Cell Genomics, Volume *4*

Supplemental information

Gene regulatory network inference from CRISPR perturbations in primary CD4^{+} T cells elucidates the genomic basis of immune disease

Joshua S. Weinstock, Maya M. Arce, Jacob W. Freimer, Mineto Ota, Alexander Marson, Alexis Battle, and Jonathan K. Pritchard

Supplemental Methods 1

Contents

Direct effects vs total effects for network reconstruction

We first simulated a cyclic gene regulatory network to define a ground truth. We chose a graph structure with five nodes connected in a cycle (Supplementary Note Figure 1) and set all the edge weights to 0.5, resulting in the following adjacency matrix:

$$\boldsymbol{\beta} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0 & 0.5 \end{bmatrix}$$

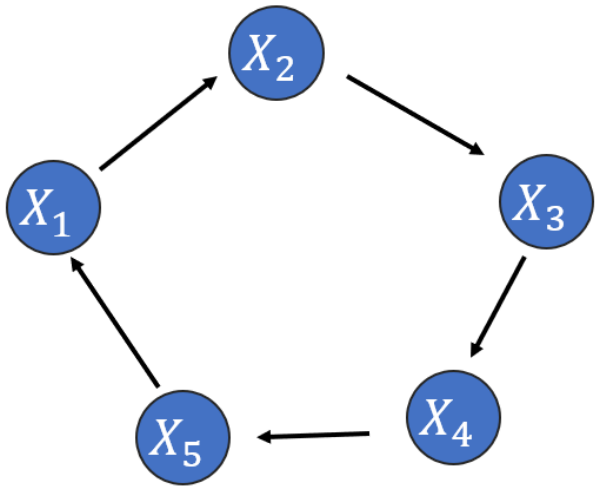
We then simulated an initial expression observation:

$$\boldsymbol{X^{(0)}} \sim LogNormal(1.00, 0.10)$$

Then, given β we model the effect of a perturbation on the kth gene as setting the kth column of β to 0, i.e., we remove all incoming connections. We denote this modified adjacency matrix as $\widetilde{\boldsymbol{\beta}}$. To simulate the effect of a "knock-out", we also set $\boldsymbol{X_k^{(0)}} = 0$. We then draw the expected "steady-state" as:

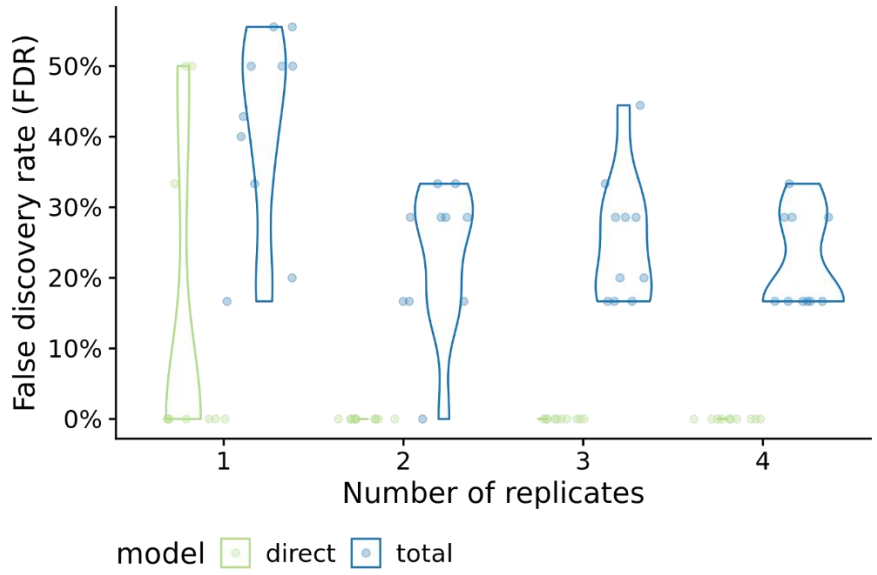
$$\lim_{t \to \infty} \boldsymbol{X^{(t)}} = \boldsymbol{X^{(0)}}\left(\boldsymbol{I} - \widetilde{\boldsymbol{\beta}}\right)^{-1}$$

We then used LLCB to estimate the network (as defined by direct effects matrix $\boldsymbol{\beta}$) and compared this estimate to a total effect derived network. To estimate total effects, we simply performed the first step of LLCB without performing the deconvolution step. We repeated this simulation 10 times, varying across the number of simulated technical replicates within each donor. We defined the graphs by simply thresholding the effect estimates above 0.3, whereas the true edges all had weights of 0.5.

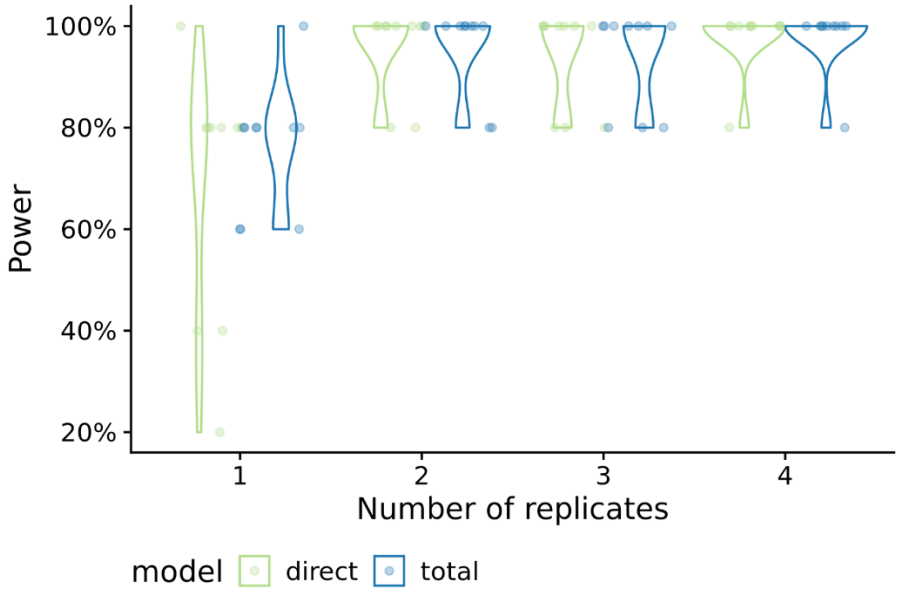


Supplementary Note Figure 1 | The ground truth graph

Overall, we observed that although the two approaches (estimating total effects or estimating direct effects) had similar power (Supplementary Figure 2), the direct effect estimates substantially reduced false discovery. For example, the total effect network had an average FDR of 28% with a single technical replicate per donor as compared to an average FDR of 3% for the direct effect estimates. Power remained comparable (89% for direct effects vs 91% for total effect, Supplementary Note Figure 2).



