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

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

[1]Department of Computer Science, Harvard University, Cambridge, MA
[2]Harvard/MIT MD/PhD Program, Boston, MA
[3]Broad Institute of MIT and Harvard, Cambridge, MA
[4]California Life Sciences Company, South San Francisco, CA
[5]Dana Farber Cancer Institute, Boston, MA
[6]Boston Children's Hospital, Boston, MA
[7]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA
[8]Program in Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA
[9]Brigham and Women's Hospital, Boston, MA
[10]Department of Statistics, University of Oxford, Oxford, UK
[11]The Massachusetts General Hospital, Boston, MA
[12]Google Brain, New York, NY
[13]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
[*]Correspondence should be addressed to YAR (yakir@seas.harvard.edu) and ALP (aprice@hsph.harvard.edu)

Abstract

Biological interpretation of GWAS data frequently involves analyzing unsigned genomic annotations comprising SNPs involved in a biological process and assessing enrichment for disease signal. However, it is often possible to generate signed annotations quantifying whether each SNP allele promotes or hinders a biological process, e.g., binding of a transcription factor (TF). Directional effects of such annotations on disease risk enable stronger statements about causal mechanisms of disease than enrichments of corresponding unsigned annotations. Here we introduce a new method, signed LD profile regression, for detecting such directional effects using GWAS summary statistics, and we apply the method using 382 signed annotations reflecting predicted TF binding. We show via theory and simulations that our method is well-powered and is well-calibrated even when TF binding sites co-localize with other enriched regulatory elements, which can confound unsigned enrichment methods. We apply our method to 12 molecular traits and recover many known relationships including positive associations between gene expression and genome-wide binding of RNA polymerase II, NF-κB, and several ETS family members, as well as between known chromatin modifiers and their respective chromatin marks. Finally, we apply our method to 46 diseases and complex traits (average $N = 289,617$) and identify 77 significant associations at per-trait FDR $< 5\%$, representing 12 independent signals. Our results include a positive association between educational attainment and genome-wide binding of BCL11A, consistent with recent work linking *BCL11A* hemizygosity to intellectual disability; a negative association between lupus risk and genome-wide binding of CTCF, which has been shown to suppress myeloid differentiation; and a positive association between Crohn's disease (CD) risk and genome-wide binding of IRF1, an immune regulator that lies inside a CD GWAS locus and has eQTLs that increase CD risk. Our method provides a new way to leverage functional data to draw inferences about causal mechanisms of disease.

Introduction

Mechanistic interpretation of GWAS data sets has become a central challenge for efforts to learn about 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 signal.[1–5] 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, thereby enabling us to assess whether there is a systematic relationship between SNP alleles' direction of effect on the process and their direction of effect on a trait. Transcription factor (TF) binding, which plays a major role in human disease,[6–8] represents an important case in which such signed functional annotations are available: because TFs 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 TF.[9–15]

Detecting genome-wide directional effects of TF binding on disease would constitute a significant 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 (e.g., of TF 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 TFs affect only a few disease genes or whether transcriptional programs comprising many target genes are responsible for TF associations; a genome-wide directional effect implies the latter model (see Discussion).

Here we introduce a new method, signed LD profile (SLDP) regression, for quantifying the genome-wide directional effect of a signed functional annotation on polygenic disease risk, and apply it in conjunction with 382 annotations each reflecting predicted binding of a particular TF in a particular cell line. Our method requires only GWAS summary statistics,[16] accounts for linkage disequilibrium and untyped causal SNPs, and is computationally efficient. We validate the method via extensive simulations, including null simulations confounded by unsigned enrichment as might arise from the co-localization of TF binding sites with other regulatory elements.[3,9] We apply the method to 12 molecular traits and 46 diseases and complex traits, demonstrating genome-wide directional effects of TF binding in both settings.

Results

Overview of methods

Our method for quantifying directional effects of signed functional annotations on disease risk, signed LD profile 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 linear 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[17] in which true causal SNP effects are correlated with a signed functional annotation, we show that

$$E(\hat{\alpha}|v) = r_f \sqrt{h_g^2} R v \tag{1}$$

where $\hat{\alpha}$ is the vector of marginal correlations between SNP alleles and a trait, v is the signed functional annotation (re-scaled to norm 1), 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*. (The value of r_f^2 cannot exceed the proportion of SNP-heritability explained by SNPs with non-zero values of v.) Equation 1, together with an estimate of h_g^2, allows us to estimate r_f by regressing $\hat{\alpha}$ on the signed LD profile Rv of v. We assess statistical significance by randomly flipping the signs of entries of v, with consecutive SNPs being flipped together in large blocks (e.g., ~ 300 blocks total), to obtain a null distribution and corresponding P-values and false discovery rates (FDRs). To improve power, we use generalized least-squares regression, incorporating weights to account for the fact that SNPs in linkage disequilibrium (LD) provide redundant information due to their correlated values of $\hat{\alpha}$. We remove the major histocompatibility complex (MHC) region from all analyses due to its unusual LD patterns. We perform a multiple regression that includes a "signed background model" quantifying 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 signed LD profile regression requires signed effect size estimates $\hat{\alpha}$ and quantifies directional effects, in contrast to stratified LD score regression,[3] which analyzes unsigned χ^2 statistics and quantifies unsigned heritability enrichment. Details of the method are described in the Online Methods section and the Supplementary Note; we have released open-source software implementing the method (see URLs).

We applied signed LD profile regression using a set of 382 signed annotations v, each quantifying the predicted effects of SNP alleles on binding of a particular TF in a particular cell line. We constructed the annotations by training a sequence-based neural network predictor of ChIP-seq peak calls, using the Basset software,[15] on the results of 382 TF binding ChIP-seq experiments from ENCODE[18] and comparing the neural network's predictions for the major and minor allele of each SNP in the ChIP-seq peaks. The 382 experiments spanned 75 distinct TFs and 84 distinct cell lines. The resulting annotations were sparse, with only 0.2% of SNPs having nonzero entries on average (see Online Methods and Table S1).

Simulations

We performed simulations with real genotypes, simulated phenotypes, and the 382 signed TF binding annotations to assess null calibration, robustness to confounding, and power. All simulations used well-imputed genome-wide genotypes from the GERA cohort,[19] corresponding to $M = 2.7$ million SNPs and $N = 47,360$ individuals of European ancestry. We simulated traits using normally distributed causal effect sizes (with annotation-dependent mean and variance in some cases), with $h_g^2 = 0.5$. Further details of the simulations are provided in the Online Methods section.

We first performed null simulations involving a heritable trait with no unsigned enrichment or directional relationship to any of the 382 annotations. In 1,000 independent simulations, we applied signed LD profile regression to test each of the 382 annotations for a directional effect. The resulting P-values were well-calibrated (see Figure 1a and Table S2). Analyses of the P-value distribution for each annotation in turn confirmed correct calibration for these annotations (see Figure S1a).

We next performed null simulations involving a trait with unsigned enrichment but no directional effects; these simulations were designed to mimic unsigned genomic confounding in which the binding sites of some TF lie in or near regulatory regions that are enriched for heritability for reasons other than binding of that TF. In 1,000 independent simulations, we randomly selected an annotation, simulated a trait in which the annotation had a 20x unsigned enrichment[3] (but no directional effect), and applied signed LD profile regression to test the annotation for a directional effect. We again observed well-calibrated P-values (see Figure 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 such as heritability partitioning, in which unsigned genomic confounders must be carefully accounted for and modeled.[3]

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 TF binding, which could arise due to genome-wide negative selection or population stratification. We simulated a trait for which 10% of heritability is explained by directional effects of minor alleles in the bottom fifth of the MAF spectrum (roughly MAF$<5\%$). In 1,000 independent simulations, we applied signed LD profile regression to test each of the 382 annotations for a directional effect. 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 (see Figure 1c). (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; see Figure S1b.) 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 TF binding.[20] Indeed, most of our annotations show a small but highly significant bias of minor alleles toward decreasing TF binding (see Figure S2) that is consistent with this explanation; however, it is also possible that this is a result of our procedure for constructing the annotations, and we do not explore it further in this work.

Finally, we performed causal simulations with true directional effects to assess the power and establish unbiasedness of signed LD profile regression. At default parameter settings, the method is well-powered to detect directional effects corresponding to a functional correlation of 2-6% (see Figure 2a and Table S3), similar to values observed in analyses of real traits (see below). Notably, the power of the method is improved dramatically by our use of generalized least-squares to account for redundant information (see Figure 2a). Our method is also much more powerful than a naive method that regresses the vector of GWAS summary statistics on the annotation rather than its signed LD profile, an approach that does not model untyped causal SNPs in linkage disequilibrium with typed SNPs (see Figure S3). The power of our method increases with sample size and SNP-heritability (see Figure S4), and is only minimally affected by within-Europe reference panel mismatch (see Figure S5). In all instances, our method produced either unbiased or nearly unbiased estimates of functional correlation and related quantities (see Figure 2b and Figure S6).

Analysis of molecular traits

TF binding is known to affect gene expression and other molecular traits.[21] We therefore applied signed LD profile regression to 12 molecular traits with an average sample size of $N = 149$. We first analyzed cis-eQTL data based on RNA-seq experiments in three blood cell types from the BLUEPRINT consortium[22] (see Online Methods). For each cell type, we collapsed eQTL summary statistics across 15,023-17,081 genes into a single vector of summary statistics for aggregate expression by summing, for each SNP, the marginal effect sizes of that SNP for the expression of all nearby genes (within 500kb). This is equivalent to analyzing one gene at a time and then performing a meta-analysis that accounts for linkage among nearby genes; it is also roughly equivalent to analyzing eQTL summary statistics for the sum of expression values of all genes, with each gene normalized to mean zero and variance one in the population (see Online Methods and Table S4).

