# Supplementary Information for: Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk

July 1, 2018

## Contents

# Supplementary Note

## Model and estimands

### The model

Let $M$ be the length of the genome. Given a genotype vector $x \in \mathbb{R}^M$ of an individual sampled randomly from some population distribution and a vector $\beta \in \mathbb{R}^M$ of causal SNP effects, we model the phenotype $y$ with a standard linear model:

$$y|\beta, x \sim \mathcal{N}(x^T\beta, \sigma_e^2). \tag{1}$$

We assume that the genotypes are standardized in the population, i.e., that $E(x_m) = 0$ and $E(x_m^2) = 1$ for all SNPs $m$. We assume the same of the phenotype: $E(y) = 0$ and $E(y^2) = 1$. Because our GWAS sample will be very large, these assumptions are for expositional convenience only.

The last ingredient of our model is the connection between $\beta$ and the signed functional annotation of interest $v \in \mathbb{R}^M$. To get this, we assume that $\beta$ is sampled from a distribution satisfying

$$E(\beta|v) = \mu v, \quad \text{cov}(\beta|v) = \sigma^2 I \tag{2}$$

where $\mu$ and $\sigma$ are scalars.

### The estimands

The first estimand we might be interested in is $\mu$, which would tell us the expected change in the per-normalized-genotype effect $\beta_m$ of SNP $m$ for every unit increase of $v_m$. However, this estimand depends on the units of $v$: if we multiply $v$ by a constant $c$, then $\mu$ is decreased by a factor of $c$. We therefore introduce a second estimand, the *functional correlation* $r_f$, which is defined as the genetic correlation between $y$ and the 100%-heritable phenotype $x^Tv$, i.e.,

$$r_f := \text{corr}(x^T\beta, x^Tv). \tag{3}$$

Under our model,

$$\begin{aligned} \text{cov}(x^T\beta, x^Tv) &= E(\beta^T x x^T v) && (4)\\ &= E(\beta)^T E(xx^T)v && (5)\\ &= \mu v^T R v && (6) \end{aligned}$$

where $R = E(xx^T) \in \mathbb{R}^{M\times M}$ is the (signed) population LD matrix of the genotypes, and $v$ is fixed and known. Since

$$\text{var}(x^Tv) = E(v^T x x^T v) = v^T R v, \tag{7}$$

we obtain

$$r_f = \frac{\text{cov}(x^T\beta, x^Tv)}{\sqrt{\text{var}(x^T\beta)\text{var}(x^Tv)}} = \mu\sqrt{\frac{v^TRv}{h_g^2}}. \tag{8}$$

where $h_g^2 = \text{var}(x^T\beta)$ is the SNP-heritability of the phenotype. Note that $r_f$ can also be derived under a model in which $v$ is also modeled as random and jointly distributed with $\beta$, in which case $r_f$ is equal to a standard random-effects genetic correlation.[1] The choice to model $v$ as fixed here arises from the fact that, since it is a complicated biological object, we wish to make as few assumptions as possible about its structure.

In addition to $\mu$ and $r_f$, we might wish to know how much total phenotypic variance is explained by the signed contribution of $v$ to $\beta$. This parameter, $h_v^2$, is defined by

$$h_v^2 := \text{var}(\mu x^T v) = \mu^2 v^T R v. \tag{9}$$

This is equal to the prediction $r^2$ that we would obtain if we tried to predict $y$ from $x^Tv$. If we scale $h_v^2$ by the total heritability of $y$, we obtain the proportion of heritability explained by the signed contribution of $v$, i.e.,

$$\frac{h_v^2}{h_g^2} = \frac{\mu^2 v^T R V}{h_g^2} = r_f^2. \tag{10}$$

We remark that for annotations with small support, $r_f$ and its associated quantities should generally expected to be small in magnitude. To see this, define $h^2_{|v|}$ to be the prediction $r^2$ that we would obtain if we predicted $y$ from an optimal predictor that was constrained to be zero outside the support of $v$. By construction we have $h^2_v \leq h^2_{|v|}$, but since $h^2_{|v|}$ is the total phenotypic variance explained by SNPs in the support of $v$, this implies that $r^2_f = h^2_v/h^2_g \leq h^2_{|v|}/h^2_g$ is at most the proportion of heritability explained by the SNPs in the support of $v$.

## Derivations and description of method

### Main derivation

Now suppose that $N$ individuals $x_1, \ldots, x_N$ have been sampled i.i.d. from the population with corresponding phenotypes $y_1, \ldots y_N$, and that we are given the vector of marginal correlations between each SNP and the trait, i.e., we are given

$$\hat{\alpha} := \frac{1}{N} \sum_{n=1}^{N} x_n y_n \in \mathbb{R}^M. \tag{11}$$

It is easily shown that $E(\hat{\alpha}|\beta) = R\beta$ (see Proposition 2 in the Appendix), from which it follows that

$$\begin{aligned} E(\hat{\alpha}|v) &= E(E(\hat{\alpha}|\beta, v)|v) && (12)\\ &= E(R\beta|v) && (13)\\ &= \mu R v. && (14) \end{aligned}$$

This means that naive regression of $\hat{\alpha}$ on the *signed LD profile* $Rv$ of $v$ is an unbiased estimator of $\mu$. However, ordinary least-squares is best powered when the observations have i.i.d. noise. In this regression, each SNP provides one observation $(\hat{\alpha}_m, (Rv)_m)$, but under our model the covariance of $\hat{\alpha}_m$ and $\hat{\alpha}_{m'}$ given $Rv$ is non-zero. Therefore, if we can model this covariance structure properly, we should be able to use generalized least-squares to reduce variance and increase power. In Theorem 1 of the Appendix, we show that indeed

$$\text{cov}(\hat{\alpha}|v) \approx \sigma^2 R^2 + \frac{R}{N} =: \Omega. \tag{15}$$

The default version of signed LD profile regression estimates $\Omega$ from the reference panel and the chi-squared statistics of the GWAS in question and then performs generalized least-squares using a pseudo-inverse of $\Omega$ to de-couple correlated errors among SNPs. 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 $\beta$ via $R^{-1}\hat{\alpha}$ and then regressing this estimate on $v$ to obtain $\mu$, 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 $\Omega$ is inaccurate due to, e.g., mis-match between the reference panel and the study population.

The point estimate arising from the regression described above is an estimate $\hat{\mu}$ of $\mu$. To obtain an estimate of $r_f$, we plug into Equation 3, estimating $h^2_g$ using the "aggregate estimator" of heritability[2] given by

$$\hat{h^2_g} := \frac{|\hat{\alpha}|_2^2 - \frac{M}{N}}{\frac{1}{M_{5,50}} \sum_m \widehat{\ell_m}} \tag{16}$$

where $|\hat{\alpha}|_2$ is the $\ell_2$-norm of $\hat{\alpha}$, $\widehat{\ell_m}$ is a reference-panel-based estimate of the LD-score $\ell_m := \sum_{m'} R^2_{mm'}$ of SNP $m$, and $M_{5,50}$ is the number of causal SNPs with MAF between 5% and 50%. Equation 3 also has a $v^T R v$ term; for convenience we approximate this term by $v^T v$; our simulations show that we do not suffer from this approximation, and it is empirically quite accurate for our annotations (data not shown).

To estimate $h^2_v/h^2_g$, we use the jackknife to estimate the sampling variance $\widehat{\tau^2}$ of the statistic $\widehat{r_f}$, and then report $\widehat{r_f}^2 - \widehat{\tau^2}$. Though this is an exactly unbiased estimate of $h^2_v$ only if $\widehat{r_f}$ is normally distributed and the jackknife provides an accurate estimate of the sampling variance of $\mu$, our simulations show that it is very close to unbiased in practice. Note that while we use a jackknife estimate of the variance of $\widehat{r_f}$ to

estimate $r_f^2$, this is not how we compute P-values for null hypothesis testing; for details of null hypothesis testing, see below.

To estimate $h_v^2$, we simply multiply our estimate of $r_f^2 = h_v^2/h_g^2$ by our estimate of $h_g^2$.

### Untyped SNPs

Typically, our set of potentially causal SNPs is much larger than the set of SNPs for which we have GWAS summary statistics. Signed LD profile works well in such scenarios: it simply uses only the entries of $Rv$ corresponding to typed SNPs in the regression. Because drastically different sets of typed SNPs require estimation of $\Omega$ anew, we estimate $\Omega$ assuming that all non-MHC HapMap3 SNPs are typed, and then restrict the summary statistics for each trait analyzed to non-MHC HapMap3 SNPs only.

### 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.[3] We then define the null distribution of our statistic as the distribution arising from independently multiplying $v$ by one 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 $\beta$ 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 $u$. This is done 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 for a "signed background model" consisting of 5 annotations $u^1, \ldots, u^5$, defined by

$$u_m^i = \mathbf{1}\left\{\text{MAF}_m \text{ is in } i\text{-th quintile}\right\} \sqrt{2\text{MAF}_m(1-\text{MAF}_m)^{1+\alpha_s}} \tag{17}$$

where $\text{MAF}_m$ is the minor allele frequency of SNP $m$ and $\alpha_s$ 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 $\alpha_s = -0.3$.[4]

## Computational considerations

We model the LD matrix $R$ as being block-diagonal, with the block endpoints defined by recombination hotspots.[3] This allows both more statistically efficient estimation of the true $Rv$ as well as more efficient computation.

For estimating $\Omega$, we use the above block-diagonal decomposition, together with a truncated singular value decomposition applied in each block. Specifically, we store enough singular vectors to capture 95% of the spectrum of each LD block. This is a pre-processing step that need only be carried out once per reference panel, and the relevant outputs of this step for the 1000G Phase 3 Europeans can be downloaded from our website.

## Additional details of analyses

We provide here additional details of analyses discussed in the main text and Online Methods.

### Comparison of blood molecular QTL results to UniProt annotations

We tested whether the set of significant positive SLDP associations for blood eQTL/chromatin QTL were enriched for (unambiguously) "activating" TFs in UniProt compared to the set of annotations as a whole, of which 45% corresponded to (unambiguously) "activating" TFs. We did this using 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 annotations for the same TF. Using this method, we determined that both sets of positive associations were highly significantly enriched for (unambiguously) "activating" TFs ($P = 7.9 \times 10^{-43}$ for eQTL results and $P = 1.9 \times 10^{-9}$ for chromatin QTL results).

### Conditional analysis for tissue-specific effects in GTEx

For cases in which a P-value for association to either $\hat{\alpha}^{(t)}$ or $\hat{\alpha}^{(t')}$ was $\leq 10^{-5}$ (one order of magnitude greater than the maximal resolution of of our empirical null hypothesis testing procedure), we replaced that P-value by a closed-form P-value computed by constructing a z-score out of the estimated value of $r_f$ and its jackknife-based standard error.

### Assessment for concordance with absolute expression levels in GTEx tissues

We obtained raw gene expression levels across the GTEx samples as in ref.[5] and filtered both the raw expression levels and our 382 TF binding annotations to the set of 68 TFs that were represented in both data sets. (This procedure excluded, e.g., POL2, which does not correspond to a single gene.) For each of the 34 GTEx tissues $t$ in which we detected significant association(s) among these 68 TFs, we then computed $p_t$, the proportion of the significant TFs in that tissue with a median transcripts per million (TPM) value greater than 5 across the GTEx samples for that tissue (following ref.[6]), and $q_t$, the proportion of the remaining TFs in that tissue with a median TPM value greater than 5 across the GTEx samples for that tissue. Supplementary Figure 7 contains a plot of $p_t$ against $q_t$ across tissues $t$. To evaluate the significance of the trend across tissues that $p_t > q_t$, we compared $p_T = \sum_t s_t p_t / \sum_t s_t$ to $q_T = \sum_t n_t q_t / \sum_t n_t$ where $s_t$ and $n_t$ are the numbers of TFs with significant associations and without significant associations, respectively, in tissue $t$. We then rejected the null hypothesis that $p_T \leq q_T$ using a one-sided two-sample z-test for difference in means.

### Statistical significance for complex trait analysis

In our complex trait analysis, as in our other analyses, we call significance using a per-trait FDR of 5%, following standard practice. However, 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[7] and the global FDR can be approximated using the formula

$$q = \frac{q_l(D+T)}{D+1} \tag{18}$$

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 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 $q_l T$.

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

$$q = \frac{q_l(D + 382T/S)}{D + 1}. \tag{19}$$

This yielded a global FDR of 9.4%.

Throughout this paper, we chose to use the Benjamini-Hochberg procedure rather than more sophisticated procedures such as the Storey-Tibshirani procedure.[8] This is because the latter procedure, while more powerful, is more difficult to analyze in a multi-trait setting in the way that we have done above for Benjamini-Hochberg, and it controls FDR more noisily when applied in situations with only hundreds (rather than thousands) of tests.

### Pruning 77 significant associations to 12 independent signals in complex trait analysis

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. Grouping the results by TF identity gives similar results (13 distinct TF-trait associations as opposed to 12 independent TF-trait associations; see Supplementary Table 21).

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

To verify empirically that our results were not driven by confounding due to directional effects of minor alleles, we re-analyzed our data using an alternate set of 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. Specifically, for each of the 382 ChIP-seq experiments represented by a set of peaks $C$, we set

$$v_m = \mathbf{1}\{m \in C\}\, u^1_m \tag{20}$$

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.

This analysis yielded only 4 significant annotation-trait associations at per-trait FDR$<$5%, implying that minor-allele-driven confounding is unlikely to explain our results. (Due to the small number of associations relative to the number of traits, these 4 minor-allele associations have a global FDR of 92.9% after accounting for 46 traits.) Furthermore, none of these 4 minor-allele associations overlapped with our set of 77 significant associations (see Supplementary Table 9b).

## Additional interpretation of results

### Additional interpretation of BCL11A-Years of education association

As stated in the main text, we observed a positive association between genome-wide binding of BCL11A and years of education, and the genes putatively regulated in cis by BCL11A to achieve its effect on years

of education were enriched for cholesterol metabolism genes and mTOR signalling genes (see Table 1).

Regarding the mTOR gene set, the *MTOR* gene is itself an intellectual disability gene that has been intimately linked to brain development.[9,10] Regarding the cholesterol metabolism gene set, the brain contains approximately 25% of the body's cholesterol (mostly as a component of the myelin sheaths that surround axons)[11,12] with defects in brain cholesterol metabolism being linked to central nervous system disease,[13,14] and BCL11A has recently been shown to influence (and be influenced by) lipid levels.[15–17] Furthermore, the cholesterol metabolism and mTOR gene-set enrichments may be related, as mTOR has been linked to cholesterol metabolism,[18] including in the developing brain.[19]

This information can be distilled into three observations: (i) mTOR is tied to intellectual disability and cholesterol metabolism, (ii) BCL11A is tied to intellectual disability and cholesterol metabolism, (iii) mTOR signalling genes and cholesterol metabolism genes are enriched among the genes modulated in cis by BCL11A to affect cognitive function All three of these observations could be parsimoniously explained by the hypothesis put forward in the main text, that mTOR exerts its effect on intellectual disability by interacting with BCL11A to influence cholesterol metabolism in the developing brain.

**Interpretation of additional results**

We discuss other associations in Table 1 that are not discussed in the main text. Three of these associations support and refine emerging theories of disease, while five are previously unknown. We begin by discussing the three associations that build on previous knowledge.

First, we detected a negative association between genome-wide binding of CCCTC-binding factor (CTCF) and risk of systemic lupus erythematosus (see Supplementary Figure 9a and Supplementary Table 12) that supports an emerging theory of disease. Although there exists anecdotal evidence linking CTCF binding to lupus risk at a few isolated loci,[20–22] these results are susceptible to the effects of LD and pleiotropy, whereas our approach is able to provide stronger evidence for a causal relationship using genome-wide evidence involving TF binding at many concordant loci (at least 100; see Table 1). We note that we do not observe a GWAS signal for lupus at the *CTCF* locus. This could be because the *CTCF* gene is under strong selective constraint (probability of loss-of-function intolerance[23] = 1.00, greater than 99.9% of genes), and/or because of the small sample size of the lupus GWAS. This association therefore demonstrates that signed LD profile regression can yield gene-disease associations in cases when GWAS is under-powered near the gene in question due to selection or small sample size. Our MSigDB gene-set enrichments shed additional light on this relationship: though CTCF has many diverse regulatory functions throughout the genome, the genomic regions driving the CTCF-Lupus association are most significantly enriched in immune gene sets, with the two strongest enrichments being targets of NF-$\kappa$B and genes differentially expressed between two different stages of myeloid differentiation under knockout of the gene *IKZF1* (but not in the presence of *IKZF1*) (see Supplementary Figure 9a). The latter gene-set enrichment, because it pertains to genes putatively regulated in cis by CTCF, suggests a detailed mechanism whereby IKZF1 (itself a transcription factor) regulates or acts in concert with CTCF to activate a broader transcriptional program that opposes myeloid differentiation and reduces lupus risk. This hypothesis makes three predictions, each of which has evidence in the literature and/or publicly available data that we analyzed: (i) It predicts that IKZF1 affects Lupus risk; indeed, the *IKZF1* gene lies inside a Lupus GWAS locus.[24,25] (ii) It predicts that CTCF affects myeloid development; indeed, CTCF has been experimentally shown to slow myeloid differentiation.[26,27] (iii) It predicts that IKZF1 modulates CTCF activity; indeed, we determined using publically available data[28,29] that IKZF1 has ChIP-seq peaks in the vicinity of the *CTCF* promoter (see Supplementary Table 16), consistent with a direct effect of IKZF1 binding on *CTCF* expression, and IKZF1 ChIP-seq peaks have also been shown to be enriched for the CTCF motif,[30] suggesting that these two TFs may also work in concert at binding sites throughout the genome. Thus, the association between CTCF binding and lupus that we detected, together with the associated MSigDB gene-set enrichments, enhances our understanding of the lupus GWAS signal at the *IKZF1* locus by providing evidence for *IKZF1* as the causal gene (out of 7 total protein coding genes within 500kb); suggests a mechanism to explain the effect of IKZF1 on lupus; and proposes a regulatory relationship between IKZF1 and CTCF that unifies disparate molecular evidence for the effects of both of these genes on myeloid development and ties them jointly to lupus risk.

Second, we detected a positive association between genome-wide binding of ELF1 and Crohn's disease (CD). ELF1 is a hematopoietic and immune regulator[31] that, as mentioned in the main text, lies in a

genome-wide significant Crohn's disease locus in a GWAS of a Japanese population,[32,33] along with 10 other protein-coding genes within 500kb. Our top significant MSigDB enrichment for this relationship was a set of genes differentially expressed following treatment with the drug MRL24, which is a PPAR$\gamma$ agonist. PPAR$\gamma$ has been linked to regulation of the colonic antimicrobial response and inflammatory bowel disease in several studies.[34] Moreover, PPAR$\gamma$ agonists have been shown to have clinical efficacy in treating inflammatory bowel disease,[35] with some agents in current clinical use theorized to act in part via this mechanism.[35]

Third, we detected a positive relationship between genome-wide binding of ETS1 and Crohn's disease. ETS1 is known to regulate genes involved in immunity[31] and, as mentioned in the main text, the *ETS1* gene was recently found to lie in a locus associated with CD[36] and IBD,[37] along with 6 other protein-coding genes within 500kb. The top significant MSigDB enrichments for this relationship point to transcriptional programs associated with EI24 and MYC, both of which play important roles in autophagy[38–40] (EI24 is also known as "autophagy-associated transmembrane protein"). These gene-set enrichments suggest that ETS1 may play a role in mediating the well-known relationship between autophagy and CD.[41]

We next discuss the five associations that have not previously been observed from GWAS data.

