Predicting genotype-specific gene regulatory networks

Deborah Weighill[1], Marouen Ben Guebila[1], Kimberly Glass[2,3], John Quackenbush[1], and John Platig*[2,3]

[1]Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[2]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA
[3]Harvard Medical School, Boston, MA 02115, USA

Abstract

The majority of disease-associated genetic variants are thought to have regulatory effects, including the disruption of transcription factor (TF) binding and the alteration of downstream gene expression. Identifying how a person's genotype affects their individual gene regulatory network has the potential to provide important insights into disease etiology and to enable improved genotype-specific disease risk assessments and treatments. However, the impact of genetic variants is generally not considered when constructing gene regulatory networks. To address this unmet need, we developed EGRET (Estimating the Genetic Regulatory Effect on TFs), which infers a genotype-specific gene regulatory network (GRN) for each individual in a study population by using message passing to integrate genotype-informed TF motif predictions - derived from individual genotype data, the predicted effects of variants on TF binding and gene expression, and TF motif predictions - with TF protein-protein interactions and gene expression. Comparing EGRET networks for two blood-derived cell lines identified genotype-associated cell-line specific regulatory differences which were subsequently validated using allele-specific expression, chromatin accessibility QTLs, and differential TF binding from ChIP-seq. In addition, EGRET GRNs for three cell types across 119 individuals captured regulatory differences associated with disease in a cell-type-specific manner. Our analyses demonstrate that EGRET networks can capture the impact of genetic variants on complex phenotypes, supporting a novel fine-scale stratification of individuals based on their genetic background.
EGRET is available through the Network Zoo R package (netZooR v0.9; netzoo.github.io).

*Corresponding author: john.platig@channing.harvard.edu

1 Introduction

The functional impact of disease-associated genetic variants remains an unresolved question in human genetics. Genome-wide association studies (GWASs) have demonstrated that these variants typically have modest effect sizes and mostly lie outside of coding regions [1]. It has been found that they likely influence gene regulation [2, 3, 4]. In particular, genetic variation in transcription factor (TF) binding sites and their flanking regions explains a substantial amount of the heritability of many diseases and complex traits [5]. This suggests an important regulatory mechanism; however, there remains a need for a method which can effectively predict genotype-specific TF regulatory network structure.

EGRET (Estimating the Genetic Regulatory Effect on TFs) infers regulatory network models by integrating an individual's genetic information with other gene regulatory data. Beginning with a naive TF motif-gene bipartite network, EGRET incorporates genetic effects by including individual genotypes, the predicted effects of genetic variants on TF binding [6], and population-level expression quantitative trait loci (eQTLs). EGRET then uses a message passing framework [7] to integrate these effects with existing knowledge of TF-TF interactions and population-level gene expression information to construct individual-specific gene regulatory networks (GRNs).

We validated EGRET by comparing networks inferred for two cell lines and found that differentially regulated genomic regions were enriched for genotype-affected chromatin accessibility, allele-specific expression, and differential TF binding as determined by ChIP-seq. These three independent validations provided evidence that EGRET networks capture biologically relevant disruptions in gene regulation caused by genetic variants. We also used EGRET to infer 357 individual and cell-type specific GRNs (three cell types, 119 individuals) and showed that EGRET networks captured cell-type specific, genetically influenced regulatory disruptions in relevant disease processes. Most notably, these disease-related regulatory disruptions affected network modularity in different cell types, indicating that the effects of genetic variants extended beyond differential regulation of genes to the alteration of regulatory network topology, and thus, broader regulatory processes. EGRET allows us to identify which genetic changes alter cellular functions and to understand the causal role of disease-associated variants. Because the only individual-specific data EGRET requires is genotype data, it can be used to understand genetic effects in many cohorts with SNP chip or whole genome sequencing data, such as TOPMED [8] and the UK biobank [9].

2 EGRET - Estimating the Genetic Regulatory Effects on TFs

2.1 The EGRET algorithm

EGRET uses several sources of information to capture the impact of genetic variants on TF to gene regulatory relationships and construct individual-specific GRNs (Table 1, Figures 1A and S1). The first is a TF-to-gene prior regulatory network M derived, for example, from FIMO motif scans of a reference genome [10], (Supplementary Note S1) that estimates which TFs bind to promoter regions to regulate their target gene expression. The second is a TF-TF interaction prior network P, derived from protein-

protein interactions (Supplementary Note S2), that reflects the fact that TFs can form complexes through protein-protein interactions to cooperatively regulate expression. Third, EGRET uses gene expression data and calculates a co-expression matrix C under the assumption that genes which are co-regulated are likely to exhibit correlated expression. These can be obtained either from the individuals for whom networks are to be generated, or from a large reference data set with expression profiling in the cell type of interest (e.g. GTEx, see Supplemental Note S2). Fourth, EGRET requires eQTL data either from the study population or from a public database, as with expression, from the cell type of interest. The fifth input to EGRET is genotype information. And finally, EGRET uses QBiC [6] predictions of the effect of genetic variants on TF binding (Supplementary Note S3). It is worth noting that the data for P, M, the co-expression matrix C, and the eQTLs, can be obtained from publicly available resources. Thus, one can construct an EGRET network for a given cell type in an individual of interest simply by providing the genotype information for that individual and relying on publicly available data for the other input data.

EGRET uses a message passing framework introduced in PANDA, a network reconstruction approach that takes three input matrices, which are identical to C, P, and M, and uses message passing identify and up-weight consistency between them based on our understanding of the way in which transcription factors regulate gene expression. EGRET extends this by modifying M in a way that recognizes that SNPs may affect the ability of TFs to bind and regulate nearby genes. In particular, for each input genoype, EGRET selects SNPs (A in Figure 1B) that (1) are within motif-based TF binding sites (TFBS) in the promoter regions of genes, and (2) have a statistically significant eQTL association (β in Figure 1B) with the expression of their adjacent gene. EGRET use QBiC [6] to predict the effect of these SNPs on the binding of the TF associated with the motif at that location, and selects variants with a significant QBiC effect (q in Figure 1B). This process selects genetic variants in each individual that are predicted to both affect gene expression and TF binding. The effect of a SNP s on TF binding is then defined as the product $|q_{s_{ij}} A_{s_{ij}} \beta_{s_{ij}}|$. Modifier weights to the motif prior are then calculated by adding these effects per TF-gene pair, allowing for the fact that a gene might have more than one variant in its promoter region affecting the binding of a particular TF. An EGRET prior E network is then constructed by subtracting the modifier from the generic motif prior:

$$E_{ij} = M_{ij} - \sum_{s} |q_{s_{ij}} A_{s_{ij}} \beta_{s_{ij}}|$$

thus penalizing the motif prior when the individual in question contains a genetic variant with sufficient evidence to suggest that they alter gene regulation (Figure 1B). This modified motif prior E is then used in place of M and subjected to message passing to find consistency between it, the TF-TF PPI prior P, and the co-expression prior C (Supplementary Note S4). The message passing process updates all three input matrices, boosting those relationships that show agreement between associations captured in E, C and P. Upon convergence, the primary output is an individual-specific, complete, bipartite GRN, E^*, that captures genotype-specific regulatory effects. EGRET repeats this process separately using genotype information for each individual, producing a collection of individual-specific GRNs (Figure 1C). These networks can then be examined to identify features that are unique to specific genotypes or are associated with particular phenotypic states.

2.2 Regulatory disruption scores

We used EGRET to quantitatively estimate the predicted regulatory changes produced by SNPs for a given gene, TF, or TF-gene relationship. A higher edge weight between a TF i and a gene j is interpreted as a higher confidence that the TF binds the promoter of and regulates the expression of gene j. To analyze EGRET networks, one can compare two networks with different genotypes, or compare a genotype-specific network E^* to the baseline genotype-agnostic network B^*.

As a means of assessing these effects, we define and calculate three different regulatory disruption scores for nodes and edges in genotype x (Figure 1D): the *edge disruption score* $d^{(E)}_{x_{ij}}$ to quantify the extent to which a TF-gene regulatory relationship is disrupted by genetic variants, the *gene disruption score* $d^{(G)}_{x_j}$ to quantify the extent to which a gene has disrupted regulation due to genetic variants in its promoter region, and the TF disruption score $d^{(TF)}_{x_i}$ to indicate the extent to which a TF's genome-wide binding sites are disrupted by genetic variants. These scores are defined per edge/node in each genotype-specific EGRET network E^* by comparing it to a baseline genotype-agnostic GRN B^* constructed by applying message-passing to C, P, and M instead of C, P, and E:

$$d^{(E)}_{x_{ij}} = |E^*_{x_{ij}} - B^*_{ij}|$$

where $E^*_{x_{ij}}$ denotes the EGRET score for the edge ij in the EGRET network for individual x for a given genotype and B^*_{ij} is the baseline edge weight for edge ij from the GRN predicted without using genotype information. This score quantifies the amount to which edges are disrupted by variants in a given cell line compared to a reference, genotype-agnostic regulatory network.

Similarly, TF disruption scores $d^{(TF)}_{x_i}$ and gene disruption scores $d^{(G)}_{x_j}$ are calculated by taking the sum of edge disruption scores around the specific TF or gene in question:

$$d^{(TF)}_{x_i} = \sum_j \left| E^*_{x_{ij}} - B^*_{ij} \right|$$

$$d^{(G)}_{x_j} = \sum_i \left| E^*_{x_{ij}} - B^*_{ij} \right|$$

3 Applications of EGRET

3.1 Constructing EGRET networks from two genetically distinct cell lines

To test the ability of EGRET to distinguish genotype-specific patterns of gene regulation, we analyzed two blood-derived cell lines, GM12878 and K562, using LCL gene expression data and eQTLs from GTEx as our population-level information (Supplementary Notes S2 and S3). We chose these cell lines because high quality genome sequence is available for both GM12878 [11] and K562 [12], providing high confidence variant calls; this was especially useful for K562 because the cell line is aneuploid [12]. In addition, both cell lines have had relatively large numbers of TFs mapped by ChIP-seq (110 TFs for GM12878 and

204 TFs for K562 in the ReMap 2018 database [13]), allowing us to validate predicted differential gene regulation as interpreted as differential TF binding.