We tested each of the 382 TF binding annotations for a directional effect on aggregate expression in each of the three blood cell types. We detected a total of 92 significant associations at a per-trait FDR of 5% (see Figure 3a and Table S5a; P-values from 5×10^{-6} to 1.0×10^{-2}). All 92 associations were positive, implying that greater binding of these TFs leads to greater aggregate expression and matching the known tendency of TF binding to promote rather than repress transcription for many TFs.[21]

Many of the associations we detected recapitulate known aspects of transcriptional regulation. For example, associated TF binding annotations included RNA polymerase II in many cell lines, along with other members of the transcription pre-initiation complex (PIC) such as TATA-associated Factor 1 (TAF1) and TATA Binding Protein (TBP). There were also associations for TFs unrelated to

the PIC but known to have activating activity, such as the ETS family members GABPA, ELF1, and PU.1,[23] as well as the immune-related transcriptional activators IRF1 and NF-κB family member RELA.[24,25] Overall, the vast majority of the positive associations (85 out of 92) involved known "activating" TFs, defined as TFs with activating activity but not repressing activity in UniProt[26] (compared with 45% of all 382 annotations; $P = 3.1 \times 10^{-7}$ for difference using one-sided binomial test; see Figure 3a and Online Methods). 52 of the 92 associations replicated (same direction of effect with nominal $P < 0.05$) in an independent set of whole-blood eQTL summary statistics based on expression array experiments from the Netherlands Twin Registry (NTR),[27] including all of the examples mentioned above except IRF1 (see Figure 3b and Table S5b). Across all 382 annotations analyzed, we observed a correlation of $r = 0.67$ between z-scores for signed annotation effects in the BLUEPRINT neutrophil and NTR data sets (see Figure 3c and Table S5c).

We next conducted a similar analysis using histone QTL (H3K27me1 and H3K27ac) and methylation QTL from the BLUEPRINT data set. We detected 16 significant associations at a per-trait FDR of 5%, all of which were positive, including 13 for H3K27me1 QTL (see Figure 3d and Table S5d; P-values from $\leq 10^{-6}$ to 4.3×10^{-4}), 3 for H3K27ac QTL (see Figure 3e and Table S5e; P-values from 1.2×10^{-5} to 2.1×10^{-4}), and 0 for methylation QTL. Many of the detected associations recover known aspects of histone mark biology. For example, TFs associated to H3K4me1 included PU.1 and CEBPB, both of which act to increase H3K4me1 in blood cells and play strong roles in differentiation of those cell types,[28–31] and binding of MYC, which has a known role as a chromatin modifier,[32,33] including of H3K4 methylation.[34] We also observed an association between EP300 binding and H3K27ac, matching the fact that EP300 is a lysine acetyltransferase with a well-documented role in creation and maintenance of this mark.[35] Finally, while we did not find significant associations to methylation QTL at FDR$<$ 5%, we found 40 results at FDR$<$ 10%, almost all of which were negative associations between CTCF binding and methylation that are consistent with the literature on the negative relationship between CTCF binding and this epigenetic mark[36–38] (see Table S5f). In our analysis of the activating marks H3K4me1 and H3K27ac, signed LD profile regression again distinguished between activating and repressing TFs: of the 239 positive associations and 19 negative associations at a nominal significance threshold of $P < 0.05$ (chosen due to limited number of FDR$<$ 5% associations) across all three cell types, 85% of the positive associations corresponded to activating TFs[26] (compared with 45% for all annotations; one-sided binomial $P = 7.4 \times 10^{-8}$ for difference); only 26% of the negative associations had this property (one-sided binomial $P = 7.0 \times 10^{-2}$ vs. 45%, $P = 4.6 \times 10^{-5}$ vs. 85%).

Analysis of 46 diseases and complex traits

We applied signed LD profile 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 trants for which we have publicly released summary statistics computed using BOLT-LMM[39] (see URLs and Table S6). We ran signed LD profile regression using each of our 382 TF annotations for each of these traits. We detected 77 significant associations at a per-trait FDR of 5%, spanning six diseases and complex traits (see Figure 4 and Table S7a). (Following standard practice, we report per-trait FDR, but we estimated the global FDR of this procedure to be 9.4%, which is larger than the per-trait FDR of 5%; see Online Methods). The 77 significant associations represent 12 independent signals after pruning correlated annotations (Table 1; see Online Methods). To verify empirically that our results were not driven by directional effects of minor alleles, we re-analyzed our data using 382 annotations defined using the same set of SNPs with non-zero effects but with the directionality of effect determined by minor allele coding rather than predicted TF binding, for SNPs in the bottom quintile of the MAF spectrum. This analysis yielded only 4 significant associations at per-trait FDR$<$ 5%. (Due to the small number of associations relative to the number of traits, this corresponds to a global FDR of 92.9% after accounting for 46 traits.) None of these 4 minor-allele associations overlapped with our set of 77 significant associations (see Online Methods and

Table S7b). We also 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 (see Figure S7).

Many of our results are supported by orthogonal genetic and non-genetic evidence and extend our understanding of the associated traits; we highlight three in particular. Our most significant result is a positive association between genome-wide binding of BCL11A in LCLs and years of education (see Figure 5a and Table S8). This result aligns with existing common and rare variant signals: *BCL11A* was one of the top genes identified in a GWAS of educational attainment[40] and *de novo* missense and loss-of-function mutations in *BCL11A* cause intellectual disability in a dosage-dependent manner.[41,42] (Additionally, our fine-mapping of the *BCL11A* GWAS locus[43] identified a putatively causal SNP in an intron of the *BCL11A* gene; see Table S9.) *BCL11A* has also been shown to be the causal gene for a microdeletion syndrome characterized by cognitive impairment.[44,45] Recent experimental studies showing that heterozygous knock-out of *Bcl11a* in mice leads to microcephaly and cognitive impairment[42] have further confirmed the causal role of BCL11A in cognitive function, with directionality consistent with our result. This association thus represents a case in which our method provides stronger evidence for a causal association than previously available from common variant data, and establishes that BCL11A causes intellectual disability via a genome-wide mechanism involving binding throughout the genome—and presumably the modulation of a transcriptional program relevant to brain function or development—rather than regulation of one key disease gene (see Discussion).

We also detected a negative association between genome-wide binding of CCCTC-binding factor (CTCF) in the myeloid cell line K562 and risk of systemic lupus erythematosus (SLE) (see Figure 5b), accompanied by similar associations for CTCF and cohesin subunit RAD21 (a CTCF binding partner) in several other cell lines. This finding is consistent with several SLE risk loci at which either fine-mapped causal SNPs have been found to modify CTCF binding experimentally[46] and bioinformatically,[47] or at which risk SNPs have been found to be in LD with SNPs modifying CTCF binding.[48,49] Additionally, CTCF has been shown experimentally to slow the rate of myeloid differentiation[50,51] and is involved in the regulation of 5-hydroxymethylcytosine (5-hmC), an epigenetic modification that is increased in promoters of immune-related genes in CD4+ T cells of patients with SLE relative to controls.[52] Finally, CTCF motifs are overrepresented among DNA regions that are more accessible in B cells from healthy controls relative to B cells from SLE patients,[53] consistent with the negative sign of the association arising from our study. We do not observe a GWAS signal for SLE at the *CTCF* locus. This could be because of the small sample size of the SLE GWAS, and/or because the *CTCF* gene is under strong selective constraint: its probability of loss-of-function intolerance (pLI) is estimated by the Exome Aggregation Consortium[54] to be the maximal value of 1.00, greater than 99.9% of genes. The association between CTCF binding and SLE therefore demonstrates the possibility of using signed LD profile regression to uncover aspects of disease mechanism that are difficult to directly observe in GWAS due to selective pressures on the underlying genes.

We also highlight a positive association between genome-wide binding of Interferon Regulatory Factor 1 (IRF1) in the myeloid cell line K562 and Crohn's disease (CD) (see Figure 5c). *IRF1* is located inside the *IBD5* locus, a 250kb region associated with CD and inflammatory bowel disease in multiple GWAS;[55,56] haplotypes containing *IRF1* variants have been shown to be more strongly correlated with CD risk than haplotypes containing variants in nearby genes;[57] CD risk SNPs have been shown to co-localize with *IRF1* alternative splicing QTLs;[22] and *IRF1* is more highly expressed in CD gastrointestinal tissue biopsies relative to control tissue.[57] In a recent large-scale fine-mapping study,[58] the causal signal at the *IBD5* locus was narrowed down to a set of 8 SNPs spanning 35kb and lying 15kb away from *IRF1*. However, despite this resolution, it remains unclear from the locus alone what the causal mechanism is: the study suggested that rs2188962, which received 0.59 of the

posterior probability of being causal, could function via an eQTL effect on *SLC22A5* in immune and gut epithelial cells, but rs2188962 is also an eQTL for *IRF1* in blood,[27] and we determined that the TWAS approach[59] assigns highly significant scores to both genes ($p \leq 4.0 \times 10^{-14}$ for *IRF1* and 3.17×10^{-18} for *SLC22A5*). In this context, our result therefore provides genome-wide evidence for a genuine causal link between *IRF1* and CD that, unlike the single-locus approaches, is not susceptible to pleiotropy and allelic heterogeneity near the *IRF1* gene (see Discussion). We note that the direction of effect inferred by our method agrees with the positive sign of the TWAS association between *IRF1* and CD, as expected in the case of a causal relationship.

We provide additional discussion of our other results in the Supplementary Note.

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. We applied this method, in conjunction with 382 annotations describing predicted effects of SNPs on TF binding, to 12 molecular traits (average $N = 149$) and 46 diseases and complex traits (average $N = 289,617$). In our analysis of molecular traits, our method recovered classical aspects of transcriptional regulation, including the pro-transcriptional effect of RNA polymerase and activating TFs such as NFκB, as well as relationships between several chromatin modifiers and their respective chromatin marks; to our knowledge, these relationships have not previously been demonstrated using eQTL data. Our analysis of complex traits yielded 77 TF-trait associations, corresponding to 12 independent associations. Some of our results, such as the positive association between IRF1 binding and Crohn's disease, provide strong causal hypotheses to explain long-standing GWAS associations; others, such as the positive association between BCL11A binding and educational attainment, provide mechanistic interpretation for a top GWAS locus for which orthogonal genetic evidence, such as rare variant and knock-out studies, already existed; and still others, such as the negative relationship between CTCF binding and SLE, have experimental support but had not previously been observed from GWAS data, possibly due to strong evolutionary constraint on some TFs. We note that although we constructed our predicted TF binding annotations using the neural-network predictor Basset,[15] there exist many other effective methods for making such signed predictions.[9–12,14,60]

Our method differs from unsigned GWAS enrichment methods by assessing whether there is a systematic genome-wide correlation 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. A substantial advantage of this approach is reduced susceptibility to confounding: for example, an unsigned GWAS enrichment for binding of an immune TF could indicate a causal role for that TF in the associated disease, or could instead be a side effect of a generic enrichment among cell-type specific regulatory elements in immune cells.[3] In contrast, if alleles that increase binding of the TF tend to increase disease risk and alleles that decrease binding of the TF tend to decrease disease risk, the set of potential confounders is smaller because a confounding process has not only to co-localize in the genome with binding of the TF but also to have the property that alleles that increase the process have a consistent directional effect on binding of the TF.

