nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0196-7

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

Yakir A. Reshef[1,2,3]*, Hilary K. Finucane[3], David R. Kelley[4], Alexander Gusev[5], Dylan Kotliar[3], Jacob C. Ulirsch[3,5,6], Farhad Hormozdiari[3,7], Joseph Nasser[3], Luke O'Connor[7,8], Bryce van de Geijn[7], Po-Ru Loh[3,9], Sharon R. Grossman[2,3], Gaurav Bhatia[7], Steven Gazal[7], Pier Francesco Palamara[3,7,10], Luca Pinello[3,11,12], Nick Patterson[3], Ryan P. Adams[13,14] and Alkes L Price[3,7,15]*

Biological interpretation of genome-wide association study data frequently involves assessing whether SNPs linked to a biological process, for example, binding of a transcription factor, show unsigned enrichment for disease signal. However, signed annotations quantifying whether each SNP allele promotes or hinders the biological process can enable stronger statements about disease mechanism. We introduce a method, signed linkage disequilibrium profile regression, for detecting genome-wide directional effects of signed functional annotations on disease risk. We validate the method via simulations and application to molecular quantitative trait loci in blood, recovering known transcriptional regulators. We apply the method to expression quantitative trait loci in 48 Genotype-Tissue Expression tissues, identifying 651 transcription factor-tissue associations including 30 with robust evidence of tissue specificity. We apply the method to 46 diseases and complex traits (average $n = 290$ K), identifying 77 annotation-trait associations representing 12 independent transcription factor-trait associations, and characterize the underlying transcriptional programs using gene-set enrichment analyses. Our results implicate new causal disease genes and new disease mechanisms.

Mechanistic interpretation of genome-wide association study (GWAS) data has become a central challenge for understanding the biological underpinnings of disease. One successful paradigm for such efforts has been GWAS enrichment, in which a genome annotation containing SNPs that affect some biological process is shown to be enriched for GWAS signals[1–7]. However, there are instances in which experimental data allow us not only to identify SNPs that affect a biological process, but also to predict which SNP alleles promote the process and which SNP alleles hinder it, enabling us to assess whether there is a systematic association between SNP alleles' direction of effect on the process and their direction of effect on a trait. Transcription factor binding, which plays a major role in human disease[1,8–12], represents an important case in which such signed functional annotations are available: because transcription factors have a tendency to bind to specific DNA sequences, it is possible to estimate whether the sequence change introduced by a SNP allele will increase or decrease binding of a transcription factor[1,13–19].

Detecting genome-wide directional effects of transcription factor binding on disease would constitute an important advance in terms of both evidence for causality and understanding of biological mechanism. Regarding causality, this is because directional effects are not confounded by simple co-localization in the genome (for example, of transcription factor binding sites with other regulatory elements), and thus provide stronger evidence for causality than is available using unsigned enrichment methods. Regarding biological mechanism, it is currently unknown whether disease-associated transcription factors regulate only a few key disease genes or whether broad transcriptional programs comprising many target genes are responsible for transcription factor associations; a genome-wide directional effect implies the latter model (see Discussion).

Here we introduce a new method, signed linkage disequilibrium profile (SLDP) regression, for quantifying the genome-wide directional effect of a signed functional annotation on polygenic disease risk, and apply it with 382 annotations each reflecting predicted binding of a particular transcription factor. Our method requires only GWAS summary statistics[20], accounts for linkage disequilibrium and untyped causal SNPs, and is computationally efficient. We validate the method via extensive simulations and further validate it by applying it to molecular quantitative trait loci (QTLs) in blood[21], recovering known transcriptional regulators. We then apply the method to expression quantitative trait loci (eQTL) in 48 tissues from the Genotype-Tissue Expression (GTEx) consortium[22] and to 46 diseases and complex traits, demonstrating genome-wide directional effects of transcription factor binding in both settings. We

[1]Department of Computer Science, Harvard University, Cambridge, MA, USA. [2]Harvard/MIT MD/PhD Program, Boston, MA, USA. [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]California Life Sciences LLC, South San Francisco, CA, USA. [5]Dana Farber Cancer Institute, Boston, MA, USA. [6]Boston Children's Hospital, Boston, MA, USA. [7]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]Program in Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA, USA. [9]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [10]Department of Statistics, University of Oxford, Oxford, UK. [11]Massachusetts General Hospital, Charlestown, MA, USA. [12]Department of Pathology, Harvard Medical School, Boston, MA, USA. [13]Google Brain, New York, NY, USA. [14]Department of Computer Science, Princeton University, Princeton, NJ, USA. [15]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. *e-mail: yakirr@mit.edu; aprice@hsph.harvard.edu

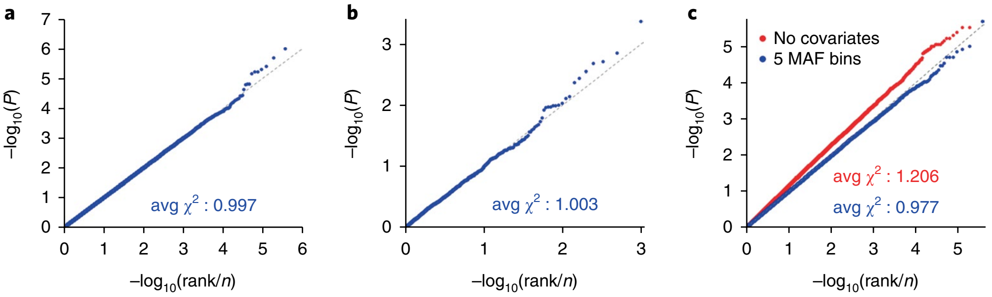

Fig. 1 | Simulations assessing null calibration. a–c, We report null calibration (*q*–*q* plots of two-sided SLDP regression −log$_{10}$(*P*) values) in simulations of no enrichment (**a**), unsigned enrichment (**b**), and directional effects of minor alleles (**c**). The *q*–*q* plots are based on 382 annotations × 1,000 simulations = 382,000 (**a**), 1,000 (**b**), and two sets of 382 × 1,000 = 382,000 (**c**) *P* values. A 5-MAF-bin signed background model is included in all cases except for the red points in **c**, which are computed with no covariates. We also report the average χ^2-statistic corresponding to each set of *P* values. Numerical results are reported in Supplementary Table 2.

further characterize the transcriptional programs underlying our complex trait associations via gene-set enrichment analyses using gene sets from the Molecular Signatures Database (MSigDB)[23,24].

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[25] in which true causal SNP effects are correlated with a signed functional annotation, we show that

$$E(\widehat{\alpha} \mid v) = r_f \sqrt{h_g^2}\, Rv \qquad (1)$$

where $\widehat{\alpha}$ 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_g^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_g^2, allows us to estimate r_f by regressing $\widehat{\alpha}$ on the signed LD profile Rv 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 $\widehat{\alpha}$ and quantifies directional effects, in contrast to stratified LD score regression[5], which analyzes unsigned χ^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[19], 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[5,13] (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[5].