First, we detected a positive association between binding of RNA polymerase II (POL2) and Crohn's disease (CD) (Supplementary Figure 9b). This association is surprising given the very broad role of POL2 throughout the genome. However, our MSigDB gene-set enrichments shed some light on the biology underlying this association, with many significant enrichments in immune and immune-related gene sets (see Supplementary Table 10). In particular, the top two significant gene sets are genes down-regulated upon immunosuppression and genes involved in cell-cycle regulation (see Supplementary Figure 9b and Supplementary Table 10). Because of the central role of POL2 in gene transcription, these results suggest that there may exist a large set of immune- or proliferation-related genes whose increased expression contributes to CD risk. Indeed, CD is an auto-immune disease, and it has been hypothesized that increased expression is a prominent component of many immune responses since it can be enacted more quickly than decreased expression.[42–44] Furthermore, acute inflammation has been associated in observational studies with CD onset,[45,46] and recent experimental work[47] has shown that the acute inflammatory response in mice is greatly attenuated by non-specific inhibition of the general-purpose transcriptional machinery containing POL2. Our result potentially links these two findings, providing evidence that the observational association between acute inflammation and CD is causal and suggesting that there exists a polygenic liability for acute inflammation that acts via increased transcription of a large set of immune- or proliferation-related genes and contributes to CD risk. To better understand the POL2-CD association, we investigated whether any of the 14 genes comprising the RNA polymerase II protein complex lie inside a CD GWAS locus. We identified a CD GWAS peak located 28kb from one of these genes, *POLR2E*. This locus is quite gene-dense (28 protein-coding genes within 500kb; 3 protein-coding genes within 28kb), and a recent large-scale CD fine-mapping effort[48] was unable to nominate any gene as potentially causal. Thus, our POL2-CD association also nominates a potential causal gene for the CD GWAS association at this gene-dense locus.

The association described above is accompanied by two additional biologically concordant but statistically approximately independent associations: a positive association between CD and genome-wide binding of POL2 in a separate cell line, and a positive association between CD and genome-wide binding of TATA-binding protein (TBP), a component of the transcription pre-initiation complex.

Fourth, we detected a positive association between genome-wide binding of FOS and HDL. In mice, liver-specific overexpression of the *FOS* gene leads to increased intrahepatic cholesterol and modulation of genes in metabolic pathways connected to cholesterol and fatty acid biosynthesis.[49] FOS has also been shown to be up-regulated when HeLa cells are grown in a sterol-depleted medium designed to activate cellular sterol homeostatic machinery,[50] and the AP-1 complex that it forms has been shown to be down-regulated by high-cholesterol diet in model organisms.[51] A different mechanism is suggested by the fact that in humans, a mutation in the *FOS* gene is associated with congenital generalized lipodystrophy, a phenotype characterized by absence of adipocytes.[52] Our top (and only) significant MSigDB gene-set enrichment for this association was genes regulated by NF-$\kappa$B in response to TNF stimulation. This is potentially consistent with emerging relationships between NF-$\kappa$B and FOS,[53] as well as between TNF and HDL.[54]

Fifth, we detected a positive association between E2F1 and Crohn's disease. E2F1 has roles in immunity, and E2f1-deficient mice challenged with lipopolysaccharide exhibit an attenuated inflammatory response.[55] Additionally, chronic colonic inflammation is associated with release of E2F1 inhibition and activation of E2F1 target genes.[56] Finally, activity of RB, an upstream regulator of the E2F1 pathway, is a highly sensitive

and specific test for distinguishing Crohn's disease from ulcerative colitis in some cases, with RB activity being elevated in Crohn's disease.[57]

**Note: suggestively significant CTCF associations** The relationships we detected between CTCF binding and both lupus and eczema (see main text) raised the question of whether any other traits had sub-significant signals of this sort. We investigated this question, with a primary goal of identifying specifically auto-immune diseases with this property and a secondary goal of identifying any traits with this property. We determined that beyond lupus and eczema no other auto-immune trait exhibited a suggestive (per-trait FDR$<$ 25%) association with CTCF binding. However, we note a suggestive positive association between CTCF binding and red blood cell count ($p = 2.7 \times 10^{-4}$; FDR$=$ 11%).

## Additional discussion points

### Relationship to existing methods

Our method differs from unsigned GWAS enrichment methods[2,58–63] 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. An 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.[2] Unsigned enrichments can also be complicated by LD, as functional elements in LD with binding sites of a TF may contribute to its enrichment if not properly modeled.[2] 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.

Our method differs from existing single-locus GWAS methods[64–66] in that it enables stronger statements about causality and mechanism. Regarding causality, this is because a consistent genome-wide directional effect of SNPs predicted to affect TF binding due to sequence change (across a large set of TF binding sites; see Table 1) is less susceptible to pleiotropy, LD, and allelic heterogeneity.[66,67] The robustness of our method to these potential confounders is also greater than that of genetic correlation and Mendelian randomization[1] (MR) analyses, which can be confounded by reverse causality and pleiotropic effects[68–70] (and which would scale poorly because they would 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.

### Local versus global disease mechanisms

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,[71,72] 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 Table 1 and Supplementary Figure 8.)

### Analysis with annotations from DeepSEA, GTRD, and HOCOMOCO

Although we constructed our predicted TF binding annotations using the neural-network predictor Basset,[73] there exist many other effective methods for making such signed predictions[58,74–80] and many other data sets on which to train them.[81–83] In an initial effort to assess these, we repeated our analyses of molecular traits in blood, gene expression in 48 GTEx tissues, and 46 diseases and complex traits using annotations generated via three other approaches: 382 annotations generated using the DeepSEA neural-network predictor[76] applied to the same ENCODE ChIP-seq data that we analyzed using Basset; 184 annotations generated using the Basset predictor trained on a larger but noisier set of meta-analyzed ChIP-seq data from the Gene Transcription Regulation Database[81] (GTRD) followed by our Basset QC procedures; and 276 annotations generated using position-weight matrices (PWMs) from the Homo sapiens Comprehensive Model Collection[82] (HOCOMOCO), which are based in part on data from the GTRD. (See below for additional detail on how these annotations were constructed.)

Results are reported in Supplementary Tables 17, 18, and 19, respectively, and summarized in Supplementary Table 20. For the 382 DeepSEA annotations, we obtained results similar to our primary set of 382 Basset annotations, including replication of many of our top results (see Supplementary Figures 10 and 11 and Supplementary Table 17); intriguingly, we also determined that the concordance between signed LD profile regression results using Basset and DeepSEA was greater than the concordance between Basset and DeepSEA at the level of annotations (see Supplementary Figure 12), suggesting that the signal that is shared between the predictions made by the two methods is indeed biological. The DeepSEA annotations produced fewer significant associations in total (see Supplementary Table 20), although this comparison was restricted to annotations passing our Basset QC procedures, including a filter on Basset prediction accuracy (see Supplementary Figure 11). The 184 GTRD annotations produced fewer significant annotations than either set of annotations created using ENCODE data, though they did identify new associations, especially in GTEx eQTL data (see Supplementary Tables 18 and 20). For the 276 PWM-based annotations from HOCOMOCO, we again observed correlation between results using PWMs and results using Basset (see Supplementary Figures 13 and 14), though this correlation was weaker than the correlation between the DeepSEA results and the Basset results. We identified fewer significant associations overall using the PWM-based annotations than we did using the more sophisticated neural-network based annotations (see Supplementary Tables 19 and 20), providing evidence that the latter methods can provide a scientifically meaningful increase in performance.

### Other potential biological sources of signed annotations

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,[73,75,76] histone modifications,[73,76] splicing,[77,84] and transcription initiation.[85] Additionally, allele-specific molecular assays, massively parallel assays, and CRISPR screens are increasingly yielding high-resolution experimental information about the effects of genetic variation on gene expression[86–91] as well as cellular processes such as growth[92–94] and inflammation.[95] Finally, perturbational differential expression experiments can yield signed predictions for the relationships of genes to a variety of biological processes such as drug response,[96] immune stimuli,[97] and many others.[98] 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.

### Additional limitations of signed LD profile regression

First, 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 Supplementary Table 9b). Second, our results are limited by the quality of the annotations we are able to produce. For example, TF binding is easier to measure in open chromatin and so it may be the case that our annotations for activating TFs are more representative of underlying biology — and therefore better powered — than our annotations for repressing TFs. 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 in magnitude from 2.4% to 8.9% (recall that $r_f$ is analogous to a genetic correlation; see Supplementary Table 9a), although signals of this size for predicted TF binding could be indicative of much stronger associations, 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. Specifically, $r_f^2$ divided by the proportion of SNPs in an annotation quantifies how much heritability the signed TF binding signal that we detect explains as compared to the total heritability explained by a random set of SNPs of the same size. This ratio is as large as 3.5x (see Supplementary Table 9c), implying that our signed TF binding signals can account — in a signed fashion — for substantial trait heritability relative to the proportion of SNPs. Fifth, our annotations are constructed by testing each minor allele in the context of the reference genome and separately from variation at all other SNPs, rather than taking into account potential non-linear interactions between nearby SNPs;[58] this is a source of reduced power but not increased false positive rates. Sixth, though we detected many significant associations overall, there were many diseases and complex traits, including schizophrenia, height, and blood cell traits, for which we did not detect any significant associations using our Basset TF annotations. We believe that three factors may contribute to this: (i) As we observed here and as others have noted as well,[99] auto-immune traits appear to have a stronger association to TFs than other traits, at least for the TFs on which we have systematic, high-quality ChIP-seq data, and these traits comprised only 8 out of 47 (17%) of the diseases and complex traits in our study; it may be that genome-wide directional effects of these TFs are not as prominent a mechanism for other traits. (ii) We construct our annotations by annotating all SNPs in the ChIP-seq peaks for the TF in question; it could be that in many cases these annotations represent multiple opposing or unrelated transcriptional programs, and that restricting them to more specific sets of SNPs would reveal additional genome-wide directional effects. (iii) Genome-wide directional effects may be contingent on annotations constructed using data generated in the "correct" cellular context (beyond the set of cell lines analyzed in this paper). It is possible that additional signed TF-trait associations will be identified as higher-quality functional data sets become more available and molecular hypotheses become more detailed.

## Creation of additional annotations using DeepSEA, GTRD, and HOCOMOCO

### Creation of additional annotations using DeepSEA

For each of the 382 ENCODE TF ChIP-seq tracks used to generate our post-QC Basset annotations, we obtained predictions for the same track using the DeepSEA method from the authors of that method. We then created 382 new annotations using the same procedure used to generate the 382 Basset annotations (see Equation (7) of Online Methods). We analyzed each of these annotations against the blood molecular QTL, the GTEx eQTL, and the 46 diseases and complex traits; for results, see Supplementary Table 17 and Supplementary Figures 10 and 12. We also obtained the reported AUPRCs of Basset and DeepSEA on all 691 of ENCODE TF ChIP-seq tracks; these are compared in Supplementary Figure 11.

### Creation of additional annotations using GTRD

We downloaded all 482 of the meta-cluster tracks from the GTRD (see URLs) and trained Basset to predict these tracks jointly with the ENCODE tracks used to train our main Basset predictor. We created 482 annotations from these tracks using the same procedure used to generate the 382 (ENCODE) Basset annotations (see Equation (7) of Online Methods). Only 149 (31%) of these annotations passed our standard QC filter (Basset prediction AUPRC $> 0.3$ and at least 5,000 SNPs with non-zero annotation values). We analyzed each of these 149 annotations against the blood molecular QTL, the GTEx eQTL, and the 46 diseases and complex traits; for results, see Supplementary Table 18.

### Creation of additional annotations using HOCOMOCO

We downloaded the 402 core human mononucleotide TF binding PWMs from the HOCOMOCO database (see URLs). We filtered these 402 PWMs to those for which the TF in question had a ChIP-seq track among the 382 post-QC ENCODE TF binding tracks used to produce our main set of annotations. For each of the

resulting 58 PWMs, we then created one new annotation for every matching ENCODE TF binding track by using the PWM to score SNPs inside the ChIP-seq peaks in the matching track. This resulted in 276 annotations. To create an annotation from a PWM and an ENCODE TF binding track, we first computed a score $t(x)$ for every SNP allele $x$ via $t(x) = \sum_{i=-l+1}^{0} \exp \mathrm{pwm}_i(x)$ where $l$ is the length of the PWM, and where $\mathrm{pwm}_i(a)$ is the PWM score given by the motif in question to the reference genome sequence with allele $x$ substituted for the SNP in question and the first position of the PWM placed $i$ bases before the SNP. (The PWM score of a sequence is the sum of the entries of the PWM specified by the bases comprising each position of the sequence.[100]) We then treated these scores as binding predictions and produced an annotation from them using the same procedure used to generate the 382 Basset annotations (see Equation (7) of Online Methods). We analyzed each of the resulting 276 annotations against the blood molecular QTL, the GTEx eQTL, and the 46 diseases and complex traits; for results, see Supplementary Table 19 and Supplementary Figures 13 and 14.

## Appendix: the distribution of GWAS summary statistics

We define the vector $\hat{\alpha}$ of marginal correlations between SNPs and trait and derive its first two moments under a variety of relevant models, building up to the signed LD profile regression model.

### Definitions

Let $M$ be the number of SNPs in the genome. Assume we have sampled $N$ genotype vectors $x_1, \ldots, x_N$ i.i.d. from some population distribution, and that the phenotypes $y_1, \ldots, y_N$ of those individuals satisfy

$$y_n = x_n^T \beta + \varepsilon_n \tag{21}$$

where $\beta \in \mathbb{R}^M$ is the vector of true causal SNP effects on trait, and $\varepsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$ are independent of the $x_n$. We assume throughout this section that genotypes are standardized in the population, i.e., $E(x_{nm}) = 0$ and $E(x_{nm}^2) = 1$ for all $n, m$. We assume the same of the phenotype: $E(y_n) = 0$ and $E(y_n^2) = 1$ for all $n$. These assumptions are for expositional convenience.

Let $X \in \mathbb{R}^{N \times M}$ be the matrix whose $n$-th row is $x_n^T$, and let $Y \in \mathbb{R}^N$ be the vector whose $n$-th entry is $y_n$. The vector

$$\hat{\alpha} = \frac{X^T Y}{N}, \tag{22}$$

which has as its $m$-th entry the in-sample marginal correlation between SNP $m$ and the trait, is the vector of *GWAS summary statistics*.

Having defined $\hat{\alpha}$, we now proceed to derive its first two moments, initially for fixed $X$ and fixed $\beta$, and then for fixed $\beta$ only. After doing so, we will impose the distributional assumption on $\beta$ used in signed LD profile regression and, by marginalizing out $\beta$ according to this distribution, we will obtain the result required for this paper.

### Derivation for fixed $X$ and fixed $\beta$

When both $X$ and $\beta$ are fixed, the following proposition[101] gives the moments of $\hat{\alpha}$.

**Proposition 1.** *Under the model defined above, $\hat{\alpha}$ satisfies*

$$\hat{\alpha} | X, \beta \sim \mathcal{N}\left(\hat{R}\beta, \sigma_e^2 \frac{\hat{R}}{N}\right) \tag{23}$$

*where $\hat{R} = X^T X / N$ is the sample covariance matrix of the genotypes.*

*Proof.* Let $\varepsilon \in \mathbb{R}^N$ be the vector whose $n$-th entry is $\varepsilon_n$. When $X$ and $\beta$ are both fixed, it is easy to see that

$$\hat{\alpha} = \frac{1}{N} X^T Y \tag{24}$$

$$= \frac{1}{N} X^T (X^T \beta + \varepsilon) \tag{25}$$

$$= \hat{R}\beta + \frac{1}{N} X^T \varepsilon. \tag{26}$$

The result follows from normality of $\varepsilon$, together with $E(\varepsilon) = 0$, and $\text{var}(X^T\varepsilon/N) = \sigma_e^2 X^T X/N^2 = \sigma_e^2 \hat{R}/N$. □

**Derivation for random $X$ and fixed $\beta$**

When working with summary statistics, it is desirable to explicitly model the relationship between the unobserved individuals and the LD reference panel by assuming the individuals were drawn from a population distribution whose LD properties we are given by the reference panel. The following result states the moments of $\hat{\alpha}$ when we do so. We prove the result assuming Gaussian genotypes, but it can be shown to be robust to this assumption provided there is a lower bound on minor allele frequency relative to sample size.

**Proposition 2.** *Under the model defined above and assuming Gaussian genotypes, $\hat{\alpha}$ satisfies*

$$\hat{\alpha}|\beta \sim \left[ R\beta, \frac{1}{N}\left(R + R\beta\beta^T R\right) \right] \tag{27}$$

*where $R = cov(x_n) \in \mathbb{R}^{M \times M}$ is the population covariance matrix of the genotypes, and the notation $[,]$ is used to specify the mean and covariance of the distribution without specifying any higher moments.*

*Proof.* Application of the law of total expectation to the result from Proposition 1 readily gives

$$E(\hat{\alpha}|\beta) = E\left(E(\hat{\alpha}|X,\beta)|\beta\right) \tag{28}$$

$$= E(\hat{R}\beta|\beta) \tag{29}$$

$$= R\beta. \tag{30}$$

Application of the law of total covariance yields

$$\text{cov}(\hat{\alpha}|\beta) = E\left(\text{cov}(\hat{\alpha}|X,\beta)|\beta\right) + \text{cov}\left(E(\hat{\alpha}|X,\beta)|\beta\right) \tag{31}$$

$$\sigma_e^2 \frac{\hat{R}}{N} + \text{cov}(\hat{R}\beta|\beta). \tag{32}$$

It is left then only to analyze $\text{cov}(\hat{R}\beta|\beta) = E(\hat{R}\beta\beta^T\hat{R}) - R\beta\beta^T R$. To do so, we note that

$$\text{cov}(\hat{R}\beta|\beta)_{mm'} = \left(E(\hat{R}\beta\beta^T\hat{R}) - R\beta\beta^T R\right)_{mm'} \tag{33}$$

$$= \sum_{i,j} \left( E\left(\hat{R}_{mi}\beta_i\beta_j\hat{R}_{jm'}\right) - R_{mi}\beta_i\beta_j R_{jm'} \right) \tag{34}$$

$$= \sum_{i,j} \beta_i\beta_j \left( E\left(\hat{R}_{mi}\hat{R}_{m'j}\right) - R_{mi}R_{m'j} \right) \tag{35}$$

$$= \frac{1}{N} \sum_{i,j} \beta_i\beta_j \left(R_{mm'}R_{ij} + R_{mj}R_{m'i}\right) \tag{36}$$

$$= \frac{1}{N} R_{mm'} \sum_{i,j} \beta_i\beta_j R_{ij} + \frac{1}{N} \sum_{i,j} \beta_i\beta_j R_{mj}R_{m'i} \tag{37}$$

$$= \frac{1}{N} R_{mm'} \beta^T R\beta + \frac{1}{N} \sum_{i,j} \beta_i\beta_j R_{mj}R_{m'i} \tag{38}$$

where Equation 36 follows from the fact that for Gaussian genotypes, Isselis' theorem implies that

$$E(\hat{R}_{mi}\hat{R}_{m'j}) = R_{mi}R_{m'j} + \frac{1}{N}(R_{mm'}R_{ij} + R_{mj}R_{m'i}). \tag{39}$$

The result of this argument can be summarized across all pairs of SNPs $m, m'$ by

$$\text{cov}(\hat{R}\beta|\beta) = \frac{1}{N}\left((\beta^T R\beta)R + R\beta\beta^T R\right), \tag{40}$$

whereupon noticing that $\beta^T R\beta + \sigma_e^2 = \text{var}(y_n) = 1$ completes the proof. □

**Corollary 1.** *Under the model defined above, $\hat{\alpha}$ approximately satisfies*

$$\hat{\alpha}|\beta \sim \left[R\beta, \frac{R}{N}\right] \tag{41}$$

*where $R = cov(x_n) \in \mathbb{R}^{M\times M}$ is the population covariance matrix of the genotypes.*

*Proof.* For a polygenic trait, $\beta_m \approx O(1/M)$, and so $\beta_m\beta_{m'} \approx O(1/M^2)$. This means that we have that $(R\beta\beta^T R)_{mm'} = O(k^2/M^2)$ where $k$ is the number of SNPs in non-zero LD with both SNP $m$ and SNP $m'$. Since $k \ll M$, $k^2/M^2$ is very small compared to $R_{mm'}$. □

We remark that the above argument does indeed require a polygenic trait. In the other extreme of a trait determined entirely by the value of one SNP, $R\beta\beta^T R$ can take on large values around the single causal SNP.

