and the signed background model. Given a signed covariate \( u \in \mathbb{R}^n \), we can perform inference on the signed effect of \( v \) conditional on \( u \) by first regressing \( Ru \) out of \( u \) and \( Rv \) using the generalized least-squares method outlined above, and then proceeding as usual with the residuals of \( u \) and \( Rv \).

Unless stated otherwise, all analyses in this paper control in this fashion for a 'signed background model' consisting of five annotations \( u', \ldots, u'_{\text{IV}} \), defined by

\[ u_m' := 1 \left\{ \text{MAF}_m < 0.05 \right\} \left( \frac{1 - \text{MAF}_m}{2} \right) \]

where \( \text{MAF}_m \) is the minor allele frequency of SNP \( m \) and \( \alpha \) is a parameter describing the MAF dependence of the signed effect of minor alleles on phenotype. Based on the literature on MAF dependence of the unsigned effects \( \text{var}(\beta) \), we set \( \alpha = -0.3118 \).

382 transcription factor annotations. Briefly, we constructed the annotations by training a sequence-based neural network predictor of ChIP-seq peak calls, using the Basset software\(^{1} \), to predict the results of 382 transcription factor binding ChIP-seq experiments from ENCODE\(^{119} \) and comparing the neural network's predictions for the major and minor allele of each SNP in the ChIP-seq peaks.

Specifically, we downloaded every ChIP-seq and DNase I hypersensitivity experiment in ENCODE and trained the Basset model to jointly predict each downloaded track on a set of held-out genomic segments. (We included tracks other than transcription factor binding tracks because training predictions using all tracks slightly improved prediction accuracy for the transcription factor binding tracks.) After training the joint predictor, we retained the predictions for every transcription factor binding track for which (i) the number of SNPs in that track's ChIP-seq peaks with non-zero difference in Basset predictions between the major and minor allele was at least 5,000 in our 1,000 G reference panel, and (ii) Basset's estimated area under the precision-recall curve was at least 0.3. This yielded a set of 382 transcription factor ChIP-seq experiments that spanned 75 distinct transcription factors and 84 distinct cell lines. For each experiment, we constructed an annotation via

\[ v_m = \mathbb{1}(m \in C)(P_m - P_m^m) \]

where \( C \) is the set of SNPs in the ChIP-seq peaks arising from the experiment, \( P_m \) is the Basset prediction for the 1,000 base-pair sequence around SNP \( m \) when the minor allele is placed at SNP \( m \), \( P_m^m \) is the Basset prediction for the 1,000 base-pair sequence around SNP \( m \) when the major allele is placed at SNP \( m \). (We always used the minor allele as the reference allele in both our transcription factor binding annotations and our GWAS summary statistics.)

Simulations. All simulations were carried out using real genotypes of individuals with European ancestry from the GERA cohort\(^{1} \) (n = 47,360). The set of \( M = 2.7 \) million causal SNPs was defined as the set of genome-wide very well imputed SNPs (INFO ≥ 0.97) that had very low missingness (<0.5%) and non-negligible MAF.
Null simulations. For the simulations in Fig. 1a, we simulated 1,000 independent null phenotypes with the architecture $\beta_1^{null} = N(0, \sigma^2)$ with $\sigma^2 = b_0^{null}/M$ and $b_0^{null} = 0.5$. For each phenotype, we computed GWAS summary statistics using plink2122 (URLs), adjusting for three principal components as well as GERA chip type as covariates. For each of our 382 transcription factor annotations, we then ran SLDP regression on each of these 1,000 phenotypes, yielding a set of 382,000 $P$ values. (Supplementary Fig. 1a shows an analysis of the $P$ value distribution for each annotation individually, confirming correct calibration for these annotations.)

For the simulations in Fig. 1b, we simulated 1,000 independent traits in which each trait had an unsigned enrichment for a randomly chosen annotation: after choosing an annotation $v$, we set $\beta_{1,v}^{null} = N(\mu_{null}, \sigma^2)$ where $\mu_{null}$ is non-zero if SNP $m$ is in the bottom quintile of the MAF spectrum of the GERA sample and 0 otherwise, as in the signed background model. We set $\mu_{null}$ such that 10% of heritability would be explained by this directional effect, and then set $\sigma^2$ to achieve $b_0^{null} = 0.5$. We then computed summary statistics as above and ran SLDP regression to assess $v$ for a genome-wide directional effect. This procedure yielded 1,000 $P$ values.