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[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),

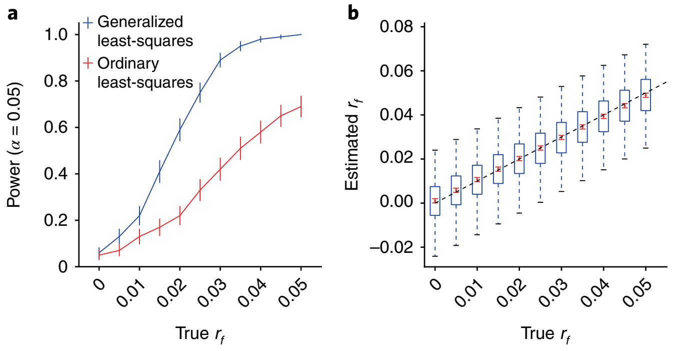

Fig. 2 | Simulations assessing power, bias, and variance. a, Power curves comparing SLDP regression using generalized least-squares (that is, weighting) to an ordinary (that is, unweighted) regression of the summary statistics on the signed LD profile. Error bars indicate standard errors of power estimates. **b**, Assessment of bias and variance of the SLDP regression estimate of r_f, across a range of values of the true r_f. 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.

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 untyped 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).

Analysis of molecular traits in blood. Transcription factor binding is known to affect gene expression and other molecular traits[27], and regulatory relationships in blood are particularly well characterized[28]. 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[21] (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[29] (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[30–32]. We obtained similar results in an independent set of whole-blood eQTLs based on expression array experiments from the Netherlands Twin Registry (NTR)[33] (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[29]. The four negative associations involved MAFK and MAFF, both of which lack a transactivation domain[34], and CTCF, which is known to act as an insulator[35,36]. Many of the detected associations recover known aspects of histone mark biology[36–44] and match a prior analysis of allelic imbalance in chromatin immunoprecipitation sequencing (ChIP-seq) data[45].

Analysis of gene expression across 48 GTEx tissues. We next applied SLDP regression to GTEx eQTL across 48 tissues[22] (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[37–39], early B-cell factor 1 (EBF-1) in lymphocytes[40,41], hepatocyte nuclear factors 4γ and 4α (HNF4G and HNF4A) in liver[42,43], PU.1 in spleen[44], and FOS in fibroblasts[45] and nerve tissue[46–48]. 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[49]. 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[50], and TAF1 was recently shown to be the causal gene in a rare form of Parkinsonism[51]. 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[52], it is expressed in pancreatic beta cells specifically under metabolic stress[52]. 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[53] (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[35,36,54], 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[55–58] has been theorized via detailed experimental work to mediate enhancer–promoter interaction[59]. However, *YY1*

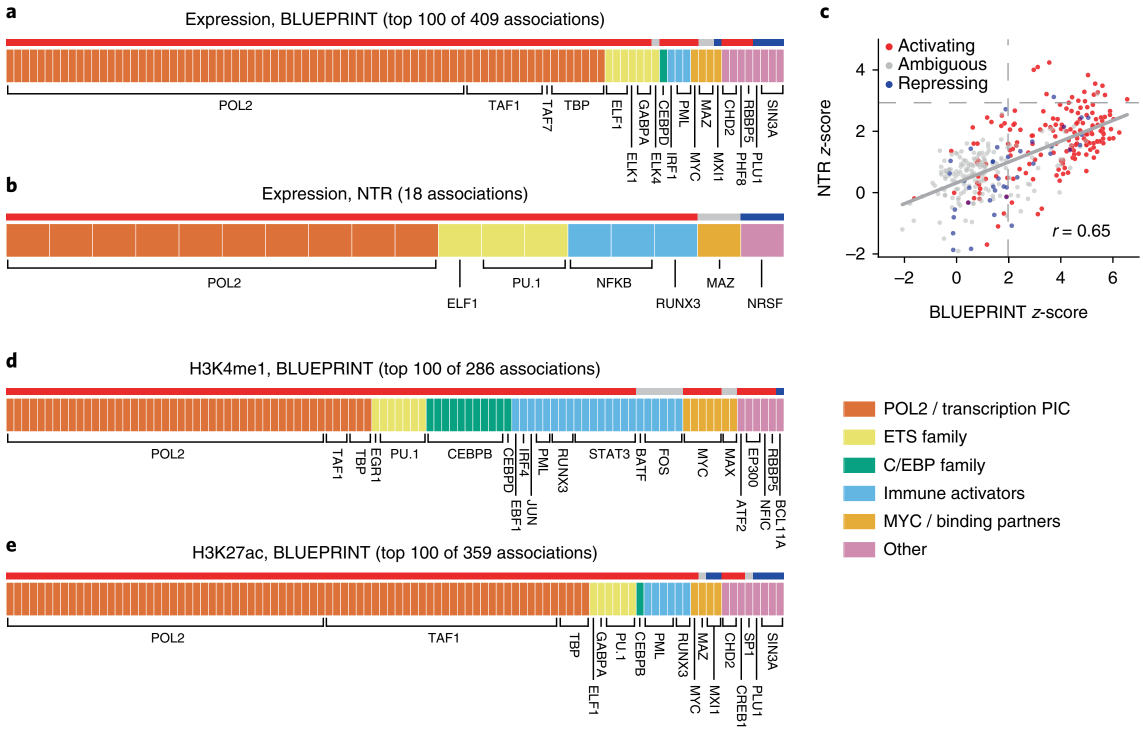

Fig. 3 | Analysis of blood molecular traits using signed LD profile regression. a–e, Each segmented bar in **a**, **b**, **d**, and **e** represents the set of significant annotations (or top 100 annotations) at a per-trait FDR < 5% for the indicated traits, with each annotation corresponding to a particular transcription factor profiled in a particular cell line. Results in **a**, **d**, and **e** are aggregated across the three BLUEPRINT cell types. The stripe above each segmented bar is colored red for transcription factors with activating activity and no repressing activity in UniProt (see main text and Methods), blue for transcription factors with repressing activity and no activating activity, and gray for ambiguous transcription factors. **c**, *Z*-scores from the analyses of expression in the NTR data set and neutrophil expression in the BLUEPRINT data set, respectively, for each of the 382 annotations tested; red, blue, and gray again indicate UniProt (unambiguously) activating transcription factors, (unambiguously) repressing transcription factors, and ambiguous transcription factors, respectively. Dashed lines represent significance thresholds for 5% FDR. GWAS data are described in Supplementary Table 4, and the statistical method and multiple comparisons adjustments are described in the Methods. Numerical results are reported in Supplementary Table 5.

knockdown experiments have shown a mix of upregulation and downregulation of many genes[59], 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 MSigDB[23,24] (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 between genome-wide binding of BCL11A and years of education (Fig. 6a) that aligns with existing evidence from educational attainment GWAS[60], rare-variant studies of intellectual disability[61–64], and experimental knockout work in mice[64], as well as our fine-mapping of the *BCL11A* 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[65,66] with links to cholesterol[67,68], defects in brain cholesterol metabolism have been linked to central nervous system disease[69,70], and BCL11A has also been linked to lipid levels[71–73]. These observations raise the possibility that mTOR causes intellectual disability by interacting with BCL11A to regulate cholesterol metabolism in the developing brain (Supplementary Note).