For network construction, we used the TF-TF interaction data reported by Sonawane and colleagues [14]; genes and TFs with low expression, defined as having non-zero values in <50 samples, were filtered out (Supplementary Note S2). We constructed a genotype-naive prior using FIMO [10] to identify TF motifs in the promoter regions of genes ([-750, +250] relative to the transcription start sites; Supplementary Note S1). We obtained expression QTLs for lymphoblastoid cell lines from GTEx [15], which define instances where a SNP is significantly associated with the expression levels of a gene (denoted the "eGene"). These were filtered to select those eQTLs in which the variant alternate allele exists in either GM12878 or K562 and maps to a TF motif in the promoter region of an eGene. We then ran QBiC [6] on the relevant eQTLs to predict SNP effects on binding q (Supplementary Note S3). We used EGRET to take the selected eQTL beta values β, QBiC effects q and individual genotype x and construct a genotype-specific "EGRET prior" E for each of GM12878 and K562 (Supplementary Note S4) and infer genotype-specific network models for both the GM12878 and K562 cell lines. We also constructed a reference genotype-agnostic "baseline" GRN for LCLs using PANDA [7], and calculated the edge disruption score for each TF-gene pair in the network. A deviation from the naive edge weight requires that an edge contain an alternate allele at a location within a TF binding motif in the promoter region of the target gene, that there is eQTL data indicating that a variant in the promoter alters the expression of the target gene, and that QBiC predicts a SNP in the promoter will alter TF binding. Comparing edges in E against M identified 1,520 genotype-altered prior edges for GM12878 and 1,182 for K562 (Figure S2) out of a total of 39,690,052 edges in the naive prior M. Thus, most edge disruption scores are very close to zero in both cell lines (Figure S3).

To assess the improved ability of EGRET networks to predict TF binding compared to baseline networks, we used ChIP-seq data [13] to construct cell line-specific reference regulatory networks, connecting a TF to a gene if there was ChIP-seq evidence for a regulatory interaction (Supplementary Note S5.1). At multiple cut-offs for the edge disruption scores, EGRET networks outperformed the genotype-agnostic baseline network prediction of TF binding for potentially variant-disrupted edges (Tables S1 and S2, Supplementary Note S5.2). Based on this, we consider potentially variant-impacted scores to be those at or above 0.35 and a "high" disruption score to be anything at or above 0.5.

We also calculated the "regulatory difference score" for each edge $R_{ij}^{(E)}$ between genotypes GM12878 (g) and K562 (k), defined as

$$R_{ij}^{(E)} = \left| d_{g_{ij}}^{(E)} - d_{k_{ij}}^{(E)} \right|.$$

The magnitude of this score is the difference in predicted edge disruption scores between GM12878 $d_{g_{ij}}^{(E)}$ and K562 $d_{k_{ij}}^{(E)}$ and reflects the assumption that genetic differences between cell lines will cause differences in predicted regulatory TF-gene interaction strength. To validate these predictions, we again used available ChIP-seq data [13] to construct cell line-specific reference regulatory networks (Supplementary Note S5.1). We used these reference networks to construct a *differential regulatory network* by taking the absolute value of the difference between the two networks. In analyzing differences in the EGRET-predicted networks, we found that edges with high differential regulation scores $R_{ij}^{(E)}$ were enriched for edges showing differential TF binding in the ChIP-seq based differential regulatory network (Fisher p-value = 2.4e-226, T-test p-value = 2.296e-07).

As an example of genotype-specific TF-gene binding, consider the edge between the TF RELA and the gene SLC16A9 (ENSG00000165449), which has an edge disruption score of 0.000256 in GM12878 and 6.1 in the K562. These scores suggest that the binding of RELA to the promoter region of SLC16A9 is disrupted in K562, but not in GM12878. The positions of eQTLs, genetic variants and ChIP-seq binding regions for RELA in both genotypes (Figure 2A), indicate an eQTL variant is present in the promoter region of SLC16A9 (purple track in Figure 2A), is associated with the expression of SLC16A9, resides within a RELA binding motif, and is predicted by QBiC to affect the binding of RELA at that location; the disrupting variant is present only in K562 (orange track in Figure 2A) and not in GM12878; this prediction is confirmed by the presence of a RELA ChIP-seq binding range in GM12878, but not in K562 (teal track in Figure 2A). As a second example, consider the edge between of the TF ARID3A and the gene PMS2CL (ENSG00000187953), with an edge disruption score of 0.0096 in GM12878 and 1.066 in K562, suggesting that the binding of ARID3A to the promoter region of PMS2CL is disrupted in K562, but not in GM12878. This prediction is confirmed when looking at ChIP-seq binding data in the region (Figure 2B).

The results from comparing these cell lines also suggests that genes with high regulatory difference scores between GM12878 and K562 may harbor variants that change gene expression. To explicitly test this, we used data from an in-vitro allele-specific expression (ASE) assay (Biallelic Targeted Self-Transcribing Active Regulatory Region sequencing — BiT-STARR-seq) performed in LCLs [16] (Supplementary Note S5.3). We defined regulatory differences scores per gene (Supplementary Note S5.3) and found that the 101 genes having highest differential regulation scores $R_j^{(G)}$ (within the top 10%) were enriched for genes harboring ASE-causing variants located within promoter region TF motifs (Fisher p-value = 2.5e-03). As a second independent validation, we also compared those genes with a regulatory difference score in the top 10% with data from a published chromatin accessibility QTL (caQTL) analysis in LCLs (Supplementary Note S5.4) [17] and found they were enriched for having caQTLs within motifs in their promoter regions (Fisher p-value = 1.4e-04), suggesting that many of these predicted regulatory SNPs alter their associated regulatory networks by affecting chromatic accessibility. It is worth noting that these results are based only on the genotypes of two cell lines; we anticipate that using a larger number of genotyped cell lines with available ChIP-seq, caQTL and ASE data would increase both the specificity and sensitivity of predicting genotype-mediated effects in gene expression.

Overall, these results from two cell lines paint a compelling picture. EGRET is capable of synthesizing diverse sources of data to model gene regulatory processes and can predict genotype-associated patterns of gene expression.

3.2 EGRET networks for a population of individuals reveal tissue-specific disease associations

A growing body of work indicates that cell-type specific gene regulatory processes affect gene expression [18, 14] and do so in a manner that depends on an individual's genotype [19, 20, 21], often producing network changes that alter the structure of functional "communities" comprised of TFs and genes enriched for tissue-specific biological processes [22]. Banovich and colleagues [17] had previously analyzed RNA-seq data derived from three cell types: lymphoblastoid cell lines (LCLs), induced pluripotent stem cells (iPSCs), and cardiomyocytes (CMs; differentiated from the iPSCs). They demonstrated that genes preferentially expressed in CMs were enriched for processes associated with coronary artery disease, and those enriched in LCLs were associated with immune-related conditions. Our working hypothesis was that these effects should be linked to tissue-specific regulatory processes affected by an individual's genetic background.

To test this, we constructed 357 individual-specific EGRET networks using expression, genotype and eQTL data from 119 Yoruba individuals for all three cell types used in the Banovich *et al.* [17] study (Supplementary Note S6). We also constructed a baseline, genotype-agnostic GRN using PANDA for each cell type (Supplementary Note S6.1). Using these, we calculated TF disruption scores for each individual EGRET network to identify TFs whose regulatory influence was disrupted by variants. TF disruption scores ($d_{x_i}^{(TF)}$) were scaled these per individual and cell type to have a mean of zero and standard deviation of one, and are denoted $d_{x_i}^{(TF)'}$ (Supplementary Note S6.2). We then downloaded a list of genes from the NHGRI-EBI GWAS catalog [23] that are associated with Crohn's disease (CD) and coronary artery disease (CAD) (Tables S3 and S4). Those TFs whose coding gene is disease-associated were then considered disease-related TFs. We tested to see if disease-associated were more likely to have significant disruption scores in relevant cell types. Using a T-test, we found that TF disruption scores were significantly higher in cardiomyocytes (CMs) for TFs associated with CAD than were the disruption scores for non-CAD related TFs. This CAD enrichment was not observed in LCLs. Similarly, we found TF disruption scores in LCLs, but not CMs, were substantially higher for TFs linked to Crohn's disease that for non CD-linked TFs (Table 2). This analysis leads to an important observation: genotype-mediated disease-related TF disruptions are cell-type specific and can be identified using networks inferred using EGRET. Indeed, we find that the highest TF disruption scores for CAD TFs occur in CMs (Figure 3A) and that the highest TF disruption scores for CD TFs occur in LCLs (Figure 3B).

Further supporting this observation, the TF disruption signal in CAD is dominated in a subset of the study population by single a TF, ERG, which is a member of the erythroblast transformation-specific (ETS) gene family and known to be involved in angiogenesis [24]. In these individuals, the high TF disruption scores for CAD genes in CMs are driven by the presence or absence of a mutation on chromosome 1 (chr1:201476815, an eQTL for CSRP1) that lies in the binding motif for the TF ERG in the promoter region of the gene CSRP1 (ENSG00000159176). While ERG is identified as a CAD-related gene in the GWAS catalog, CSRP1 (alias CRP1) is not. However, CSRP1 is a known smooth muscle marker [25] and has been found by GTEx [15] to be highly expressed in smooth muscles, especially in arteries (Figure S4). CSRP1 has also been associated with the bundling of actin filaments [26], cardiovascular development [27], and with response to arterial injury [28]. Further, knockdown of CSRP1 in zebrafish caused cardiac bifida [29] and a frameshift mutation CSRP1 has been linked to congenital cardiac defects in a large human pedigree [30]. The results of our EGRET analysis support a previously unreported mechanism of action for ERG in heart disease—that ERG regulates the expression of CSRP1.