### Derivation for random $X$ and random $\beta$

We now assume the full signed LD profile regression model, i.e., we fix some signed annotation $v \in \mathbb{R}^M$, and let $\beta \sim [\mu v, \sigma^2]$. Under this model, we have the following result.

**Theorem 1.** *If $\beta \sim [\mu v, \sigma^2]$ for some $v \in \mathbb{R}^M$ and $\sigma^2 > 0$, then $\hat{\alpha}$ approximately satisfies*

$$\hat{\alpha}|v \sim \left[\mu Rv, \sigma^2 R^2 + \frac{R}{N}\right] \tag{42}$$

*where $R = cov(x_n) \in \mathbb{R}^{M\times M}$ is the population covariance matrix of the genotypes.*

*Proof.* The law of total expectation applied to the result of Corollary 1 yields $E(\hat{\alpha}|v) = \mu Rv$ as desired. The law of total covariance yields

$$\begin{aligned}
\text{cov}(\hat{\alpha}|v) &\approx E\left(\text{cov}(\hat{\alpha}|\beta)|v\right) + \text{cov}\left(E(\hat{\alpha}|\beta)|v\right) && (43)\\
&= \frac{R}{N} + \text{cov}(R\beta|v) && (44)\\
&= \frac{R}{N} + R\text{cov}(\beta|v)R && (45)\\
&= \frac{R}{N} + \sigma^2 R^2 && (46)
\end{aligned}$$

as desired. □

# References


[1] 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/.

[2] 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).

[3] 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/.

[4] 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).

[5] Hilary K. Finucane et al. "Heritability Enrichment of Specifically Expressed Genes Identifies Disease-Relevant Tissues and Cell Types". In: *Nature Genetics* 50.4 (Apr. 2018), pp. 621–629. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0081-4. URL: https://www.nature.com/articles/s41588-018-0081-4 (visited on 04/16/2018).

[6] Abraham S. Weintraub et al. "YY1 Is a Structural Regulator of Enhancer-Promoter Loops". In: *Cell* 171.7 (Dec. 14, 2017), 1573–1588.e28. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2017.11.008. pmid: 29224777. URL: http://www.cell.com/cell/abstract/S0092-8674(17)31317-X (visited on 04/12/2018).

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

[8] 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).

[9] Jonathan O. Lipton and Mustafa Sahin. "The Neurology of mTOR". In: *Neuron* 84.2 (Oct. 22, 2014), pp. 275–291. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2014.09.034. pmid: 25374355. URL: http://www.cell.com/neuron/abstract/S0896-6273(14)00892-7 (visited on 04/12/2018).

[10] M. R. F. Reijnders et al. "Variation in a Range of mTOR-Related Genes Associates with Intracranial Volume and Intellectual Disability". In: *Nature Communications* 8 (Oct. 20, 2017). ISSN: 2041-1723. DOI: 10.1038/s41467-017-00933-6. pmid: 29051493. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5648772/.

[11] John M. Dietschy and Stephen D. Turley. "Thematic Review Series: Brain Lipids. Cholesterol Metabolism in the Central Nervous System during Early Development and in the Mature Animal". In: *Journal of Lipid Research* 45.8 (Jan. 8, 2004), pp. 1375–1397. ISSN: 0022-2275, 1539-7262. DOI: 10.1194/jlr.R400004-JLR200. pmid: 15254070. URL: http://www.jlr.org/content/45/8/1375 (visited on 04/12/2018).

[12] Frank W. Pfrieger and Nicole Ungerer. "Cholesterol Metabolism in Neurons and Astrocytes". In: *Progress in Lipid Research* 50.4 (Oct. 1, 2011), pp. 357–371. ISSN: 0163-7827. DOI: 10.1016/j.plipres.2011.06.002. URL: http://www.sciencedirect.com/science/article/pii/S0163782711000312.

[13] Alexei R. Koudinov and Natalia V. Koudinova. "Cholesterol Homeostasis Failure as a Unifying Cause of Synaptic Degeneration". In: *Journal of the Neurological Sciences* 229 (Mar. 15, 2005), pp. 233–240. ISSN: 0022-510X, 1878-5883. DOI: 10.1016/j.jns.2004.11.036. pmid: 15760645. URL: http://www.jns-journal.com/article/S0022-510X(04)00457-5/fulltext (visited on 04/12/2018).

[14] Juan Zhang and Qiang Liu. "Cholesterol Metabolism and Homeostasis in the Brain". In: *Protein & Cell* 6.4 (Apr. 2015), pp. 254–264. ISSN: 1674-800X. DOI: 10.1007/s13238-014-0131-3. pmid: 25682154. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383754/.

[15] Elizabeth R. Macari et al. "Simvastatin and T-Butylhydroquinone Suppress KLF1 and BCL11A Gene Expression and Additively Increase Fetal Hemoglobin in Primary Human Erythroid Cells". In: *Blood* 121.5 (Jan. 31, 2013), pp. 830–839. ISSN: 0006-4971. DOI: 10.1182/blood-2012-07-443986. pmid: 23223429. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3563366/.

[16] LINLIN TANG et al. "BCL11A Gene DNA Methylation Contributes to the Risk of Type 2 Diabetes in Males". In: *Experimental and Therapeutic Medicine* 8.2 (Aug. 2014), pp. 459–463. ISSN: 1792-0981. DOI: 10.3892/etm.2014.1783. pmid: 25009601. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4079426/.

[17] Shan Li et al. "Transcription Factor CTIP1/ BCL11A Regulates Epidermal Differentiation and Lipid Metabolism During Skin Development". In: *Scientific Reports* 7.1 (Oct. 18, 2017), p. 13427. ISSN: 2045-2322. DOI: 10.1038/s41598-017-13347-7. URL: https://www.nature.com/articles/s41598-017-13347-7 (visited on 04/12/2018).

[18] Mathieu Laplante and David M. Sabatini. "An Emerging Role of mTOR in Lipid Biosynthesis". In: *Current biology: CB* 19.22 (Dec. 1, 2009), R1046–1052. ISSN: 1879-0445. DOI: 10.1016/j.cub.2009.09.058. pmid: 19948145.

[19] Emily S. Mathews and Bruce Appel. "Cholesterol Biosynthesis Supports Myelin Gene Expression and Axon Ensheathment through Modulation of P13K/Akt/mTor Signaling". In: *Journal of Neuroscience* 36.29 (July 20, 2016), pp. 7628–7639. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.0726-16.2016. pmid: 27445141. URL: http://www.jneurosci.org/content/36/29/7628 (visited on 04/12/2018).

[20] 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.

[21] 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).

[22] 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).

[23] 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).

[24] Jian-Wen Han et al. "Genome-Wide Association Study in a Chinese Han Population Identifies Nine New Susceptibility Loci for Systemic Lupus Erythematosus". In: *Nature Genetics* 41.11 (Nov. 2009), pp. 1234–1237. ISSN: 1546-1718. DOI: 10.1038/ng.472. pmid: 19838193.

[25] Deborah S. Cunninghame Graham et al. "Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with Systemic Lupus Erythematosus". In: *PLOS Genetics* 7.10 (Oct. 27, 2011), e1002341. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1002341. URL: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002341 (visited on 04/12/2018).

[26] 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.

[27] 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.

[28] Carrie A. Davis et al. "The Encyclopedia of DNA Elements (ENCODE): Data Portal Update". In: *Nucleic Acids Research* 46 (D1 Jan. 4, 2018), pp. D794–D801. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1081. pmid: 29126249.

[29] Hilde Schjerven et al. "Genetic Analysis of Ikaros Target Genes and Tumor Suppressor Function in BCR-ABL1+ Pre-B ALL". In: *The Journal of Experimental Medicine* 214.3 (Mar. 6, 2017), pp. 793–814. ISSN: 1540-9538. DOI: 10.1084/jem.20160049. pmid: 28190001.

[30] Jiangwen Zhang et al. "Harnessing of the Nucleosome Remodeling Deacetylase Complex Controls Lymphocyte Development and Prevents Leukemogenesis". In: *Nature immunology* 13.1 (Nov. 13, 2011), pp. 86–94. ISSN: 1529-2908. DOI: 10.1038/ni.2150. pmid: 22080921. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3868219/.

[31] Sarah Gallant and Gary Gilkeson. "ETS Transcription Factors and Regulation of Immunity". In: *Archivum Immunologiae et Therapiae Experimentalis* 54.3 (June 1, 2006), pp. 149–163. ISSN: 0004-069X, 1661-4917. DOI: 10.1007/s00005-006-0017-z. URL: https://link-springer-com/article/10.1007/s00005-006-0017-z (visited on 10/12/2017).

[32] K Yamazaki et al. "A Genome-Wide Association Study Identifies 2 Susceptibility Loci for Crohn's Disease in a Japanese Population." In: *Gastroenterology* 144.4 (Apr. 2013), pp. 781–788. ISSN: 0016-5085. DOI: 10.1053/j.gastro.2012.12.021. pmid: 23266558. URL: http://europepmc.org/abstract/MED/23266558 (visited on 05/03/2017).

[33] Yuta Fuyuno et al. "Genetic Characteristics of Inflammatory Bowel Disease in a Japanese Population". In: *Journal of Gastroenterology* 51.7 (July 2016), pp. 672–681. ISSN: 1435-5922. DOI: 10.1007/s00535-015-1135-3. pmid: 26511940.

[34] Tamas Varga, Zsolt Czimmerer, and Laszlo Nagy. "PPARs Are a Unique Set of Fatty Acid Regulated Transcription Factors Controlling Both Lipid Metabolism and Inflammation". In: *Biochimica et Biophysica Acta* 1812.8 (Aug. 2011), pp. 1007–1022. ISSN: 0006-3002. DOI: 10.1016/j.bbadis.2011.02.014. pmid: 21382489. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117990/.

[35] Joana Torres, Silvio Danese, and Jean-Frédéric Colombel. "New Therapeutic Avenues in Ulcerative Colitis: Thinking out of the Box". In: *Gut* 62.11 (Nov. 1, 2013), pp. 1642–1652. ISSN: 0017-5749, 1468-3288. DOI: 10.1136/gutjnl-2012-303959. pmid: 24104885. URL: http://gut.bmj.com/content/62/11/1642 (visited on 04/12/2018).

[36] David Ellinghaus et al. "Analysis of Five Chronic Inflammatory Diseases Identifies 27 New Associations and Highlights Disease-Specific Patterns at Shared Loci". In: *Nature Genetics* 48.5 (May 2016), pp. 510–518. ISSN: 1546-1718. DOI: 10.1038/ng.3528. URL: https://www.nature.com/articles/ng.3528 (visited on 04/12/2018).

[37] Katrina M. de Lange et al. "Genome-Wide Association Study Implicates Immune Activation of Multiple Integrin Genes in Inflammatory Bowel Disease". In: *Nature Genetics* 49.2 (Feb. 2017), pp. 256–261. ISSN: 1546-1718. DOI: 10.1038/ng.3760. URL: https://www.nature.com/articles/ng.3760 (visited on 04/12/2018).

[38] Ye Tian et al. "C. Elegans Screen Identifies Autophagy Genes Specific to Multicellular Organisms". In: *Cell* 141.6 (June 11, 2010), pp. 1042–1055. ISSN: 1097-4172. DOI: 10.1016/j.cell.2010.04.034. pmid: 20550938.

[39] Pearl P. C. Toh et al. "Myc Inhibition Impairs Autophagosome Formation". In: *Human Molecular Genetics* 22.25 (Dec. 20, 2013), pp. 5237–5248. ISSN: 0964-6906. DOI: 10.1093/hmg/ddt381. URL: https://academic.oup.com/hmg/article/22/25/5237/576587 (visited on 04/12/2018).

[40] Franz X. Schaub et al. "Myc-Directed Suppression of Autophagy Provides Therapeutic Vulnerabilities Targeting Amino Acid Homeostasis". In: *Blood* 126.23 (Dec. 3, 2015), pp. 2450–2450. ISSN: 0006-4971, 1528-0020. URL: http://www.bloodjournal.org/content/126/23/2450 (visited on 04/12/2018).

[41] Paul Henderson and Craig Stevens. "The Role of Autophagy in Crohn's Disease". In: *Cells* 1.3 (Aug. 3, 2012), pp. 492–519. ISSN: 2073-4409. DOI: 10.3390/cells1030492. pmid: 24710487. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3901108/.

[42] Kathrin Davari et al. "Rapid Genome-Wide Recruitment of RNA Polymerase II Drives Transcription, Splicing, and Translation Events during T Cell Responses". In: *Cell Reports* 19.3 (Apr. 18, 2017), pp. 643–654. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.03.069. pmid: 28423325.

[43] Ginger W. Muse et al. "RNA Polymerase Is Poised for Activation across the Genome". In: *Nature Genetics* 39.12 (Dec. 2007), pp. 1507–1511. ISSN: 1061-4036. DOI: 10.1038/ng.2007.21. URL: http://www.nature.com/ng/journal/v39/n12/full/ng.2007.21.html?foxtrotcallback=true (visited on 10/12/2017).

[44] Jie Xu et al. "Transcriptional Pausing Controls A Rapid Antiviral Innate Immune Response In Drosophila". In: *Cell host & microbe* 12.4 (Oct. 18, 2012), pp. 531–543. ISSN: 1931-3128. DOI: 10.1016/j.chom.2012.08.011. pmid: 23084920. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3479682/.

[45] Luis Alberto García Rodríguez, Ana Ruigómez, and Julián Panés. "Acute Gastroenteritis Is Followed by an Increased Risk of Inflammatory Bowel Disease". In: *Gastroenterology* 130.6 (May 2006), pp. 1588–1594. ISSN: 0016-5085. DOI: 10.1053/j.gastro.2006.02.004. pmid: 16697722.

[46] Chad K. Porter et al. "Infectious Gastroenteritis and Risk of Developing Inflammatory Bowel Disease". In: *Gastroenterology* 135.3 (Sept. 2008), pp. 781–786. ISSN: 1528-0012. DOI: 10.1053/j.gastro.2008.05.081. pmid: 18640117.

[47] Alex Rialdi et al. "Topoisomerase 1 Inhibition Suppresses Inflammatory Genes and Protects from Death by Inflammation". In: *Science* 352.6289 (May 27, 2016), aad7993. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aad7993. pmid: 27127234. URL: http://science.sciencemag.org.ezp-prod1.hul.harvard.edu/content/352/6289/aad7993 (visited on 04/12/2018).

[48] 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).

[49] Latifa Bakiri et al. "Liver Carcinogenesis by FOS-Dependent Inflammation and Cholesterol Dysregulation". In: *Journal of Experimental Medicine* (Mar. 29, 2017), jem.20160935. ISSN: 0022-1007, 1540-9538. DOI: 10.1084/jem.20160935. pmid: 28356389. URL: http://jem.rupress.org/content/early/2017/03/28/jem.20160935 (visited on 10/12/2017).

[50] Fabian Bartz et al. "Identification of Cholesterol-Regulating Genes by Targeted RNAi Screening". In: *Cell Metabolism* 10.1 (July 8, 2009), pp. 63–75. ISSN: 1550-4131. DOI: 10.1016/j.cmet.2009.05.009. URL: http://www.sciencedirect.com/science/article/pii/S1550413109001570.

[51] J. Kálmán et al. "High Cholesterol Diet down Regulates the Activity of Activator Protein-1 but Not Nuclear Factor-Kappa B in Rabbit Brain". In: *Life Sciences* 68.13 (Feb. 16, 2001), pp. 1495–1503. ISSN: 0024-3205. pmid: 11253166.

[52] Birgit Knebel et al. "A Mutation in the C-Fos Gene Associated with Congenital Generalized Lipodystrophy". In: *Orphanet Journal of Rare Diseases* 8 (Aug. 7, 2013), p. 119. ISSN: 1750-1172. DOI: 10.1186/1750-1172-8-119. URL: https://doi.org/10.1186/1750-1172-8-119.

[53] Shuichi Fujioka et al. "NF-κB and AP-1 Connection: Mechanism of NF-κB-Dependent Regulation of AP-1 Activity". In: *Molecular and Cellular Biology* 24.17 (Sept. 2004), pp. 7806–7819. ISSN: 0270-7306. DOI: 10.1128/MCB.24.17.7806-7819.2004. pmid: 15314185. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC507000/.

[54] F. Jeffrey Field, Kim Watt, and Satya N. Mathur. "TNF-Alpha Decreases ABCA1 Expression and Attenuates HDL Cholesterol Efflux in the Human Intestinal Cell Line Caco-2". In: *Journal of Lipid Research* 51.6 (June 2010), pp. 1407–1415. ISSN: 1539-7262. DOI: 10.1194/jlr.M002410. pmid: 20103810.

[55] Laura A. Warg et al. "The Role of the E2F1 Transcription Factor in the Innate Immune Response to Systemic LPS". In: *American Journal of Physiology. Lung Cellular and Molecular Physiology* 303.5 (Sept. 2012), pp. L391–400. ISSN: 1522-1504. DOI: 10.1152/ajplung.00369.2011. pmid: 22707615.

[56] Lei Ying et al. "Chronic Inflammation Promotes Retinoblastoma Protein Hyperphosphorylation and E2F1 Activation". In: *Cancer Research* 65.20 (Oct. 15, 2005), pp. 9132–9136. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-05-1358. pmid: 16230367.

[57] Rossana Colla Soletti et al. "Immunohistochemical Analysis of Retinoblastoma and β-Catenin as an Assistant Tool in the Differential Diagnosis between Crohn's Disease and Ulcerative Colitis". In: *PLOS ONE* 8.8 (Aug. 14, 2013), e70786. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0070786. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0070786 (visited on 10/13/2017).

[58] Richard Cowper-Sal lari et al. "Breast Cancer Risk-Associated SNPs Modulate the Affinity of Chromatin for FOXA1 and Alter Gene Expression". In: *Nature Genetics* 44.11 (Nov. 2012), pp. 1191–1198. ISSN: 1546-1718. DOI: 10.1038/ng.2416. pmid: 23001124.