Supplementary Note Figure 2 | False discovery rate comparison from simulated data



Supplementary Note Figure 3 | Power comparison from simulated data

Overall, this is consistent with an intuition where total effect analyses are a very powerful approach; total effects analyses will discover the majority of true connections. However, they are quite prone to the false discovery of edges because the true underlying graph structure is ignored, resulting in several redundant correlations.

On the association between upstream gene groups and downstream non-perturbed genes

We used the following model:

$$Y_{ik} \sim NB(g^{-1}(\eta_{ik}), \psi_k)$$
$$g\big(E(Y_{ik})\big) = \eta_{ik} = X_i \beta_k$$

NB represents a mean-variance parameterization of the negative binomial distribution, Y_{ik} is the number of incoming connections from the kth gene group (i.e., Background TFs / IEI TFs / IL2RA Regulators / IEI or IL2RA Regulators) for a given ith non-perturbed gene, and X_i is a design matrix with an indicator for the intercept and a series of covariates describing the ith non-perturbed gene including the following terms:

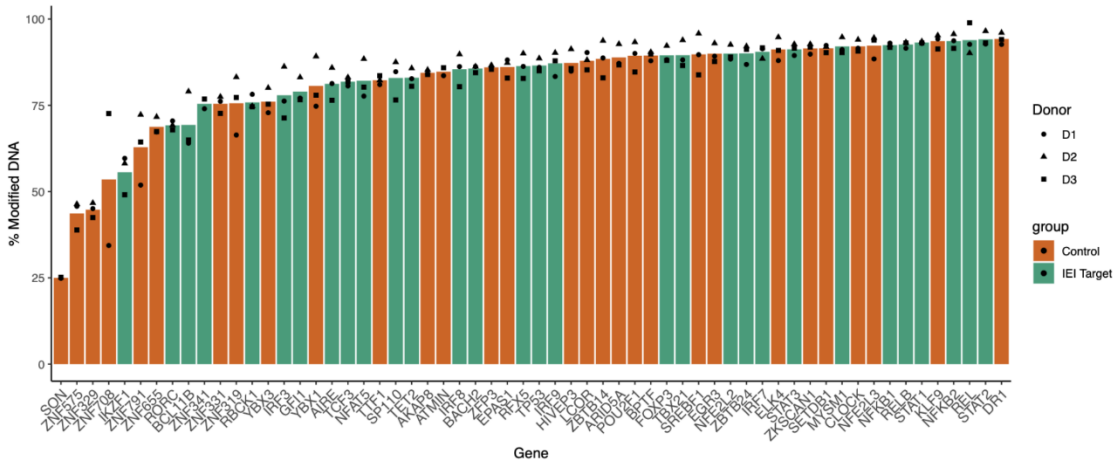
1. The S_{het} estimate of the ith gene
2. An indicator for whether the ith gene is an IEI gene (Reference = Not an IEI gene)
3. Expression of the ith gene measured in control samples (i.e., CRISPR K/O targeted to AAVS1 locus)
4. The number of incoming connections to the ith gene from background TFs
5. An indicator for whether the ith gene is a trans-eGene in eQTLgen
6. An indicator for whether the ith gene is an immune GWAS gene, has reported through PICS
7. An indicator for whether the ith gene is only expressed in blood as defined through analysis of GTEx samples

Each of the four sub-panels in Figure 3D plots the estimated β_k along with the respective uncertainties for the four gene groups.

Supplementary Figures

Figure S1 | CRISPR editing efficiency by gene group, related to Figure 2. A Percent of reads with indels, stratified by individual gene. **B** Percent of reads with indels, aggregating by gene group.

A



B

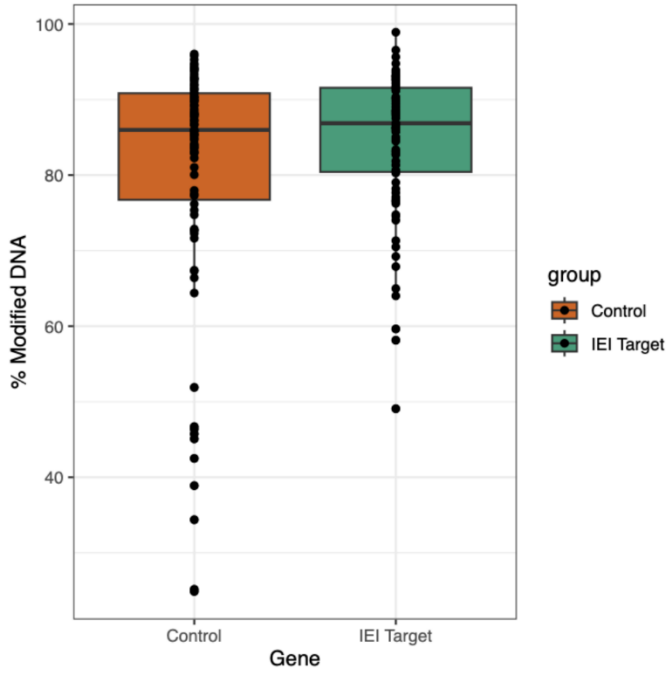


Figure S2 | Enrichment of LLCB posterior mean edges in the ABC-DAC validation network, related to Figure 2