We also tested the hypothesis that the network effects of genetic variants have the potential to subtly change the modular structure of genotype-specific networks, altering the functional network communities (modules) active in an individual. ALPACA [22] is a method that compares the community structure of two networks and identifies community that differ between the networks. The resulting differential modularity (DM) scores indicate which genes have undergone the greatest change in their "modular environment." We used ALPACA to compare the community structure of the individual/tissue genotype-specific GRNs with the baseline genotype-agnostic GRN, and calculated the DM score for each network node (Supplementary Note S6.3, Figure S5).

Given that individual 18 had the greatest TF disruption score for ERG in CMs, we further investigated cellular processes that may be variant-perturbed within this individual's cell-type specific EGRET networks. We ranked individual 18's genes by DM scores from highest to lowest in each cell type reflecting their predicted impact on altering the modular structure of each cell-type specific network. We used GORILLA [31] with these ranked lists to identify GO biological process functions associated with communities altered by the presence of specific SNP variants. Several GO terms relevant to CMs and cardiovascular functioning and development, including "regulation of actomyosin structure organizarion," "prepulse inhibition," "ephrin receptor signaling pathway," "maintenance of postsynaptic specialization structure," and "actin cytoskeleton reorganization" are enriched in CMs from Individual 18 (Figure 4, Table S5) but not in their LCLs or iPSCs (Figure 4, Tables S6 and S7). For full enrichment results, see Figure S6, Figure S7, and S8, Figure S9.) Further evidence of tissue-specific alteration of functional modules can be seen by examining the DM scores of disease-associated genes (as annotated by the NHGRI-EBI GWAS catalog [23]). Coronary artery disease genes with high DM scores in CMs had low DM scores in iPSCs and LCLs (Figure 5A). In contrast, genes associated with Crohn's disease, which has a strong immune component, that had high DM scores in LCLs had low DM scores in iPSCs and CMs (Figure 5B).

EGRET also predicts dosage effects of regulatory SNP variants on network structure. Consider CSRP1, which we previously discussed as having a regulatory SNP in its promoter region that can affect binding of the transcription factor ERG. EGRET shows that the presence of a variant SNP in the promoter region of CSRP1 affects not only regulation by ERG (as seen by a substantial TF disruption score) but also the role that CSRP1 plays in altering the functional modules in cardiomyocyte GRN models. As seen by CSRP1's DM score in Figure S10, EGRET predicts that the variant SNP exerts its influence on network structure in a dosage-specific manner; individuals homozygous for the disrupting variant are predicted to exhibit the greatest impact on the modularity, those who are heterozygous to have an intermediate effect, and those homozygous for the wild-type to exhibit minimal or no effect on modularity.

Collectively, these results suggest that phenotype- and disease-associated variants can act through disruption of TF binding leading to regulatory changes that manifest themselves both through altered expression of specific target genes and the modification of GRN functional community structure.

4 Discussion

After more than a decade of GWAS, it has become clear that many, if not most, phenotype- and disease-associate genetic variants influence gene regulation and TF binding. While progress has been made in predicting the effects of genetic variants on transcription factor binding [32, 6], our goal was to take such analyses further by estimating both the effect of these genetic variants on the *regulatory* relationships between TFs and their target genes and the way in which the collective TF-gene changes alter the overall gene regulatory network. EGRET infers individual-specific GRNs by combining multiple lines of evidence (Figures 1 and S1) to predict the effects of an individual's genetic variants on TF-to-gene edges and to construct a complete, individual-specific bipartite GRN.

To initiate EGRET, TF motifs are used to construct a prior bipartite network of the presence or absence of TFs in the promoter regions of genes. This prior serves as an initial "guess" as to which TFs bind within the promoter regions of genes, and thus potentially regulate their expression. This prior is then modified to account for individual-specific genetic information using the individual's genotype combined with publicly available eQTL data [15], as well as computational predictions of the effects of variants on TF binding using QBiC [6]. This produces an individual-specific regulatory prior that depends on genotype. For a given individual and a given prior edge connecting TF i to gene j, the edge weight is penalized if the individual has a genetic variant meeting three conditions: the individual must have (1) an alternate allele at a location within a TF binding motif in the promoter region of a gene, which (2) is an eQTL affecting the expression of the gene adjacent to the promoter, and (3) that variant allele must be predicted by QBiC to affect the binding of the TF corresponding to the motif at that location. Each of these data types is essential to the accurate capturing of variant-derived regulatory disruptions (Supplementary Note S7, Figure S11).

The altered prior is then integrated with gene expression data and TF-TF protein-protein interaction information to further refine the edge weights using message passing [7] to seek consistency among the various input data. This iterative optimization process reflects our understanding of the regulatory process: TFs preferentially bind in canonical sites defined by their known sequence motifs and exert control on the regulatory process; if two genes are co-expressed, they are more likely to be co-regulated and thus are more likely to be regulated by a similar set of TFs; if two TFs physically interact, they can form regulatory complexes that bind promoter regions and allow "indirect" regulation of genes using sites not captured in the regulatory prior. This refinement of edge weights through message passing has been found to improve the prediction accuracy of GRNs [7] and has been successful in the construction of meaningful GRNs in the study of various diseases, including chronic obstructive pulmonary disease [33], asthma [34], ovarian cancer [35] and colorectal cancer [18, 36]. In our analysis, message passing among the EGRET prior E, co-expression matrix C and PPI prior P improved the structure of the overall network by refining the edge weights and improving the ability of the edge weights to predict ChIP-seq binding information (Supplementary Note S7).

In our applications and demonstrations described below, EGRET was able to identify regulatory differences between two genotypes of cell lines, and, when applied to a population over multiple cell types, was able to identify cell-type specific disease-associated regulatory disruptions.

We demonstrated applications of EGRET first using two blood-derived cell lines, GM12878 and K562, that have been sequenced and extensively characterized with data including ChIP-seq TF binding data. Using genotype data from these cell lines, and LCL gene expression and eQTL data from GTEx, we inferred GRNs specific to these cell lines. Overall, we found that the predicted GRNs improved on the baseline networks inferred without using genotype information. In comparing the networks for these cell lines we found, as expected, that most of the regulatory edges are identical. However, we were able to identify edges specific to each cell line in which EGRET predicted disrupted TF-gene binding, a result that we were able to confirm using ChIP-seq data. We showed that genes that had high regulatory difference scores between the two cell lines were enriched for QTLs associated with chromatin accessibility and enriched for allele-specific expression.

We then applied EGRET to infer 357 GRNs using RNA-seq data generated from three cell lines, lymphoblastoid cell lines (LCLs), induced pluripotent stem cells (iPSCs), and cardiomyocytes (CMs; differentiated from the iPSCs), derived from 119 Yoruba individuals for whom genotype data were also available. The 357 networks, one network in each cell type for each individual, captured individual and cell-type specific variant-disrupted regulatory relationships, as well as cell-type specific disease associations, and provided insight into disease-associated regulatory mechanisms. We found that TF regulatory relationships with genes linked to Crohn's disease were disrupted in LCLs for individuals carrying specific genetic variants. We also saw that TFs implicated in coronary artery disease had edges that were disrupted in iPSC-derived cardiomyocytes for a subset of individuals and identified a variant-perturbed relationship between the TF ERG and its target gene CSRP1 (both of which have been associated with cardiovascular disease). This variant alters the modular structure of EGRET networks in a dosage-dependent manner, and may be influencing the risk for coronary artery disease.

These results demonstrate that EGRET is able to synthesize genetic and gene expression data in a way that, for the first time, allows verifiable, disease-associated regulatory changes to be inferred for individual research subjects. However, EGRET can be applied to any individual for whom genotype data are available, and without associated gene expression data, provided there is expression and eQTL data from a relevant cell type obtained from a sufficiently large population to infer accurate regulatory network models. This implies that EGRET can be used to retrospectively analyze large cohort GWAS studies to tease out mechanistic associations for phenotype-linked genetic variants, as well as in the context of new studies that seek to understand disease mechanisms and the regulatory role of noncoding genetic variants.

Data and Code Availability

EGRET is available through the Network Zoo R package (netZooR v0.9; netzoo.github.io) with a step-by-step tutorial.

5 Funding

DW, MBG, and JQ are supported through a grant from the US National Cancer Institute, 1R35CA220523. MBG and JQ also supported by 1U24CA231846. JP is supported by a grant from the US National Heart, Lung and Blood Institute, K25HL140186.

6 Author Contributions

DW and JP developed the EGRET method, DW implemented, tested and applied the method, DW and JP interpreted the results, MBG assisted with software distribution and manuscript preparation, JQ and KG provided input into all stages of the project, JP conceived of the study.

References

[1] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. An expanded view of complex traits: From polygenic to omnigenic. *Cell*, 169(7):1177 – 1186, 2017.

[2] Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K. Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Steven McCarroll, Benjamin M. Neale, Roel A. Ophoff, Michael C. O'Donovan, Gregory E. Crawford, Daniel H. Geschwind, Nicholas Katsanis, Patrick F. Sullivan, Bogdan Pasaniuc, Alkes L. Price, and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics*, 50(4):538–548, 2018.

[3] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, Stephan Ripke, Felix R Day, Shaun Purcell, Eli Stahl, Sara Lindstrom, John R B Perry, Yukinori Okada, Soumya Raychaudhuri, Mark J Daly, Nick Patterson, Benjamin M Neale, Alkes L Price, ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, and The RACI Consortium. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, 2015.

[4] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, and Jian Yang. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nat Genet*, 48(5):481–487, May 2016.