When applied to TF binding, our method enables stronger statements about causality and mechanism than were previously possible with genome-wide methods. Regarding causality, this is because a consistent directional effect throughout the genome of SNPs predicted to affect binding due to sequence change supports stronger causal statements than i) single-locus methods, which are susceptible to pleiotropy and allelic heterogeneity,[59] ii) unsigned heritability enrichment methods, which can be confounded by co-localization in the genome of TF binding sites with other enriched regulatory elements as described above,[3] and iii) genetic correlation and Mendelian randomization (MR), which can be confounded by reverse causality and pleiotropic effects[61–63] and which scale poorly

because they require TF ChIP-seq in many individuals for every TF/cell-type pair studied. The reason that our method is not confounded by reverse causality is that each of our annotations is produced in a cell population that is isogenic and therefore does not have variance in genetic liability for any trait. In other words, our annotations provide ideal instrumental variables for the effect of TF binding on the trait of interest because they are created not by naively correlating SNPs with TF binding but rather by examining the effect of each SNP on local DNA sequence.

Regarding mechanism, our method sheds light on the question of whether TFs affect traits via coordinated regulation of gene expression throughout the genome[64] (a "genome-wide" model) or via regulation of one or a small number of key disease genes[65] (a "local" model). Since the associations we find involve a consistent net direction of effect of TF binding on a trait throughout the genome, they cannot be explained by a local model and therefore represent evidence for the existence of transcriptional programs and their relevance to complex traits. This is of basic interest, but it also has therapeutic relevance: if a TF causally affects a trait but the TF is not druggable due to its nuclear localization or large DNA- and protein-binding domains,[66,67] then the local model suggests targeting a downstream gene, whereas the genome-wide model instead suggests targeting an upstream regulator since the causal link between TF and trait is mediated through a large number of downstream genes. (We emphasize that a significant result for our method does not imply that all binding events of the TF in question affect disease via activation of a single transcriptional program; rather, it implies that there exists a program that is widespread enough that we observe its effect on disease in a large number of locations in the genome; see Figure S7.)

Our method could be used to link disease to biological processes beyond TF binding. For example, sequence-based models can also produce signed predictions of DNase I hypersensitivity,[10,11,15] histone modifications,[11,15] splicing,[12,68] and transcription initiation.[69] Additionally, massively parallel assays and CRISPR screens are increasingly yielding high-resolution experimental information about the effects of genetic variation on gene expression[21,70–72] as well as cellular processes such as growth[73–75] and inflammation.[76] Finally, perturbational differential expression experiments can yield signed predictions for the relationships of genes to a variety of biological processes such as drug response,[77] immune stimuli,[78] and many others.[79] Though converting such data to signed functional annotations will require care, doing so could allow us to leverage them to make detailed statements about disease mechanism.

We note several limitations of signed LD profile regression. First, though 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 TFs that are close binding partners and thus share sequence motifs. Second, although we have shown our method to be robust in a wide range of scenarios, we cannot rule out the possibility of un-modeled directional effects of minor alleles on both trait and TF binding as a confounder; however, our empirical analysis of real traits with minor-allele-based signed annotations suggests that directional effects of minor alleles are very unlikely to explain our results (see Table S7b). Third, our method is not well-powered to detect instances in which a TF affects trait in different directions via multiple heterogeneous programs. Fourth, the effect sizes of the associations to diseases and complex traits that we report are small in terms of the estimated values of r_f, which range from 2.4% to 8.9% (see Table S7a), although signals of this size for predicted TF binding could be indicative of much stronger relationships, e.g., with true TF binding, TF expression, TF phosphorylation, or TF binding in specific subsets of the genome. We further note that the magnitude of the signals that we detect is commensurate with the very small number of SNPs in our annotations, together with the fact that r_f^2 is bounded by the proportion of SNP-heritability explained by those SNPs (see Table S7c). Fifth, though we detected many significant associations overall, there were many traits, such as schizophrenia, height, and blood cell traits, for which we did not detect any significant associations using our TF annotations. We believe that this limitation is partially due to the set of TF ChIP-seq annotations available through the ENCODE project, and in particular to the bias of those experiments toward core regulatory proteins such as RNA polymerase II and CTCF as well as their use of cell lines rather than primary tissue samples;

we expect this to become clearer as more diverse functional data sets become available.

Despite these limitations, signed LD profile regression is a powerful new way to leverage functional genomics data to draw causal and mechanistic conclusions from GWAS about both diseases and underlying cellular processes.

Acknowledgements

We thank C de Boer, L Dicker, J Engreitz, N Friedman, X Liu, M Mitzenmacher, J Perry, S Reilly, D Reshef, S Raychaudhuri, A Schoech, P Sabeti, R Tewhey, P Turley, and the CGTA discussion group for helpful discussions. This research was conducted using the UK Biobank Resource under Application #16549 and was supported by US National Institutes of Health grants U01 HG009379, R01 MH101244 and R01 MH107649. L.P. is supported by National Institutes of Health award R00HG008399. Computational analyses were performed on the Orchestra High Performance Compute Cluster at Harvard Medical School, which is partially supported by grant NCRR 1S10RR028832-01.

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/`
TWAS weights for NTR data: `https://data.broadinstitute.org/alkesgroup/FUSION/WGT/NTR.BLOOD.RNAARR.tar.bz2`

Online Methods

Signed LD profile regression

Model and estimands

Let M be the number of SNPs in the genome. We assume a linear model:

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

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

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

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.

Though 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

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

where $h_g^2 = \text{var}(x^T\beta)$ is the SNP-heritability of the phenotype and $R = E(xx^T) \in \mathbb{R}^{M\times M}$ is the (signed) population LD matrix of the genotypes. (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 $v^T Rv \approx |v|^2$.) We additionally estimate $h_v^2 = r_f^2 h_g^2$, 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 $X \in \mathbb{R}^{N\times M}$ be the genotype matrix in a GWAS of N individuals, with standardized columns, and let $Y \in \mathbb{R}^N$ be the phenotype vector. In the Supplementary Note, we show that under the above model the following identity approximately holds:

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

where $\hat{\alpha} := X^T Y/N$ is a vector whose m-th entry contains the marginal correlation of SNP m to the phenotype and $R \in \mathbb{R}^{M\times M}$ is the population LD matrix. Equation 1 from the main text can be derived from Equation 4 by re-scaling v so that $v^T Rv = 1$, then substituting for μ.

We call Rv the *signed LD profile* of v. Equation 5, together with central limit theorem considerations, implies that it is nearly optimal to estimate μ by regressing $\hat{\alpha}$ on the signed LD profile using generalized least-squares with $\Omega := \sigma^2 R^2 + R/N$ as the inverse weight matrix. 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 $R^{-1}\hat{\alpha}$ and then regressing this estimate on v to obtain μ, which is the optimal 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, e.g., 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.[80] We then define the null distribution of our statistic as the distribution arising from independently multiplying v by an independent random sign for each 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 TF 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 relationship of our annotation to the LD structure of the genome. In choosing how many blocks to use for this procedure, we took into account that i) the fewer blocks we use the fewer assumptions we make about LD structure and the faster we can compute P-values, and ii) the more

blocks we use the higher the precision of the P-values that we can obtain. Our choice to use 300 blocks is a compromise between these two considerations.

Controlling for covariates and the signed background model

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

Unless stated otherwise, all analyses in this paper are done controlling in this fashion for a "signed background model" consisting of 5 annotations $u^1, \ldots, u^5$, defined by

$$u^i_m = \mathbf{1}\left\{\text{MAF}_m \text{ is in } i\text{-th quintile}\right\} \sqrt{2\text{MAF}_m(1-\text{MAF}_m)^{1+\alpha_s}} \tag{6}$$

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 $\text{var}(\beta_m)$, we set $\alpha_s = -0.3$.[81]

382 TF annotations

We downloaded every ChIP-seq and DNase I hypersensitivity experiment in ENCODE and trained the sequence-based predictor of peak presence/absence, Basset,[15] to jointly predict each downloaded track on a set of held-out genomic segments. (We included tracks other than TF binding tracks because training predictions using all tracks slightly improved prediction accuracy for the TF binding tracks.) After training the joint predictor, we retained the predictions for every TF binding track for which a) the set of ChIP-seq peaks spanned at least 5,000 SNPs in our 1000G reference panel, and b) Basset's estimated area under the precision-recall curve was at least 0.3. This yielded a set of 382 TF ChIP-seq experiments. For each experiment, we constructed an annotation via

$$v_m = \mathbf{1}\{m \in C\}(P^a_m - P^A_m) \tag{7}$$

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

Simulations

All simulations were carried out using real genotypes from the GERA cohort[19] ($N = 47,360$). The set of $M = 2.7$ million causal SNPs was defined as the set of very well imputed SNPs (INFO ≥ 0.97) that had very low missingness ($< 0.5\%$), non-negligible MAF (MAF $\geq 0.1\%$) in the GERA data set, and were represented in our 1000G Phase 3 European reference panel.[82,83]

Null simulations

For the simulations in Figure 1a, we simulated 1,000 independent null phenotypes with the architecture $\beta_m \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = h^2_g/M$ and $h^2_g = 0.5$. For each phenotype, we computed GWAS summary statistics using plink2[84] (see URLs), adjusting for 3 principal components as well as GERA chip type as covariates. For each of our 382 TF annotations, we then ran signed LD profile regression on each of these 1,000 phenotypes, yielding a set of 382,000 P-values. For the simulations in Figure 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 $\beta_m \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 + \tau^2 \mathbf{1}\{v_m \neq 0\})$ where σ^2 and τ^2 were set to achieve $h_g^2 = 0.5$ and a 20x unsigned enrichment for the SNPs with non-zero values of v. We then computed summary statistics as above and ran signed LD profile regression to assess v for a genome-wide directional effect. This procedure yielded 1,000 P-values. For the simulations in Figure 1c, we simulated 1,000 independent phenotypes with a directional effect of minor alleles: we set $\beta_m \overset{iid}{\sim} \mathcal{N}(\mu u_m^1, \sigma^2)$ 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 signed LD profile 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. Finally, we repeated the same computation but running signed LD profile 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 Figure 2, we fixed a representative annotation v (binding of IRF4 in GM12878), and simulated traits using $\beta_m \overset{iid}{\sim} \mathcal{N}(\mu v_m, \sigma^2)$, 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.