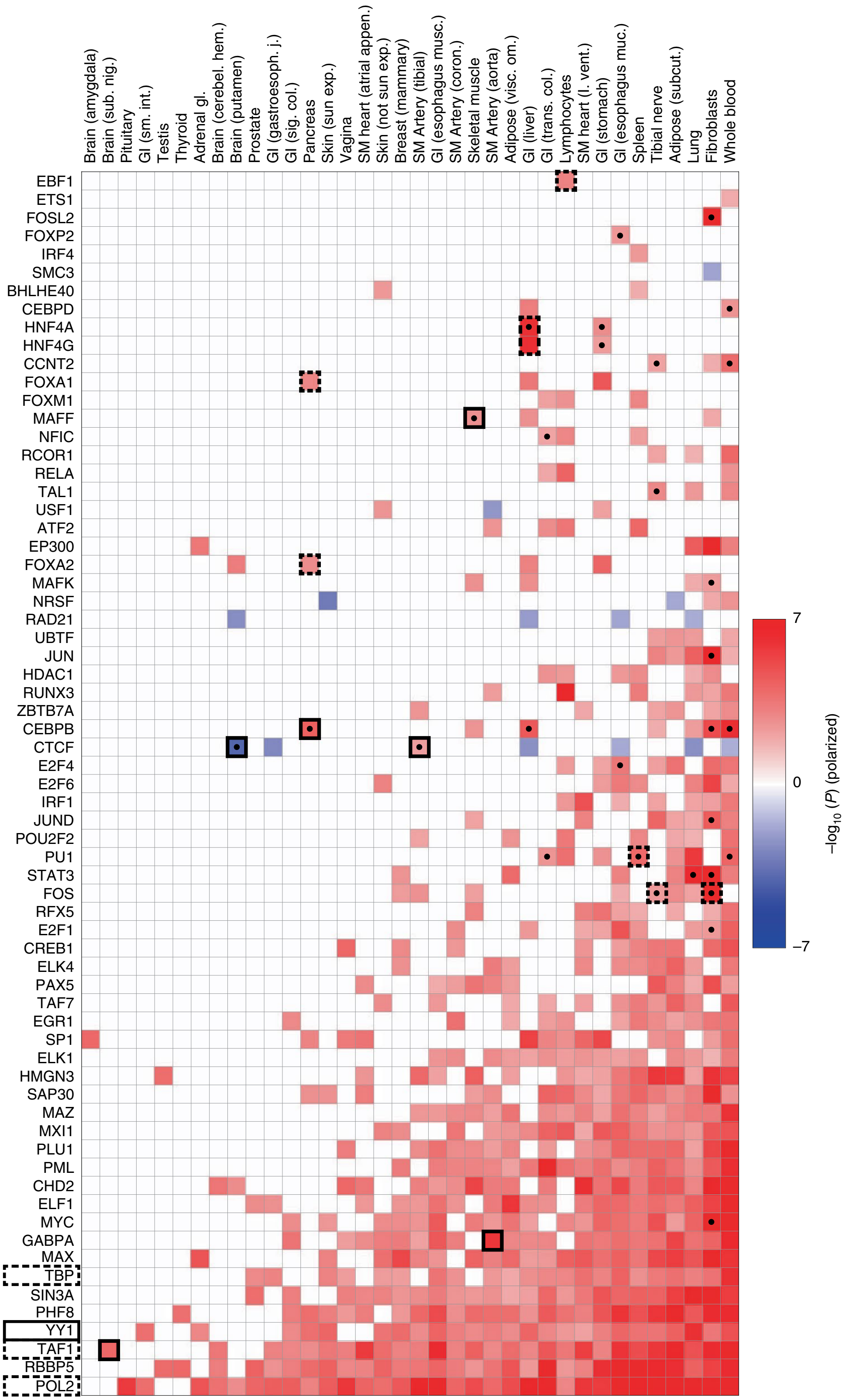

Fig. 4 | Analysis of GTEx eQTL using signed LD profile regression. We plot polarized $-\log_{10}(P)$ values for all significant associations as a heatmap. Columns denote the 36 GTEx tissues (of 48 GTEx tissues tested) with significant associations. Rows denote the 67 transcription factors (of 75 transcription 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%	3.9×10^{-5}	1.5×10^{-2}	104	• Cholesterol homeostasis • ↑ on mTOR inhibition
Crohn's	POL2[a] (20)	5.3%	4.8×10^{-5}	1.5×10^{-2}	74	• ↓ on immunosuppression • regulation of reproductive process
Anorexia	SP1 (1)	−8.9%	1.1×10^{-4}	4.0×10^{-2}	30	• ↑ on mTOR inhibition • Androgen response
HDL	FOS (1)	4.8%	1.2×10^{-4}	4.6×10^{-2}	19	• Regulated by NF-κB in response to TNF -
Eczema	CTCF (12)	2.7%	1.4×10^{-4}	3.4×10^{-2}	106	• ↑ on *BCL6* knockout • ↑ on IL21 stimulation
Crohn's	ELF1 (1)	4.9%	1.6×10^{-4}	1.5×10^{-2}	58	• ↓ on PPARγ activation • Transcription co-repressor activity
Crohn's	POL2 (1)	4.4%	2.6×10^{-4}	1.5×10^{-2}	50	• ↓ in fibroblast early serum response • ↓ on *ALK* knockdown
Lupus	CTCF[b] (36)	−5.0%	3.6×10^{-4}	4.4×10^{-2}	100	• Targets of NF-κB • ↓ in LMPP versus GMP cells on *IKZF1* knockout
Crohn's	TBP (1)	5.4%	4.9×10^{-4}	1.5×10^{-2}	54	• Late estrogen response -
Crohn's	E2F1 (1)	4.3%	6.4×10^{-4}	2.7×10^{-2}	90	• Cancer module 323 (immune) • Targets of *miR-17-3p*
Crohn's	IRF1 (1)	4.7%	9.8×10^{-4}	1.5×10^{-2}	90	• Regulation of nuclear division • Regulation of type I interferon production
Crohn's	ETS1 (1)	6.1%	1.4×10^{-3}	1.5×10^{-2}	114	• Neighborhood of autophagy-associated EI24 • Targets of MYC

For each of 12 independent associations at per-trait FDR < 5% after pruning correlated annotations ($R^2 \geq 0.25$), 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: [a]TAF1, TBP; [b]RAD21. See Supplementary Table 10 for full gene set names and enrichment *q* values (all $<5\times10^{-2}$). 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. LMPP, lymphoid-primed pluripotent progenitor; GMP, granulocyte-monocyte precursor.