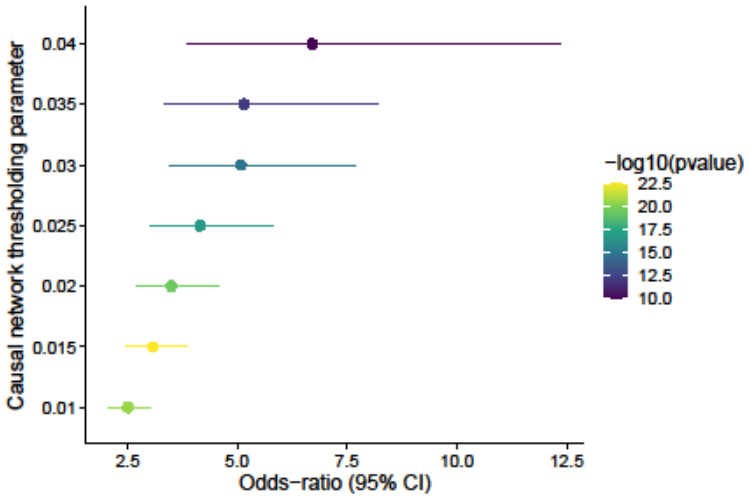


Figure S3 | Enrichment of LLCB posterior mean edges in the HBase T-cell network, related to Figure 2

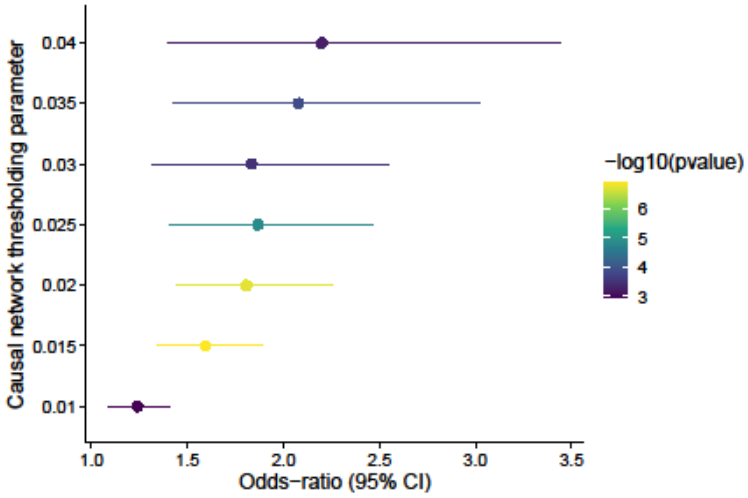


Figure S4 | Enrichment of LLCB posterior mean edges in the ABC-ChIP validation network, related to Figure 2. The ABC-ChIP network was defined through intersecting ChIP-seq peaks of TFs in CD4+ cells with enhancer-gene predictions from the ABC model.

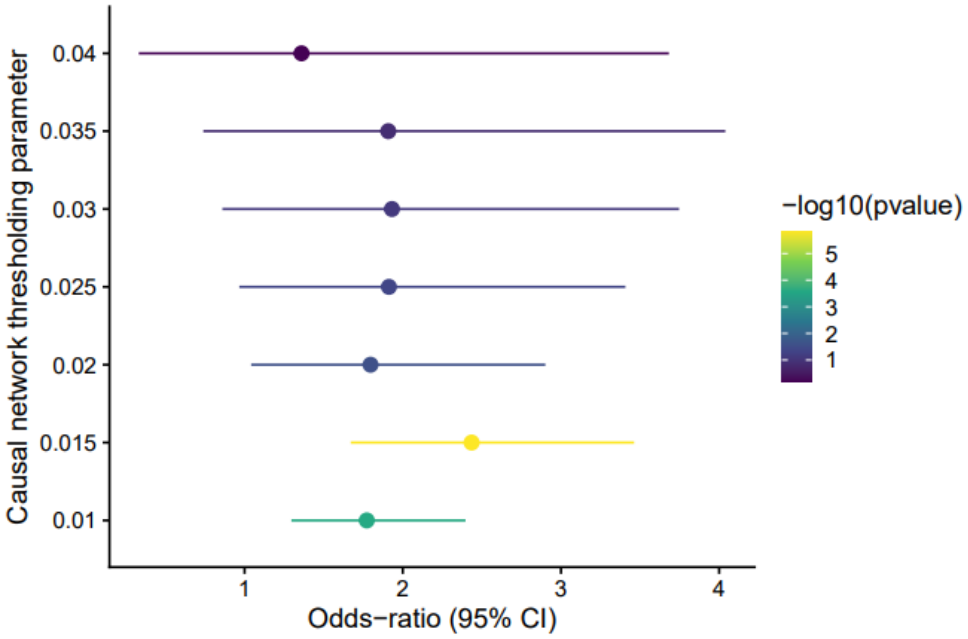


Figure S5 | Permutation test of the number of edges between genes that share the same gene group, related to Figure 2. A set of 2,000 null permutations of the network were generated by using the rewiring algorithm to preserve the node degree. Within each permutation, the number of edges with the same gene group were counted. The observed value is denoted by the red vertical line, and the empirical 2.5% and 97.5% quantiles from the permuted data are denoted by vertical dashed lines.