Analysis of molecular traits

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[22] (see Table S5 and URLs). For each of the 3 gene expression traits, we constructed one summary statistics vector $\hat{\alpha}$ by setting

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

where G_m is the set of all genes within 500kb of SNP m, and $\hat{\alpha}_m^{(k)}$ is the marginal correlation of SNP m to the expression of gene k. Assuming a) infinite sample size, and b) zero correlation between every SNP and any gene not cis to that SNP, this procedure is equivalent up to a scalar to performing a GWAS of total relative expression. To see this, let $y^{(k)}$ denote expression of gene k after standardization to mean zero and unit variance in the population, and let γ be the GWAS summary statistics arising from a GWAS of total relative expression $\sum_k y^{(k)}$ at infinite sample size. By linearity, we have

$$\begin{aligned}
\gamma_m &\propto \sum_k \alpha_m^{(k)} && (9)\\
&= \sum_{k \in G_m} \alpha_m^{(k)} + \sum_{k \notin G_m} \alpha_m^{(k)} && (10)\\
&= \sum_{k \in G_m} \alpha_m^{(k)} && (11)\\
&= \alpha_m && (12)
\end{aligned}$$

where α_m^k denotes the large-sample limit of $\hat{\alpha}_m^{(k)}$ and α denotes the large-sample limit of $\hat{\alpha}$ defined in Equation 8.

Applying the same procedure to the two histone marks and to methylation in addition to gene expression yielded a total of 12 sets of summary statistics (see Table S4). We ran signed LD profile regression using each of our 382 TF annotations for each of these 12 traits. We obtained results at FDR$<5\%$ using the Benjamini-Hochberg procedure[85] within each of the 12 traits (see discussion of Benjamini-Hochberg versus other alternatives below), and reported the union of significant results across cell types for each trait.

For our replication analysis, we used expression array-based whole blood eQTL data from the NTR,[27] which we obtained by downloading the set of TWAS weights[59] computed for that data set (see URLs). We then proceeded as above.

Enrichment analysis for activating TFs

For each TF represented in our annotations, we queried the UniProt database[26] to establish whether the TF was annotated as having activating activity and, separately, whether it was annotated as having repressing activity. We then defined as "activating" any TF with the former but not the latter. To estimate whether the set of significant positive signed LD profile associations with gene expression were enriched for activating TFs compared to the set of annotations as a whole, we conducted a one-sided binomial test. To account for the correlated nature of our annotations, we assumed independence only among distinct TFs but not among distinct cell lines for the same TF. We used the same scheme to test for enrichment and depletion of activating TFs among the positive and negative associations detected by signed LD profile regression in our analysis of histone marks.

Analysis of 46 diseases and complex traits

We applied signed LD profile 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 trants for which we have publicly released summary statistics computed using BOLT-LMM[39] (see URLs and Table S6). We ran signed LD profile regression using each of our 382 TF annotations for each of these traits. We obtained results at per-trait FDR$<5\%$ using the Benjamini-Hochberg procedure.[85] We chose to use the Benjamini-Hochberg procedure rather than more sophisticated procedures such as the Storey-Tibshirani procedure[86] because the latter procedure, while more powerful, is more difficult to analyze in a multi-trait setting (see below) and controls FDR more noisily when applied in situations with only hundreds (rather than thousands) of tests.

Estimation of global FDR for complex trait analysis

When many traits are analyzed, per-trait FDR control does not imply global FDR control. This is because in the case of a completely null trait, the guarantee of FDR control does not imply that there will never be any rejections but rather only that there will be a non-zero number of rejections at most 5% of the time. Therefore, if enough null traits are analyzed the set of results may be contaminated by these spurious findings. In the case of independent tests (i.e., uncorrelated annotations) with FDR controlled by the Benjamini-Hochberg procedure, this can be taken into account[87] and the global FDR can be approximated using the formula

$$q = \frac{q_\ell (D + T)}{D + 1} \tag{13}$$

where q is the estimated global FDR, q_ℓ is the per-trait FDR, D is the observed total number of discoveries at per-trait FDR q_ℓ, and T is the number of traits. This correction is based on the intuition that for a null trait with independent tests, the Benjamini-Hochberg procedure behaves very similarly to a Bonferroni correction, and so the expected number of rejections per null trait is approximately q_ℓ, and the expected number of rejections for T null traits would be approximately $q_\ell T$.

Applying this correction to our results yields a global FDR estimate of 7.9%. However, since our annotations are dependent, this estimate can be anti-conservative. To see this, imagine a null trait with 100 perfectly correlated tests. The Benjamini-Hochberg procedure will give more than zero rejections only 5% of the time, but whenever it rejects it will yield 100 rejections rather than 1. Therefore, the expected number of rejections is not 0.05 but rather 5. We heuristically corrected for this using the intuition that under dependent tests, the expected number of false discoveries in a null stratum is not q_ℓ but rather q_ℓ times the number of tests conducted per single "independent" test. We estimated the number of independent tests as in the GWAS literature, by simulating 1,000 independent null traits with a heritability of 0.5, testing each trait against our 382 annotations, and asking for what S we see at least one p-value $\leq 0.05/S$ in approximately 5% of the 1,000 null traits. This procedure gave us $S = 250$. We then estimated the global FDR using the equation

$$q = \frac{q_\ell(D + 382T/S)}{D + 1}. \tag{14}$$

This yielded the reported global FDR of 9.4%.

Pruning 77 significant associations to 12 independent signals

To prune our set of 77 significant associations to a set of approximately independent results, we used the following iterative greedy approach for each trait: we chose the pair of associations whose annotations had the most strongly correlated signed LD profiles, removed the annotation with the less significant p-value, and repeated until no annotations in the result set had signed LD profiles that were correlated at $R^2 > 0.25$. We used correlation between signed LD profiles rather than between the annotations themselves because, since our method regresses the summary statistics on the signed LD profile rather than the raw annotation, correlation between signed LD profiles most accurately represents the correlation between the test statistics for the two annotations.

Analysis of diseases and complex traits with annotations corresponding to directional effects of minor alleles

We constructed an alternate set of 382 annotations as follows. For each of the 382 ChIP-seq experiments represented by a set of peaks C, we set

$$v_m = \mathbf{1}\{m \in C\}u^1_m \tag{15}$$

where u^1 is the signed background annotation corresponding to SNPs in the bottom quintile of the MAF spectrum. We then used signed LD profile regression to test for association between each of these 382 annotations and each of our 46 traits, assessing significance as above.

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/`.

Code availability

Open-source software implementing our approach is available at `http://www.github.com/yakirr/sldp`.

References

[1] Matthew T. Maurano et al. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA". In: *Science (New York, N.Y.)* 337.6099 (Sept. 7, 2012), pp. 1190–1195. ISSN: 0036-8075. DOI: 10.1126/science.1222794. pmid: 22955828. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3771521/.

[2] Gosia Trynka et al. "Chromatin Marks Identify Critical Cell Types for Fine Mapping Complex Trait Variants". In: *Nature Genetics* 45.2 (Feb. 2013), pp. 124–130. ISSN: 1546-1718. DOI: 10.1038/ng.2504. pmid: 23263488.

[3] Hilary K. Finucane et al. "Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics". In: *Nature Genetics* 47.11 (Nov. 2015), pp. 1228–1235. ISSN: 1061-4036. DOI: 10.1038/ng.3404. URL: http://www.nature.com/ng/journal/v47/n11/full/ng.3404.html#/introduction (visited on 06/17/2017).

[4] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. "An Expanded View of Complex Traits: From Polygenic to Omnigenic". In: *Cell* 169.7 (June 15, 2017), pp. 1177–1186. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2017.05.038. pmid: 28622505. URL: http://www.cell.com/cell/abstract/S0092-8674(17)30629-3 (visited on 10/02/2017).

[5] Xiang Zhu and Matthew Stephens. "A Large-Scale Genome-Wide Enrichment Analysis Identifies New Trait-Associated Genes, Pathways and Tissues across 31 Human Phenotypes". In: *bioRxiv* (July 8, 2017), p. 160770. DOI: 10.1101/160770. URL: https://www.biorxiv.org/content/early/2017/07/08/160770 (visited on 10/02/2017).

[6] Konrad J. Karczewski et al. "Systematic Functional Regulatory Assessment of Disease-Associated Variants". In: *Proceedings of the National Academy of Sciences* 110.23 (Apr. 6, 2013), pp. 9607–9612. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1219099110. pmid: 23690573. URL: http://www.pnas.org/content/110/23/9607 (visited on 10/02/2017).

[7] Anthony Mathelier, Wenqiang Shi, and Wyeth W. Wasserman. "Identification of Altered Cis-Regulatory Elements in Human Disease". In: *Trends in genetics: TIG* 31.2 (Feb. 2015), pp. 67–76. ISSN: 0168-9525. DOI: 10.1016/j.tig.2014.12.003. pmid: 25637093.

[8] Alkes L. Price, Chris C. A. Spencer, and Peter Donnelly. "Progress and Promise in Understanding the Genetic Basis of Common Diseases". In: *Proc. R. Soc. B* 282.1821 (Dec. 22, 2015), p. 20151684. ISSN: 0962-8452, 1471-2954. DOI: 10.1098/rspb.2015.1684. pmid: 26702037. URL: http://rspb.royalsocietypublishing.org/content/282/1821/20151684 (visited on 10/02/2017).