[59] 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/.

[60] 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.

[61] Joseph K. Pickrell. "Joint Analysis of Functional Genomic Data and Genome-Wide Association Studies of 18 Human Traits". In: *American Journal of Human Genetics* 94.4 (Apr. 3, 2014), pp. 559–573. ISSN: 0002-9297. DOI: 10.1016/j.ajhg.2014.03.004. pmid: 24702953. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3980523/.

[62] 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).

[63] 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).

[64] Thomas Whitington et al. "Gene Regulatory Mechanisms Underpinning Prostate Cancer Susceptibility". In: *Nature Genetics* 48.4 (Apr. 2016), pp. 387–397. ISSN: 1546-1718. DOI: 10.1038/ng.3523. pmid: 26950096.

[65] Yunxian Liu et al. "Identification of Breast Cancer Associated Variants That Modulate Transcription Factor Binding". In: *PLoS genetics* 13.9 (Sept. 2017), e1006761. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1006761. pmid: 28957321.

[66] 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).

[67] Michael Wainberg et al. "Vulnerabilities of Transcriptome-Wide Association Studies". In: *bioRxiv* (Oct. 20, 2017), p. 206961. DOI: 10.1101/206961. URL: https://www.biorxiv.org/content/early/2017/10/20/206961 (visited on 04/12/2018).

[68] 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).

[69] 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).

[70] 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).

[71] 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.

[72] 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).

[73] 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).

[74] 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/.

[75] 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).

[76] 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).

[77] 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).

[78] 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).

[79] 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.

[80] Sunil Kumar, Giovanna Ambrosini, and Philipp Bucher. "SNP2TFBS – a Database of Regulatory SNPs Affecting Predicted Transcription Factor Binding Site Affinity". In: *Nucleic Acids Research* 45 (Database issue Jan. 4, 2017), pp. D139–D144. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1064. pmid: 27899579. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210548/.

[81] Ivan Yevshin et al. "GTRD: A Database of Transcription Factor Binding Sites Identified by ChIP-Seq Experiments". In: *Nucleic Acids Research* 45 (D1 Jan. 4, 2017), pp. D61–D67. ISSN: 0305-1048. DOI: 10.1093/nar/gkw951. URL: https://academic.oup.com/nar/article/45/D1/D61/2290924 (visited on 04/12/2018).

[82] Ivan V. Kulakovskiy et al. "HOCOMOCO: Towards a Complete Collection of Transcription Factor Binding Models for Human and Mouse via Large-Scale ChIP-Seq Analysis". In: *Nucleic Acids Research* 46 (D1 Jan. 4, 2018), pp. D252–D259. ISSN: 0305-1048. DOI: 10.1093/nar/gkx1106. URL: https://academic.oup.com/nar/article/46/D1/D252/4616875 (visited on 04/12/2018).

[83] Anand Venkataraman et al. "A Toolbox of Immunoprecipitation-Grade Monoclonal Antibodies to Human Transcription Factors". In: *Nature Methods* (Mar. 19, 2018). ISSN: 1548-7105. DOI: 10.1038/nmeth.4632. URL: https://www-nature-com.ezp-prod1.hul.harvard.edu/articles/nmeth.4632 (visited on 04/15/2018).

[84] 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).

[85] 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).

[86] 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).

[87] Graham McVicker et al. "Identification of Genetic Variants That Affect Histone Modifications in Human Cells". In: *Science (New York, N.Y.)* 342.6159 (Nov. 8, 2013), pp. 747–749. ISSN: 1095-9203. DOI: 10.1126/science.1242429. pmid: 24136359.

[88] Ashley K. Tehranchi et al. "Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk". In: *Cell* 165.3 (Apr. 21, 2016), pp. 730–741. ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.03.041. pmid: 27087447.

[89] 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.

[90] 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).

[91] 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).

[92] 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).

[93] 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.

[94] 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.

[95] 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.

[96] 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).

[97] 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).

[98] 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.

[99] Samuel A. Lambert et al. "The Human Transcription Factors". In: *Cell* 172.4 (Feb. 8, 2018), pp. 650–665. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.01.029. pmid: 29425488.

[100] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. "FIMO: Scanning for Occurrences of a given Motif". In: *Bioinformatics (Oxford, England)* 27.7 (Apr. 1, 2011), pp. 1017–1018. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btr064. pmid: 21330290.

[101] Huwenbo Shi, Gleb Kichaev, and Bogdan Pasaniuc. "Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data". In: *The American Journal of Human Genetics* 99.1 (July 7, 2016), pp. 139–153. ISSN: 0002-9297, 1537-6605. DOI: 10.1016/j.ajhg.2016.05.013. pmid: 27346688. URL: http://www.cell.com/ajhg/abstract/S0002-9297(16)30148-3 (visited on 06/17/2017).

[102] Katerina Hatzi et al. "A Hybrid Mechanism of Action for BCL6 in B Cells Defined by Formation of Functionally Distinct Complexes at Enhancers and Promoters". In: *Cell Reports* 4.3 (Aug. 15, 2013), pp. 578–588. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2013.06.016. pmid: 23911289.

[103] Chuanxin Huang, Katerina Hatzi, and Ari Melnick. "Lineage-Specific Functions of Bcl-6 in Immunity and Inflammation Are Mediated by Distinct Biochemical Mechanisms". In: *Nature Immunology* 14.4 (Apr. 2013), pp. 380–388. ISSN: 1529-2916. DOI: 10.1038/ni.2543. pmid: 23455674.

[104] Christian Hurtz et al. "BCL6-Mediated Repression of P53 Is Critical for Leukemia Stem Cell Survival in Chronic Myeloid Leukemia". In: *The Journal of Experimental Medicine* 208.11 (Oct. 24, 2011), pp. 2163–2174. ISSN: 1540-9538. DOI: 10.1084/jem.20110304. pmid: 21911423.

[105] Srividya Swaminathan et al. "BACH2 Mediates Negative Selection and P53-Dependent Tumor Suppression at the Pre-B Cell Receptor Checkpoint". In: *Nature Medicine* 19.8 (Aug. 2013), pp. 1014–1022. ISSN: 1546-170X. DOI: 10.1038/nm.3247. pmid: 23852341.

[106] Shenglin Mei et al. "Cistrome Data Browser: A Data Portal for ChIP-Seq and Chromatin Accessibility Data in Human and Mouse". In: *Nucleic Acids Research* 45 (D1 Jan. 4, 2017), pp. D658–D662. ISSN: 1362-4962. DOI: 10.1093/nar/gkw983. pmid: 27789702.

[107] Chunhua Song et al. "Targeting Casein Kinase II Restores Ikaros Tumor Suppressor Activity and Demonstrates Therapeutic Efficacy in High-Risk Leukemia". In: *Blood* 126.15 (Oct. 8, 2015), pp. 1813–1822. ISSN: 1528-0020. DOI: 10.1182/blood-2015-06-651505. pmid: 26219304.

[108] R John Simes. "An Improved Bonferroni Procedure for Multiple Tests of Significance". In: *Biometrika* 73.3 (1986), pp. 751–754.

# Supplementary Tables

**Supplementary Table 1: Summary information about ChIP-seq annotations used in analyses.** $v$ denotes annotation, $M$ denotes the total number of SNPs in the reference panel, $|v|_0$ denotes the number of SNPs with non-zero values of $v$, and $|v|_2$ denotes the 2-norm of $v$.

| Lab | Cell line | Experiment | BASSET AUPRC | $\|v\|_0$ | $\|v\|_0/M$ (%) | $\|v\|_2$ |
|---|---|---|---|---|---|---|
| HAIB | SKNSHRA | CTCF | 0.880098 | 18646 | 0.19 | 13.20 |
| BROAD | NHA | CTCF | 0.869841 | 27912 | 0.28 | 12.68 |
| HAIB | A549 | CTCFSC5916 | 0.866840 | 21517 | 0.22 | 12.73 |
| UW | NB4 | CTCF | 0.866150 | 25419 | 0.25 | 13.23 |
| UW | HRE | CTCF | 0.864149 | 28846 | 0.29 | 13.64 |
| HAIB | A549 | CTCFSC5916 | 0.863801 | 21011 | 0.21 | 13.41 |
| UTA | HUVEC | CTCF | 0.861944 | 21000 | 0.21 | 14.18 |
| BROAD | HUVEC | CTCF | 0.859699 | 29576 | 0.30 | 12.68 |
| UW | HFF | CTCF | 0.859124 | 25034 | 0.25 | 11.61 |
| UW | RPTEC | CTCF | 0.858547 | 44995 | 0.45 | 17.53 |
| BROAD | HMEC | CTCF | 0.858372 | 27488 | 0.27 | 12.58 |
| UW | HASP | CTCF | 0.858100 | 29663 | 0.30 | 14.75 |
| UW | GM12878 | CTCF | 0.858056 | 25981 | 0.26 | 13.11 |
| UW | A549 | CTCF | 0.857446 | 35097 | 0.35 | 15.54 |
| UW | HFFMYC | CTCF | 0.857241 | 38004 | 0.38 | 14.93 |
| UTA | GM12878 | CTCF | 0.856204 | 24907 | 0.25 | 15.67 |
| UW | GM06990 | CTCF | 0.855834 | 33120 | 0.33 | 14.51 |
| UW | HMF | CTCF | 0.854815 | 35825 | 0.36 | 16.13 |
| UW | HCFAA | CTCF | 0.854650 | 26214 | 0.26 | 13.36 |
| UW | GM12874 | CTCF | 0.854489 | 24822 | 0.25 | 12.73 |
| UW | HEK293 | CTCF | 0.854351 | 31140 | 0.31 | 15.48 |
| UTA | HEPG2 | CTCF | 0.853428 | 17547 | 0.18 | 13.62 |
| UW | MCF7 | CTCF | 0.852776 | 40427 | 0.40 | 17.06 |
| UW | NHEK | CTCF | 0.852312 | 31784 | 0.32 | 13.27 |
| HAIB | H1HESC | CTCFSC5916 | 0.852040 | 30644 | 0.31 | 18.33 |
| UW | HVMF | CTCF | 0.851735 | 33859 | 0.34 | 14.79 |
| UW | GM12875 | CTCF | 0.851254 | 26436 | 0.26 | 13.21 |
| UW | HCT116 | CTCF | 0.851195 | 36485 | 0.36 | 15.57 |
| UW | GM12865 | CTCF | 0.850843 | 29599 | 0.30 | 14.14 |
| HAIB | HEPG2 | CTCFSC5916 | 0.850684 | 29285 | 0.29 | 17.25 |
| UW | HRPE | CTCF | 0.850296 | 33503 | 0.34 | 16.27 |
| BROAD | H1HESC | CTCF | 0.849116 | 47350 | 0.47 | 20.96 |
| UW | GM12872 | CTCF | 0.847288 | 34212 | 0.34 | 15.09 |
| SYDH | H1HESC | RAD21 | 0.846410 | 35780 | 0.36 | 17.12 |
| UW | BE2C | CTCF | 0.846211 | 41476 | 0.41 | 15.80 |
| UW | HPF | CTCF | 0.845889 | 29441 | 0.29 | 14.13 |
| UW | NHLF | CTCF | 0.845237 | 24971 | 0.25 | 11.64 |
| BROAD | NHDFAD | CTCF | 0.844702 | 33708 | 0.34 | 14.84 |
| UW | SAEC | CTCF | 0.843178 | 27722 | 0.28 | 13.59 |
| BROAD | HSMMT | CTCF | 0.843109 | 39253 | 0.39 | 14.10 |
| BROAD | GM12878 | CTCF | 0.842508 | 39752 | 0.40 | 14.28 |
| BROAD | NHLF | CTCF | 0.842394 | 30215 | 0.30 | 12.99 |
| UW | HELAS3 | CTCF | 0.842036 | 24028 | 0.24 | 11.95 |
| UW | GM12864 | CTCF | 0.841830 | 33480 | 0.33 | 14.86 |

Continued on next page

| Lab | Cell line | Experiment | BASSET AUPRC | $\|v\|_0$ | $\|v\|_0/M$ (%) | $\|v\|_2$ |
|---|---|---|---|---|---|---|
| UW | SKNSHRA | CTCF | 0.841702 | 26551 | 0.27 | 13.96 |
| UW | HCM | CTCF | 0.839966 | 42907 | 0.43 | 15.57 |
| UTA | GLIOBLA | CTCF | 0.839859 | 37388 | 0.37 | 18.58 |
| UTA | K562 | CTCF | 0.838050 | 27610 | 0.28 | 16.98 |
| UW | HUVEC | CTCF | 0.837666 | 23780 | 0.24 | 12.51 |
| UW | K562 | CTCF | 0.835751 | 30678 | 0.31 | 14.23 |
| UW | GM12873 | CTCF | 0.834805 | 36107 | 0.36 | 15.83 |
| UW | HMEC | CTCF | 0.834803 | 36092 | 0.36 | 14.96 |
| BROAD | HEPG2 | CTCF | 0.834631 | 36924 | 0.37 | 14.72 |
| BROAD | HSMM | CTCF | 0.833446 | 34415 | 0.34 | 15.13 |
| UW | HEPG2 | CTCF | 0.831350 | 31010 | 0.31 | 15.52 |
| UW | HPAF | CTCF | 0.830419 | 40688 | 0.41 | 16.57 |
| UW | AG09309 | CTCF | 0.830321 | 31862 | 0.32 | 13.56 |
| BROAD | HELAS3 | CTCF | 0.828969 | 49347 | 0.49 | 15.31 |
| UW | BJ | CTCF | 0.828852 | 32555 | 0.33 | 13.39 |
| BROAD | NHEK | CTCF | 0.828230 | 37413 | 0.37 | 14.19 |
| UW | HEE | CTCF | 0.828217 | 33823 | 0.34 | 13.55 |
| UW | HAC | CTCF | 0.828210 | 36662 | 0.37 | 13.83 |
| UTA | HELAS3 | CTCF | 0.828109 | 25915 | 0.26 | 16.07 |
| UW | AG04450 | CTCF | 0.827331 | 32761 | 0.33 | 13.88 |
| UTA | PROGFIB | CTCF | 0.826811 | 22840 | 0.23 | 14.38 |
| HAIB | ECC1 | CTCFC | 0.826438 | 15251 | 0.15 | 8.81 |
| BROAD | DND41 | CTCF | 0.824320 | 38541 | 0.39 | 13.81 |
| HAIB | H1HESC | RAD21 | 0.823698 | 47411 | 0.47 | 22.20 |
| SYDH | IMR90 | CTCFB | 0.820777 | 26982 | 0.27 | 13.99 |
| UW | AG09319 | CTCF | 0.820556 | 33669 | 0.34 | 14.46 |
| UW | HBMEC | CTCF | 0.819613 | 41152 | 0.41 | 16.62 |
| UW | WI38 | CTCF | 0.819609 | 25725 | 0.26 | 10.62 |
| UTA | H1HESC | CTCF | 0.818739 | 22472 | 0.22 | 15.80 |
| UTA | A549 | CTCF | 0.817553 | 32700 | 0.33 | 17.81 |
| UW | AG10803 | CTCF | 0.817006 | 29517 | 0.30 | 13.69 |
| BROAD | OSTEOBL | CTCF | 0.816996 | 53644 | 0.54 | 16.04 |
| UW | HCPE | CTCF | 0.816798 | 42276 | 0.42 | 16.83 |
| SYDH | GM12878 | CTCFSC15914C20 | 0.815991 | 30691 | 0.31 | 15.49 |
| UTA | MCF7 | CTCF | 0.815467 | 49073 | 0.49 | 22.63 |
| BROAD | K562 | CTCF | 0.815351 | 52427 | 0.52 | 15.60 |
| UW | WERIRB1 | CTCF | 0.815231 | 30972 | 0.31 | 15.58 |
| UTA | MCF7 | CTCF | 0.814259 | 37438 | 0.37 | 18.94 |
| UW | AOAF | CTCF | 0.810198 | 25402 | 0.25 | 12.89 |
| UW | CACO2 | CTCF | 0.808883 | 28146 | 0.28 | 12.68 |
| UW | AG04449 | CTCF | 0.808085 | 24368 | 0.24 | 14.42 |
| SYDH | K562 | CTCFB | 0.807922 | 34266 | 0.34 | 15.56 |
| HAIB | HEPG2 | RAD21 | 0.806753 | 31414 | 0.31 | 14.66 |
| UW | NHDFNEO | CTCF | 0.805912 | 34150 | 0.34 | 13.07 |
| UTA | FIBROBL | CTCF | 0.802580 | 24917 | 0.25 | 14.54 |
| HAIB | K562 | CTCFC | 0.800330 | 29034 | 0.29 | 14.24 |
| SYDH | HEPG2 | RAD21 | 0.795326 | 24061 | 0.24 | 10.74 |
| SYDH | GM12878 | RAD21 | 0.793772 | 22165 | 0.22 | 9.93 |
| UTA | GM19240 | CTCF | 0.787095 | 24254 | 0.24 | 14.44 |
| UTA | GM19238 | CTCF | 0.784621 | 28109 | 0.28 | 15.19 |