[5] Bryce van de Geijn, Hilary Finucane, Steven Gazal, Farhad Hormozdiari, Tiffany Amariuta, Xuanyao Liu, Alexander Gusev, Po-Ru Loh, Yakir Reshef, Gleb Kichaev, Soumya Raychauduri, and Alkes L Price. Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability. *Human Molecular Genetics*, 29(7):1057–1067, 10 2019.

[6] Vincentius Martin, Jingkang Zhao, Ariel Afek, Zachery Mielko, and Raluca Gordân. Qbic-pred: quantitative predictions of transcription factor binding changes due to sequence variants. *Nucleic acids research*, 47(W1):W127–W135, 2019.

[7] Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. Passing messages between biological networks to refine predicted interactions. *PloS one*, 8(5):e64832, 2013.

[8] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *BioRxiv*, page 563866, 2019.

[9] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[10] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[11] Michael A Eberle, Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L Moore, Mitchell A Bekritsky, Zamin Iqbal, Han-Yu Chuang, Sean J Humphray, Aaron L Halpern, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome research*, 27(1):157–164, 2017.

[12] Bo Zhou, Steve S Ho, Stephanie U Greer, Xiaowei Zhu, John M Bell, Joseph G Arthur, Noah Spies, Xianglong Zhang, Seunggyu Byeon, Reenal Pattni, et al. Comprehensive, integrated, and phased whole-genome analysis of the primary encode cell line k562. *Genome research*, 29(3):472–484, 2019.

[13] Jeanne Chèneby, Marius Gheorghe, Marie Artufel, Anthony Mathelier, and Benoit Ballester. Remap 2018: an updated atlas of regulatory regions from an integrative analysis of dna-binding chip-seq experiments. *Nucleic acids research*, 46(D1):D267–D275, 2018.

[14] Abhijeet Rajendra Sonawane, John Platig, Maud Fagny, Cho-Yi Chen, Joseph Nathaniel Paulson, Camila Miranda Lopes-Ramos, Dawn Lisa DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer. Understanding tissue-specific gene regulation. *Cell reports*, 21(4):1077–1088, 2017.

[15] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature genetics*, 45(6):580, 2013.

[16] Cynthia A Kalita, Christopher D Brown, Andrew Freiman, Jenna Isherwood, Xiaoquan Wen, Roger Pique-Regi, and Francesca Luca. High-throughput characterization of genetic effects on dna–protein binding and gene transcription. *Genome research*, 28(11):1701–1708, 2018.

[17] Nicholas E Banovich, Yang I Li, Anil Raj, Michelle C Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, Jonathan E Burnett, Marsha Myrthil, Samantha M Thomas, et al. Impact of regulatory variation across human ipscs and differentiated cells. *Genome research*, 28(1):122–131, 2018.

[18] Camila M Lopes-Ramos, Marieke L Kuijjer, Shuji Ogino, Charles S Fuchs, Dawn L DeMeo, Kimberly Glass, and John Quackenbush. Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. *Cancer research*, 78(19):5538–5547, 2018.

[19] Maud Fagny, Joseph N Paulson, Marieke L Kuijjer, Abhijeet R Sonawane, Cho-Yi Chen, Camila M Lopes-Ramos, Kimberly Glass, John Quackenbush, and John Platig. Exploring regulation in tissues with eqtl networks. *Proceedings of the National Academy of Sciences*, 114(37):E7841–E7850, 2017.

[20] GTEx Consortium et al. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[21] Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, Valentin Wucher, Stephane E Castel, Andrew R Hamel, Ana Viñuela, Amy L Roberts, et al. Cell type–specific genetic regulation of gene expression across human tissues. *Science*, 369(6509), 2020.

[22] Megha Padi and John Quackenbush. Detecting phenotype-driven transitions in regulatory network structure. *NPJ systems biology and applications*, 4(1):1–12, 2018.

[23] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research*, 47(D1):D1005–D1012, 2019.

[24] Aarti V Shah, Graeme M Birdsey, and Anna M Randi. Regulation of endothelial homeostasis, vascular development and angiogenesis by the transcription factor erg. *Vascular pharmacology*, 86:3–13, 2016.

[25] James R Henderson, Teresita Macalma, Doris Brown, James A Richardson, Eric N Olson, and Mary C Beckerle. The lim protein, crp1, is a smooth muscle marker. *Developmental dynamics: an official publication of the American Association of Anatomists*, 214(3):229–238, 1999.

[26] Thuan C Tran, CoreyAyne Singleton, Tamara S Fraley, and Jeffrey A Greenwood. Cysteine-rich protein 1 (crp1) regulates actin filament bundling. *BMC cell biology*, 6(1):45, 2005.

[27] David F Chang, Narasimhaswamy S Belaguli, Dinakar Iyer, Wilmer B Roberts, San-Pin Wu, Xiu-Rong Dong, Joseph G Marx, Mary Shannon Moore, Mary C Beckerle, Mark W Majesky, et al. Cysteine-rich lim-only proteins crp1 and crp2 are potent smooth muscle differentiation cofactors. *Developmental cell*, 4(1):107–118, 2003.

[28] Brenda Lilly, Kathleen A Clark, Masaaki Yoshigi, Stephen Pronovost, Meng-Ling Wu, Muthu Periasamy, Mei Chi, Richard J Paul, Shaw-Fang Yet, and Mary C Beckerle. Loss of the serum response factor cofactor, cysteine-rich protein 1, attenuates neointima formation in the mouse. *Arteriosclerosis, thrombosis, and vascular biology*, 30(4):694–701, 2010.

[29] Kota Y Miyasaka, Yasuyuki S Kida, Takayuki Sato, Mari Minami, and Toshihiko Ogura. Csrp1 regulates dynamic cell movements of the mesendoderm and cardiac mesoderm through interactions with dishevelled and diversin. *Proceedings of the National Academy of Sciences*, 104(27):11274–11279, 2007.

[30] Amina Kamar, Akl C. Fahed, Kamel Shibbani, Nehme El-Hachem, Salim Bou-Slaiman, Mariam Arabi, Mazen Kurban, Jonathan G. Seidman, Christine E. Seidman, Rachid Haidar, Elias Baydoun, Georges Nemer, and Fadi Bitar. A novel role for csrp1 in a lebanese family with congenital cardiac defects. *Frontiers in Genetics*, 8:217, 2017.

[31] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):1–7, 2009.

[32] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[33] Kimberly Glass, John Quackenbush, Edwin K Silverman, Bartolome Celli, Stephen I Rennard, Guo-Cheng Yuan, and Dawn L DeMeo. Sexually-dimorphic targeting of functionally-related genes in copd. *BMC systems biology*, 8(1):1–17, 2014.

[34] Weiliang Qiu, Feng Guo, Kimberly Glass, Guo Cheng Yuan, John Quackenbush, Xiaobo Zhou, and Kelan G Tantisira. Differential connectivity of gene regulatory networks distinguishes corticosteroid response in asthma. *Journal of Allergy and Clinical Immunology*, 141(4):1250–1258, 2018.

[35] Kimberly Glass, John Quackenbush, Dimitrios Spentzos, Benjamin Haibe-Kains, and Guo-Cheng Yuan. A network model for angiogenesis in ovarian cancer. *BMC bioinformatics*, 16(1):1–17, 2015.

[36] Ashley J Vargas, John Quackenbush, and Kimberly Glass. Diet-induced weight loss leads to a switch in gene regulatory network control in the rectal mucosa. *Genomics*, 108(3-4):126–133, 2016.

[37] Muredach P Reilly, Mingyao Li, Jing He, Jane F Ferguson, Ioannis M Stylianou, Nehal N Mehta, Mary Susan Burnett, Joseph M Devaney, Christopher W Knouff, John R Thompson, et al. Identification of adamts7 as a novel locus for coronary atherosclerosis and association of abo with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies. *The Lancet*, 377(9763):383–392, 2011.

[38] Martin Dichgans, Rainer Malik, Inke R König, Jonathan Rosand, Robert Clarke, Solveig Gretarsdottir, Gudmar Thorleifsson, Braxton D Mitchell, Themistocles L Assimes, Christopher Levi, et al. Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants. *Stroke*, 45(1):24–36, 2014.

[39] Majid Nikpay, Anuj Goel, Hong-Hee Won, Leanne M Hall, Christina Willenborg, Stavroula Kanoni, Danish Saleheen, Theodosios Kyriakou, Christopher P Nelson, Jemma C Hopewell, et al. A comprehensive 1000 genomes–based genome-wide association meta-analysis of coronary artery disease. *Nature genetics*, 47(10):1121, 2015.

[40] Salma M Wakil, Ramesh Ram, Nzioka P Muiya, Munish Mehta, Editha Andres, Nejat Mazhar, Batoul Baz, Samya Hagos, Maie Alshahid, Brian F Meyer, et al. A genome-wide association study reveals susceptibility loci for myocardial infarction/coronary artery disease in saudi arabs. *Atherosclerosis*, 245:62–70, 2016.

[41] Derek Klarin, Qiuyu Martin Zhu, Connor A Emdin, Mark Chaffin, Steven Horner, Brian J McMillan, Alison Leed, Michael E Weale, Chris CA Spencer, François Aguet, et al. Genetic analysis in

uk biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nature genetics*, 49(9):1392, 2017.

[42] Pim van der Harst and Niek Verweij. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research*, 122(3):433–443, 2018.

[43] Yi Han, Rajkumar Dorajoo, Xuling Chang, Ling Wang, Chiea-Chuen Khor, Xueling Sim, Ching-Yu Cheng, Yuan Shi, Yih Chung Tham, Wanting Zhao, et al. Genome-wide association study identifies a missense variant at apoa5 for coronary artery disease in multi-ethnic cohorts from southeast asia. *Scientific reports*, 7(1):1–11, 2017.