For the simulations in Fig. 1c, we simulated 1,000 independent phenotypes with a directional effect of minor alleles: we set $\beta_{1,m}^{null} = N(0, \sigma^2)$ with $\sigma^2 = b_0^{null}/M$ and $b_0^{null} = 0.5$ in each case. For each value of $v$, we simulated 100 independent traits, computed summary statistics using plink2, and then ran each of the methods under consideration using the annotation $v$. In addition to the findings stated in the main text, our simulations also show that the power of our method increases with sample size and SNP heritability (Supplementary Fig. 4), and is only minimally affected by within-Europe reference panel mismatch (Supplementary Fig. 5).

Analysis of molecular traits in blood. We downloaded BLUEPRINT consortium GTQ data for gene expression, H3K4me1, H3K27ac, and methylation in three different blood cell types with sample sizes of $n = 158$, 165, and 125 for monocytes, neutrophils, and T cells, respectively19 (Supplementary Table 4 and URLs). For each of the three gene expression traits, we constructed one summary statistics vector $\mathbf{g}$ by meta-analyzing, for each SNP, the marginal effect sizes of that SNP for the expression of all nearby genes. Specifically, we set

$$g_{m} = \frac{1}{\sqrt{\sigma_{m}}} \sum_{c \in C_m} \tilde{g}_{m,c}^{(1)}$$

where $C_m$ is the set of all genes within 500kb of SNP $m$, and $\tilde{g}_{m,c}^{(1)}$ is the marginal correlation of SNP $m$ to the expression of gene $k$. Assuming independence of expression across genes this is analogous to a fixed-effects meta-analysis across genes at every SNP to determine that SNP’s effect on aggregate expression, although our results do not rely on this theoretical characterization because of the empirical, signed nature of our null hypothesis testing procedure. Since in practice gene expression is not independent across genes, the scale of the resulting vector $\mathbf{g}$ is arbitrary. Therefore, we placed all such vectors on the same scale by scaling them so that they have an estimated SNP heritability of 0.5. (This only affects the regression weights used by SLDP regression.) Applying the same procedure to the two histone marks and to methylation in addition to gene expression yielded 12 sets of summary statistics (Supplementary Table 4).

We ran SLDP regression using each of our 382 transcription factor annotations for each of these 12 traits. We obtained results at $FDR < 5\%$ using the Benjamini–Hochberg procedure20–22 within each of the 12 traits and reported the union of significant results across cell types for each trait. We determined the top 100 associations to display in Fig. 3a by choosing the significant associations with the highest estimated values of $\mathbf{g}$.

For our replication analysis, we used expression array-based whole blood eQTL data from the NTR18, which we obtained by downloading the set of TWAS weights14 computed for that data set (Supplementary Table 4 and URLs). We then proceeded as above. 196 of the 409 BLUEPRINT gene expression associations replicated (same direction of effect with nominal $P < 0.05$). We note, however, that because TWAS weights were only available for genes with a significantlyheritable cis-expression in NTR, we only had data for 2,454 genes compared with 15,023–17,081 genes for the BLUEPRINT traits, thereby lowering our power in this analysis.

Comparison to UniProt annotations. For each transcription factor represented in our annotations, we queried the UniProt database23 to establish whether the transcription factor was: (unambiguously) ‘activating’, defined as all transcription factors annotated as having activating activity but not repressing activity in UniProt; (unambiguously) repressing, defined as all transcription factors with repressing activity but not activating activity; or ‘ambiguous’, defined as all transcription factors with both activating and repressing activity. 78% and 82% of our positive associations in the BLUEPRINT eQTLs and chromatin QTLs, respectively, were unambiguous activators. The set of significant positive SLDP associations for eQTLs/chromatin QTLs were enriched for (unambiguously) ‘activating’ transcription factors compared to the set of annotations as a whole ($P = 7.9 \times 10^{-4}$ QTL results and $P = 1.9 \times 10^{-5}$ for chromatin QTL results). For additional details, see Supplementary Note.

Analysis of gene expression across 48 GTEx tissues. We downloaded GTEx v7 eQTLs for all 48 tissues for which data were available and processed them using the same procedure described for the blood molecular traits, resulting in one vector of summary statistics per GTEx tissue (Supplementary Table 6 and URLs). We ran SLDP regression using each of our 382 transcription factor annotations for each of these tissues. We obtained results at $FDR < 5\%$ using the Benjamini–Hochberg procedure20–22 within each of the 48 tissues.