[9] Roger Pique-Regi et al. "Accurate Inference of Transcription Factor Binding from DNA Sequence and Chromatin Accessibility Data". In: *Genome Research* 21.3 (Mar. 2011), pp. 447–455. ISSN: 1088-9051. DOI: 10.1101/gr.112623.110. pmid: 21106904. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3044858/.

[10] Dongwon Lee et al. "A Method to Predict the Impact of Regulatory Variants from DNA Sequence". In: *Nature Genetics* 47.8 (Aug. 2015), pp. 955–961. ISSN: 1061-4036. DOI: 10.1038/ng.3331. URL: http://www.nature.com/ng/journal/v47/n8/full/ng.3331.html (visited on 10/02/2017).

[11] Jian Zhou and Olga G. Troyanskaya. "Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model". In: *Nature Methods* 12.10 (Oct. 2015), pp. 931–934. ISSN: 1548-7091. DOI: 10.1038/nmeth.3547. URL: http://www.nature.com/nmeth/journal/v12/n10/full/nmeth.3547.html (visited on 10/02/2017).

[12] Babak Alipanahi et al. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning". In: *Nature Biotechnology* 33.8 (Aug. 2015), pp. 831–838. ISSN: 1087-0156. DOI: 10.1038/nbt.3300. URL: http://www.nature.com/nbt/journal/v33/n8/full/nbt.3300.html?foxtrotcallback=true (visited on 10/02/2017).

[13] Roadmap Epigenomics Consortium et al. "Integrative Analysis of 111 Reference Human Epigenomes". In: *Nature* 518.7539 (Feb. 19, 2015), pp. 317–330. ISSN: 0028-0836. DOI: 10.1038/nature14248. URL: http://www.nature.com/nature/journal/v518/n7539/full/nature14248.html (visited on 10/10/2017).

[14] Haoyang Zeng et al. "GERV: A Statistical Method for Generative Evaluation of Regulatory Variants for Transcription Factor Binding". In: *Bioinformatics* 32.4 (Feb. 15, 2016), pp. 490–496. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv565. URL: https://academic.oup.com/bioinformatics/article/32/4/490/1743515/GERV-a-statistical-method-for-generative (visited on 10/10/2017).

[15] David R. Kelley, Jasper Snoek, and John Rinn. "Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks". In: *Genome Research* (May 3, 2016), gr.200535.115. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.200535.115. pmid: 27197224. URL: http://genome.cshlp.org/content/early/2016/05/03/gr.200535.115 (visited on 06/17/2017).

[16] Bogdan Pasaniuc and Alkes L. Price. "Dissecting the Genetics of Complex Traits Using Summary Association Statistics". In: *Nature Reviews Genetics* 18.2 (Feb. 2017), pp. 117–127. ISSN: 1471-0056. DOI: 10.1038/nrg.2016.142. URL: https://www.nature.com/nrg/journal/v18/n2/abs/nrg.2016.142.html (visited on 10/02/2017).

[17] Jian Yang et al. "Common SNPs Explain a Large Proportion of the Heritability for Human Height". In: *Nature Genetics* 42.7 (July 2010), pp. 565–569. ISSN: 1061-4036. DOI: 10.1038/ng.608. URL: http://www.nature.com/ng/journal/v42/n7/abs/ng.608.html (visited on 09/01/2017).

[18] The ENCODE Project Consortium. "An Integrated Encyclopedia of DNA Elements in the Human Genome". In: *Nature* 489.7414 (Sept. 6, 2012), pp. 57–74. URL: http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html (visited on 06/17/2017).

[19] Yambazi Banda et al. "Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort". In: *Genetics* 200.4 (Aug. 1, 2015), pp. 1285–1295. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.115.178616. pmid: 26092716. URL: http://www.genetics.org/content/200/4/1285 (visited on 09/01/2017).

[20] Leonardo Arbiza et al. "Genome-Wide Inference of Natural Selection on Human Transcription Factor Binding Sites". In: *Nature Genetics* 45.7 (July 2013), pp. 723–729. ISSN: 1061-4036. DOI: 10.1038/ng.2658. URL: http://www.nature.com/ng/journal/v45/n7/abs/ng.2658.html?foxtrotcallback=true (visited on 09/01/2017).

[21] Jason Ernst et al. "Genome-Scale High-Resolution Mapping of Activating and Repressive Nucleotides in Regulatory Regions". In: *Nature Biotechnology* 34.11 (Nov. 2016), pp. 1180–1190. ISSN: 1087-0156. DOI: 10.1038/nbt.3678. URL: https://www.nature.com/nbt/journal/v34/n11/full/nbt.3678.html (visited on 09/05/2017).

[22] Lu Chen et al. "Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells". In: *Cell* 167.5 (Nov. 17, 2016), 1398–1414.e24. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.10.026. pmid: 27863251. URL: http://www.cell.com/cell/abstract/S0092-8674(16)31446-5 (visited on 09/01/2017).

[23] Andrew D. Sharrocks et al. "The ETS-Domain Transcription Factor Family". In: *The International Journal of Biochemistry & Cell Biology* 29.12 (Dec. 1, 1997), pp. 1371–1387. ISSN: 1357-2725. DOI: 10.1016/S1357-2725(97)00086-1. URL: http://www.sciencedirect.com/science/article/pii/S1357272597000861.

[24] S K Hansen, P A Baeuerle, and F Blasi. "Purification, Reconstitution, and I Kappa B Association of the c-Rel-P65 (RelA) Complex, a Strong Activator of Transcription." In: *Molecular and Cellular Biology* 14.4 (Apr. 1994), pp. 2593–2603. ISSN: 0270-7306. pmid: 8139561. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC358627/.

[25] Tohru Kimura et al. "Involvement of the IRF-1 Transcription Factor in Antiviral Responses to Interferons". In: *Science* 264.5167 (1994), pp. 1921–1924. ISSN: 0036-8075. DOI: 10.2307/2883984. JSTOR: 2883984.

[26] The UniProt Consortium. "UniProt: The Universal Protein Knowledgebase". In: *Nucleic Acids Research* 45 (D1 Jan. 4, 2017), pp. D158–D169. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1099. URL: https://academic.oup.com/nar/article/45/D1/D158/2605721/UniProt-the-universal-protein-knowledgebase (visited on 09/01/2017).

[27] Fred A. Wright et al. "Heritability and Genomics of Gene Expression in Peripheral Blood". In: *Nature Genetics* 46.5 (May 2014), pp. 430–437. ISSN: 1061-4036. DOI: 10.1038/ng.2951. URL: http://www.nature.com/ng/journal/v46/n5/abs/ng.2951.html (visited on 05/03/2017).

[28] Catherine V. Laiosa, Matthias Stadtfeld, and Thomas Graf. "Determinants of Lymphoid-Myeloid Lineage Diversification". In: *Annual Review of Immunology* 24 (2006), pp. 705–738. ISSN: 0732-0582. DOI: 10.1146/annurev.immunol.24.021605.090742. pmid: 16551264.

[29] Chamutal Bornstein et al. "A Negative Feedback Loop of Transcription Factors Specifies Alternative Dendritic Cell Chromatin States". In: *Molecular cell* 56.6 (Dec. 18, 2014), pp. 749–762. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2014.10.014. pmid: 25453760. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4412443/.

[30] Chris van Oevelen et al. "C/EBPα Activates Pre-Existing and De Novo Macrophage Enhancers during Induced Pre-B Cell Transdifferentiation and Myelopoiesis". In: *Stem Cell Reports* 5.2 (July 30, 2015), pp. 232–247. ISSN: 2213-6711. DOI: 10.1016/j.stemcr.2015.06.007. pmid: 26235892. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4618662/.

[31] Branko Cirovic et al. "C/EBP-Induced Transdifferentiation Reveals Granulocyte-Macrophage Precursor-like Plasticity of B Cells". In: *Stem Cell Reports* 8.2 (Feb. 14, 2017), pp. 346–359. ISSN: 2213-6711. DOI: 10.1016/j.stemcr.2016.12.015. pmid: 28111277.

[32] Francesca Martinato et al. "Analysis of Myc-Induced Histone Modifications on Target Chromatin". In: *PLOS ONE* 3.11 (Nov. 5, 2008), e3650. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0003650. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003650 (visited on 09/04/2017).

[33] Stefano Amente, Luigi Lania, and Barbara Majello. "Epigenetic Reprogramming of Myc Target Genes". In: *American Journal of Cancer Research* 1.3 (Feb. 6, 2011), pp. 413–418. ISSN: 2156-6976. pmid: 21969221. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3180057/.

[34] Candace J. Poole and Jan van Riggelen. "MYC—Master Regulator of the Cancer Epigenome and Transcriptome". In: *Genes* 8.5 (May 13, 2017). ISSN: 2073-4425. DOI: 10.3390/genes8050142. pmid: 28505071. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5448016/.

[35] V. V. Ogryzko et al. "The Transcriptional Coactivators P300 and CBP Are Histone Acetyltransferases". In: *Cell* 87.5 (Nov. 29, 1996), pp. 953–959. ISSN: 0092-8674. pmid: 8945521.

[36] Jennifer E. Phillips and Victor G. Corces. "CTCF: Master Weaver of the Genome". In: *Cell* 137.7 (June 26, 2009), pp. 1194–1211. ISSN: 0092-8674. DOI: 10.1016/j.cell.2009.06.001. URL: http://www.sciencedirect.com/science/article/pii/S0092867409006990.

[37] Hao Wang et al. "Widespread Plasticity in CTCF Occupancy Linked to DNA Methylation". In: *Genome Research* 22.9 (Jan. 9, 2012), pp. 1680–1688. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.136101.111. pmid: 22955980. URL: http://genome.cshlp.org/content/22/9/1680 (visited on 09/06/2017).

[38] Matthew T. Maurano et al. "Role of DNA Methylation in Modulating Transcription Factor Occupancy". In: *Cell Reports* 12.7 (Aug. 18, 2015), pp. 1184–1195. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2015.07.024. URL: http://www.sciencedirect.com/science/article/pii/S2211124715007640.