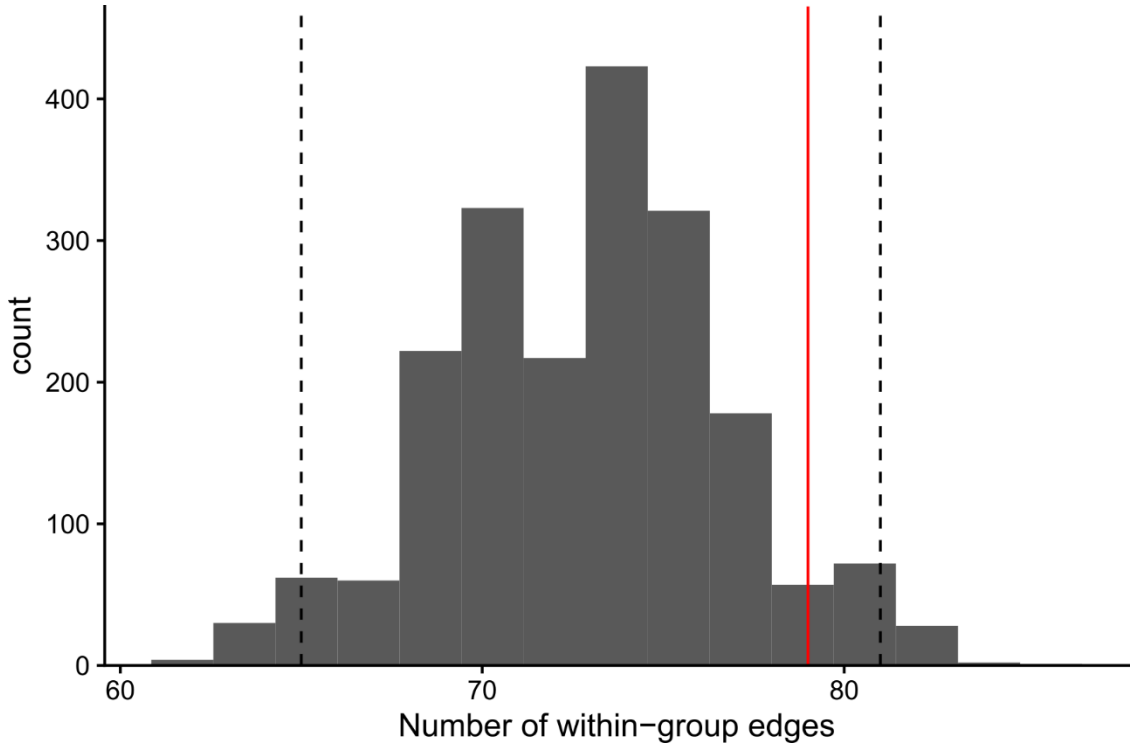
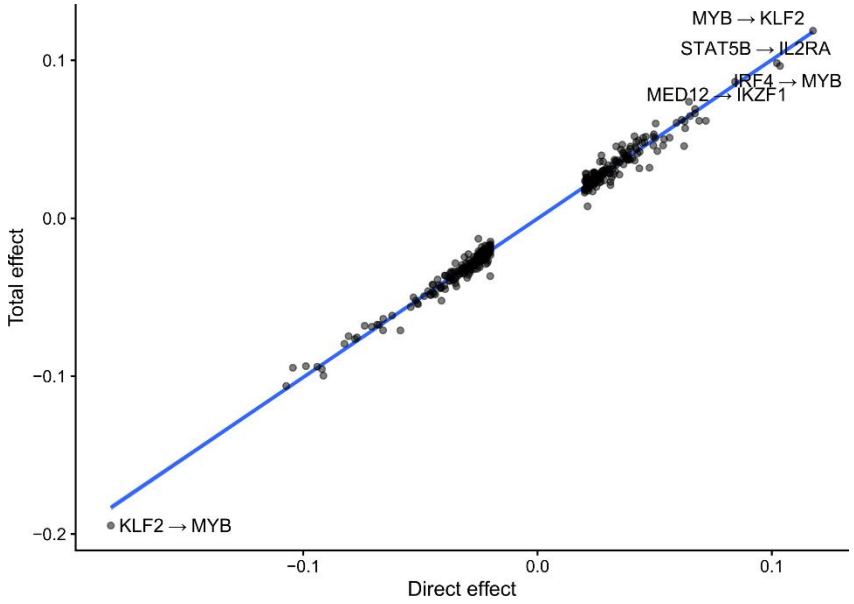


Figure S6 | Comparison of direct to total effects among the 84 KO'd genes, related to Figure 2. The x-axis is defined as the posterior mean estimates of the adjacency matrix estimated by LLCB. Units are in terms of standard deviations of normalized gene expression. The y-axis is estimated through the processing procedure described in Methods.



S7 | The largest indirect effects are mediated by cycles of short length, related to Figure 2

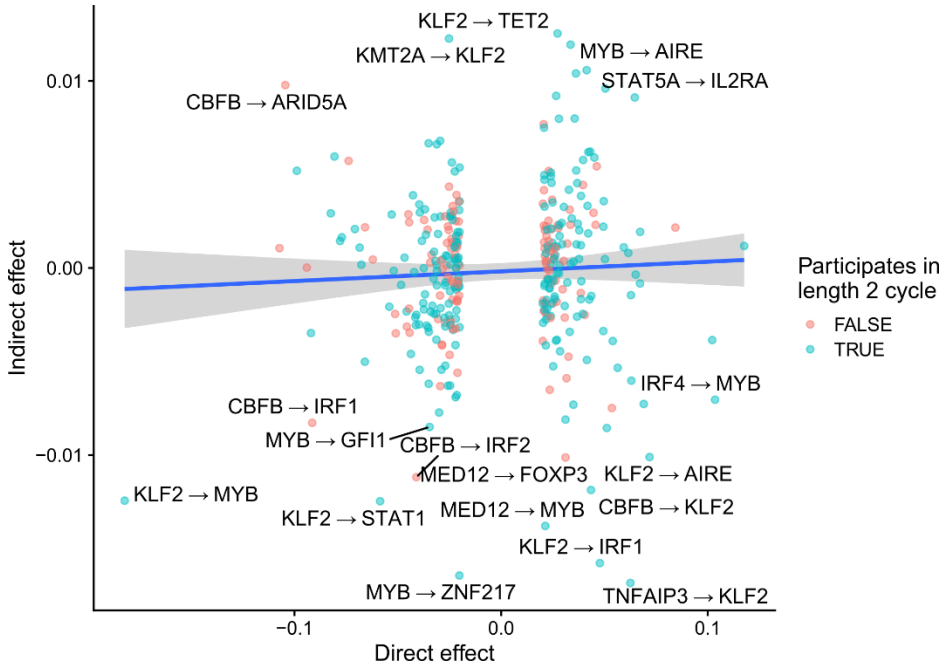


Figure S8 | Enrichment of module effects on KEGG signaling pathways, related to Figure 5.
Enrichment analyses were performed with pathfindR.