Conditional analysis for tissue-specific effects. We obtained a set of eQTL summary statistics for a fixed-effect meta-analysis across the GTEx tissues126 and processed these via the procedure described above into a single vector $\mathbf{a}_{Ti}$. For each tissue $t$, we then residualized $\mathbf{a}_{Ti}$ out of the vector $\mathbf{a}_{T}$ of eQTL data for tissue $t$ to obtain a residualized vector $\mathbf{a}_{Tt}$. This amounts to subtracting a scalar multiple of $\mathbf{a}_{Tt}$ from $\mathbf{a}_{T}$, with the scalar determined to remove as much signal as possible from $\mathbf{a}_{Tt}$. For each significant association between an annotation $v$ and a vector $\mathbf{a}_{T}$ from our main GTEx analysis, we then compared the $P$ value of the association for the annotated eQTL to the residualized vector $\mathbf{a}_{Tt}$, declaring as tissue-specific any association for which the latter was at least as significant as the former. For additional details, see Supplementary Note.

This criterion for tissue-specificity is conservative and stands in contrast to, for example, reporting associations that remain significant at a specified threshold after conditioning. The latter approach is susceptible to the fact that conditioning on a noisily measured confounder can produce false positives125; associations meeting the former criterion are likely to be robustly tissue-specific.

Assessment for concordance with absolute expression levels in GTEx tissues. Briefly, we assessed whether the proportion of significant transcription factor associations in which the transcription factor was expressed above a minimum threshold in the associated GTEx tissues was greater than the corresponding proportion for non-significant transcription factors. This held in 32 out of the 34 tissues for which we could perform the comparison ($P = 2.1 \times 10^{-7}$ for trend across tissues; see Supplementary Fig. 7 for breakdown by tissue.) For additional details, see Supplementary Note.

Analysis of 46 diseases and complex traits. We applied SLDP regression to 46 diseases and complex traits with an average sample size of 289,617, including 16 traits with publicly available summary statistics and 30 UK Biobank traits for which we have published supplementary summary statistics computed using HOFT-LMM v2.3 (Supplementary Table 8 and URLs). We ran SLDP regression using each of our 382 transcription factor annotations for each of these traits. We obtained results at per-trait $FDR < 5\%$ using the Benjamini–Hochberg procedure20–22. We report as significant results at a per-trait $FDR < 5\%$, following standard practice. However, when many traits are analyzed, per-trait FDR control does not imply global FDR control, and we estimate the global FDR of our results to be $9.4\%$ (Supplementary Note).

MSigDB gene-set enrichment analysis of results on diseases and complex traits. We also downloaded a set of LD blocks in Europeans derived from estimated recombination hotspots17 and converted each gene set into a length-1693 vector $\mathbf{g}$ whose entry equaled the number of genes from the set that are present in the ith LD block. We then converted each significant SLDP regression association between an annotation $v$ and a trait summary statistics vector $\mathbf{a}$ into a length-1693 vector $q$ whose ith entry equaled the covariance between $\mathbf{a}$ and the signed LD profile $R_v$ within the ith LD block. To assess the SLDP result for...
enrichment of a gene-set vector $s$, we computed a weighted mean of the $q_i$ whose weights were given by $s$. That is, we computed

$$a(v, \hat{\alpha}, s) = \frac{\sum q_i}{\sum s_i}.$$

The idea is that if the LD blocks in which $s$ is large correspond to the LD blocks in which the SLDP regression signal is the strongest, the weighted mean $a$ should be large in magnitude and have the same sign as the overall SLDP regression association. We assess this via an empirical null distribution constructed by permuting the LD blocks to obtain ‘shuffled’ versions of $s$ and $q$. This enrichment method is more conservative than ordinary gene-set enrichment methods for two reasons. First, by permuting only LD blocks and not genes, it accounts for correlations induced by LD as well as co-regulation of nearby genes and gene overlap in the genome. Second, because a significant SLDP regression association cannot arise as a result of a strong signal in only one genomic location, this method is more robust to outliers and cannot, for example, produce a rejection simply because of a very strong signal at just one gene. In comparison to gene-set enrichment methods for GWAS data, this method also has the advantage that it will not cause gene sets containing large genes to produce signals of enrichment. Separately from null hypothesis testing, we computed heuristic standard errors for use in Figs. 6 and 7 by computing the closed-form standard deviation of $a(v, \hat{\alpha}, s)$ assuming that the $s_i$ are fixed and the $q_i$ are i.i.d.