[44] Yang Li, Dao Wen Wang, Yundai Chen, Can Chen, Jian Guo, Shu Zhang, Zhijun Sun, Hu Ding, Yan Yao, Lei Zhou, et al. Genome-wide association and functional studies identify scml4 and thsd7a as novel susceptibility genes for coronary artery disease. *Arteriosclerosis, thrombosis, and vascular biology*, 38(4):964–975, 2018.

[45] Wei Zhou, Jonas B Nielsen, Lars G Fritsche, Rounak Dey, Maiken E Gabrielsen, Brooke N Wolford, Jonathon LeFaive, Peter VandeHaar, Sarah A Gagliano, Aliya Gifford, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature genetics*, 50(9):1335–1341, 2018.

[46] Yoshiji Yamada, Yoshiki Yasukochi, Kimihiko Kato, Mitsutoshi Oguri, Hideki Horibe, Tetsuo Fujimaki, Ichiro Takeuchi, and Jun Sakuma. Identification of 26 novel loci that confer susceptibility to early-onset coronary artery disease in a japanese population. *Biomedical reports*, 9(5):383–404, 2018.

[47] John D Rioux, Ramnik J Xavier, Kent D Taylor, Mark S Silverberg, Philippe Goyette, Alan Huett, Todd Green, Petric Kuballa, M Michael Barmada, Lisa Wu Datta, et al. Genome-wide association study identifies new susceptibility loci for crohn disease and implicates autophagy in disease pathogenesis. *Nature genetics*, 39(5):596–604, 2007.

[48] Cécile Libioulle, Edouard Louis, Sarah Hansoul, Cynthia Sandor, Frédéric Farnir, Denis Franchimont, Séverine Vermeire, Olivier Dewit, Martine De Vos, Anna Dixon, et al. Novel crohn disease locus identified by genome-wide association maps to a gene desert on 5p13. 1 and modulates expression of ptger4. *PLoS Genet*, 3(4):e58, 2007.

[49] Miles Parkes, Jeffrey C Barrett, Natalie J Prescott, Mark Tremelling, Carl A Anderson, Sheila A Fisher, Roland G Roberts, Elaine R Nimmo, Fraser R Cummings, Dianne Soars, et al. Sequence variants in the autophagy gene irgm and multiple other replicating loci contribute to crohn's disease susceptibility. *Nature genetics*, 39(7):830–832, 2007.

[50] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

[51] Andre Franke, Jochen Hampe, Philip Rosenstiel, Christian Becker, Florian Wagner, Robert Häsler, Randall D Little, Klaus Huse, Andreas Ruether, Tobias Balschun, et al. Systematic association mapping identifies nell1 as a novel ibd disease gene. *PloS one*, 2(8):e691, 2007.

[52] John V Raelson, Randall D Little, Andreas Ruether, Hélene Fournier, Bruno Paquin, Paul Van Eerdewegh, WEC Bradley, Pascal Croteau, Quynh Nguyen-Huu, Jonathan Segal, et al. Genome-wide association study for crohn's disease in the quebec founder population identifies multiple validated disease loci. *Proceedings of the National Academy of Sciences*, 104(37):14747–14752, 2007.

[53] Jeffrey C Barrett, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, Mark S Silverberg, Kent D Taylor, M Michael Barmada, et al. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nature genetics*, 40(8):955–962, 2008.

[54] Dermot PB McGovern, Michelle R Jones, Kent D Taylor, Kristin Marciante, Xiaofei Yan, Marla Dubinsky, Andy Ippoliti, Eric Vasiliauskas, Dror Berel, Carrie Derkowski, et al. Fucosyltransferase 2 (fut2) non-secretor status is associated with crohn's disease. *Human molecular genetics*, 19(17):3468–3476, 2010.

[55] Andre Franke, Dermot PB McGovern, Jeffrey C Barrett, Kai Wang, Graham L Radford-Smith, Tariq Ahmad, Charlie W Lees, Tobias Balschun, James Lee, Rebecca Roberts, et al. Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci. *Nature genetics*, 42(12):1118–1125, 2010.

[56] Jie Huang, David Ellinghaus, Andre Franke, Bryan Howie, and Yun Li. 1000 genomes-based imputation identifies novel and refined associations for the wellcome trust case control consortium phase 1 data. *European Journal of Human Genetics*, 20(7):801–805, 2012.

[57] Eimear E Kenny, Itsik Pe'er, Amir Karban, Laurie Ozelius, Adele A Mitchell, Sok Meng Ng, Monica Erazo, Harry Ostrer, Clara Abraham, Maria T Abreu, et al. A genome-wide scan of ashkenazi jewish crohn's disease suggests novel susceptibility loci. *PLoS Genet*, 8(3):e1002559, 2012.

[58] Antonio Julià, Eugeni Domènech, Elena Ricart, Raül Tortosa, Valle García-Sánchez, Javier P Gisbert, Pilar Nos Mateu, Ana Gutiérrez, Fernando Gomollón, Juan Luís Mendoza, et al. A genome-wide association study on a southern european population identifies a new crohn's disease susceptibility locus at rbx1-ep300. *Gut*, 62(10):1440–1445, 2013.

[59] Luke Jostins, Stephan Ripke, Rinse K Weersma, Richard H Duerr, Dermot P McGovern, Ken Y Hui, James C Lee, L Philip Schumm, Yashoda Sharma, Carl A Anderson, et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124, 2012.

[60] Keiko Yamazaki, Junji Umeno, Atsushi Takahashi, Atsushi Hirano, Todd Andrew Johnson, Natsuhiko Kumasaka, Takashi Morizono, Naoya Hosono, Takaaki Kawaguchi, Masakazu Takazoe, et al. A genome-wide association study identifies 2 susceptibility loci for crohn's disease in a japanese population. *Gastroenterology*, 144(4):781–788, 2013.

[61] Suk-Kyun Yang, Myunghee Hong, Wanting Zhao, Yusun Jung, Jiwon Baek, Naeimeh Tayebi, Kyung Mo Kim, Byong Duk Ye, Kyung-Jo Kim, Sang Hyoung Park, et al. Genome-wide association study of crohn's disease in koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. *Gut*, 63(1):80–87, 2014.

[62] Suk-Kyun Yang, Myunghee Hong, Hyunchul Choi, Wanting Zhao, Yusun Jung, Talin Haritunians, Byong Duk Ye, Kyung-Jo Kim, Sang Hyoung Park, Inchul Lee, et al. Immunochip analysis identification of 6 additional susceptibility loci for crohn's disease in koreans. *Inflammatory bowel diseases*, 21(1):1–7, 2015.

[63] Jimmy Z Liu, Suzanne Van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, James C Lee, Luke Jostins, Tejas Shah, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics*, 47(9):979–986, 2015.

[64] Chengrui Huang, Talin Haritunians, David T Okou, David J Cutler, Michael E Zwick, Kent D Taylor, Lisa W Datta, Joseph C Maranville, Zhenqiu Liu, Shannon Ellis, et al. Characterization of genetic loci that affect susceptibility to inflammatory bowel diseases in african americans. *Gastroenterology*, 149(6):1575–1586, 2015.

[65] Eun Suk Jung, Jae Hee Cheon, Ji Hyun Lee, Soo Jung Park, Hui Won Jang, Sook Hee Chung, Myoung Hee Park, Tai-Gyu Kim, Heung-Bum Oh, Suk-Kyun Yang, et al. Hla-c* 01 is a risk factor for crohn's disease. *Inflammatory bowel diseases*, 22(4):796–806, 2016.

[66] Jerzy Ostrowski, Agnieszka Paziewska, Izabella Lazowska, Filip Ambrozkiewicz, Krzysztof Goryca, Maria Kulecka, Tomasz Rawa, Jakub Karczmarski, Michalina Dabrowska, Natalia Zeber-Lubecka, et al. Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the polish population. *Scientific reports*, 6:39831, 2016.

[67] Katrina M De Lange, Loukas Moutsianas, James C Lee, Christopher A Lamb, Yang Luo, Nicholas A Kennedy, Luke Jostins, Daniel L Rice, Javier Gutierrez-Achury, Sun-Gou Ji, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256–261, 2017.

[68] Yoichi Kakuta, Yosuke Kawai, Takeo Naito, Atsushi Hirano, Junji Umeno, Yuta Fuyuno, Zhenqiu Liu, Dalin Li, Takeru Nakano, Yasuhiro Izumiyama, et al. A genome-wide association study identifying rap1a as a novel susceptibility gene for crohn's disease in japanese individuals. *Journal of Crohn's and Colitis*, 13(5):648–658, 2019.