[39] Po-Ru Loh et al. "Mixed Model Association for Biobank-Scale Data Sets". In: *bioRxiv* (Sept. 27, 2017), p. 194944. DOI: 10.1101/194944. URL: https://www.biorxiv.org/content/early/2017/09/27/194944 (visited on 09/28/2017).

[40] Aysu Okbay et al. "Genome-Wide Association Study Identifies 74 Loci Associated with Educational Attainment". In: *Nature* 533.7604 (May 26, 2016), pp. 539–542. ISSN: 0028-0836. DOI: 10.1038/nature17671. URL: https://www.nature.com/nature/journal/v533/n7604/full/nature17671.html (visited on 09/25/2017).

[41] Deciphering Developmental Disorders Study. "Large-Scale Discovery of Novel Genetic Causes of Developmental Disorders". In: *Nature* 519.7542 (Mar. 12, 2015), pp. 223–228. ISSN: 1476-4687. DOI: 10.1038/nature14135. pmid: 25533962.

[42] Cristina Dias et al. "BCL11A Haploinsufficiency Causes an Intellectual Disability Syndrome and Dysregulates Transcription". In: *The American Journal of Human Genetics* 99.2 (Aug. 4, 2016), pp. 253–274. ISSN: 0002-9297, 1537-6605. DOI: 10.1016/j.ajhg.2016.05.030. pmid: 27453576. URL: http://www.cell.com/ajhg/abstract/S0002-9297(16)30201-4 (visited on 09/01/2017).

[43] Farhad Hormozdiari et al. "Identifying Causal Variants at Loci with Multiple Signals of Association". In: *Genetics* 198.2 (Oct. 1, 2014), pp. 497–508. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.114.167908. pmid: 25104515. URL: http://www.genetics.org/content/198/2/497 (visited on 10/03/2017).

[44] Anindita Basak et al. "*BCL11A* Deletions Result in Fetal Hemoglobin Persistence and Neurodevelopmental Alterations". In: *The Journal of Clinical Investigation* 125.6 (June 1, 2015), pp. 2363–2368. ISSN: 0021-9738. DOI: 10.1172/JCI81163. URL: https://www.jci.org/articles/view/81163 (visited on 09/06/2017).

[45] Alister P. W. Funnell et al. "2p15-P16.1 Microdeletions Encompassing and Proximal to BCL11A Are Associated with Elevated HbF in Addition to Neurologic Impairment". In: *Blood* 126.1 (July 2, 2015), pp. 89–93. ISSN: 0006-4971, 1528-0020. DOI: 10.1182/blood-2015-04-638528. pmid: 26019277. URL: http://www.bloodjournal.org/content/126/1/89 (visited on 09/06/2017).

[46] Ming Zhao et al. "Increased 5-Hydroxymethylcytosine in CD4(+) T Cells in Systemic Lupus Erythematosus". In: *Journal of Autoimmunity* 69 (May 2016), pp. 64–73. ISSN: 1095-9157. DOI: 10.1016/j.jaut.2016.03.001. pmid: 26984631.

[47] Prithvi Raj et al. "Regulatory Polymorphisms Modulate the Expression of HLA Class II Molecules and Promote Autoimmunity". In: *eLife* 5 (Feb. 15, 2016), e12089. ISSN: 2050-084X. DOI: 10.7554/eLife.12089. pmid: 26880555. URL: https://elifesciences.org/content/5/e12089v2 (visited on 05/03/2017).

[48] Zhonghui Tang et al. "CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription". In: *Cell* 163.7 (Dec. 2015), pp. 1611–1627. ISSN: 00928674. DOI: 10.1016/j.cell.2015.11.024. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867415015044 (visited on 05/03/2017).

[49] Benjamin Joachim Schmiedel et al. "17q21 Asthma-Risk Variants Switch CTCF Binding and Regulate IL-2 Production by T Cells". In: *Nature Communications* 7 (Nov. 16, 2016), p. 13426. ISSN: 2041-1723. DOI: 10.1038/ncomms13426. URL: http://www.nature.com/ncomms/2016/161116/ncomms13426/full/ncomms13426.html (visited on 05/03/2017).

[50] Verónica Torrano et al. "CTCF Regulates Growth and Erythroid Differentiation of Human Myeloid Leukemia Cells". In: *The Journal of Biological Chemistry* 280.30 (July 29, 2005), pp. 28152–28161. ISSN: 0021-9258. DOI: 10.1074/jbc.M501481200. pmid: 15941718.

[51] Lylia Ouboussad, Sarah Kreuz, and Pascal F. Lefevre. "CTCF Depletion Alters Chromatin Structure and Transcription of Myeloid-Specific Factors". In: *Journal of Molecular Cell Biology* 5.5 (Oct. 2013), pp. 308–322. ISSN: 1759-4685. DOI: 10.1093/jmcb/mjt023. pmid: 23933634.

[52] Jian Zhao et al. "Preferential Association of a Functional Variant in Complement Receptor 2 with Antibodies to Double-Stranded DNA". In: *Annals of the Rheumatic Diseases* 75.1 (Jan. 1, 2016), pp. 242–252. ISSN: 0003-4967, 1468-2060. DOI: 10.1136/annrheumdis-2014-205584. pmid: 25180293. URL: http://ard.bmj.com/content/75/1/242 (visited on 05/03/2017).

[53] Christopher D. Scharer et al. "ATAC-Seq on Biobanked Specimens Defines a Unique Chromatin Accessibility Structure in Naïve SLE B Cells". In: *Scientific Reports* 6 (June 1, 2016), p. 27030. ISSN: 2045-2322. DOI: 10.1038/srep27030. URL: http://www.nature.com/srep/2016/160601/srep27030/full/srep27030.html (visited on 05/03/2017).

[54] Monkol Lek et al. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans". In: *Nature* 536.7616 (Aug. 18, 2016), pp. 285–291. ISSN: 0028-0836. DOI: 10.1038/nature19057. URL: http://www.nature.com/nature/journal/v536/n7616/full/nature19057.html (visited on 10/05/2017).

[55] Andre Franke et al. "Genome-Wide Meta-Analysis Increases to 71 the Number of Confirmed Crohn's Disease Susceptibility Loci". In: *Nature genetics* 42.12 (2010), pp. 1118–1125.

[56] Luke Jostins et al. "Host-Microbe Interactions Have Shaped the Genetic Architecture of Inflammatory Bowel Disease". In: *Nature* 491.7422 (2012), pp. 119–124.

[57] Chad D. Huff et al. "Crohn's Disease and Genetic Hitchhiking at IBD5". In: *Molecular Biology and Evolution* 29.1 (Jan. 2012), pp. 101–111. ISSN: 1537-1719. DOI: 10.1093/molbev/msr151. pmid: 21816865.

[58] Hailiang Huang et al. "Fine-Mapping Inflammatory Bowel Disease Loci to Single-Variant Resolution". In: *Nature* 547.7662 (July 13, 2017), pp. 173–178. ISSN: 0028-0836. DOI: 10.1038/nature22969. URL: http://www.nature.com/nature/journal/v547/n7662/full/nature22969.html (visited on 10/05/2017).

[59] Alexander Gusev et al. "Integrative Approaches for Large-Scale Transcriptome-Wide Association Studies". In: *Nature Genetics* 48.3 (Mar. 2016), pp. 245–252. ISSN: 1061-4036. DOI: 10.1038/ng.3506. URL: http://www.nature.com/ng/journal/v48/n3/full/ng.3506.html (visited on 05/03/2017).

[60] Haoyang Zeng et al. "Convolutional Neural Network Architectures for Predicting DNA-Protein Binding". In: *Bioinformatics (Oxford, England)* 32.12 (June 15, 2016), pp. i121–i127. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw255. pmid: 27307608.

[61] Brendan Bulik-Sullivan et al. "An Atlas of Genetic Correlations across Human Diseases and Traits". In: *Nature genetics* 47.11 (Nov. 2015), pp. 1236–1241. ISSN: 1061-4036. DOI: 10.1038/ng.3406. pmid: 26414676. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4797329/.

[62] George Davey Smith and Gibran Hemani. "Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies". In: *Human Molecular Genetics* 23 (R1 Sept. 15, 2014), R89–R98. ISSN: 0964-6906. DOI: 10.1093/hmg/ddu328. URL: https://academic.oup.com/hmg/article/23/R1/R89/2900899/Mendelian-randomization-genetic-anchors-for-causal (visited on 10/10/2017).

[63] Marie Verbanck et al. "Widespread Pleiotropy Confounds Causal Relationships between Complex Traits and Diseases Inferred from Mendelian Randomization". In: *bioRxiv* (June 30, 2017), p. 157552. DOI: 10.1101/157552. URL: https://www.biorxiv.org/content/early/2017/06/30/157552 (visited on 10/10/2017).

[64] Alan M. Michelson. "Deciphering Genetic Regulatory Codes: A Challenge for Functional Genomics". In: *Proceedings of the National Academy of Sciences* 99.2 (Jan. 22, 2002), pp. 546–548. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.032685999. pmid: 11805309. URL: http://www.pnas.org/content/99/2/546 (visited on 10/02/2017).

[65] Bart Deplancke, Daniel Alpern, and Vincent Gardeux. "The Genetics of Transcription Factor DNA Binding Variation". In: *Cell* 166.3 (July 28, 2016), pp. 538–554. ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.07.012. pmid: 27471964.

[66] David A. Frank. "Targeting Transcription Factors for Cancer Therapy". In: *IDrugs: the investigational drugs journal* 12.1 (Jan. 2009), pp. 29–33. ISSN: 2040-3410. pmid: 19127502.

[67] Panagiotis A. Konstantinopoulos and Athanasios G. Papavassiliou. "Seeing the Future of Cancer-Associated Transcription Factor Drug Targets". In: *JAMA* 305.22 (June 8, 2011), pp. 2349–2350. ISSN: 0098-7484. DOI: 10.1001/jama.2011.727. URL: http://jamanetwork.com/journals/jama/fullarticle/900536 (visited on 10/03/2017).

[68] Hui Y. Xiong et al. "The Human Splicing Code Reveals New Insights into the Genetic Determinants of Disease". In: *Science* 347.6218 (Jan. 9, 2015), p. 1254806. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1254806. pmid: 25525159. URL: http://science.sciencemag.org/content/347/6218/1254806 (visited on 10/12/2017).