Continued on next page

| Lab | Cell line | Experiment | BASSET AUPRC | $\|v\|_0$ | $\|v\|_0/M$ (%) | $\|v\|_2$ |
|---|---|---|---|---|---|---|
| UTA | NHEK | CTCF | 0.782123 | 28029 | 0.28 | 15.70 |
| HAIB | T47D | CTCFSC5916 | 0.780735 | 20119 | 0.20 | 9.44 |
| UTA | GM12891 | CTCF | 0.776692 | 23165 | 0.23 | 13.77 |
| SYDH | GM12878 | SMC3AB9263 | 0.775055 | 22604 | 0.23 | 9.36 |
| HAIB | GM12878 | RAD21 | 0.773313 | 19232 | 0.19 | 10.90 |
| UTA | MCF7 | CTCF | 0.771586 | 32289 | 0.32 | 17.54 |
| SYDH | IMR90 | RAD21 | 0.771096 | 21035 | 0.21 | 10.62 |
| UTA | GM19239 | CTCF | 0.770649 | 21921 | 0.22 | 12.29 |
| UTA | GM12892 | CTCF | 0.764533 | 27003 | 0.27 | 14.40 |
| SYDH | K562 | SMC3AB9263 | 0.764408 | 17833 | 0.18 | 8.29 |
| HAIB | K562 | RAD21 | 0.762473 | 17349 | 0.17 | 10.54 |
| UW | HL60 | CTCF | 0.760612 | 11834 | 0.12 | 6.43 |
| SYDH | HEPG2 | MAFKAB50322 | 0.756003 | 36764 | 0.37 | 16.31 |
| SYDH | HEK293 | POL2 | 0.750713 | 11423 | 0.11 | 2.57 |
| HAIB | SKNSHRA | RAD21 | 0.748781 | 34221 | 0.34 | 14.81 |
| UTA | MCF7 | CTCF | 0.744677 | 33804 | 0.34 | 16.07 |
| UTA | A549 | POL2 | 0.743474 | 13317 | 0.13 | 2.99 |
| UTA | MCF7 | CTCF | 0.737779 | 31703 | 0.32 | 15.80 |
| SYDH | HELAS3 | RAD21 | 0.732822 | 23726 | 0.24 | 9.90 |
| UTA | GLIOBLA | POL2 | 0.730622 | 12444 | 0.12 | 2.89 |
| SYDH | A549 | RAD21 | 0.726374 | 15727 | 0.16 | 8.17 |
| SYDH | GM10847 | POL2 | 0.725536 | 11162 | 0.11 | 2.82 |
| SYDH | K562 | RAD21 | 0.719791 | 11216 | 0.11 | 5.92 |
| UTA | HUVEC | POL2 | 0.710965 | 9848 | 0.10 | 2.62 |
| SYDH | GM18526 | POL2 | 0.704244 | 15927 | 0.16 | 3.59 |
| SYDH | HELAS3 | SMC3AB9263 | 0.703877 | 25410 | 0.25 | 9.28 |
| SYDH | MCF10AES | CFOS | 0.695666 | 52371 | 0.52 | 14.00 |
| SYDH | GM15510 | POL2 | 0.692228 | 18641 | 0.19 | 3.92 |
| SYDH | GM12878 | ZNF143166181AP | 0.691695 | 16121 | 0.16 | 6.52 |
| SYDH | MCF10AES | CFOS | 0.689921 | 41778 | 0.42 | 11.91 |
| SYDH | HEPG2 | SMC3AB9263 | 0.683574 | 21539 | 0.22 | 8.17 |
| SYDH | MCF10AES | CFOS | 0.678308 | 49334 | 0.49 | 12.33 |
| SYDH | MCF10AES | CFOS | 0.672546 | 37719 | 0.38 | 10.03 |
| SYDH | H1HESC | ZNF143 | 0.665846 | 25229 | 0.25 | 8.50 |
| SYDH | GM18951 | POL2 | 0.662339 | 23305 | 0.23 | 4.19 |
| SYDH | K562 | NFYB | 0.661296 | 9570 | 0.10 | 3.91 |
| HAIB | GM12878 | GABP | 0.660956 | 5625 | 0.06 | 2.43 |
| HAIB | ECC1 | POL2 | 0.657365 | 19849 | 0.20 | 3.32 |
| UTA | MCF7 | POL2 | 0.652882 | 18193 | 0.18 | 3.05 |
| HAIB | HEPG2 | TAF1 | 0.650101 | 16181 | 0.16 | 2.94 |
| SYDH | K562 | IRF1 | 0.649426 | 12976 | 0.13 | 3.16 |
| SYDH | K562 | POL2 | 0.647737 | 16308 | 0.16 | 3.35 |
| SYDH | GM12892 | POL2 | 0.645338 | 23295 | 0.23 | 4.12 |
| SYDH | HEPG2 | MAFKSC477 | 0.643218 | 24770 | 0.25 | 9.07 |
| UTA | MCF7 | POL2 | 0.642949 | 15229 | 0.15 | 2.94 |
| SYDH | NB4 | POL2 | 0.641432 | 16158 | 0.16 | 3.31 |
| SYDH | K562 | POL2 | 0.640277 | 15063 | 0.15 | 2.99 |
| SYDH | K562 | POL2 | 0.635903 | 17161 | 0.17 | 3.22 |
| SYDH | K562 | ZNF143 | 0.634772 | 23343 | 0.23 | 7.50 |
| SYDH | HEPG2 | MAFFM8194 | 0.634067 | 25009 | 0.25 | 8.93 |

Continued on next page

| Lab | Cell line | Experiment | BASSET AUPRC | $\|v\|_0$ | $\|v\|_0/M$ (%) | $\|v\|_2$ |
|---|---|---|---|---|---|---|
| HAIB | GM12878 | ELF1SC631 | 0.631869 | 20946 | 0.21 | 5.37 |
| HAIB | H1HESC | TAF1 | 0.627966 | 21837 | 0.22 | 3.08 |
| HAIB | HEPG2 | GABP | 0.627412 | 9290 | 0.09 | 3.07 |
| SYDH | HEPG2 | CEBPB | 0.625633 | 34970 | 0.35 | 14.15 |
| SYDH | K562 | POL2 | 0.624054 | 15843 | 0.16 | 3.14 |
| SYDH | IMR90 | MAFK | 0.620883 | 25154 | 0.25 | 8.57 |
| SYDH | GM18505 | POL2 | 0.618220 | 24625 | 0.25 | 3.97 |
| UTA | HELAS3 | POL2 | 0.617348 | 19384 | 0.19 | 3.25 |
| UTA | PROGFIB | POL2 | 0.617226 | 14761 | 0.15 | 2.91 |
| SYDH | GM19099 | POL2 | 0.606235 | 22799 | 0.23 | 4.01 |
| SYDH | GM19193 | POL2 | 0.604915 | 24050 | 0.24 | 3.91 |
| SYDH | K562 | POL2 | 0.602457 | 15110 | 0.15 | 2.90 |
| HAIB | SKNSH | TAF1 | 0.601160 | 11185 | 0.11 | 2.76 |
| SYDH | HCT116 | POL2 | 0.598756 | 17455 | 0.17 | 2.72 |
| SYDH | PBDE | POL2 | 0.596470 | 22492 | 0.22 | 3.29 |
| HAIB | K562 | TAF1 | 0.594640 | 13400 | 0.13 | 3.11 |
| UTA | MCF7 | POL2 | 0.587761 | 14677 | 0.15 | 2.73 |
| SYDH | MCF10AES | POL2 | 0.581721 | 22034 | 0.22 | 3.45 |
| BROAD | K562 | PLU1 | 0.578953 | 19126 | 0.19 | 2.78 |
| SYDH | IMR90 | CEBPB | 0.577892 | 44228 | 0.44 | 14.66 |
| HAIB | A549 | CREB1SC240 | 0.576054 | 13155 | 0.13 | 3.07 |
| UTA | K562 | POL2 | 0.575441 | 19966 | 0.20 | 3.30 |
| HAIB | GM12878 | PU1 | 0.574256 | 27757 | 0.28 | 9.34 |
| SYDH | GM12878 | POL2 | 0.573648 | 23803 | 0.24 | 3.93 |
| UTA | GM12878 | POL2 | 0.572056 | 17552 | 0.18 | 3.00 |
| HAIB | GM12878 | NRSF | 0.568899 | 5888 | 0.06 | 3.82 |
| BROAD | K562 | PHF8A301772A | 0.566331 | 27457 | 0.27 | 2.88 |
| SYDH | RAJI | POL2 | 0.564973 | 21621 | 0.22 | 3.36 |
| SYDH | HEPG2 | POL2 | 0.563102 | 18212 | 0.18 | 2.71 |
| HAIB | K562 | YY1 | 0.558414 | 10704 | 0.11 | 2.79 |
| HAIB | A549 | POL2 | 0.555363 | 31308 | 0.31 | 3.68 |
| HAIB | A549 | POL2 | 0.553825 | 29976 | 0.30 | 3.58 |
| HAIB | GM12878 | YY1SC281 | 0.553334 | 26103 | 0.26 | 5.34 |
| SYDH | GM12878 | POL2 | 0.552473 | 11117 | 0.11 | 2.41 |
| HAIB | GM12891 | PU1 | 0.551608 | 28912 | 0.29 | 9.97 |
| HAIB | GM12878 | TAF1 | 0.551273 | 12105 | 0.12 | 2.98 |
| SYDH | A549 | CEBPB | 0.551046 | 26389 | 0.26 | 9.72 |
| SYDH | HUVEC | CFOS | 0.550936 | 42775 | 0.43 | 7.57 |
| HAIB | A549 | TAF1 | 0.550319 | 11038 | 0.11 | 2.08 |
| HAIB | GM12892 | POL2 | 0.548292 | 23439 | 0.23 | 3.42 |
| HAIB | HELAS3 | TAF1 | 0.547530 | 14406 | 0.14 | 2.81 |
| HAIB | HEPG2 | POL24H8 | 0.547414 | 18782 | 0.19 | 3.01 |
| SYDH | HEPG2 | JUND | 0.545643 | 23439 | 0.23 | 5.68 |
| SYDH | HELAS3 | HAE2F1 | 0.544870 | 9314 | 0.09 | 1.47 |
| SYDH | HELAS3 | POL2 | 0.543185 | 29222 | 0.29 | 3.11 |
| HAIB | GM12892 | TAF1 | 0.542027 | 8249 | 0.08 | 2.23 |
| SYDH | K562 | MAZAB85725 | 0.541193 | 33691 | 0.34 | 6.34 |
| SYDH | MCF10AES | POL2 | 0.541022 | 25900 | 0.26 | 3.53 |
| SYDH | H1HESC | MAFK | 0.540650 | 8262 | 0.08 | 2.09 |
| HAIB | A549 | ETS1 | 0.539878 | 6635 | 0.07 | 2.60 |

Continued on next page

| Lab | Cell line | Experiment | BASSET AUPRC | $\|v\|_0$ | $\|v\|_0/M$ (%) | $\|v\|_2$ |
|---|---|---|---|---|---|---|
| SYDH | GM12891 | POL2 | 0.538971 | 24040 | 0.24 | 3.79 |
| HAIB | K562 | GABP | 0.535852 | 12143 | 0.12 | 3.59 |
| HAIB | K562 | E2F6 | 0.535787 | 20429 | 0.20 | 2.89 |
| HAIB | HEPG2 | YY1SC281 | 0.535256 | 17564 | 0.18 | 3.27 |
| HAIB | HCT116 | POL24H8 | 0.534399 | 29439 | 0.29 | 4.18 |
| SYDH | HELAS3 | ELK4 | 0.533836 | 6984 | 0.07 | 2.00 |
| HAIB | U87 | NRSF | 0.533645 | 10740 | 0.11 | 3.53 |
| SYDH | H1HESC | TBP | 0.533586 | 17933 | 0.18 | 3.13 |
| SYDH | GM12878 | ELK112771 | 0.532557 | 5585 | 0.06 | 1.90 |
| UTA | H1HESC | POL2 | 0.528904 | 15666 | 0.16 | 2.28 |
| HAIB | HEPG2 | POL2 | 0.527603 | 26528 | 0.27 | 3.51 |
| HAIB | GM12878 | PMLSC71910 | 0.523565 | 21007 | 0.21 | 3.16 |
| HAIB | HEPG2 | NRSF | 0.522989 | 11697 | 0.12 | 3.82 |
| HAIB | K562 | ELF1SC631 | 0.521651 | 20676 | 0.21 | 5.35 |
| SYDH | GM12878 | NFYB | 0.521437 | 14633 | 0.15 | 3.58 |
| HAIB | GM12891 | TAF1 | 0.520083 | 10825 | 0.11 | 2.70 |
| HAIB | HUVEC | POL2 | 0.519612 | 24168 | 0.24 | 3.11 |
| HAIB | A549 | ELF1 | 0.516848 | 8792 | 0.09 | 2.24 |
| HAIB | PFSK1 | FOXP2 | 0.514938 | 15908 | 0.16 | 2.79 |
| SYDH | MCF10AES | E2F4 | 0.514526 | 12559 | 0.13 | 2.58 |
| SYDH | HELAS3 | NFYA | 0.513807 | 5483 | 0.05 | 1.98 |
| SYDH | K562 | HMGN3 | 0.513410 | 18241 | 0.18 | 2.26 |
| SYDH | HELAS3 | NFYB | 0.512540 | 6653 | 0.07 | 2.22 |
| SYDH | HUVEC | CJUN | 0.510520 | 20080 | 0.20 | 4.26 |
| HAIB | HUVEC | POL24H8 | 0.509722 | 35149 | 0.35 | 4.72 |
| HAIB | HEPG2 | ELF1SC631 | 0.509441 | 13489 | 0.13 | 3.73 |
| SYDH | K562 | MAFKAB50322 | 0.508412 | 13001 | 0.13 | 3.37 |
| HAIB | GM12891 | POL2 | 0.505543 | 17852 | 0.18 | 2.78 |
| SYDH | H1HESC | USF2 | 0.503572 | 5202 | 0.05 | 2.27 |
| HAIB | H1HESC | GABP | 0.501419 | 5292 | 0.05 | 1.53 |
| SYDH | K562 | E2F4 | 0.500739 | 7900 | 0.08 | 1.74 |
| SYDH | K562 | MAFF | 0.499311 | 17035 | 0.17 | 4.41 |
| SYDH | IMR90 | POL2 | 0.499139 | 21099 | 0.21 | 2.57 |
| HAIB | H1HESC | USF1 | 0.498243 | 16631 | 0.17 | 6.39 |
| HAIB | K562 | MAX | 0.494249 | 42934 | 0.43 | 5.98 |
| SYDH | HELAS3 | POL2S2 | 0.492278 | 14434 | 0.14 | 2.32 |
| HAIB | H1HESC | NRSF | 0.491469 | 8454 | 0.08 | 5.74 |
| SYDH | HELAS3 | MAZAB85725 | 0.489070 | 16019 | 0.16 | 2.24 |
| HAIB | HELAS3 | NRSF | 0.488734 | 6360 | 0.06 | 4.97 |
| HAIB | GM12891 | YY1SC281 | 0.487772 | 11490 | 0.11 | 2.73 |
| HAIB | HEPG2 | SIN3AK20 | 0.487522 | 17653 | 0.18 | 2.53 |
| HAIB | HELAS3 | POL2 | 0.487393 | 28715 | 0.29 | 3.64 |
| HAIB | K562 | POL2 | 0.486825 | 36854 | 0.37 | 3.37 |
| SYDH | HEPG2 | MAX | 0.486481 | 11059 | 0.11 | 1.92 |
| HAIB | GM12878 | SP1 | 0.486260 | 15317 | 0.15 | 3.48 |
| SYDH | HEPG2 | POL2 | 0.484689 | 20477 | 0.20 | 2.83 |
| HAIB | GM12892 | POL24H8 | 0.483645 | 20500 | 0.21 | 2.59 |
| HAIB | K562 | ETS1 | 0.483398 | 10444 | 0.10 | 2.37 |
| SYDH | GM12878 | MAZAB85725 | 0.483322 | 22411 | 0.22 | 3.16 |
| SYDH | HELAS3 | CJUN | 0.478779 | 16492 | 0.16 | 2.98 |

Continued on next page

| Lab | Cell line | Experiment | BASSET AUPRC | $\|v\|_0$ | $\|v\|_0/M$ (%) | $\|v\|_2$ |
|---|---|---|---|---|---|---|
| SYDH | K562 | CFOS | 0.478299 | 5481 | 0.05 | 2.17 |
| SYDH | HEPG2 | MXI1 | 0.477728 | 21106 | 0.21 | 3.26 |
| HAIB | H1HESC | POL2 | 0.476246 | 26239 | 0.26 | 2.59 |
| SYDH | K562 | CEBPB | 0.474134 | 28505 | 0.29 | 9.12 |
| HAIB | U87 | POL24H8 | 0.473137 | 23582 | 0.24 | 3.29 |
| SYDH | K562 | MAX | 0.471849 | 29516 | 0.30 | 4.86 |
| HAIB | A549 | GABP | 0.471447 | 13855 | 0.14 | 3.02 |
| SYDH | HELAS3 | CHD2 | 0.471053 | 19320 | 0.19 | 3.33 |
| SYDH | K562 | E2F6 | 0.470723 | 16483 | 0.16 | 2.33 |
| HAIB | GM12878 | EGR1 | 0.468941 | 10841 | 0.11 | 2.08 |
| SYDH | HUVEC | MAX | 0.466519 | 6425 | 0.06 | 1.93 |
| HAIB | GM12878 | RUNX3SC101553 | 0.466113 | 56840 | 0.57 | 8.61 |
| HAIB | GM12878 | USF1 | 0.465793 | 7272 | 0.07 | 2.57 |
| HAIB | K562 | USF1 | 0.464692 | 12871 | 0.13 | 4.61 |
| BROAD | K562 | RBBP5A300109A | 0.463994 | 20083 | 0.20 | 1.84 |
| SYDH | K562 | TBP | 0.463143 | 17767 | 0.18 | 3.22 |
| HAIB | K562 | SIN3AK20 | 0.463116 | 8897 | 0.09 | 1.77 |
| SYDH | K562 | CMYC | 0.462873 | 32161 | 0.32 | 5.06 |
| SYDH | A549 | MAX | 0.461439 | 9266 | 0.09 | 1.72 |
| SYDH | HELAS3 | MAX | 0.458337 | 29171 | 0.29 | 4.12 |
| HAIB | HEPG2 | USF1 | 0.457588 | 12887 | 0.13 | 3.90 |
| SYDH | K562 | CCNT2 | 0.456697 | 21697 | 0.22 | 2.94 |
| SYDH | GM12878 | MXI1 | 0.456679 | 19923 | 0.20 | 2.77 |
| HAIB | GM12892 | YY1 | 0.456003 | 12740 | 0.13 | 2.83 |
| HAIB | GM12891 | POL24H8 | 0.455418 | 17929 | 0.18 | 2.50 |
| SYDH | HELAS3 | CEBPB | 0.450802 | 39105 | 0.39 | 7.92 |
| SYDH | NB4 | MAX | 0.449059 | 28193 | 0.28 | 4.72 |
| SYDH | HEPG2 | TBP | 0.448004 | 13778 | 0.14 | 2.88 |
| HAIB | HCT116 | YY1SC281 | 0.447206 | 9601 | 0.10 | 2.36 |
| UTA | MCF7 | CMYC | 0.446932 | 17429 | 0.17 | 2.52 |
| SYDH | K562 | CMYC | 0.446684 | 26346 | 0.26 | 3.95 |
| HAIB | SKNSHRA | YY1SC281 | 0.445929 | 13128 | 0.13 | 2.71 |
| HAIB | H1HESC | YY1SC281 | 0.445242 | 15591 | 0.16 | 2.65 |
| SYDH | HELAS3 | JUND | 0.444612 | 22640 | 0.23 | 4.23 |
| SYDH | HEPG2 | MAZAB85725 | 0.444409 | 12934 | 0.13 | 1.88 |
| UTA | MCF7 | CMYC | 0.443654 | 24235 | 0.24 | 3.51 |
| HAIB | A549 | USF1 | 0.441291 | 7881 | 0.08 | 2.59 |
| SYDH | HEPG2 | CJUN | 0.440671 | 8890 | 0.09 | 1.91 |
| HAIB | SKNSHRA | USF1SC8983 | 0.439829 | 12682 | 0.13 | 3.64 |
| SYDH | GM12878 | MAX | 0.439437 | 14531 | 0.15 | 2.21 |
| HAIB | K562 | POL24H8 | 0.438629 | 19971 | 0.20 | 3.52 |
| HAIB | PFSK1 | NRSF | 0.435981 | 9928 | 0.10 | 4.63 |
| SYDH | H1HESC | SIN3ANB6001263 | 0.433869 | 26283 | 0.26 | 2.93 |
| UTA | HEPG2 | POL2 | 0.432243 | 21612 | 0.22 | 2.23 |
| HAIB | A549 | FOSL2 | 0.430795 | 23494 | 0.24 | 3.95 |
| HAIB | SKNSH | POL24H8 | 0.427949 | 22879 | 0.23 | 3.35 |
| SYDH | HUVEC | POL2 | 0.427119 | 11883 | 0.12 | 1.94 |
| HAIB | K562 | YY1 | 0.426097 | 19380 | 0.19 | 3.54 |
| UCHICAGO | K562 | EFOS | 0.425453 | 6855 | 0.07 | 1.91 |
| SYDH | H1HESC | CHD2 | 0.424343 | 6252 | 0.06 | 1.25 |