Figures

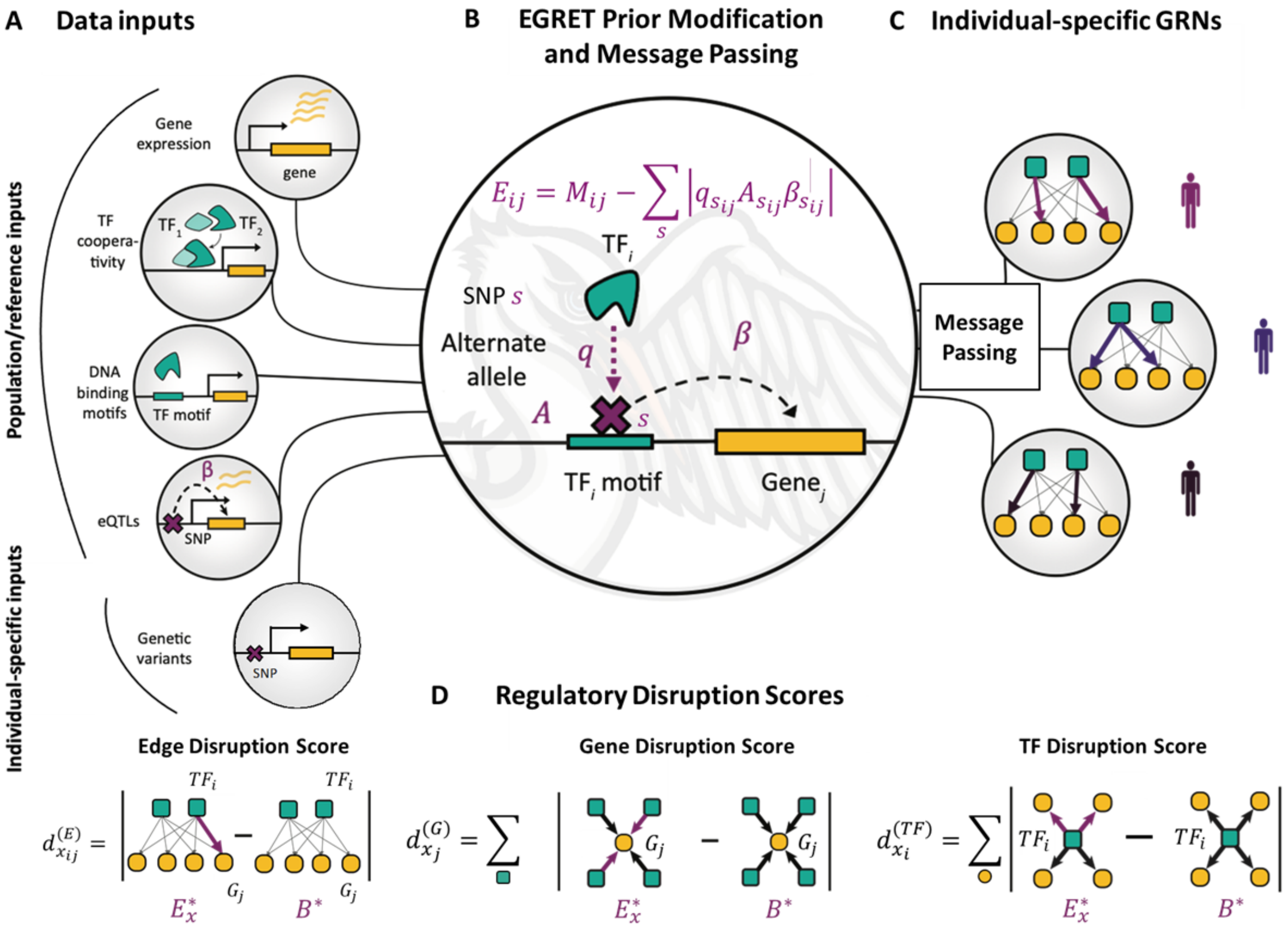

Figure 1: **EGRET integrates multiple data types to construct individual-specific GRNs.** **(A)** EGRET takes as input several data types: gene expression to estimate a gene co-expression matrix (C), PPI data as an estimate of TF-TF co-operativity (P), an initial estimate of the binding locations of TFs in the form of a TF motif-gene prior (M_{ij}), the beta values of an eQTL association between an "eSNP" and an "eGene" (β), and the genetic variants (s) harbored by the individual in question. **(B)** An individual's genetic variants are incorporated into the motif prior, penalizing motif-gene connections in which that individual carries a variant allele (A) in the relevant promoter-region motif such that the variant is an eQTL for the adjacent gene (β) and the variant is predicted by QBiC to affect TF binding at that location (q). Message passing is used to integrate the expression and PPI data with the modified motif prior, **(C)** resulting in a final, unique GRN per individual. **(D)** Regulatory disruption scores can be calculated to quantify the extent to which an edge or node in the network is disrupted by variants. Edge disruption scores are calculated by subtracting a baseline genotype-agnostic network from the individual EGRET network and taking the absolute value. TF or gene disruption scores are calculated taking the sum of the edge disruption scores around the node in question.

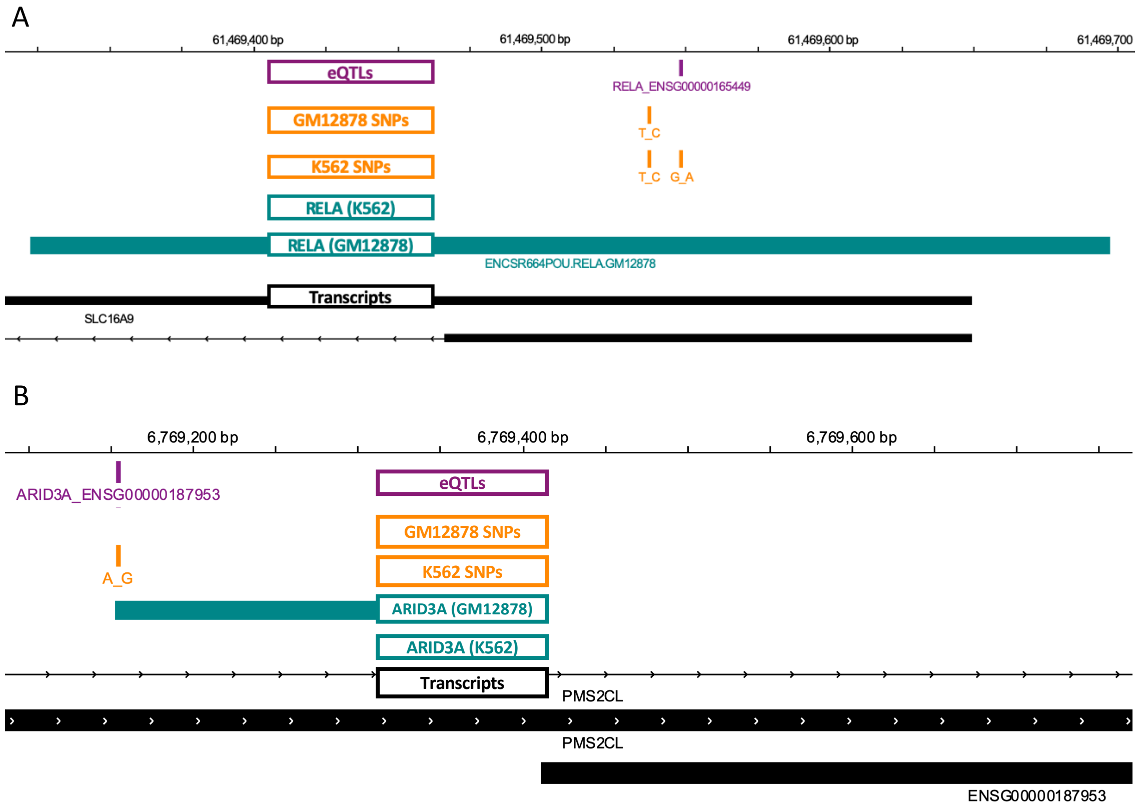

Figure 2: **EGRET identifies variant-impacted TF binding disruptions.** (**A**) Example of variant-disrupted RELA binding in K562 but not in GM12878. Positions of eQTLs (purple track), genetic variants (orange tracks), ChIP-seq binding regions (teal tracks) and genes (black track) are shown in the region of SLC16A9. (**B**) Example of variant-disrupted ARID3A binding in K562 but not in GM12878. Positions of eQTLs (purple track), genetic variants (orange tracks), ChIP-seq binding regions (teal tracks) and genes (black track) are shown in the region of PMS2CL. The eQTL track is labeled according to the TF motif in which the eSNP resides as well as the adjacent eGene.

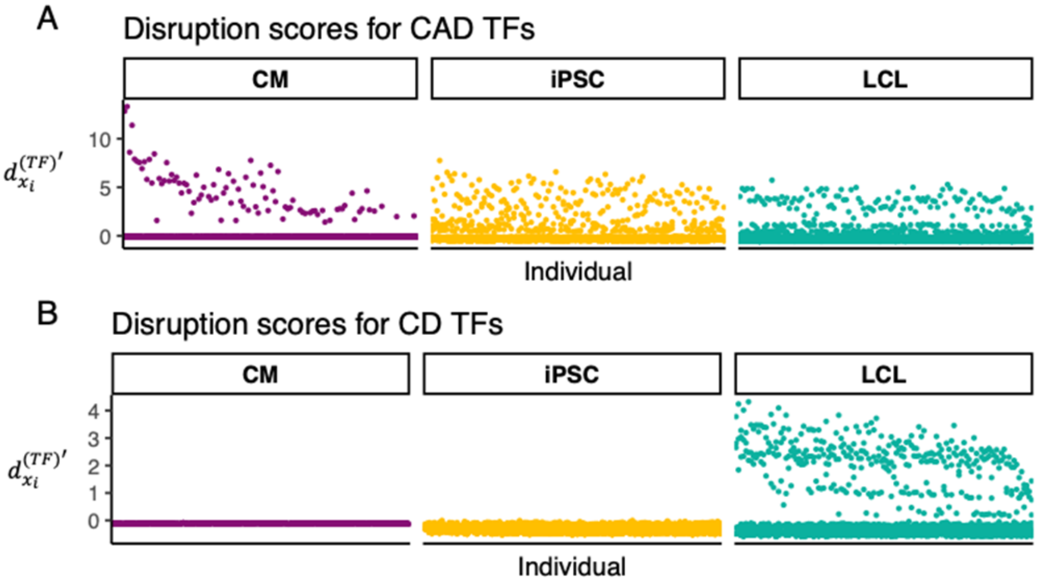

Figure 3: **Disease-related TFs in a population are disrupted in the relevant cell types.** Scaled TF disruption scores $d_{x_i}^{(TF)'}$ are shown for 119 Yoruba individuals for TFs associated with coronary artery disease (CAD) or Crohn's disease (CD). Each point represents the TF disruption score for a disease-related (CD or CAD) TF, for a given individual for a given cell type (LCL, CM or iPSC). Disease-related TFs were identified using the GWAS catalog [23]. (**A**) TF disruption scores for CAD-related TFs are highest in the cardiac-related cell type, CMs. (**B**) TF disruption scores for CD-related TFs are highest in the immune cell type, LCLs.