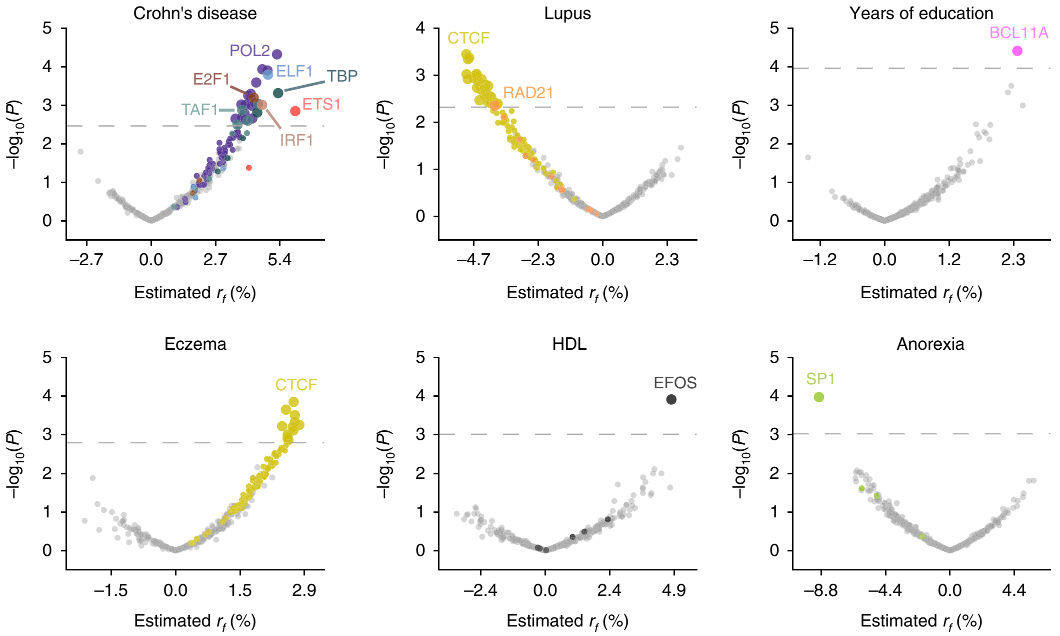

Fig. 5 | Analysis of diseases and complex traits using signed LD profile regression. For each disease or complex trait with at least one significant result, we plot $-\log_{10}(P)$ against estimated effect size for each of the 382 annotations analyzed. Points are colored by transcription factor identity, 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; eczema, 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.

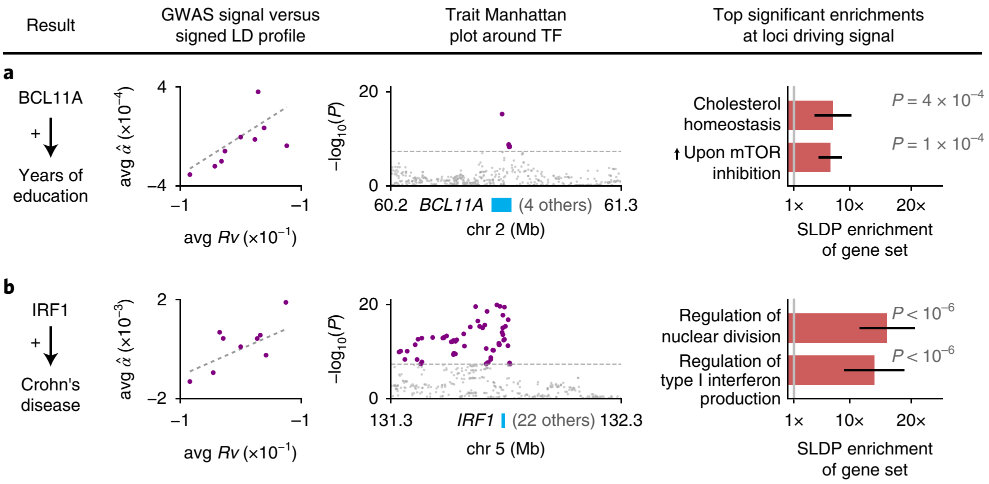

Fig. 6 | Highlighted transcription factor binding-complex trait associations that refine emerging theories of disease. a,**b**, For each of BCL11A-Years of education (**a**) and IRF1-Crohn's disease (**b**), we display plots of the marginal correlation $\hat{\alpha}$ of SNP to trait versus the signed LD profile Rv of the annotation in question, with SNPs averaged in bins of 4,000 SNPs with similar Rv values and a larger bin around $Rv = 0$; 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.

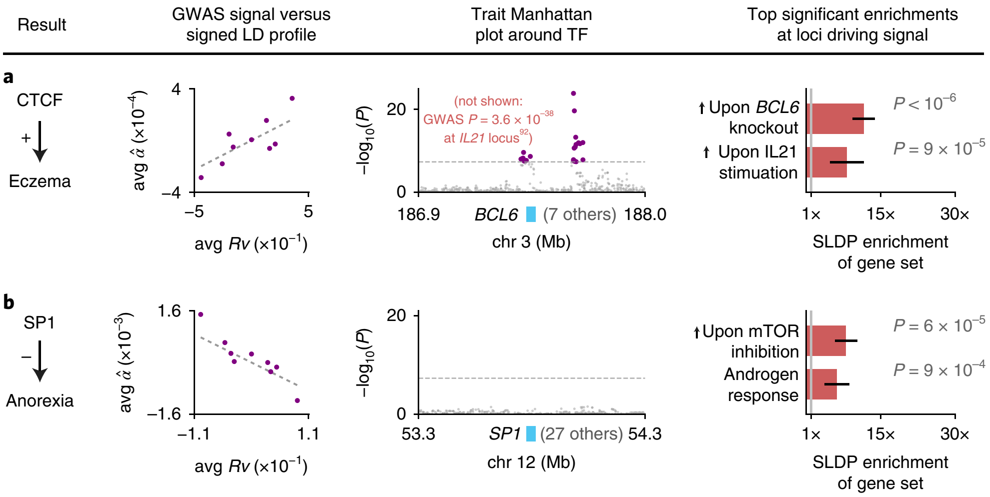

Fig. 7 | Highlighted previously unknown transcription factor binding-complex trait associations. a,**b**, For each of CTCF-Eczema (**a**) and SP1-Anorexia (**b**), we display plots of the marginal correlation $\hat{\alpha}$ of SNP to trait versus the signed LD profile Rv of the annotation in question, with SNPs averaged in bins of 4,000 SNPs with similar Rv values and a larger bin around $Rv = 0$; Manhattan plots of the trait GWAS signal near the associated transcription factor or, in the case of CTCF-Eczema, the *BCL6* gene (see main text; there is no GWAS peak at *CTCF*); 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 12.

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 GWAS[74–76] (one of the earliest Crohn's disease associations[77]), this locus remains mysterious. Strong LD makes it challenging to determine which variant(s) are 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 genes[74,78,79] being previously nominated as potentially causal. For example, a recent large-scale fine-mapping study[80] narrowed the causal signal to eight SNPs including rs2188962, an eQTL for *SLC22A5* in immune and gut epithelial cells[22,80] but also for *IRF1* in blood[33]. Transcriptome-wide association studies have also been inconclusive[81–83]. 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 IRF1[84,85]. 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[86] =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 T_{reg} 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[87,88] (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[69,89–91] (Supplementary Table 14), and the *IL21* and *BCL6* genes each fall in eczema GWAS loci[92–94] (in each case along with seven other protein-coding genes within 500 kb). Thus, the association between CTCF binding and eczema that we detected nominates 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[95]. SP1 levels observationally correlate negatively with psychiatric conditions such as schizophrenia[96,97] (which is positively genetically correlated with anorexia[98]), 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[98]; 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[99]. The mTOR signaling result is noteworthy given the well-established connections between mTOR, caloric restriction, and growth[100]; it also suggests that a link between SP1 and mTOR could explain prior observations tying SP1 to insulin[101,102], appetite[103,104], and energy metabolism[105]. mTOR has also been shown to play an important role in androgen signaling[106], 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[1–7] 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[11,12,81] 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[81,82]. Finally, our method differs from genetic correlation and Mendelian randomization[98] analyses, which can be confounded by reverse causality and pleiotropic effects[107–109]; 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[110] (a 'genome-wide' model) rather than via regulation of one or a small number of key disease genes[111] (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[1,13–16,18,112,113] and many potential data sets on which to train them[114–116]. 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.