Continued on next page

| Lab | Cell line | Experiment | BASSET AUPRC | $\|v\|_0$ | $\|v\|_0/M$ (%) | $\|v\|_2$ |
|---|---|---|---|---|---|---|
| SYDH | MCF7 | HAE2F1 | 0.423359 | 27514 | 0.28 | 2.20 |
| HAIB | K562 | SP1 | 0.422803 | 6215 | 0.06 | 1.58 |
| SYDH | K562 | JUND | 0.420900 | 30409 | 0.30 | 5.93 |
| SYDH | HELAS3 | ZNF143 | 0.420784 | 5406 | 0.05 | 2.13 |
| HAIB | A549 | YY1C | 0.420411 | 11293 | 0.11 | 2.20 |
| SYDH | GM12878 | POL2S2 | 0.420026 | 12996 | 0.13 | 1.84 |
| HAIB | GM12878 | POL2 | 0.419133 | 48007 | 0.48 | 3.33 |
| HAIB | PFSK1 | TAF1 | 0.415078 | 6236 | 0.06 | 1.35 |
| HAIB | K562 | PU1 | 0.411073 | 15386 | 0.15 | 4.70 |
| SYDH | GM12878 | CHD2AB68301 | 0.410210 | 16016 | 0.16 | 2.63 |
| SYDH | NB4 | CMYC | 0.406744 | 23774 | 0.24 | 3.73 |
| HAIB | H1HESC | TAF7SC101167 | 0.406696 | 10442 | 0.10 | 1.54 |
| SYDH | H1HESC | CEBPB | 0.405410 | 11800 | 0.12 | 3.73 |
| SYDH | MCF10AES | STAT3 | 0.404351 | 33486 | 0.33 | 5.08 |
| HAIB | GM12878 | POL24H8 | 0.402366 | 31663 | 0.32 | 2.85 |
| HAIB | SKNSH | NRSF | 0.401931 | 7233 | 0.07 | 3.71 |
| HAIB | K562 | ZBTB7ASC34508 | 0.399912 | 19683 | 0.20 | 2.16 |
| HAIB | K562 | EGR1 | 0.399163 | 24881 | 0.25 | 3.28 |
| SYDH | MCF10AES | STAT3 | 0.398512 | 29538 | 0.30 | 4.81 |
| SYDH | K562 | CHD2AB68301 | 0.398431 | 7834 | 0.08 | 2.01 |
| HAIB | SKNMC | POL24H8 | 0.393543 | 21485 | 0.21 | 2.96 |
| HAIB | H1HESC | POL24H8 | 0.391510 | 19419 | 0.19 | 1.99 |
| HAIB | K562 | CTCFLSC98982 | 0.391258 | 5891 | 0.06 | 2.85 |
| SYDH | MCF10AES | STAT3 | 0.388008 | 31591 | 0.32 | 4.98 |
| HAIB | A549 | USF1 | 0.387810 | 6778 | 0.07 | 1.84 |
| HAIB | HEPG2 | FOXA1SC6553 | 0.386906 | 33656 | 0.34 | 5.34 |
| SYDH | MCF10AES | STAT3 | 0.385338 | 25848 | 0.26 | 4.56 |
| HAIB | SKNSH | NRSF | 0.385146 | 14169 | 0.14 | 3.45 |
| SYDH | GM12891 | NFKB | 0.383466 | 29206 | 0.29 | 4.56 |
| HAIB | H1HESC | SP1 | 0.380258 | 12393 | 0.12 | 2.05 |
| SYDH | MCF10AES | CMYC | 0.379656 | 27000 | 0.27 | 4.33 |
| SYDH | HEPG2 | CEBPB | 0.379397 | 11572 | 0.12 | 4.10 |
| HAIB | K562 | NRSF | 0.379106 | 9598 | 0.10 | 4.30 |
| SYDH | GM12878 | USF2 | 0.377835 | 6661 | 0.07 | 2.16 |
| SYDH | HELAS3 | TBP | 0.376722 | 17555 | 0.18 | 3.06 |
| UTA | K562 | CMYC | 0.372061 | 5833 | 0.06 | 1.68 |
| HAIB | K562 | ATF3 | 0.371010 | 10360 | 0.10 | 2.78 |
| SYDH | HELAS3 | MXI1AF4185 | 0.368398 | 12174 | 0.12 | 1.83 |
| HAIB | HEPG2 | FOSL2 | 0.367104 | 16407 | 0.16 | 3.44 |
| SYDH | K562 | CMYC | 0.366773 | 21209 | 0.21 | 3.20 |
| SYDH | HELAS3 | MAFK | 0.366364 | 9993 | 0.10 | 1.82 |
| SYDH | HELAS3 | P300SC584SC584 | 0.364830 | 18694 | 0.19 | 2.54 |
| HAIB | HEPG2 | SP1 | 0.364172 | 21711 | 0.22 | 3.58 |
| HAIB | K562 | PMLSC71910 | 0.362038 | 18655 | 0.19 | 2.75 |
| HAIB | K562 | FOSL1SC183 | 0.359258 | 6436 | 0.06 | 2.20 |
| HAIB | GM12878 | BCL11A | 0.358333 | 12360 | 0.12 | 2.80 |
| SYDH | GM12878 | SIN3ANB6001263 | 0.356799 | 13694 | 0.14 | 1.61 |
| SYDH | K562 | CJUN | 0.354626 | 5656 | 0.06 | 1.98 |
| SYDH | GM12878 | TBP | 0.353883 | 15238 | 0.15 | 2.78 |
| HAIB | HEPG2 | FOXA1SC101058 | 0.353734 | 29596 | 0.30 | 4.83 |

Continued on next page

| Lab | Cell line | Experiment | BASSET AUPRC | $\|v\|_0$ | $\|v\|_0/M$ (%) | $\|v\|_2$ |
|---|---|---|---|---|---|---|
| HAIB | HEPG2 | CEBPBSC150 | 0.348724 | 9795 | 0.10 | 3.67 |
| HAIB | A549 | NRSF | 0.348252 | 12999 | 0.13 | 3.65 |
| HAIB | GM12878 | BATF | 0.347600 | 18755 | 0.19 | 3.78 |
| HAIB | A549 | USF1 | 0.347257 | 8140 | 0.08 | 2.22 |
| BROAD | H1HESC | RBBP5A300109A | 0.343881 | 25833 | 0.26 | 1.35 |
| HAIB | GM12892 | PAX5C20 | 0.343844 | 8182 | 0.08 | 1.34 |
| BROAD | K562 | POL2B | 0.341811 | 15495 | 0.15 | 1.86 |
| HAIB | GM12878 | NFICSC81335 | 0.341187 | 33737 | 0.34 | 3.76 |
| SYDH | HELAS3 | RFX5200401194 | 0.341053 | 15994 | 0.16 | 2.36 |
| HAIB | GM12878 | IRF4SC6059 | 0.340861 | 14517 | 0.15 | 2.83 |
| HAIB | GM12878 | POU2F2 | 0.336826 | 18566 | 0.19 | 2.97 |
| HAIB | HEPG2 | FOXA2SC6554 | 0.336085 | 27428 | 0.27 | 4.48 |
| HAIB | SKNSH | SIN3AK20 | 0.336066 | 13855 | 0.14 | 1.95 |
| HAIB | GM12878 | ATF2SC81188 | 0.335843 | 26054 | 0.26 | 3.55 |
| SYDH | HELAS3 | USF2 | 0.329562 | 8429 | 0.08 | 1.85 |
| SYDH | HELAS3 | E2F1 | 0.328842 | 5081 | 0.05 | 0.74 |
| SYDH | MCF10AES | CMYC | 0.327448 | 19677 | 0.20 | 2.88 |
| HAIB | HEPG2 | HNF4ASC8987 | 0.325563 | 13192 | 0.13 | 3.15 |
| SYDH | K562 | UBTFSAB1404509 | 0.325086 | 14930 | 0.15 | 1.59 |
| UCHICAGO | K562 | EJUND | 0.323401 | 26489 | 0.26 | 3.49 |
| UTA | GM12878 | CMYC | 0.322020 | 5627 | 0.06 | 0.63 |
| BROAD | K562 | SAP3039731 | 0.320382 | 11693 | 0.12 | 1.16 |
| SYDH | K562 | CMYC | 0.318111 | 11312 | 0.11 | 2.06 |
| HAIB | H1HESC | EGR1 | 0.317297 | 7071 | 0.07 | 0.68 |
| HAIB | K562 | CEBPBSC150 | 0.311232 | 18052 | 0.18 | 3.71 |
| HAIB | H1HESC | SIN3AK20 | 0.310984 | 7354 | 0.07 | 1.48 |
| SYDH | GM15510 | NFKB | 0.309530 | 13887 | 0.14 | 2.14 |
| BROAD | K562 | HDAC1SC6298 | 0.308889 | 15009 | 0.15 | 1.08 |
| SYDH | GM19099 | NFKB | 0.308646 | 6705 | 0.07 | 1.71 |
| HAIB | GM12878 | FOXM1SC502 | 0.307947 | 26561 | 0.27 | 2.91 |
| HAIB | PANC1 | POL24H8 | 0.306956 | 11954 | 0.12 | 1.43 |
| HAIB | HEPG2 | HNF4GSC6558 | 0.305644 | 14815 | 0.15 | 2.92 |
| HAIB | HEPG2 | JUND | 0.305335 | 14409 | 0.14 | 2.61 |
| SYDH | K562 | TAL1SC12984 | 0.304212 | 18090 | 0.18 | 4.50 |
| HAIB | HEPG2 | CEBPDSC636 | 0.303716 | 8698 | 0.09 | 1.82 |
| SYDH | K562 | CORESTSC30189 | 0.303011 | 28293 | 0.28 | 3.98 |
| SYDH | K562 | BHLHE40NB100 | 0.301552 | 19955 | 0.20 | 2.77 |
| HAIB | GM12878 | EBF1SC137065 | 0.301285 | 24230 | 0.24 | 3.43 |

See Excel file

**Supplementary Table 2: Numerical results for Figure 1.** We list all P-values used for the simulations of a) no enrichment, b) unsigned enrichment, and c) directional effects of minor alleles, with and without the 5-MAF-bin signed background model. Simulation details are described in the Methods, and the statistical method is described in the "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 3: Numerical results for Figure 2.** We list a) estimated power, with standard errors, for both methods analyzed in Figure 2a, b) mean estimate of $r_f$, with standard error, for all values of $r_f$ simulated, together with quantiles of the sampling distribution of our estimator. Simulation details are described in the Methods, and the statistical method is described in the "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 4: List of traits analyzed in BLUEPRINT/NTR analysis.** We list the set of traits analyzed in the BLUEPRINT/NTR analysis, with number of typed SNPs for each trait.

See Excel file

**Supplementary Table 5: Details of results of BLUEPRINT/NTR analysis.** We list a) the set of 409 significant associations at per-trait FDR$< 5\%$ for the BLUEPRINT gene expression analysis, with laboratory, cell line, and TF listed for each significant annotation, along with estimated $r_f$, P-value, and whether the TF is known to be activating; b) the set of 18 significant associations at per-trait FDR $< 5\%$ for the NTR gene expression analysis; c) the side-by-side comparison of z-scores from the BLUEPRINT neutrophil expression analysis and the NTR analysis; d) the set of 286 significant associations at per-trait FDR $< 5\%$ for the BLUEPRINT H3K4me1 analysis; and e) the set of 359 significant associations at per-trait FDR $< 5\%$ for the BLUEPRINT H3K27ac analysis. Note that because the QTL summary statistics analyzed here are processed in a way that places different SNPs on different scales, the relative values of $r_f$ in these results are interpretable but the absolute magnitudes are not. GWAS data are described in Supplementary Table 4, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 6: List of GTEx traits analyzed.** We list the set of GTEx traits analyzed, with number of typed SNPs and average sample size for each trait.

See Excel file

**Supplementary Table 7: Results of GTEx analysis.** We list a) the set of 2,330 significant associations at per-trait FDR$<5\%$ for the GTEx gene expression analysis, with laboratory, cell line, and TF listed for each significant annotation, along with estimated $r_f$ and P-value; and b) the same information for the subset of results whose significance did not decrease in the conditional analysis. Note that because the QTL summary statistics analyzed here are processed in a way that places different SNPs on different scales, the relative values of $r_f$ in these results are interpretable but the absolute magnitudes are not. GWAS data are described in Supplementary Table 6, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 8: List of diseases and complex traits analyzed.** We list the set of diseases and complex traits analyzed, with sample size, number of typed SNPs, and estimated SNP-heritability for each trait.

See Excel file

**Supplementary Table 9: Results of SLDP analysis of 46 diseases and complex traits.** We list a) the set of 77 significant associations at per-trait FDR$<5\%$ for the TF annotations, with laboratory, cell line, and transcription factor listed for each significant annotation, along with estimated $r_f$ and P-value; b) the set of 4 significant associations at per-trait FDR $<5\%$ for the alternate set of 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; c) quantification of unsigned heritability explained by signed enrichments reported in (a). Specifically: because $r_f^2$ for an annotation can never exceed the total proportion of heritability explained by the SNPs with nonzero values of the annotation, we computed for each association the ratio of estimated $r_f^2$ to the proportion of SNPs with nonzero values of the annotation. We found that in some cases the signed signal was not only non-trivially different from zero but also substantial enough to imply an unsigned enrichment. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 10: Results of enrichment analysis of signed LD profile regression disease/complex trait results.** We list significant gene-set enrichments for the 77 significant signed LD profile regression associations to diseases and complex traits. For (a) each of the top significant enrichments listed in Table 1 and (b) all of the significant enrichments at per-stratum FDR$<5\%$, we list: details of the annotation and phenotype underlying the signed LD profile regression result, the full name of the enriched gene set, the enrichment, the average signed LD profile covariance among LD blocks containing genes in the set (with standard error), the average signed LD profile covariance among LD block not containing genes in the set (with standard error), a p-value generated by permuting LD blocks, and a q-value calculated among the tests conducted for each signed LD profile result within each MSigDB database. 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 "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 11: Numerical results for Figure 6.** For each result in the figure, we list i) the numerical values used to make the plot of $\hat{\alpha}$ against $Rv$, ii) the association summary statistics used to make the Manhattan plot, and iii) the numerical results underlying the two displayed gene-set enrichments. 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 "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 12: Numerical results for Figure 7.** For each result in the figure, we list i) the numerical values used to make the plot of $\hat{\alpha}$ against $Rv$, ii) the association summary statistics used to make the Manhattan plot, and iii) the numerical results underlying the two displayed gene-set enrichments. 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 "Overview of methods" section and the Methods.

| SNP | P(in causal set) | Causal post. prob. | Z |
|---|---|---|---|
| rs10189857 | 0.25 | 1 | 8.0933 |
| rs356991 | 0.128176 | 0.512705 | 6.03 |
| rs168565 | 0.0366951 | 0.14678 | 5.9928 |
| rs6545816 | 0.154972 | 0.619888 | 5.4231 |
| rs6545817 | 0.0950247 | 0.380099 | 5.3862 |
| rs243071 | 0.25 | 1 | -5.2992 |

**Supplementary Table 13: Fine mapping of EDU signal at *BCL11A* locus**. We list the six SNPs in the 95% credible set when running the CAVIAR method with the parameter $c = 4$. rs10189857 is an intronic SNP in the *BCL11A* gene. (Results with $c = 2$ and $c = 3$ were similar.) GWAS data are described in Supplementary Table 8.

| Cistrome ID | cell type/line | position on chr12 (kb) | TSS | body | reference |
|---|---|---|---|---|---|
| 35517 | OCI-Ly1 (B lymph) | 67552.542-67552.694 | | | Hatzi et al.[102] |
| 35517 | OCI-Ly1 (B lymph) | 67561.828-67561.989 | * | | Hatzi et al.[102] |
| 35517 | OCI-Ly1 (B lymph) | 67562.145-67562.414 | * | * | Hatzi et al.[102] |
| 35517 | OCI-Ly1 (B lymph) | 67563.131-67563.598 | * | * | Hatzi et al.[102] |
| 35517 | OCI-Ly1 (B lymph) | 67644.691-67644.898 | | | Hatzi et al.[102] |
| 35517 | OCI-Ly1 (B lymph) | 67645.077-67645.307 | | | Hatzi et al.[102] |
| 52774 | T lymphocyte | 67562.240-67562.531 | * | * | Hatzi et al.[102] |
| 52774 | T lymphocyte | 67562.729-67562.889 | * | * | Hatzi et al.[102] |
| 52303 | T lymphocyte | 67561.830-67561.976 | * | | Hatzi et al.[102] |
| 52303 | T lymphocyte | 67562.145-67562.304 | * | | Hatzi et al.[102] |
| 52303 | T lymphocyte | 67563.760-67563.957 | * | * | Hatzi et al.[102] |
| 35085 | B lymphocyte | 67554.235-67554.385 | | | Huang et al.[103] |
| 35085 | B lymphocyte | 67562.163-67562.501 | * | * | Huang et al.[103] |
| 35085 | B lymphocyte | 67563.134-67563.543 | * | * | Huang et al.[103] |
| 35085 | B lymphocyte | 67563.826-67564.064 | * | * | Huang et al.[103] |
| 35085 | B lymphocyte | 67644.751-67645.243 | | | Huang et al.[103] |
| 1958 | B JURL-MK1 (myeloid) | 67561.844-67562.437 | * | * | Hurtz et al.[104] |
| 1958 | B JURL-MK1 (myeloid) | 67562.740-67562.886 | * | * | Hurtz et al.[104] |
| 1958 | B JURL-MK1 (myeloid) | 67563.293-67563.487 | * | * | Hurtz et al.[104] |
| 1958 | B JURL-MK1 (myeloid) | 67644.723-67644.898 | | | Hurtz et al.[104] |
| 39572 | B OCI-Ly1 (B lymph) | 67562.242-67562.397 | * | | Swaminathan et al.[105] |
| 39572 | B OCI-Ly1 (B lymph) | 67563.257-67563.556 | * | * | Swaminathan et al.[105] |

**Supplementary Table 14:** BCL6 ChIP-seq peaks within 10kb of the *CTCF* gene body in publicly available ChIP-seq data processed by the cistrome database. Peaks located within 2kb of the *CTCF* TSS and located within the *CTCF* gene body are indicated. Raw data were found using the Cistrome data browser.[106]

See Excel file

**Supplementary Table 15: Numerical results for Supplementary Figure 9.** For each result in the figure, we list i) the numerical values used to make the plot of $\hat{\alpha}$ against $Rv$, ii) the association summary statistics used to make the Manhattan plot, and iii) the numerical results underlying the two displayed gene-set enrichments. 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 "Overview of methods" section and the Methods.

| Cistrome ID | cell type/line | position on chr12 (kb) | TSS | body | reference |
|---|---|---|---|---|---|
| 63463 | K562 (myeloid) | 67561.047-67561.370 | * | | Davis et al.[28] |
| 63463 | K562 (myeloid) | 67644.456-67644.877 | | | Davis et al.[28] |
| 64734 | GM12878 (LCL) | 67566.529-67566.998 | | * | Davis et al.[28] |
| 64734 | GM12878 (LCL) | 67644.381-67644.827 | | | Davis et al.[28] |
| 64919 | K562 (myeloid) | 67561.765-67562.191 | * | | Davis et al.[28] |
| 64919 | K562 (myeloid) | 67601.114-67601.351 | | * | Davis et al.[28] |
| 64919 | K562 (myeloid) | 67644.587-67644.893 | | | Davis et al.[28] |
| 64735 | GM12878 (LCL) | 67561.943-67562.241 | * | | Davis et al.[28] |
| 64735 | GM12878 (LCL) | 67566.547-67567.093 | | * | Davis et al.[28] |
| 64735 | GM12878 (LCL) | 67644.406-67644.874 | | | Davis et al.[28] |
| 73238 | B cell precursor | 67562.052-67562.249 | * | | Schjerven et al.[29] |
| 57640 | Nalm6 (B cell precursor) | 67552.499-67552.751 | | | Song et al.[107] |

**Supplementary Table 16:** IKZF1 ChIP-seq peaks within 10kb of the *CTCF* gene body (chr16:67562.406kb-67639.185kb) in publicly available ChIP-seq data processed by the cistrome database. Peaks located within 2kb of the *CTCF* TSS and located within the *CTCF* gene body are indicated. Raw data were found using the Cistrome data browser.[106]

See Excel file

**Supplementary Table 17: Results of signed LD profile regression using DeepSEA-based annotations.** We list significant results at per-trait FDR< 5% for (a) the BLUEPRINT blood molecular traits, (b) the NTR whole blood eQTLs, (c) the GTEx tissue eQTLs, and (d) the diseases and complex traits analyzed. For each significant annotation, we list TF name, laboratory, and cell line, along with estimated $r_f$ and P-value. The number of significant results identified by these 382 annotations was BLUEPRINT: 810; NTR: 0; GTEx: 1298; complex traits: 7. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 18: Results of signed LD profile regression using GTRD-based annotations.** We list significant results at per-trait FDR< 5% for (a) the BLUEPRINT blood molecular traits, (b) the NTR whole blood eQTLs, (c) the GTEx tissue eQTLs, and (d) the diseases and complex traits analyzed. For each significant annotation, we list the GTRD TF name, along with estimated $r_f$ and P-value. The number of significant results identified by these 149 annotations was BLUEPRINT: 313; NTR: 27; GTEx: 242; complex traits: 7. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.