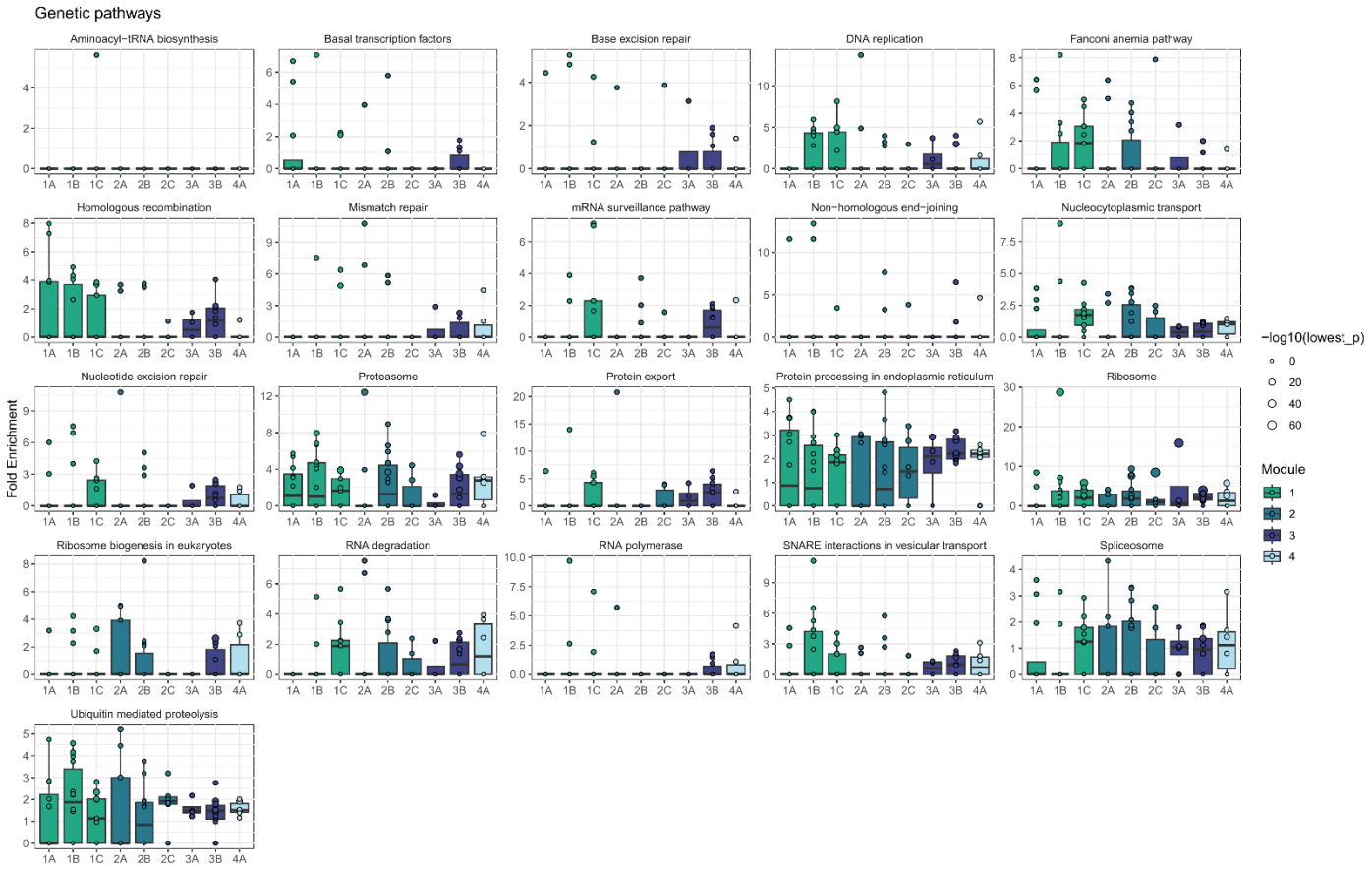


Figure S9 | Enrichment of module effects on KEGG signaling pathways, related to Figure 5.
Enrichment analyses were performed with pathfindR.