[69] David R. Kelley and Yakir A. Reshef. "Sequential Regulatory Activity Prediction across Chromosomes with Convolutional Neural Networks". In: *bioRxiv* (July 10, 2017), p. 161851. DOI: 10.1101/161851. URL: https://www.biorxiv.org/content/early/2017/07/10/161851 (visited on 10/12/2017).

[70] Ryan Tewhey et al. "Direct Identification of Hundreds of Expression-Modulating Variants Using a Multiplexed Reporter Assay". In: *Cell* 165.6 (June 2, 2016), pp. 1519–1529. ISSN: 0092-8674. DOI: 10.1016/j.cell.2016.04.027. URL: http://www.sciencedirect.com/science/article/pii/S0092867416304214.

[71] Atray Dixit et al. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens". In: *Cell* 167.7 (Dec. 15, 2016), 1853–1866.e17. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.11.038. URL: http://www.cell.com/cell/abstract/S0092-8674(16)31610-5 (visited on 10/10/2017).

[72] Charles P. Fulco et al. "Systematic Mapping of Functional Enhancer-Promoter Connections with CRISPR Interference". In: *Science* (Sept. 29, 2016), aag2445. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aag2445. pmid: 27708057. URL: http://science.sciencemag.org/content/early/2016/10/05/science.aag2445 (visited on 10/12/2017).

[73] Glenn S. Cowley et al. "Parallel Genome-Scale Loss of Function Screens in 216 Cancer Cell Lines for the Identification of Context-Specific Genetic Dependencies". In: *Scientific Data* 1 (Sept. 30, 2014), sdata201435. ISSN: 2052-4463. DOI: 10.1038/sdata.2014.35. URL: https://www.nature.com/articles/sdata201435 (visited on 10/11/2017).

[74] Aviad Tsherniak et al. "Defining a Cancer Dependency Map". In: *Cell* 170.3 (July 27, 2017), 564–576.e16. ISSN: 0092-8674. DOI: 10.1016/j.cell.2017.06.010. URL: http://www.sciencedirect.com/science/article/pii/S0092867417306517.

[75] Tariq M. Rana et al. "Genome-Wide CRISPR Screen for Essential Cell Growth Mediators in Mutant KRAS Colorectal Cancers". In: *Cancer Research* (Sept. 27, 2017). ISSN: 1538-7445. DOI: 10.1158/0008-5472.CAN-17-2043. pmid: 28954733.

[76] Oren Parnas et al. "A Genome-Wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks". In: *Cell* 162.3 (July 30, 2015), pp. 675–686. ISSN: 1097-4172. DOI: 10.1016/j.cell.2015.06.059. pmid: 26189680.

[77] Aravind Subramanian et al. "A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles". In: *bioRxiv* (May 10, 2017), p. 136168. DOI: 10.1101/136168. URL: https://www.biorxiv.org/content/early/2017/05/10/136168 (visited on 10/10/2017).

[78] Jernej Godec et al. "Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation". In: *Immunity* 44.1 (Jan. 19, 2016), pp. 194–206. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2015.12.006. pmid: 26795250. URL: http://www.cell.com/immunity/abstract/S1074-7613(15)00532-4 (visited on 10/12/2017).

[79] Arthur Liberzon et al. "Molecular Signatures Database (MSigDB) 3.0". In: *Bioinformatics (Oxford, England)* 27.12 (June 15, 2011), pp. 1739–1740. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr260. pmid: 21546393.

[80] Tomaz Berisa and Joseph K. Pickrell. "Approximately Independent Linkage Disequilibrium Blocks in Human Populations". In: *Bioinformatics* 32.2 (Jan. 15, 2016), pp. 283–285. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv546. pmid: 26395773. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4731402/.

[81] Armin Schoech et al. "Quantification of Frequency-Dependent Genetic Architectures and Action of Negative Selection in 25 UK Biobank Traits". In: *bioRxiv* (Sept. 13, 2017), p. 188086. DOI: 10.1101/188086. URL: https://www.biorxiv.org/content/early/2017/09/13/188086 (visited on 10/02/2017).

[82] 1000 Genomes Project Consortium et al. "A Global Reference for Human Genetic Variation". In: *Nature* 526.7571 (Oct. 1, 2015), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393. pmid: 26432245.

[83] Brendan K. Bulik-Sullivan et al. "LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies". In: *Nature Genetics* 47.3 (Mar. 2015), pp. 291–295. ISSN: 1061-4036. DOI: 10.1038/ng.3211. URL: http://www.nature.com/ng/journal/v47/n3/full/ng.3211.html (visited on 10/02/2017).

[84] Christopher C. Chang et al. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets". In: *GigaScience* 4 (Feb. 25, 2015), p. 7. ISSN: 2047-217X. DOI: 10.1186/s13742-015-0047-8. URL: https://doi.org/10.1186/s13742-015-0047-8.

[85] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (1995), pp. 289–300. ISSN: 0035-9246. JSTOR: 2346101.

[86] John D. Storey and Robert Tibshirani. "Statistical Significance for Genomewide Studies". In: *Proceedings of the National Academy of Sciences* 100.16 (2003), pp. 9440–9445. URL: http://www.pnas.org/content/100/16/9440.short (visited on 10/04/2016).

[87] Daniel Yekutieli. "Hierarchical False Discovery Rate–Controlling Methodology". In: *Journal of the American Statistical Association* 103.481 (2008), pp. 309–316.

Tables

Trait	Top TF (num)	Top cell line	r_f	p	q
Years of ed.	BCL11A (1)	GM12878 (LCL)	2.4%	3.9×10^{-5}	1.5×10^{-2}
Crohn's	POL2* (20)	GM18951 (LCL)	5.3%	4.8×10^{-5}	1.5×10^{-2}
Anorexia	SP1 (1)	HEPG2 (hepatocyte)	-8.9%	1.1×10^{-4}	4.0×10^{-2}
HDL	FOS (1)	K562 (myeloid)	4.8%	1.2×10^{-4}	4.6×10^{-2}
Eczema	CTCF (12)	MCF7 (mammary)	2.7%	1.4×10^{-4}	3.4×10^{-2}
Crohn's	ELF1 (1)	GM12878 (LCL)	4.9%	1.6×10^{-4}	1.5×10^{-2}
Crohn's	POL2 (1)	U87 (glioblast)	4.4%	2.6×10^{-4}	1.5×10^{-2}
Lupus	CTCF** (36)	K562 (myeloid)	-5.0%	3.6×10^{-4}	4.4×10^{-2}
Crohn's	TBP (1)	HEPG2 (hepatocyte)	5.4%	4.9×10^{-4}	1.5×10^{-2}
Crohn's	E2F1 (1)	HELAS3 (cervical epithelium)	4.3%	6.4×10^{-4}	2.7×10^{-2}
Crohn's	IRF1 (1)	K562 (myeloid)	4.7%	9.8×10^{-4}	1.5×10^{-2}
Crohn's	ETS1 (1)	K562 (myeloid)	6.1%	1.4×10^{-3}	1.5×10^{-2}

Table 1: Independent associations from analysis of diseases and complex traits using signed LD profile regression. 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 TF corresponding to the most significant annotation and the total number of correlated annotations that produced a significant result; the cell line corresponding to the most significant annotation; and the estimate of the functional correlation r_f, the P-value, and the per-trait q-value for the most significant annotation. Linked TFs also producing significant associations include (*) TAF1, TBP, and (**) RAD21.

Figures

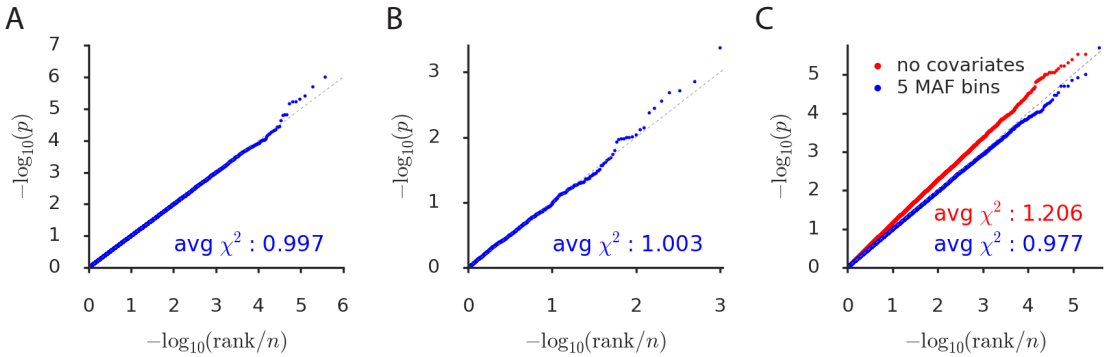

Figure 1: Simulations assessing null calibration. We report null calibration (q-q plots of $-\log_{10}$ P-values) in simulations of (a) no enrichment, (b) unsigned enrichment, and (c) directional effects of minor alleles. The q-q plots are based on (a) 382 annotations $\times 1,000$ simulations $= 382,000$, (b) $1,000$, and (c) two sets of $382 \times 1,000 = 382,000$ P-values. A 5-MAF-bin signed background model is included in all cases except for the red points in part (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 Table S2.

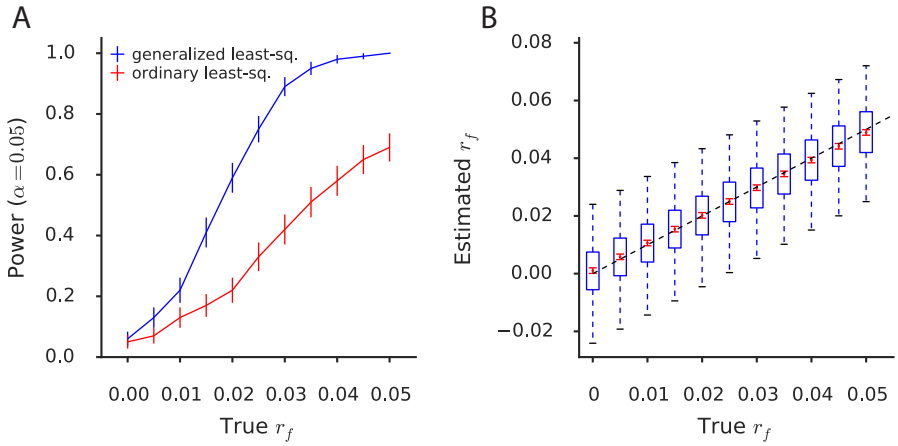

Figure 2: Simulations assessing power, bias, and variance. (a) Power curves under simulation scenarios comparing signed LD profile regression using generalized least-squares (i.e., weighting) to an ordinary (i.e., unweighted) regression of the summary statistics on the signed LD profile. Error bars indicate standard errors of power estimates. (b) Assessment of bias and variance of the signed LD profile regression estimate of r_f at realistic sample size $(47,360)$ and heritability (0.5), 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. Numerical results are reported in Table S3.