URLs. Signed LD profile regression, open-source software is available at http://www.github.com/yakirr/sldp. Plink2, https://www.cog-genomics.org/plink2/. BLUEPRINT consortium data, ftp://ftp.ebi.ac.uk/pub/databases/blueprint/blueprint_Epivar/qtl_as/QTL_RESULTS/Transciptome-wide association study (TWAS) weights for Netherlands Twin Registry (NTR) data, https://data.broadinstitute.org/alkesgroup/FUSION/WGT/NTR.BLOOD.RNAARR.tar.bz2. GTEx eQTL data, https://www.gtexportal.org/home/datasets. MSigDB data, http://software.broadinstitute.org/gsea/msigdb. GTRD data, http://gtrd.biouml.org/. HOCOMOCO motif data, http://hocomoco11.autosome.ru/

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0196-7.

Received: 19 October 2017; Accepted: 11 July 2018;
Published online: 3 September 2018

References

1. Cowper-Sal lari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
2. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
4. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
5. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
6. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
7. Zhu, X. & Stephens, M. A large-scale genome-wide enrichment analysis identifies new trait-associated genes, pathways and tissues across 31 human phenotypes. *bioRxiv* 160770 (2017).
8. Karczewski, K. J. et al. Systematic functional regulatory assessment of disease-associated variants. *Proc. Natl Acad. Sci., USA* **110**, 9607–9612 (2013).
9. Mathelier, A., Shi, W. & Wasserman, W. W. Identification of altered cis-regulatory elements in human disease. *Trends Genet.* **31**, 67–76 (2015).
10. Price, A. L., Spencer, C. C. A. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proc. R. Soc. B* **282**, 20151684 (2015).
11. Whitington, T. et al. Gene regulatory mechanisms underpinning prostate cancer susceptibility. *Nat. Genet.* **48**, 387–397 (2016).
12. Liu, Y. et al. Identification of breast cancer associated variants that modulate transcription factor binding. *PLoS Genet.* **13**, e1006761 (2017).
13. Pique-Regi, R. et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).
14. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
15. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
16. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
17. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
18. Zeng, H., Hashimoto, T., Kang, D. D. & Gifford, D. K. GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* **32**, 490–496 (2016).
19. Kelley, D. R., Snoek, J. & Rinn, J. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
20. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
21. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
22. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
23. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci., USA* **102**, 15545–15550 (2005).
24. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
25. Yang, W. et al. Genome-wide association study in Asian populations identifies variants in *ETS1* and *WDFY4* associated with systemic lupus erythematosus. *PLoS Genet.* **6**, e1000841 (2010).
26. Arbiza, L. et al. Genome-wide inference of natural selection on human transcription factor binding sites. *Nat. Genet.* **45**, 723–729 (2013).
27. Ernst, J. et al. Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).
28. Bodine, D. M. Introduction to a review series on transcription factors in hematopoiesis and hematologic disease. *Blood* **129**, 2039 (2017).
29. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
30. Sharrocks, A. D., Brown, A. L., Ling, Y. & Yates, P. R. The ETS-domain transcription factor family. *Int. J. Biochem. Cell Biol.* **29**, 1371–1387 (1997).
31. Kimura, T. et al. Involvement of the IRF-1 transcription factor in antiviral responses to interferons. *Science* **264**, 1921–1924 (1994).
32. Kakizuka, A. et al. Chromosomal translocation t(15;17) in human acute promyelocytic leukemia fuses RARα with a novel putative transcription factor, PML. *Cell* **66**, 663–674 (1991).
33. Wright, F. A. et al. Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
34. Friedman, J. S. et al. The minimal transactivation domain of the basic motif-leucine zipper transcription factor NRL interacts with TATA-binding protein. *J. Biol. Chem.* **279**, 47233–47241 (2004).
35. Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387–396 (1999).
36. Xie, X. et al. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci., USA* **104**, 7145–7150 (2007).
37. Gao, N. et al. Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. *Genes Dev.* **22**, 3435–3448 (2008).
38. Song, Y., Washington, M. K. & Crawford, H. C. Loss of FOXA1/2 is essential for the epithelial-to-mesenchymal transition in pancreatic cancer. *Cancer Res.* **70**, 2115–2125 (2010).
39. Gao, N. et al. Foxa1 and Foxa2 maintain the metabolic and secretory features of the mature β-cell. *Mol. Endocrinol.* **24**, 1594–1604 (2010).
40. Hagman, J., Ramírez, J. & Lukin, K. B lymphocyte lineage specification, commitment and epigenetic control of transcription by early B cell factor 1. *Curr. Top. Microbiol. Immunol.* **356**, 17–38 (2012).
41. Somasundaram, R., Prasad, M. A. J., Ungerbäck, J. & Sigvardsson, M. Transcription factor networks in B-cell differentiation link development to acute lymphoid leukemia. *Blood* **126**, 144–152 (2015).
42. Odom, D. T. et al. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**, 1378–1381 (2004).
43. Bonzo, J. A., Ferry, C. H., Matsubara, T., Kim, J.-H. & Gonzalez, F. J. Suppression of hepatocyte proliferation by hepatocyte nuclear factor 4α in adult mice. *J. Biol. Chem.* **287**, 7345–7356 (2012).
44. Wolff, L. & Ruscetti, S. The spleen focus-forming virus (SFFV) envelope gene, when introduced into mice in the absence of other SFFV genes, induces acute erythroleukemia. *J. Virol.* **62**, 2158–2163 (1988).
45. Angel, P. E. & Herrlich, P. *The FOS and JUN Families of Transcription Factors*. (CRC Press, Boca Raton, FL, USA 1994).
46. Bullitt, E. Expression of c-fos-like protein as a marker for neuronal activity following noxious stimulation in the rat. *J. Comp. Neurol.* **296**, 517–530 (1990).
47. Velazquez, F. N. et al. Brain development is impaired in c-fos -/- mice. *Oncotarget* **6**, 16883–16901 (2015).
48. Zhang, J. et al. c-*fos* regulates neuronal excitability and survival. *Nat. Genet.* **30**, 416–420 (2002).
49. Nischan, J. et al. Binding sites for ETS family of transcription factors dominate the promoter regions of differentially expressed genes in abdominal aortic aneurysms. *Circ. Genomic Precis. Med.* **2**, 565–572 (2009).
50. Triarhou, L. C. *Dopamine and Parkinson's Disease*. (Landes Bioscience, Austin, TX, USA, 2013).
51. Aneichyk, T. et al. Dissecting the causal mechanism of X-linked dystonia-parkinsonism by integrating genome and transcriptome assembly. *Cell* **172**, 897–909.e21 (2018).
52. Davis, F. P. & Eddy, S. R. Transcription factors that convert adult cell identity are differentially Polycomb repressed. *PLoS One* **8**, e63407 (2013).
53. Popov, D. V., Lysenko, E. A., Makhnovskii, P. A., Kurochkina, N. S. & Vinogradova, O. L. Regulation of *PPARGC1A* gene expression in trained and untrained human skeletal muscle. *Physiol. Rep.* **5**, e13543 (2017).
54. Kim, S., Yu, N.-K. & Kaang, B.-K. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.* **47**, e166 (2015).