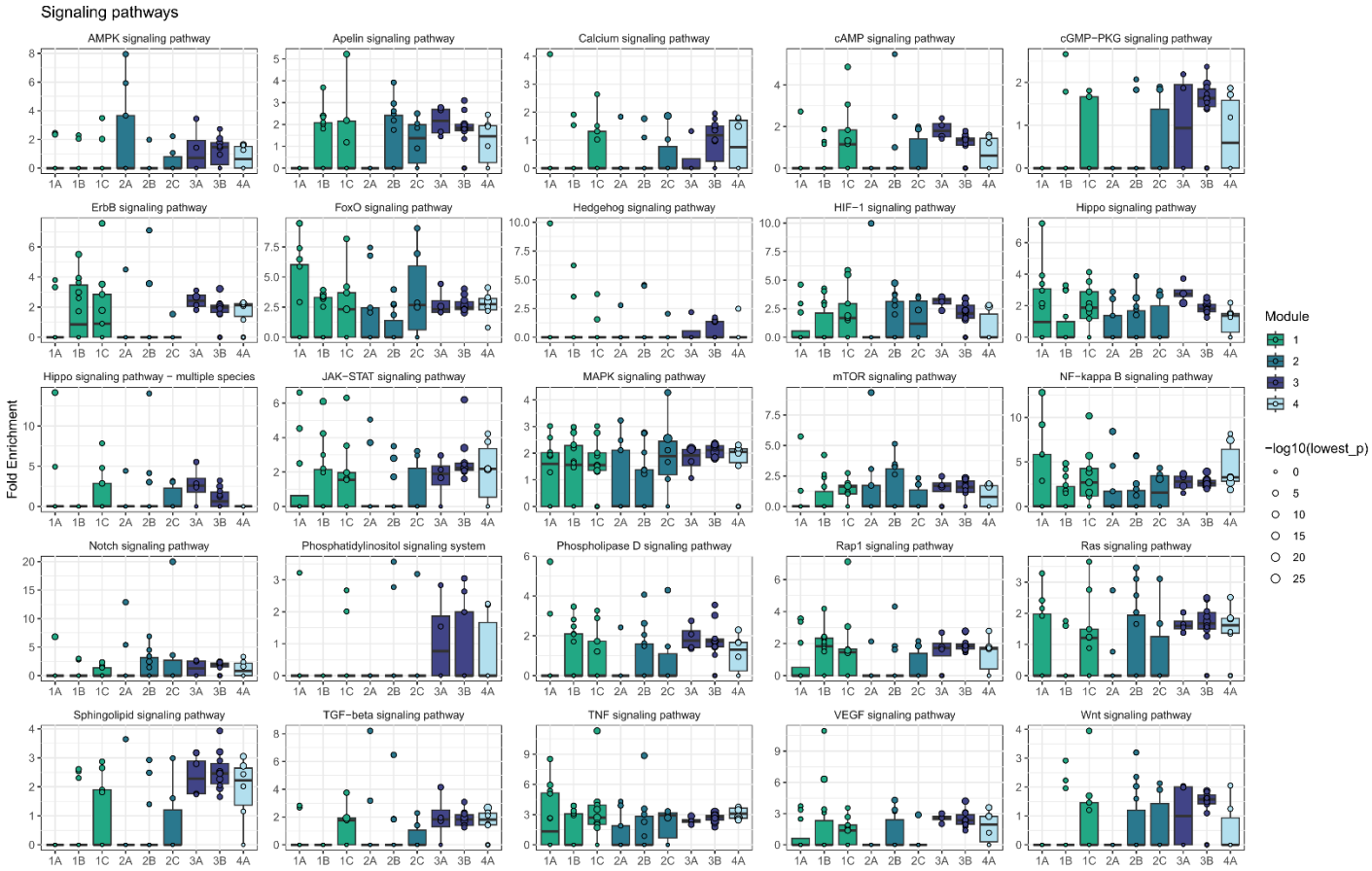


Figure S10 | Enrichment of module effects on KEGG immune pathways, related to Figure 5. Enrichment analyses were performed with pathfindR.

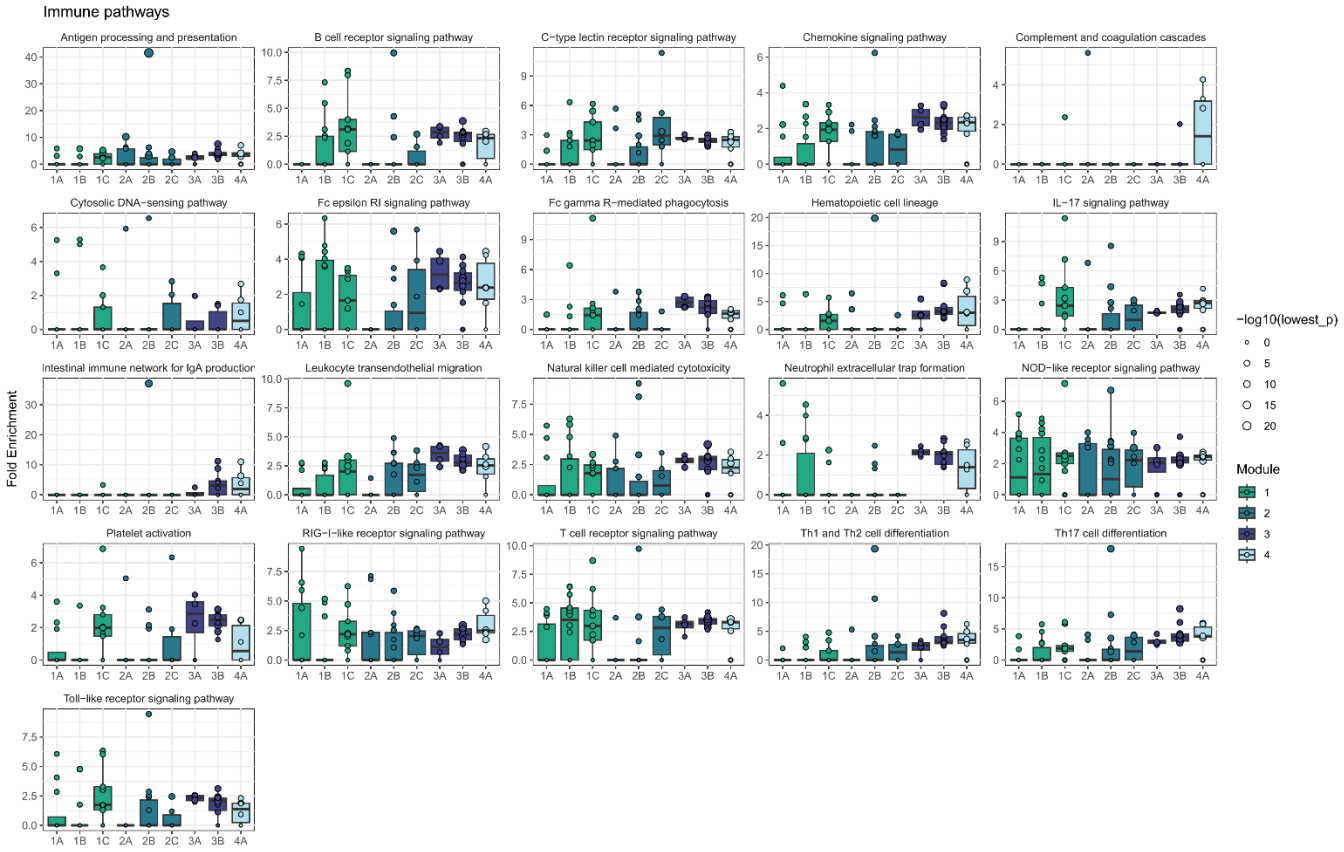


Figure S11 | Network plot demonstrates the effect of the cluster 2A upstream regulators on cell-cycle genes, related to Figure 5. The network using edges estimated from the BG model are plotted. Colors indicate the effect size and arrows indicate the direction of effect. The genes on the left-hand side are among the 84 KO'd genes, and the genes on the right are genes that are listed among the KEGG cell cycle pathway genes.