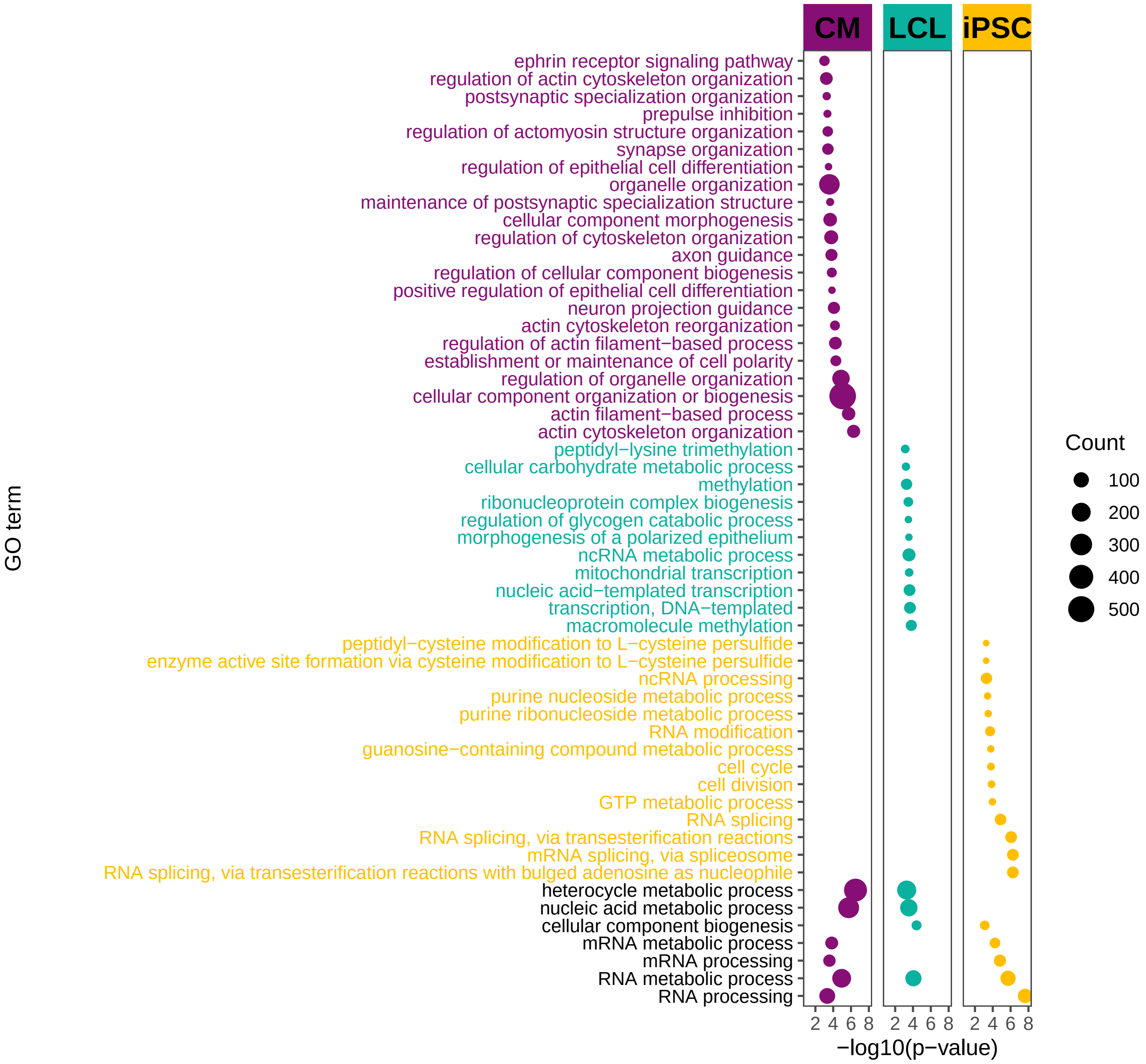

Figure 4: **Variant-disrupted genes affecting network modularity are enriched in coronary/heart related functions in CMs, for an individual with a CAD disruption signature.** GO terms enriched in genes with high DM scores for individual 18, the individual with the highest TF disruption score for ERG. Several GO terms related to coronary/cardiac function are enriched in high-DM genes in CMs but not in LCLs and iPSCs. Point size corresponds to the the number of high-DM genes annotated with the corresponding GO term. For display purposes, several generic GO terms enriched only in CMs were omitted in this figure. The entire set of enriched GO terms can be seen in Figure S9, as well as Tables S5, S8 and S7.

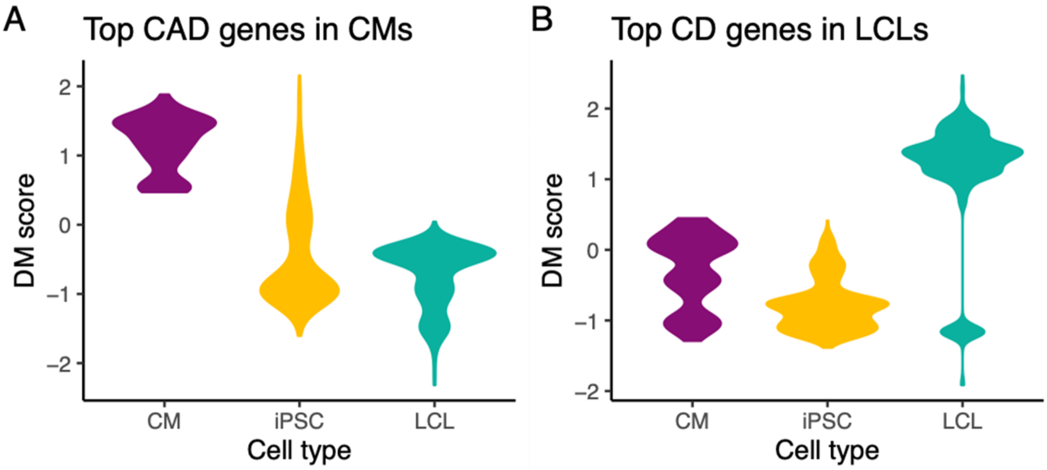

Figure 5: **Variant-disrupted disease genes affect the modularity of the individual's regulatory network in the relevant cell type.** Differential modularity (DM) scores indicate the extent to which a gene's modular environment in the network changes between the genotype-specific EGRET network and the genotype-agnostic network. **(A)** CAD-related genes with high DM scores in cardiomyocytes (CMs) have low scores in the other cell types; **(B)** CD-related genes with high DM scores in LCLs have low scores in the other cell types.

Tables

Table 1: Data types and sources used as input to EGRET. Note that the application of EGRET to the 119 Yoruba individuals uses the same motif and PPI priors as used in the cell line analysis.

Data type	Tissue/Genotype	Source
	Genotype cell line comparison	
eQTLs	LCL	https://gtexportal.org/home/datasets [15]
Gene expression	LCL	https://gtexportal.org/home/datasets [15]
Genotype	GM12878	https://www.illumina.com/platinumgenomes.html [11]
Genotype	K562	https://www.encodeproject.org/files/ENCFF538YDL/ [12]
ChIP-seq	GM12878, K562	http://pedagogix-tagc.univ-mrs.fr/remap/ [13]
PPI	N/A	https://sites.google.com/a/channing.harvard.edu/kimberlyglass/home [14]
Motif	N/A	constructed using FIMO [10]
	Population application: 119 Yoruba individuals	
eQTLs	LCL	http://eqtl.uchicago.edu/jointLCL/ [17]
eQTLs	iPSC/iPSC-CM	http://eqtl.uchicago.edu/yri_ipsc/ [17]
Gene expression	LCL	http://eqtl.uchicago.edu/jointLCL/ [17]
Gene expression	iPSC/iPSC-CM	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107654 [17]
Genotype	Yoruba individuals	http://eqtl.uchicago.edu/yri_ipsc/ [17]

Table 2: T-test p-values of differences between the TF disruption scores of disease (CD or CAD related TFs, determined from the GWAS catalog) vs. non-disease TFs in difference cell types. CAD TFs have significantly higher TF disruption scores than non-CAD TFs in CMs, but not in LCLs. CD TFs have significantly higher TF disruption scores than non-CD TFs in LCLs, but not in CMs.

Disease	Cell type	P-value
CAD	LCL	0.99831
CAD	CM	4.5256e-06
CD	LCL	5.3374e-16
CD	CM	1