55. Kleiman, E., Jia, H., Loguercio, S., Su, A. I. & Feeney, A. J. YY1 plays an essential role at all stages of B-cell differentiation. *Proc. Natl Acad. Sci., USA* **113**, E3911–E3920 (2016).
56. Hwang, S. S. et al. YY1 inhibits differentiation and function of regulatory T cells by blocking Foxp3 expression and activity. *Nat. Commun.* **7**, 10789 (2016).
57. Kwon, H.-K., Chen, H.-M., Mathis, D. & Benoist, C. Different molecular complexes that mediate transcriptional induction and repression by FoxP3. *Nat. Immunol.* **18**, 1238–1248 (2017).
58. Gabriele, M. et al. YY1 haploinsufficiency causes an intellectual disability syndrome featuring transcriptional and chromatin dysfunction. *Am. J. Hum. Genet.* **100**, 907–925 (2017).
59. Weintraub, A. S. et al. YY1 is a structural regulator of enhancer-promoter loops. *Cell* **171**, 1573–1588.e28 (2017).
60. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
61. Basak, A. et al. *BCL11A* deletions result in fetal hemoglobin persistence and neurodevelopmental alterations. *J. Clin. Invest.* **125**, 2363–2368 (2015).
62. Funnell, A. P. W. et al. 2p15-p16.1 microdeletions encompassing and proximal to *BCL11A* are associated with elevated HbF in addition to neurologic impairment. *Blood* **126**, 89–93 (2015).
63. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
64. Dias, C. et al. BCL11A haploinsufficiency causes an intellectual disability syndrome and dysregulates transcription. *Am. J. Hum. Genet.* **99**, 253–274 (2016).
65. Lipton, J. O. & Sahin, M. The neurology of mTOR. *Neuron* **84**, 275–291 (2014).
66. Reijnders, M. R. F. et al. Variation in a range of mTOR-related genes associates with intracranial volume and intellectual disability. *Nat. Commun.* **8**, 1052 (2017).
67. Laplante, M. & Sabatini, D. M. An emerging role of mTOR in lipid biosynthesis. *Curr. Biol.* **19**, R1046–R1052 (2009).
68. Mathews, E. S. & Appel, B. Cholesterol biosynthesis supports myelin gene expression and axon ensheathment through modulation of P13K/Akt/mTor signaling. *J. Neurosci.* **36**, 7628–7639 (2016).
69. Koudinov, A. R. & Koudinova, N. V. Cholesterol homeostasis failure as a unifying cause of synaptic degeneration. *J. Neurol. Sci.* **229**, 233–240 (2005).
70. Zhang, J. & Liu, Q. Cholesterol metabolism and homeostasis in the brain. *Protein Cell* **6**, 254–264 (2015).
71. Macari, E. R., Schaeffer, E. K., West, R. J. & Lowrey, C. H. Simvastatin and t-butylhydroquinone suppress *KLF1* and *BCL11A* gene expression and additively increase fetal hemoglobin in primary human erythroid cells. *Blood* **121**, 830–839 (2013).
72. TANG, L. et al. *BCL11A* gene DNA methylation contributes to the risk of type 2 diabetes in males. *Exp. Ther. Med.* **8**, 459–463 (2014).
73. Li, S. et al. Transcription factor CTIP1/BCL11A regulates epidermal differentiation and lipid metabolism during skin development. *Sci. Rep.* **7**, 13427 (2017).
74. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
75. Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
76. Lange, K. Mde et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
77. Rioux, J. D. et al. Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* **29**, 223–228 (2001).
78. Silverberg, M. S. OCTNs: Will the real IBD5 gene please stand up? *World J. Gastroenterol.* **12**, 3678–3681 (2006).
79. Brant, S. R. IBD5: the second Crohn's disease gene? *Inflamm. Bowel Dis.* **8**, 371–372 (2002).
80. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
81. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
82. Wainberg, M. *et al.* Vulnerabilities of transcriptome-wide association studies. *bioRxiv* 206961 (2017).
83. Mancuso, N. et al. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
84. Romeo, G. et al. IRF-1 as a negative regulator of cell proliferation. *J. Interferon Cytokine Res.* **22**, 39–47 (2002).
85. Honda, K., Takaoka, A. & Taniguchi, T. Type I interferon gene induction by the interferon regulatory factor family of transcription factors. *Immunity* **25**, 349–360 (2006).
86. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
87. Linterman, M. A. et al. IL-21 acts directly on B cells to regulate Bcl-6 expression and germinal center responses. *J. Exp. Med.* **207**, 353–363 (2010).
88. Chevrier, S., Kratina, T., Emslie, D., Tarlinton, D. M. & Corcoran, L. M. IL4 and IL21 cooperate to induce the high Bcl6 protein level required for germinal center formation. *Immunol. Cell Biol.* **95**, 925–932 (2017).
89. Hurtz, C. et al. BCL6-mediated repression of p53 is critical for leukemia stem cell survival in chronic myeloid leukemia. *J. Exp. Med.* **208**, 2163–2174 (2011).
90. Hatzi, K. et al. A hybrid mechanism of action for BCL6 in B cells defined by formation of functionally distinct complexes at enhancers and promoters. *Cell Rep.* **4**, 578–588 (2013).
91. Huang, C., Hatzi, K. & Melnick, A. Lineage-specific functions of Bcl-6 in immunity and inflammation are mediated by distinct biochemical mechanisms. *Nat. Immunol.* **14**, 380–388 (2013).
92. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* (2018).
93. Ek, W. E., Rask-Andersen, M., Karlsson, T. & Johansson, A. Genome-wide association analysis identifies 26 novel loci for asthma, hay fever and eczema. *bioRxiv* 195933 (2017).
94. Portelli, M. A., Hodge, E. & Sayers, I. Genetic risk factors for the development of allergic disease identified by genome-wide association. *Clin. Exp. Allergy* **45**, 21–31 (2015).
95. Boraska, V. et al. A genome-wide association study of anorexia nervosa. *Mol. Psychiatry* **19**, 1085–1094 (2014).
96. Ben-Shachar, D. & Karry, R. Sp1 expression is disrupted in schizophrenia; a possible mechanism for the abnormal expression of mitochondrial complex I genes, *NDUFV1* and *NDUFV2*. *PLoS One* **2**, e817 (2007).
97. Fusté, M. et al. Reduced expression of SP1 and SP4 transcription factors in peripheral blood mononuclear cells in first-episode psychosis. *J. Psychiatr. Res.* **47**, 1608–1614 (2013).
98. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
99. Striegel-Moore, R. H. et al. Gender difference in the prevalence of eating disorder symptoms. *Int. J. Eat. Disord.* **42**, 471–474 (2009).
100. Colman, R. J. et al. Caloric restriction delays disease onset and mortality in rhesus monkeys. *Science* **325**, 201–204 (2009).
101. Pan, X., Solomon, S. S., Borromeo, D. M., Martinez-Hernandez, A. & Raghow, R. Insulin deprivation leads to deficiency of Sp1 transcription factor in H-411E hepatoma cells and in streptozotocin-induced diabetic ketoacidosis in the rat. *Endocrinology* **142**, 1635–1642 (2001).
102. Yasui, D., Peedicayil, J. & Grayson, D. R. *Neuropsychiatric Disorders and Epigenetics*. (Academic Press, Cambridge, MA, USA 2016).
103. Zhang, X. et al. Hypermethylation of Sp1 binding site suppresses hypothalamic POMC in neonates and may contribute to metabolic disorders in adults: impact of maternal dietary CLAs. *Diabetes* **63**, 1475–1487 (2014).
104. Yang, G. et al. FoxO1 inhibits leptin regulation of pro-opiomelanocortin promoter activity by blocking STAT3 interaction with specificity protein 1. *J. Biol. Chem.* **284**, 3719–3727 (2009).
105. Moreno-Aliaga, M. J. et al. Sp1-mediated transcription is involved in the induction of leptin by insulin-stimulated glucose metabolism. *J. Mol. Endocrinol.* **38**, 537–546 (2007).
106. Audet-Walsh, É. et al. Nuclear mTOR acts as a transcriptional integrator of the androgen signaling pathway in prostate cancer. *Genes Dev.* **31**, 1228–1242 (2017).
107. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
108. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–R98 (2014).
109. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
110. Michelson, A. M. Deciphering genetic regulatory codes: A challenge for functional genomics. *Proc. Natl Acad. Sci., USA* **99**, 546–548 (2002).
111. Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554 (2016).
112. Zeng, H., Edwards, M. D., Liu, G. & Gifford, D. K. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* **32**, i121–i127 (2016).
113. Kumar, S., Ambrosini, G. & Bucher, P. SNP2transcription factorBS – a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).
114. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.* **45**, D61–D67 (2017).