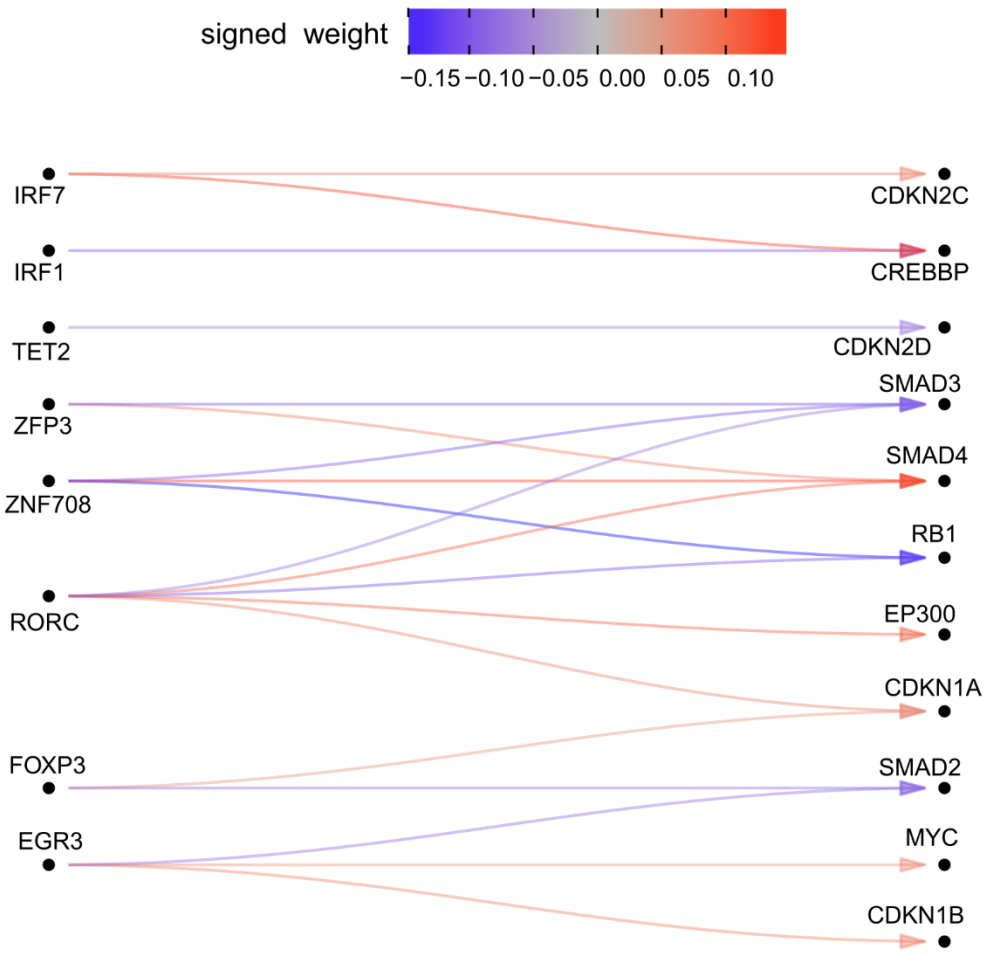


Figure S12 | Marginal heritability estimates from LD score regression, related to Figure 6. LD score regression was used to estimate the heritability enrichment of SNPs linked to genes in each module for each phenotype. SNPs were linked to genes using the ABC predictions in T cells.