See Excel file

**Supplementary Table 19: Results of signed LD profile regression using HOCOMOCO motif-based annotations.** We list significant results at per-trait FDR< 5% for (a) the BLUEPRINT blood molecular traits, (b) the NTR whole blood eQTLs, (c) the GTEx tissue eQTLs, and (d) the diseases and complex traits analyzed. For each significant annotation, we list the TF name together with the laboratory and cell line of the ENCODE ChIP-seq track used to determine which SNPs to include in the annotation, along with estimated $r_f$ and P-value. The number of significant results identified by these 276 annotations was BLUEPRINT: 9; NTR: 0; GTEx: 298; complex traits: 103. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.

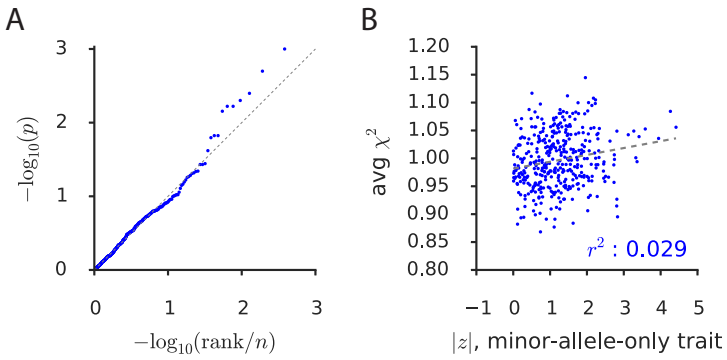
| Source (# annotations) | Blood QTL | GTEx | Diseases/complex traits | Total (per annotation) |
|---|---|---|---|---|
| Basset (382) | 1072 | 2330 | 77 | 3479 (9.1) |
| DeepSEA (382) | 810 | 1298 | 7 | 2115 (5.5) |
| GTRD (149) | 350 | 242 | 7 | 589 (3.2) |
| HOCOMOCO (276) | 9 | 298 | 103 | 410 (1.5) |

**Supplementary Table 20:** For each source of annotations, we report the number of associations at per-trait FDR< 5% obtained upon analysis of: the blood molecular QTL data, the GTEx eQTL data, the disease/complex trait data, and all traits combined. To facilitate comparison across differently sized sets of annotations, we additionally report the total number of results per annotation for each source of annotations. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.

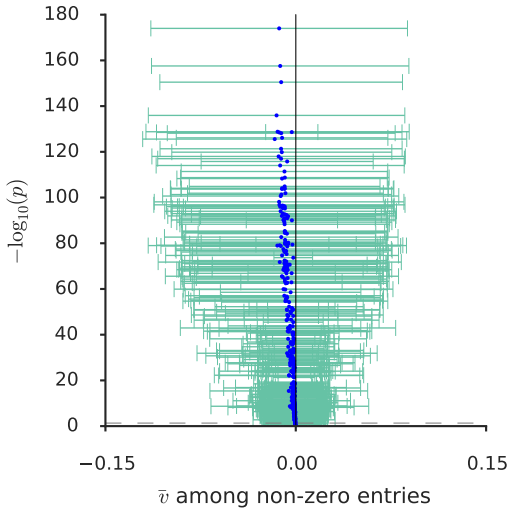
| Trait | TF (num) | $r_f$ | $p$ | $q$ |
|---|---|---|---|---|
| Years of ed. | BCL11A (1) | 2.4% | $3.9 \times 10^{-5}$ | $1.5 \times 10^{-2}$ |
| Crohn's | POL2 (16) | 5.3% | $4.8 \times 10^{-5}$ | $1.5 \times 10^{-2}$ |
| Anorexia | SP1 (1) | -8.9% | $1.1 \times 10^{-4}$ | $4.0 \times 10^{-2}$ |
| HDL | FOS (1) | 4.8% | $1.2 \times 10^{-4}$ | $4.6 \times 10^{-2}$ |
| Eczema | CTCF (12) | 2.7% | $1.4 \times 10^{-4}$ | $3.4 \times 10^{-2}$ |
| Crohn's | ELF1 (1) | 4.9% | $1.6 \times 10^{-4}$ | $1.5 \times 10^{-2}$ |
| Lupus | CTCF (35) | -5.0% | $3.6 \times 10^{-4}$ | $4.4 \times 10^{-2}$ |
| Crohn's | TBP (2) | 5.4% | $4.9 \times 10^{-4}$ | $1.5 \times 10^{-2}$ |
| Crohn's | E2F1 (1) | 4.3% | $6.4 \times 10^{-4}$ | $2.7 \times 10^{-2}$ |
| Crohn's | TAF1 (4) | 4.5% | $9.2 \times 10^{-4}$ | $2.5 \times 10^{-2}$ |
| Crohn's | IRF1 (1) | 4.7% | $9.8 \times 10^{-4}$ | $1.5 \times 10^{-2}$ |
| Crohn's | ETS1 (1) | 6.1% | $1.4 \times 10^{-3}$ | $1.5 \times 10^{-2}$ |
| Lupus | RAD21 (1) | -3.9% | $4.4 \times 10^{-3}$ | $4.1 \times 10^{-2}$ |

**Supplementary Table 21: Distinct TF-trait associations from analysis of diseases and complex traits using signed LD profile regression.** For each of 13 distinct TF-trait associations at per-trait FDR $< 5\%$, we report the associated trait, the associated TF and the total number of annotations for that TF that produced a significant result, the estimate of the functional correlation $r_f$, and the P-value for the most significant annotation. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.

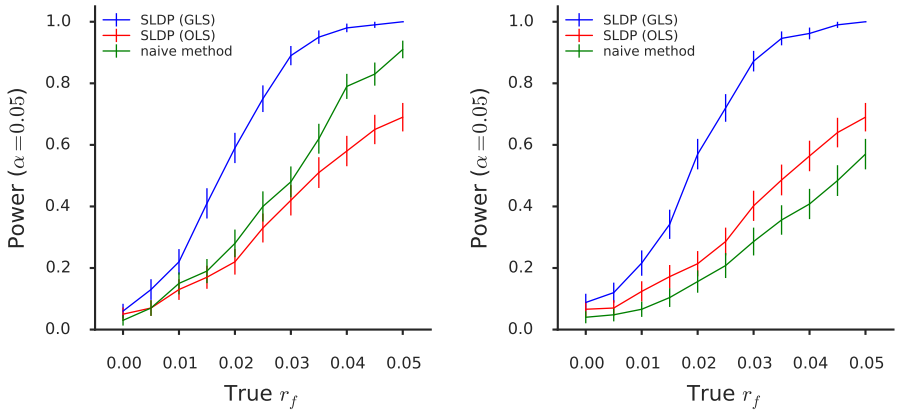
# Supplementary Figures



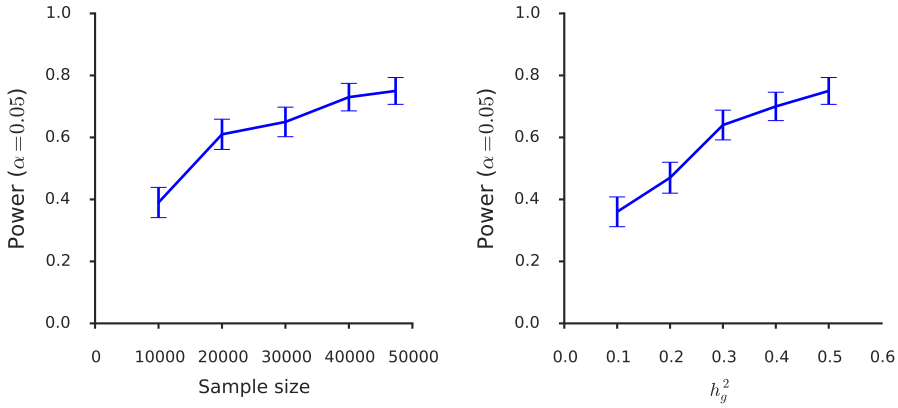

**Supplementary Figure 1: Per-annotation analyses of null calibration.** (a) For each annotation, we used the Simes test[108] to assess the p-value threshold at which the Benjamini-Hochberg procedure would lead to any rejections among 1000 simulated phenotypes with no unsigned enrichment or functional correlation, and we visualized the resulting set of 382 p-values using a q-q plot. These p-values appear uniformly distributed, as would be expected in the scenario of proper calibration. (b) For each annotation, we plot the average $\chi^2$ statistic for that annotation across the 1000 null simulations containing confounding due to genome-wide directional effects of minor alleles on disease risk, against the magnitude of that annotation's z-score for correlation with a 100%-heritable trait whose causal SNPs are exactly the bottom fifth of the MAF spectrum with minor alleles always being trait-increasing. (Statistical significance of the trend is difficult to assess because many annotations are correlated, inducing a complex dependence structure among the 382 points on the plot.) Simulation details are described in the Methods, and the statistical method is described in the "Overview of methods" section and the Methods.



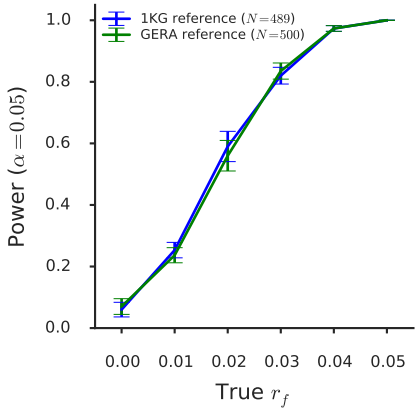

**Supplementary Figure 2: Relationship of annotations to minor alleles.** For each annotation, we computed the mean and standard deviation of the predicted effect of the minor allele among all SNPs with non-zero values of the annotation. We then performed a chi-squared test for the mean being non-zero and plotted $-\log_{10}(p)$ against the mean for each annotation. The green intervals show the standard deviation, in order to give a sense for the scale on which to interpret the mean-shift. The dotted gray line indicates the threshold for FDR significance. 373 of the 382 annotations exceeded this threshold. The number of SNPs in each annotation is specified in Supplementary Table 1.



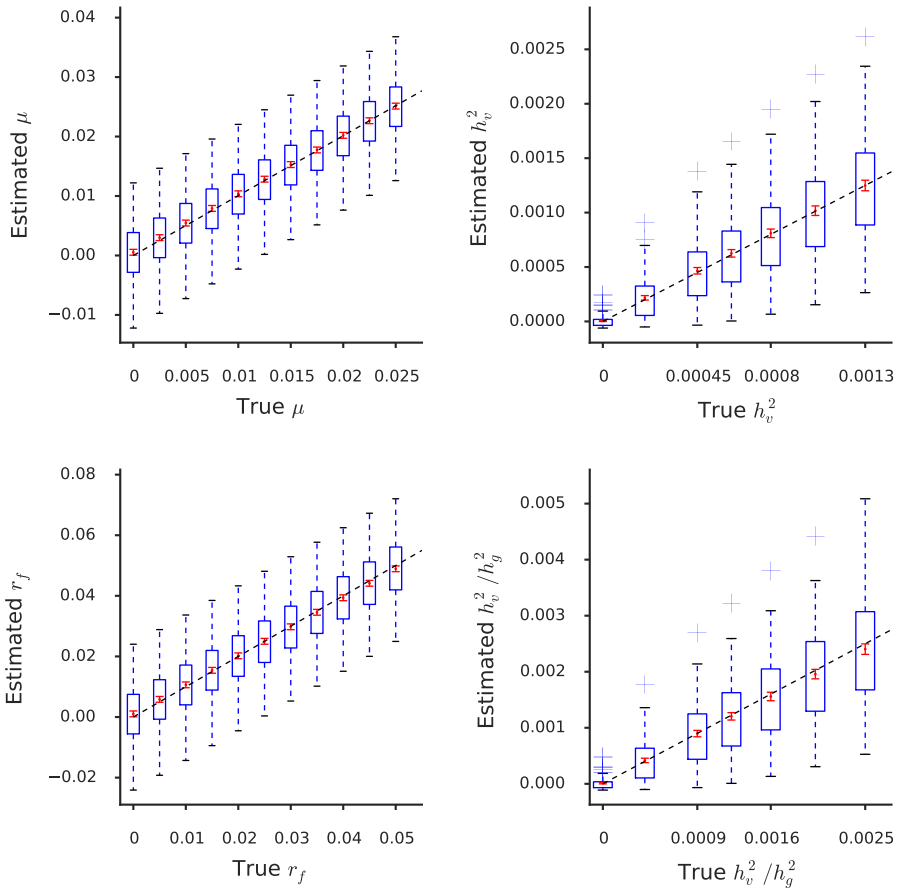

**Supplementary Figure 3: Power comparison of signed LD profile regression to additional methods.** Power curves comparing signed LD profile regression using generalized least-squares (GLS; i.e., weighting) to both ordinary (i.e., unweighted) regression of the GWAS summary statistics on the signed LD profile as well as to a naive method that simply regresses the GWAS summary statistics on the raw annotation. (Left) power comparison with 19.5% of causal SNPs typed, (Right) power comparison with only 9.75% of causal SNPs typed. The real phenotypes analyzed all have at most 11.9% of causal SNPs typed. SLDP regression with default weights is the most powerful method in both regimes. Additionally, the power of the naive method suffers when fewer SNPs are typed, while the power of SLDP regression is far less sensitive to this change. Error bares indicate standard errors of power estimates. Simulation details are described in the Methods, and the statistical method is described in the "Overview of methods" section and the Methods.



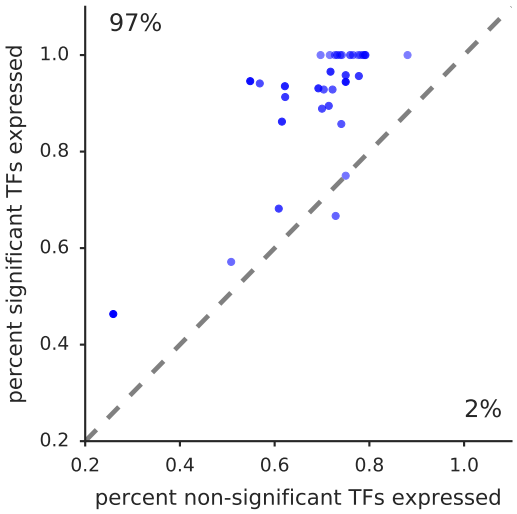

**Supplementary Figure 4: Effect of sample size and heritability on power.** Power of signed LD profile regression as a function of (left) sample size, and (right) overall trait heritability, when proportion of heritability explained by the signed effect is held constant. Error bars indicate standard errors of power estimates. Simulation details are described in the Methods, and the statistical method is described in the "Overview of methods" section and the Methods.



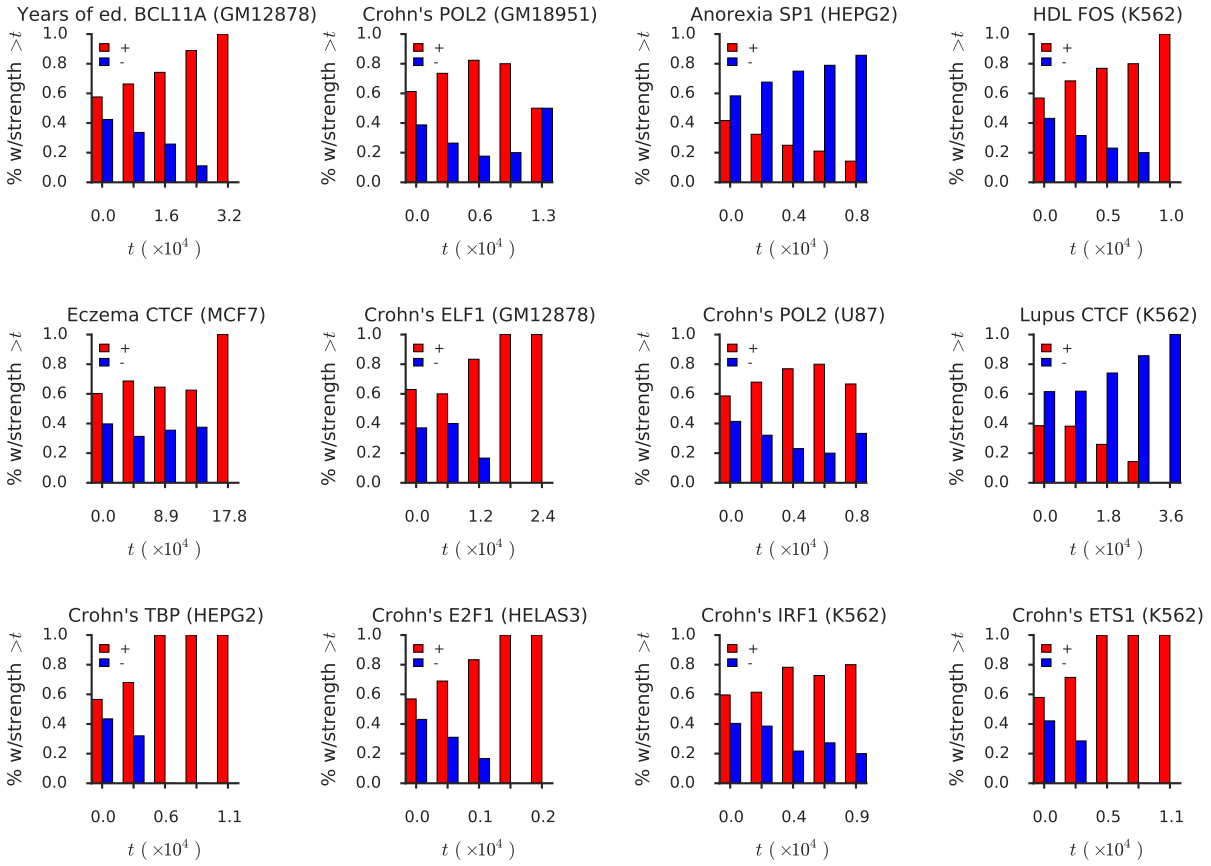

**Supplementary Figure 5: Effect of reference panel on power.** Power of signed LD profile regression as a function of effect size as measured by $r_f$, with either a 1000G reference panel or a randomly chosen in-sample reference panel of comparable size. Error bars indicate standard errors of power estimates. Simulation details are described in the Methods, and the statistical method is described in the "Overview of methods" section and the Methods.