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. Sabeti, 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 mth entry is the effect of SNP m on binding of some transcription factor. If we knew the length-M vector β of the true causal effects of the same SNPs on a trait, we could simply regress β on v to evaluate whether there is a non-trivial signed association across SNPs m between v_m and β_m. In reality, we cannot do this because we do not observe β; instead we observe a vector, denoted $\widehat{\alpha}$, of GWAS summary statistics describing the marginal correlation of every SNP to our trait of interest. This vector differs from β because it includes both causal and tagging effects, plus statistical noise. Specifically, it can be shown mathematically that, in expectation, $\widehat{\alpha}$ will equal the matrix-vector product $R\beta$ where R is the $M\times M$ LD matrix. Therefore, just as β would be proportional to v in the presence of a signed effect, $\widehat{\alpha}(\approx R\beta)$ 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 β on v (which is impossible since we do not observe β), we can regress $\widehat{\alpha}$ 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 estimands. Let M be the number of SNPs in the genome. We assume a linear model:

$$y \mid \beta, x \sim \mathcal{N}(x^T\beta, \sigma_e^2) \tag{2}$$

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

$$\beta \mid v \sim [\mu v, \sigma^2 I] \tag{3}$$

where the scalar μ represents the genome-wide directional effect of v on β, σ^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 μ, 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_f, which re-scales μ to be dimensionless and is defined as

$$r_f := \mathrm{corr}(x^T\beta, x^Tv) = \mu\sqrt{\frac{v^TRv}{h_g^2}} \tag{4}$$

where $h_g^2 = \mathrm{var}(x^T\beta)$ is the SNP heritability of the phenotype and $R = E(xx^T) \in \mathrm{R}^{M\times M}$ is the (signed) population LD matrix of the genotypes. The quantity r_f can be interpreted as a form of genetic correlation; the value of r_f^2 cannot exceed the proportion of SNP heritability explained by SNPs with non-zero values of v. (Note that r_f can also be defined as a correlation between β and v; this definition is approximately equivalent in expectation under our random effects model, provided $v^TRv\approx |v|^2$.) We additionally estimate $h_v^2 = r_f^2h_g^2$, the total phenotypic variance explained by the signed contribution of v to β, as well as $h_v^2\ /\ h_g^2 = r_f^2$. For annotations with small support, these quantities are expected to be small in magnitude. To see this, notice that h_v^2 cannot exceed the total (unsigned) phenotypic variance explained by SNPs with non-zero values of v. It follows that r_f^2 cannot exceed the proportion of (unsigned) SNP heritability explained by SNPs with non-zero values of v. For more detail on the model and estimands, see the Supplementary Note.

Main derivation. Let $X\in \mathrm{R}^{N\times M}$ be the genotype matrix in a GWAS of N individuals, with standardized columns, and let $Y\in \mathrm{R}^N$ be the phenotype vector. In the Supplementary Note, we show that under the above model the following identity approximately holds:

$$\widehat{\alpha} \mid v \sim \left[\mu Rv, \sigma^2R^2 + \frac{R}{N}\right] \tag{5}$$

where $\widehat{\alpha} := X^TY/N$ is a vector whose mth entry contains the marginal correlation of SNP m to the phenotype and $R\in \mathrm{R}^{M\times M}$ is the population LD matrix. Equation (1) from the main text can be derived from equation (5) by re-scaling v so that $v^TRv = 1$, then substituting for μ.

We call Rv the *signed LD profile* of v. Equation (5) means that we can estimate μ by regressing $\widehat{\alpha}$ on the signed LD profile using generalized least-squares regression with $\Omega := \sigma^2R^2 + R/N$ as the inverse weight matrix. (We assign a regression weight of zero to SNPs in the major histocompatibility complex region.) It can be shown that if (a) all causal SNPs are typed, (b) sample size is infinite, and (c) R is invertible, this method is equivalent to estimating β via $R^{-1}\widehat{\alpha}$ and then regressing this estimate on v to obtain μ, which is the optimal regression-based approach in that setting. Note that because we generate P values for hypothesis testing empirically (see below), we are guaranteed that our generalized least-squares scheme will remain well calibrated even if our estimate of the matrix Ω is inaccurate due to, for example, mis-match between the reference panel and the study population. Once we have estimated μ, we re-scale this estimate to yield an estimate of r_f and other estimands of interest. 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, $H_0{:}r_f = 0$), we split the genome into approximately 300 blocks of approximately the same size with the block boundaries constrained to fall on estimated recombination hotspots[117]. 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. Our use of sign-flipping ensures that any true positives found by our method are the result of genuine first-moment effects; if in contrast we estimated standard errors using least-squares theory or a re-sampling method such as the jackknife or bootstrap, our method might inappropriately reject the null hypothesis only because the variance of β is higher in parts of the genome where Rv is large in magnitude. This would make our method susceptible to confounding due to unsigned enrichments, as might arise from the co-localization of transcription factor binding sites with enriched regulatory elements such as enhancer regions. Additionally, the fact that we flip the signs of SNPs in each block together ensures that our null distribution preserves any potential association of our annotation to the LD structure of the genome. See the Supplementary Note for more details.

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

Unless stated otherwise, all analyses in this paper control in this fashion for a 'signed background model' consisting of five annotations $u^1,\ldots, u^5$, defined by

$$u_m^i = \mathbf{1}\{\mathrm{MAF}_m\text{is in}i\text{th quintile}\}\sqrt{2\mathrm{MAF}_m(1-\mathrm{MAF}_m)}^{\,1+\alpha_s} \tag{6}$$

where MAF_m is the minor allele frequency of SNP m and α_s 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 $\mathrm{var}(\beta_m)$, we set $\alpha_s = -0.3$[118].