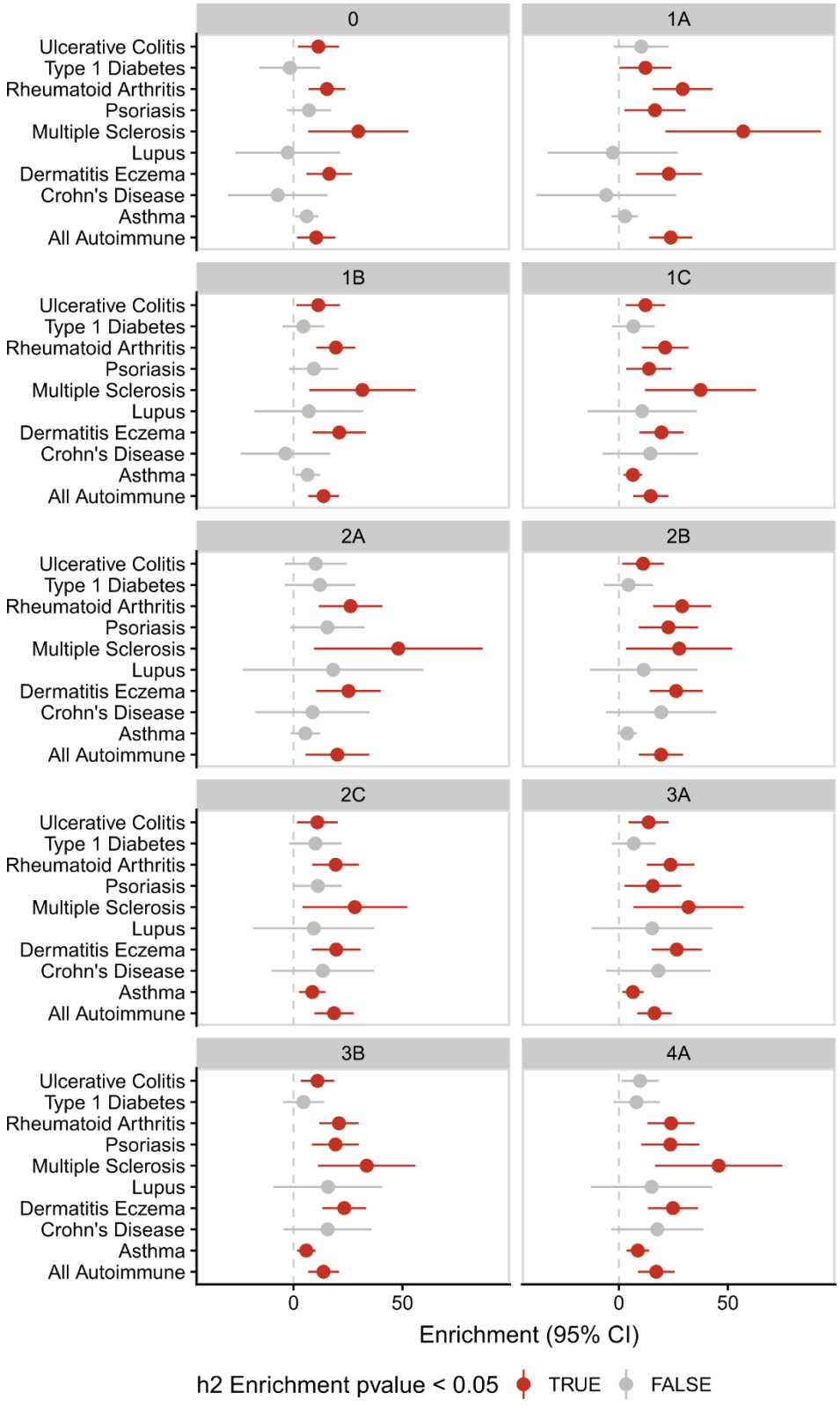
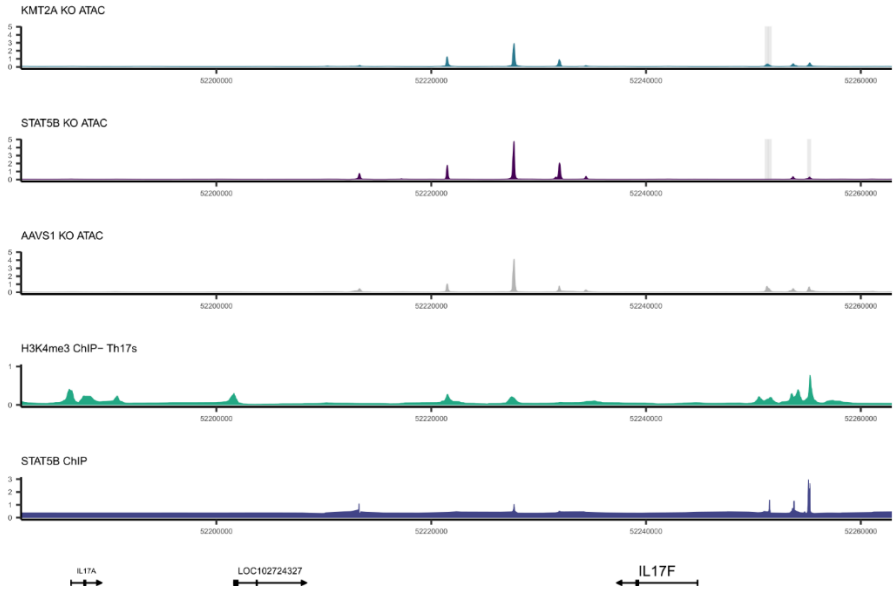


Figure S13 | KMT2A and STAT5B jointly regulate chromatin accessibility at the *IL17F* locus (A) and *IL21* locus (B), related to Figure 7. *For A and B, locus plot including tracks describing the functional characteristics of the region. Each track is constructed from publicly available ChIPseq data (methods) or ATAC-seq data from Freimer et al. Grey boxes indicate significantly different regions between the respective KO and AAVS1 control KO ATAC data (padj < 0.05, n = 3 donors per KO). The Y-axis displays normalized counts.*

A



B

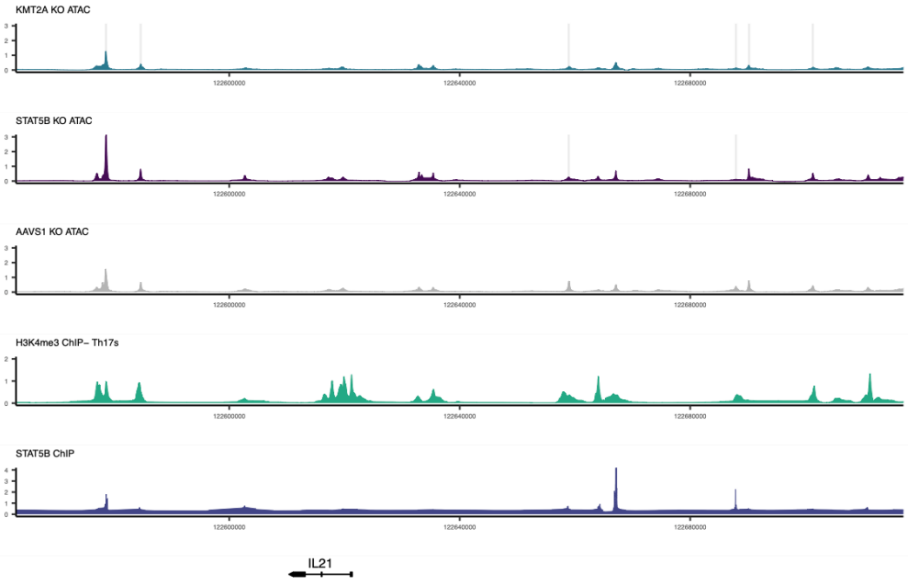


Figure S14 | Meta-analysis of autoimmune GWAS from Shirai et al. and Finngen v8, related to Figure 7. The *KMT2A* locus plot is displayed with a chromHMM[75] track from Th17 cells. The predicted enhancers of *KMT2A* from the ABC model in CD4+ T cells are shown in red arcs at the bottom.