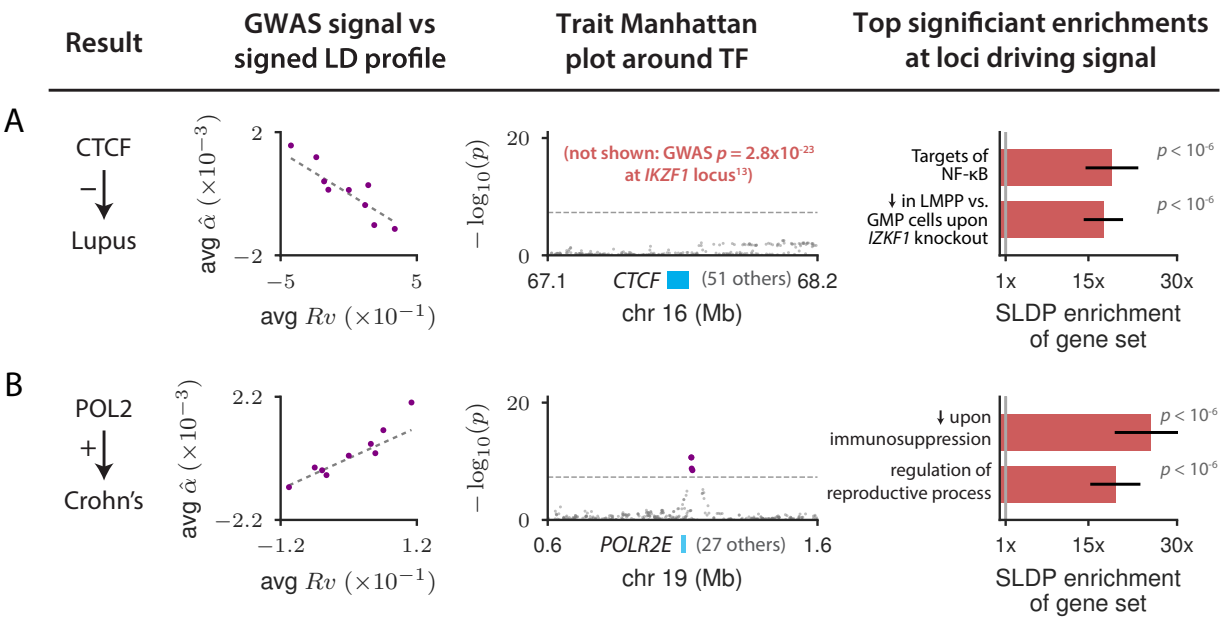

**Supplementary Figure 6: Bias in estimation of additional estimands.** Assessments of the bias of signed LD profile regression with an out-of-sample reference panel in estimating $\mu$, $h_v^2$, $r_f$, and $h_v^2/h_g^2$. For definitions of these quantities, see Supplementary Note. Blue box and whisker 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. Simulation details are described in the Methods, and the statistical method is described in the "Overview of methods" section and the Methods.



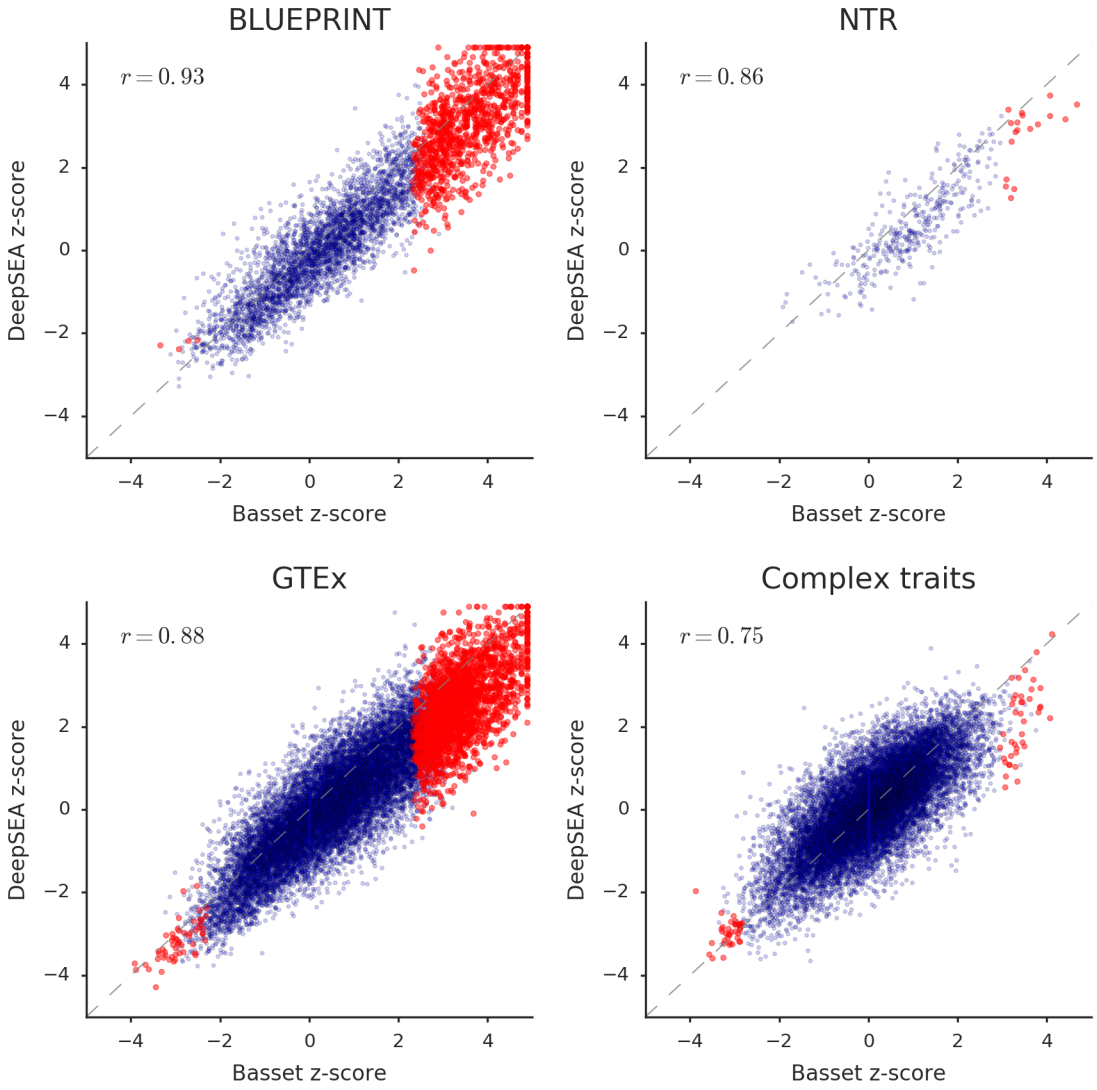

**Supplementary Figure 7: Comparison across tissues of expression levels of TFs identified by signed LD profile regression in each tissue to expression levels of TFs not identified.** For each GTEx tissue in which we found significant TF expression associations, we plot the fraction of significant TFs that are expressed (TPM>5, following ref.[6]) against the fraction of non-significant TFs that are expressed. Points are colored in proportion to the number of significant results in each tissue. GTEx data are described in Supplementary Table 6, the statistical method used to determine significant TF-expression associations and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods, and the statistical method used to assess significance of the trend in this plot is described in the Supplementary Note.



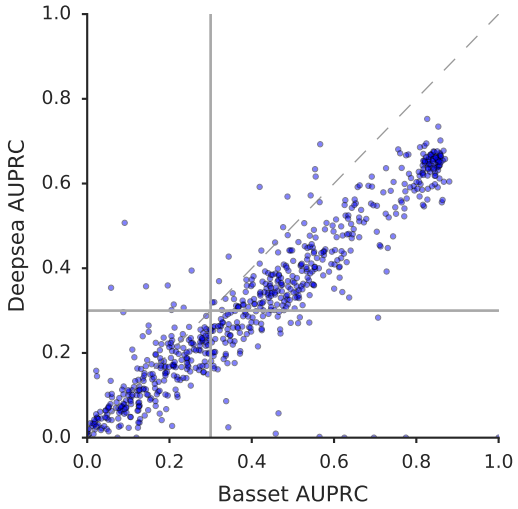

**Supplementary Figure 8: Distribution of covariance between GWAS summary statistics and signed LD profile.** For each of our twelve independent results, we plot, for a variety of thresholds $t$, the fraction of the approximately 300 independent genomic blocks with $|\mathrm{cov}(\hat{\alpha}, Rv)| > t$ in which the covariance is positive versus negative. There is an excess of blocks in which sign of the covariance matches the genome-wide direction of effect. (We note that, as this figure illustrates, our results do not imply that the sign of the covariance matches the genome-wide direction of effect in *all* blocks.) GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the Methods.



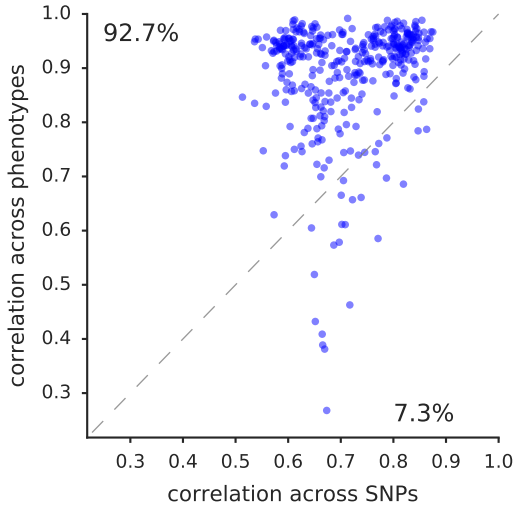

**Supplementary Figure 9: Additional highlighted TF binding-complex trait associations.** For each of (a) CTCF-Lupus and (b) POL2-Crohn's disease, we display plots of the marginal correlation $\hat{\alpha}$ of SNP to trait versus the signed LD profile $Rv$ of the annotation in question, with SNPs averaged in bins of 4,000 SNPs of similar $Rv$ values and a larger bin around Rv=0; Manhattan plots of the trait GWAS signal near the associated TF; 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 "Overview of methods" section and the Methods. Numerical results are reported in Supplementary Table 12. IKZF1 ChIP-seq peaks at the *CTCF* promoter are presented in Supplementary Table 16. LMPP: lymphoid-primed pluripotent progenitor; GMP: granulocyte-monocyte precursor.



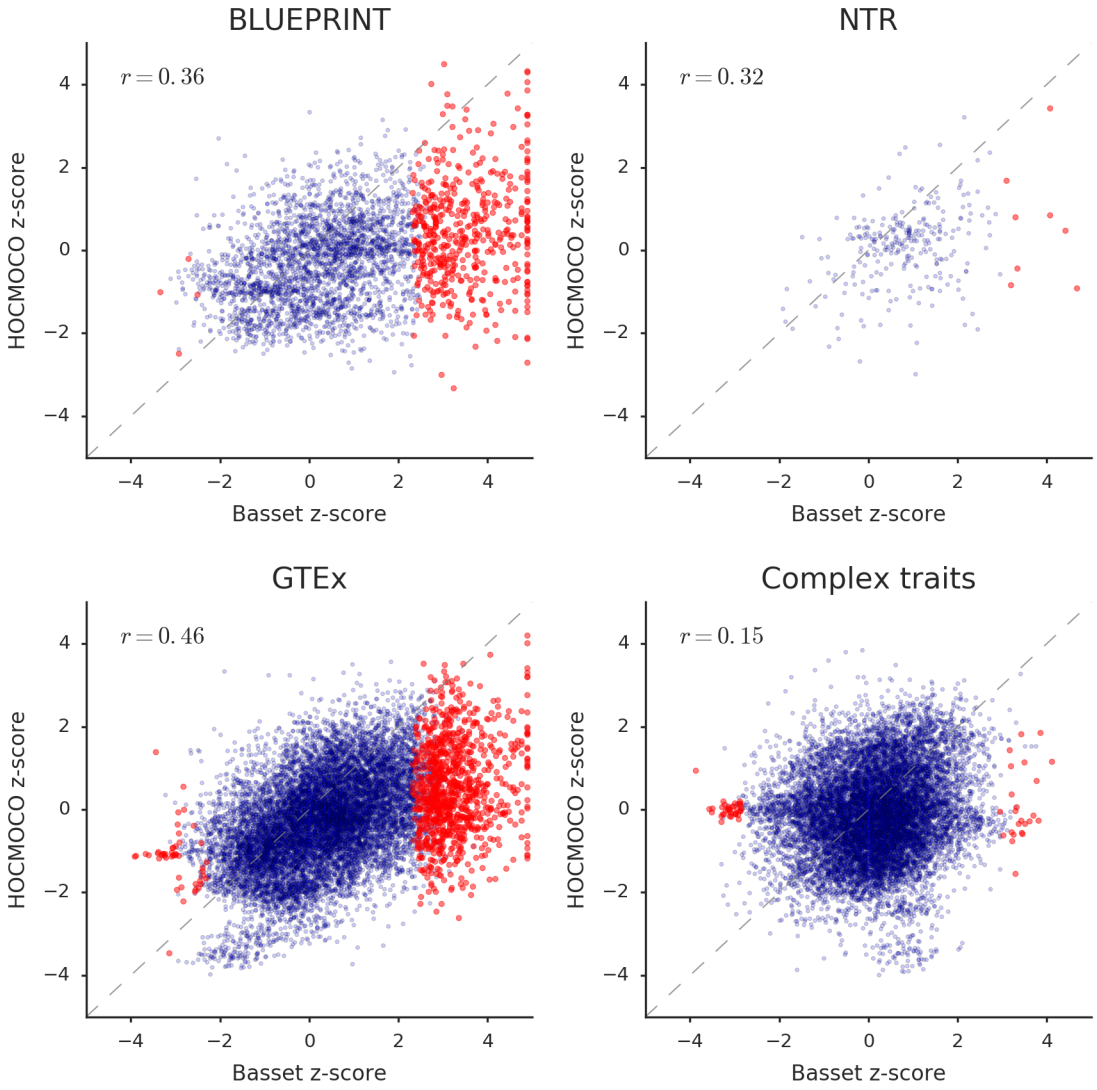

**Supplementary Figure 10: Comparison of signed LD profile regression using Basset to results using DeepSEA.** For each phenotype and each of the 382 ENCODE ChIP-seq tracks used in our main analyses, we plot the SLDP z-score of the DeepSEA-derived annotation from that track on that phenotype against SLDP z-scores of the Basset-derived annotation from that same track on that same phenotype. We display separate plots for the four sets of phenotypes analyzed in the paper; red dots indicate significant results from our main analyses using the Basset-derived annotations; correlations between the two sets of z-scores are indicated on each plot. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.



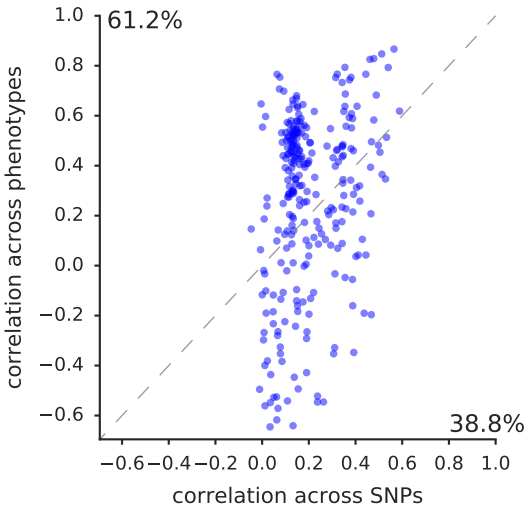

**Supplementary Figure 11: Comparison of Deepsea prediction accuracy to Basset prediction accuracy.** For each of the 691 ENCODE TF ChIP-seq tracks for which we had AUPRC information using both Basset and DeepSEA, we plot the DeepSEA AUPRC for that track against the Basset AUPRC for that track. The dashed line indicates $y = x$, and the solid lines indicate our QC threshold of AUPRC$> 0.3$.




**Supplementary Figure 12: Basset and Deepsea converge on biological signal.** For each of the 382 ENCODE ChIP-seq tracks used in our main analyses, we plot (i) the correlation *across SNPs* between the Basset-derived annotation for that track and the DeepSEA-derived annotation for that track, against (ii) the correlation *across phenotypes* between the z-scores of the Basset-derived annotation for that track and the z-scores of the DeepSEA-derived annotation for that track. The dashed line indicates $y = x$, and the percentages indicate the fraction of annotations in which either (i)>(ii) or (i)<(ii). The fact that the vast majority of annotations are more correlated when the correlation is measured across phenotypes indicates that the signal that is shared between Basset and DeepSEA is strongly reflected in GWAS data.




**Supplementary Figure 13: Comparison of signed LD profile regression using Basset to results using motifs from HOCOMOCO database.** For each phenotype and each of the 276 ENCODE ChIP-seq tracks used in our main analyses that had a corresponding motif in HOCOMOCO, we plot the SLDP z-score of the HOCOMOCO-derived annotation from that track on that phenotype against SLDP z-scores of the Basset-derived annotation from that same track on that same phenotype. We display separate plots for the four sets of phenotypes analyzed in the paper; red dots indicate significant results from our main analyses using the Basset-derived annotations; correlations between the two sets of z-scores are indicated on each plot. GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.




**Supplementary Figure 14: Basset and HOCOMOCO motifs converge weakly on biological signal.** For each of the 276 ENCODE ChIP-seq tracks used in our main analyses that had a corresponding motif in HOCOMOCO, we plot (i) the correlation *across SNPs* between the Basset-derived annotation for that track and the HOCOMOCO-derived annotation for that track, against (ii) the correlation *across phenotypes* between the z-scores of the Basset-derived annotation for that track and the z-scores of the HOCOMOCO-derived annotation for that track. The dashed line indicates $y = x$, and the percentages indicate the fraction of annotations in which either (i)>(ii) or (i)<(ii). The majority of annotations are more correlated when the correlation is measured across phenotypes, indicating that the signal that is shared between Basset and HOCOMOCO is reflected in GWAS data. However, the trend is considerably weaker than it is when Basset and DeepSEA are compared (see Supplementary Figure 12). GWAS data are described in Supplementary Table 8, and the statistical method and multiple comparisons adjustments are described in the "Overview of methods" section and the Methods.