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[19], 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 1000 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

$$v_m = \mathbf{1}\{m\in C\}\,(P_m^a - P_m^A) \tag{7}$$

where C is the set of SNPs in the ChIP-seq peaks arising from the experiment, P_m^a is the Basset prediction for the 1,000 base-pair sequence around SNP m when the minor allele is placed at SNP m, and P_m^A 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[120] ($n = 47{,}360$). 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

(MAF ≥0.1%) in the GERA data set, and were represented in our 1000 G Phase 3 European reference panel[107,121]. We simulated traits using normally distributed causal effect sizes (with annotation-dependent mean and variance in some cases), with $h_g^2 = 0.5$.

Null simulations. For the simulations in Fig. 1a, we simulated 1,000 independent null phenotypes with the architecture $\beta_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = h_g^2 / M$ and $h_g^2 = 0.5$. For each phenotype, we computed GWAS summary statistics using plink2[122] (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 $\beta_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 + \tau^2 \mathbf{1}\{v_m \neq 0\})$ where σ^2 and τ^2 were set to achieve $h_g^2 = 0.5$ and a 20× unsigned enrichment for the SNPs with non-zero values of *v*. 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 $\beta_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu u_m^1, \sigma^2)$ where u_m^1 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 μ such that 10% of heritability would be explained by this directional effect, and then set σ^2 to achieve $h_g^2 = 0.5$. We then computed summary statistics as above and ran SLDP regression to assess for a directional effect of each of our 382 annotations on each of the 1,000 phenotypes, yielding a set of 382,000 *P* values. (We note that this represents a best-case scenario in which the background model exactly matches the confounding being simulated, up to differences in MAF between the reference panel and the GWAS sample, and we caution that our method may not be appropriate for annotations with much stronger correlations to minor alleles than the annotations that we analyze here; Supplementary Fig. 1b.) Finally, we repeated the same computation but running SLDP regression without the 5-MAF-bin signed background model to obtain an additional set of 382,000 *P* values.

Causal simulations. For the simulations in Fig. 2, we fixed a representative annotation *v* (binding of IRF4 in GM12878), and simulated traits using $\beta_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu v_m, \sigma^2)$, with μ set to achieve $r_f = \{0, 0.005, 0.01, \ldots, 0.05\}$ and σ^2 set to achieve $h_g^2 = 0.5$ in each case. For each value of r_f, 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 QTL 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, respectively[21] (Supplementary Table 4 and URLs). For each of the three gene expression traits, we constructed one summary statistics vector $\hat{\alpha}$ by meta-analyzing, for each SNP, the marginal effect sizes of that SNP for the expression of all nearby genes. Specifically, we set

$$\hat{\alpha}_m = \frac{1}{\sqrt{|G_m|}} \sum_{k \in G_m} \hat{\alpha}_m^{(k)} \tag{8}$$

where G_m is the set of all genes within 500 kb of SNP *m*, and $\hat{\alpha}_m^{(k)}$ 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 $\hat{\alpha}$ 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 procedure[123] 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 r_f.

For our replication analysis, we used expression array-based whole blood eQTL data from the NTR[33], which we obtained by downloading the set of TWAS weights[81] 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 significantly heritable *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 database[29] 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 or neither activities. 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 ($P = 7.9 \times 10^{-43}$ for eQTL results and $P = 1.9 \times 10^{-9}$ 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 procedure[123] 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 tissues[124] and processed these via the procedure described above into a single vector $\hat{\alpha}^{(T)}$. For each tissue *t*, we then residualized $\hat{\alpha}^{(T)}$ out of the vector $\hat{\alpha}^{(t)}$ of eQTL data for tissue *t* to obtain a residualized vector $\hat{\alpha}^{(t')}$. This amounts to subtracting a scalar multiple of $\hat{\alpha}^{(T)}$ from $\hat{\alpha}^{(t)}$, with the scalar determined to remove as much signal as possible from $\hat{\alpha}^{(t)}$. For each significant association between an annotation *v* and a vector $\hat{\alpha}^{(t)}$ from our main GTEx analysis, we then compared the *P* value of that association to the *P* value obtained for the association between *v* and the residualized vector $\hat{\alpha}^{(t')}$, 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 positives[125]; 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 tissue 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 ($P = 2.1 \times 10^{-15}$ 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 publicly released summary statistics computed using BOLT-LMM v2.3[92] (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 procedure[123]. 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 downloaded all 10,325 MSigDB gene sets, which are organized into eight distinct tranches based on their origin, from the MSigDB online portal (URLs). We also downloaded a set of LD blocks in Europeans derived from estimated recombination hotspots[117] and converted each gene set into a length-1693 vector *s* with one entry per LD block whose *i*th entry equaled the number of genes from the set that are present in the *i*th LD block. We then converted each significant SLDP regression association between an annotation *v* and a trait summary statistics vector $\hat{\alpha}$ into a length-1693 vector *q* whose *i*th entry equaled the covariance between $\hat{\alpha}$ and the signed LD profile *Rv* within the *i*th 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

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

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 $a(v,\widehat{\alpha},s)$ 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 $\frac{\sum_i s_i q_i}{\sum_i s_i}$ by the average value of q at LD blocks containing no genes. That is the enrichment is defined as

$$e(v,\widehat{\alpha},s) = \frac{a(v,\widehat{\alpha},s)}{\mathrm{mean}(\{q_i : s_i = 0\})}.$$

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 at FDR < 5% using the Benjamini–Hochberg procedure[123]. This detected 6,379 significant enrichments in total (0.8% of all 795,025 tests conducted). We ranked these enrichments by q value, except for the 15 enrichments whose P values were less than the resolution of our empirical null hypothesis testing procedure, which we ranked by fold-enrichment.

Autoimmune enrichment among complex trait associations. Of the 12 independent transcription factor-complex trait associations, 9 involve an autoimmune disease, representing a 4.3× enrichment ($P = 1.9 \times 10^{-5}$ using one-sided binomial test) and providing additional evidence for the relevance of transcription factor binding to these phenotypes in particular[126].

Estimation of lower bound on number of independent transcription factor binding sites contributing to each association. For intuition, we first examined, for each annotation, the estimated covariance between the GWAS summary statistics and the signed LD profile in each of 300 independent genomic blocks, finding agreement with the genome-wide direction of association in 59% of the blocks on average across our 12 independent associations, and in 85% of the blocks with estimated covariances of large magnitude (Supplementary Fig. 8).

For further quantification, we then converted each of the 12 independent transcription factor–trait associations reported in Table 1 into a vector q of length ~300 whose ith entry equaled the covariance between the GWAS in question and the signed LD profile in question within the ith of the ~300 independent genomic blocks used for our null hypothesis testing. For every threshold $a(v,\widehat{\alpha},s)$, we then computed the number K_t of the entries of q with magnitude at least t, as well as the number S_t of those entries whose sign agreed with that of the genome-wide trend. Our estimated lower bound on the number of independent transcription factor binding sites contributing to the association was then given by

$$\max_t(2S_t - K_t) \tag{9}$$

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 $2S_0 - K_0$. 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.

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

Code availability. Open-source software implementing our approach is available at http://www.github.com/yakirr/sldp. Code used to make all figures is available at http://www.github.com/yakirr/sldp-display.

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**, 283–285 (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).

nature research

Corresponding author(s): Yakir Reshef, Alkes Price

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	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 SLDP 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