To quantify effect size, we computed a fold-enrichment by dividing $\frac{\sum q_i}{\sum s_i}$ by the average value of $q$ at LD blocks containing no genes. That is the enrichment is defined as

$$e(v, \hat{\alpha}, s) = \frac{a(v, \hat{\alpha}, s)}{\text{mean}(q; s_i = 0)}.$$

This quantity $e$ is the number reported in Figs. 6 and 7.

We conducted our hypothesis test for gene-set enrichment for each of our 77 significant transcription factor–complex trait associations against each of the 10,325 MSigDB gene sets. For every transcription factor–complex trait association and every tranche of gene-sets from MSigDB, we assessed significance under the null hypothesis by computing the closed-form standard deviation of $a(v, \hat{\alpha}, s)$ assuming that the $s_i$ are fixed and the $q_i$ are i.i.d.

$$\text{max}(2S_t - K_t)$$

The intuition is that the distribution of the signs of the entries of $q$ can be modeled as a mixture of a uniform distribution (for genomic chunks with no signal) and a distribution with all of its mass on the sign of the genome-wide trend (for genomic chunks with signal). The number of entries drawn from the latter distribution gives the number of independent genomic blocks contributing to the association, which is a lower bound on the number of independent transcription factor binding sites contributing to the association. Estimating this number naively without thresholding yields the expression $2S_t - K_t$. However, this is an underestimate in the presence of noise in $q$. We therefore repeat this argument considering only the subset of entries of $q$ with magnitude at least $t$ for a small number of thresholds $t$ and retain the largest estimate.

### Data availability

We have released all genome annotations we analyzed, as well as regression weight matrices for our 1000 Genomes reference panel, at http://data.broadinstitute.org/alkesgroup/SLDP/.

### References

117. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 32, 283–288 (2016).
118. Schoech, A. et al. Quantification of frequency-dependent genetic architectures and action of negative selection in 25 UK Biobank traits. bioRxiv 188086 (2017).
119. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
120. Banda, Y. et al. Characterizing race/ethnicity and genetic ancestry for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort. Genetics 200, 1285–1295 (2015).
121. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).
122. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, 7 (2015).
123. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol 57, 289–300 (1995).
124. Hormozdiari, F. et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. Nat. Genet. 50, 1041–1047 (2017).
125. Carroll, R. J. Measurement Error in Epidemiologic Studies. in Wiley StatsRef: Statistics Reference Online (Wiley, Hoboken, NJ, USA, 2014).
126. Lambert, S. A. et al. The human transcription factors. Cell 172, 650–665 (2018).
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

- [ ] n/a
- [x] Confirmed

[ ] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

[ ] An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

[ ] The statistical test(s) used AND whether they are one- or two-sided

Only common tests should be described solely by name; describe more complex techniques in the Methods section.

[ ] A description of all covariates tested

[ ] A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

[ ] A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

[ ] For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted

Give P values as exact values whenever suitable.

[ ] For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

[ ] For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

[ ] Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

[ ] Clearly defined error bars

State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection

No software was used to collect data, as no data were generated.

Data analysis

We used the SLOP package, available on GitHub. We also ran Basset and Plink2, using the 2017 versions of each tool.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No data were generated in this study.
Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Sample size | We analyzed available GWAS data, and did not do a new experiment in which we determined the sample size. |
|-------------|-------------------------------------------------------------------------------------------------------|
| Data exclusions | We excluded the HLA from all analyses and analyzed only autosomes. We excluded signed functional annotations for which Basset AUPRC was <0.3 or for which <5000 reference panel SNPs had non-zero effects. Otherwise, no data were excluded. |
| Replication | There were no experimental findings. Where possible, we validated computational results using gene expression data, orthogonal GWAS signals, and gene sets from MSigDB. |
| Randomization | We did not allocate samples into experimental groups. |
| Blinding | There was no group allocation. |

Reporting for specific materials, systems and methods

Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☒ Unique biological materials |
| ☒ Antibodies |
| ☒ Eukaryotic cell lines |
| ☒ Palaeontology |
| ☒ Animals and other organisms |
| ☒ Human research participants |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☒ ChiP-seq |
| ☒ Flow cytometry |
| ☒ MRI-based neuroimaging |