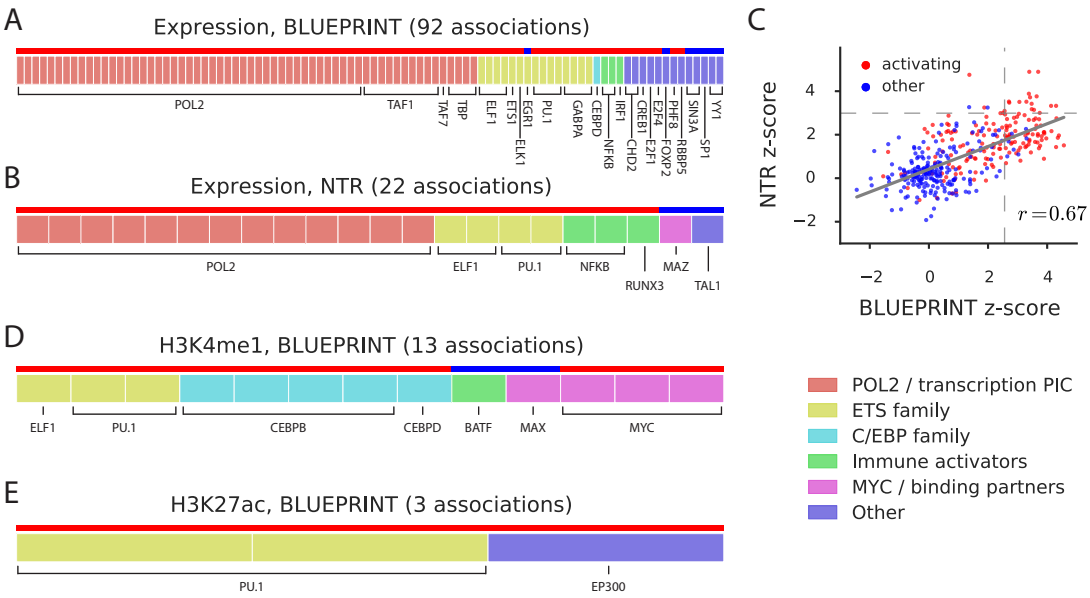

Figure 3: Analysis of molecular traits using signed LD profile regression. Each segmented bar in (a,b,d,e) represents the set of significant annotations at a per-trait FDR of 5% for the indicated traits, with each annotation corresponding to a particular TF profiled in a particular cell line. Results in (a,d,e) are aggregated across the 3 BLUEPRINT cell types. The stripe above each segmented bar is colored red for UniProt activating TFs (see main text) and blue for other TFs. (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 and blue again indicate UniProt activating TFs and other TFs, respectively. Dashed lines represent significance thresholds for 5% FDR. Numerical results are reported in Table S5.

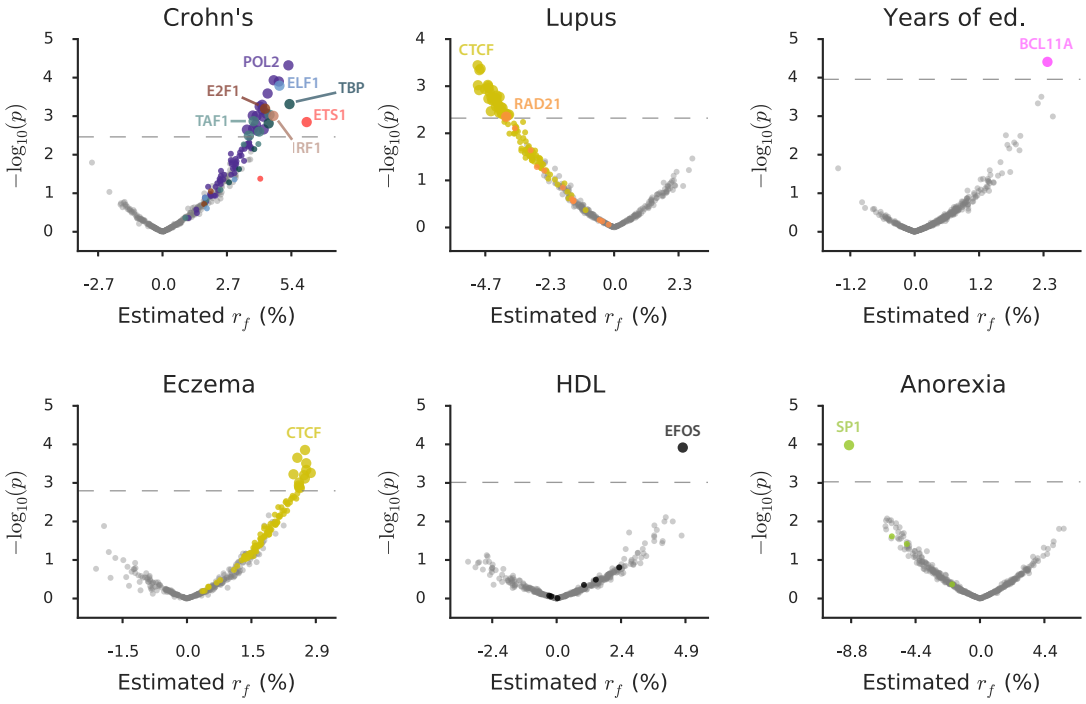

Figure 4: Analysis of diseases and complex traits using signed LD profile regression. For each disease or complex trait with at least one significant result, we plot $-\log_{10}(p)$ against estimated effect size for each of the 382 annotations analyzed. Points are colored by TF identity, with TFs with no significant associations for the trait colored in gray. Larger points denote significant results. The number of significant results for each trait is: Crohn's, 26; Lupus, 36; Years of education, 1; Eczema, 12; HDL, 1; Anorexia, 1. Numerical results are reported in Table S7a.

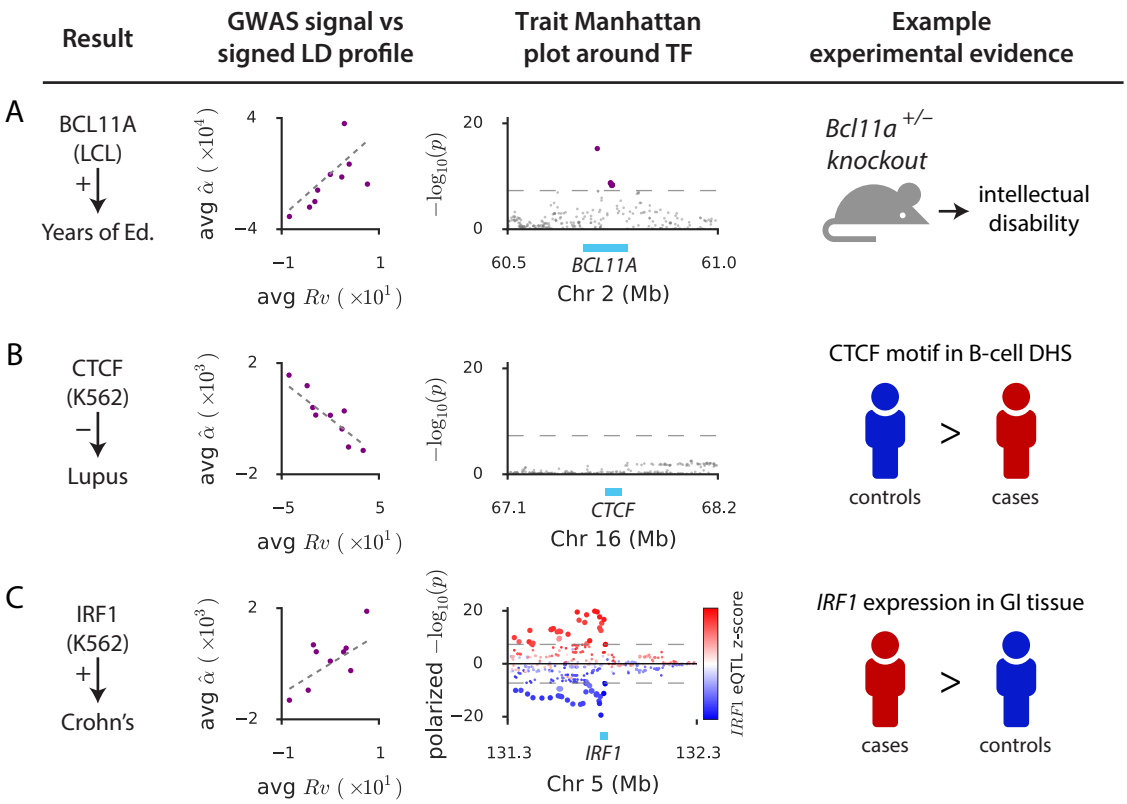

Figure 5: Genetic and non-genetic evidence for three TF binding-complex trait associations. For each of (a) BCL11A-years of education, (b) CTCF-lupus, (c) IRF1-Crohn's disease associations, we display plots of the marginal correlation $\hat{\alpha}$ of SNP to trait versus the signed LD profile Rv of the annotation in question, with SNPs collapsed into bins of $4,000$ SNPs and a larger bin around $Rv = 0$; Manhattan plots of the trait GWAS signal near the associated TF; and example experimental evidence from the literature. For Crohn's disease, the GWAS signal is polarized by direction of effect on disease and points are colored by direction and magnitude of association of each SNP to expression of *IRF1*. Additional experimental evidence relevant to each association is summarized in the main text. GI: gastrointestinal. Numerical results are reported in Table S8.