Supplementary Figures

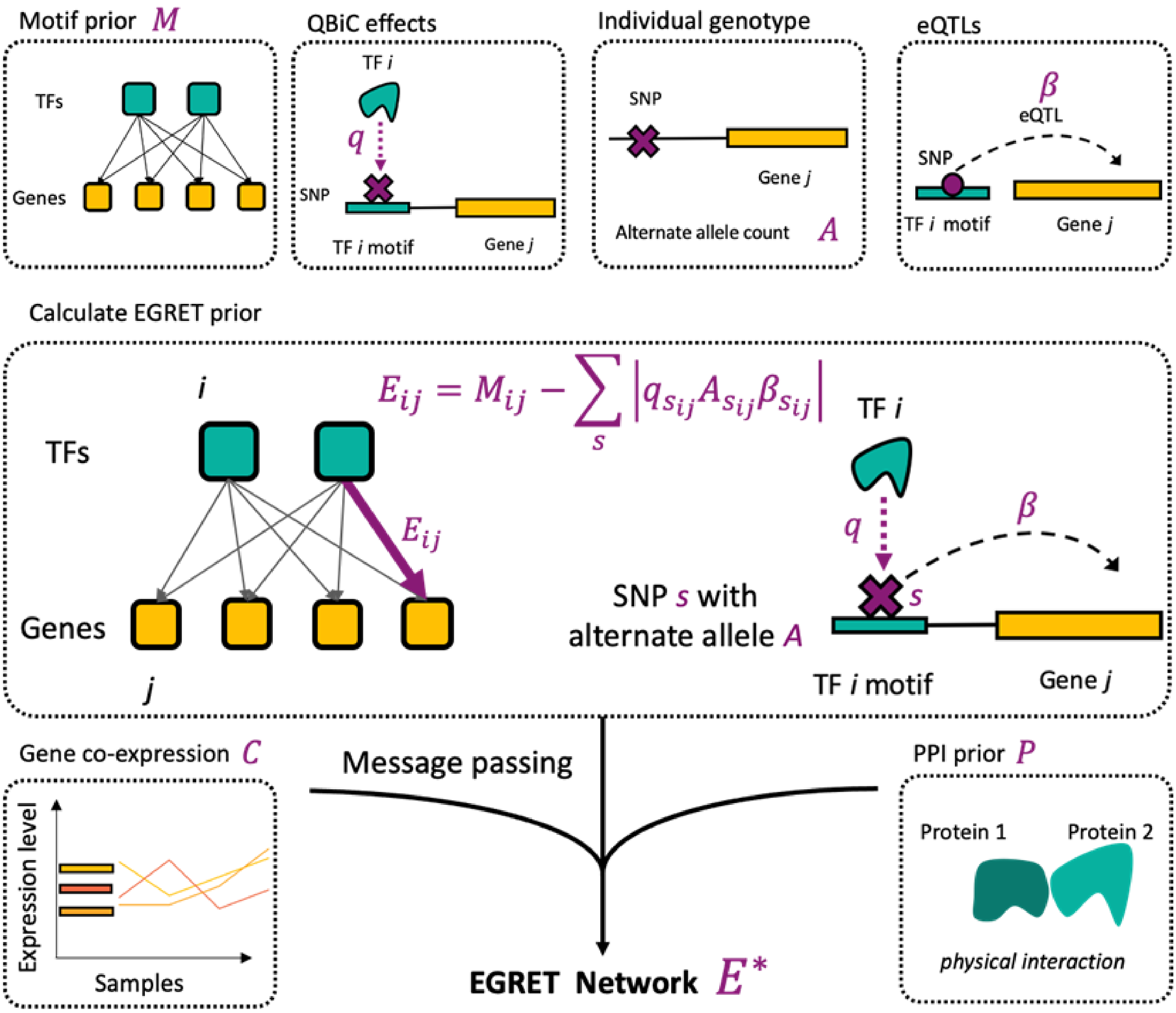

Figure S1: **Diagram illustrating the process and datatypes required for EGRET network construction.** EGRET begins with a motif-gene prior representing the presence/absence of TF motifs in the promoter regions of genes. This is then modified by the individual's genetic mutations, penalizing motif-gene edges in which there exists a variant within the TF motif for which the individual has the alternate allele (A), the variant is an eQTL for the adjacent gene (β) and the variant is predicted through QBiC to affect TF binding at that location (q). These prior edges are then penalized by the absolute value of the product of the alternate allele count, the QBiC effect and the eQTL beta value. Message passing is then integrates the co-expression data $C$and PPI data P with the modified motif prior, resulting in a final, unique GRN per individual.

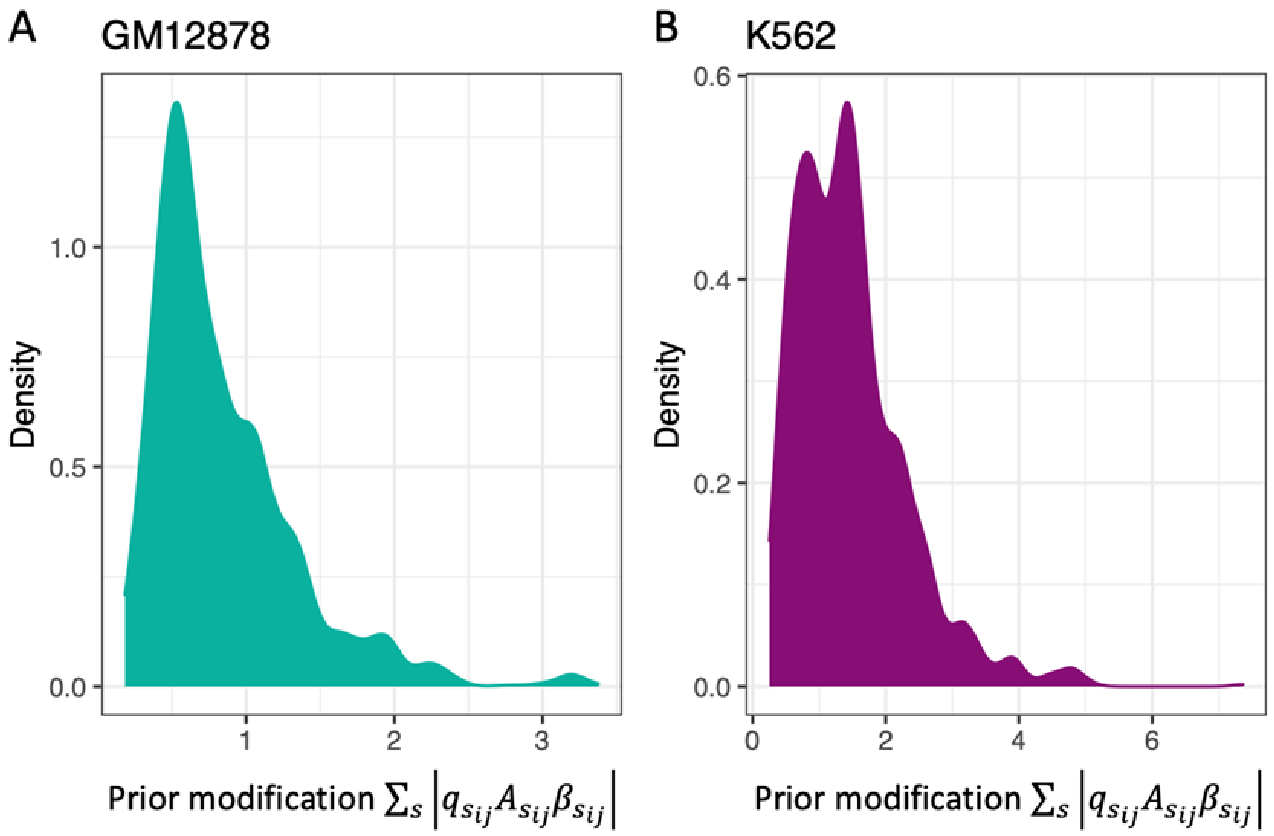

Figure S2: **Distribution of non-zero prior modifications** $\sum_s |q_{s_{ij}} A_{s_{ij}} \beta_{s_{ij}}|$ for (**A**) GM12878 and (**B**) K562.

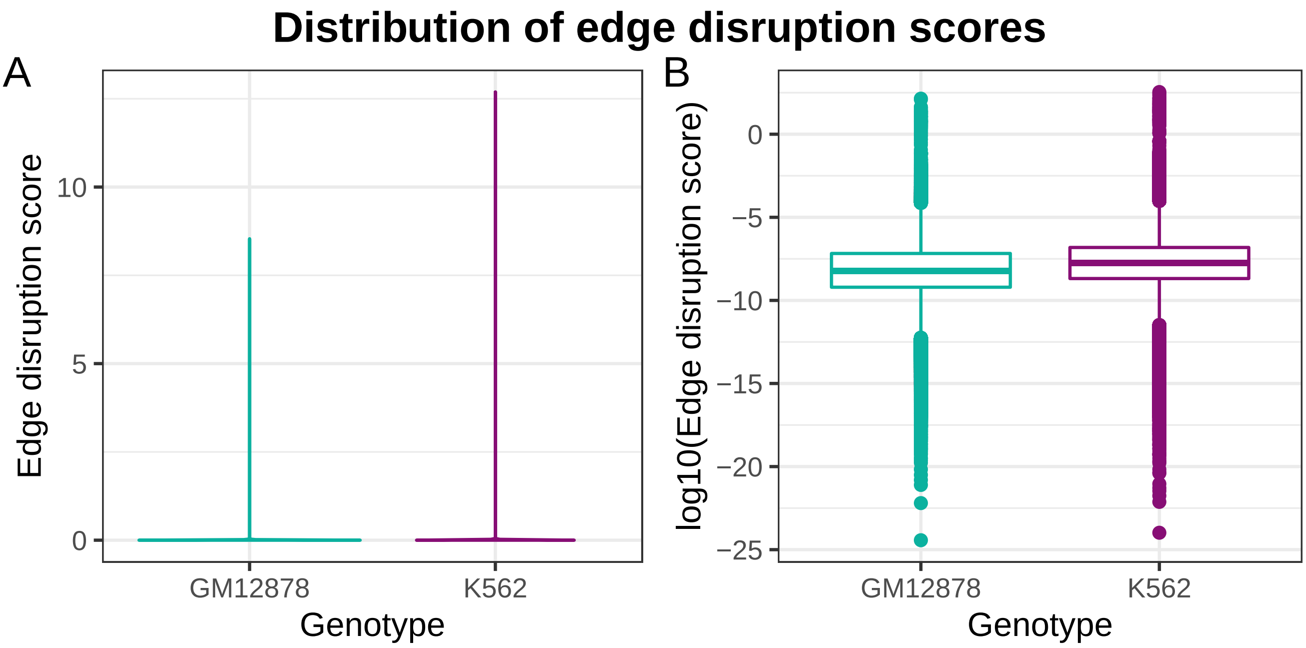

Figure S3: **Distribution of edge disruption scores** $d^{(E)}_{x_{ij}}$ **for GM12878 and K562.** (**A**) Violin plot of edge disruption scores. (**B**) Boxplot of $\log_{10}$ disruption scores.

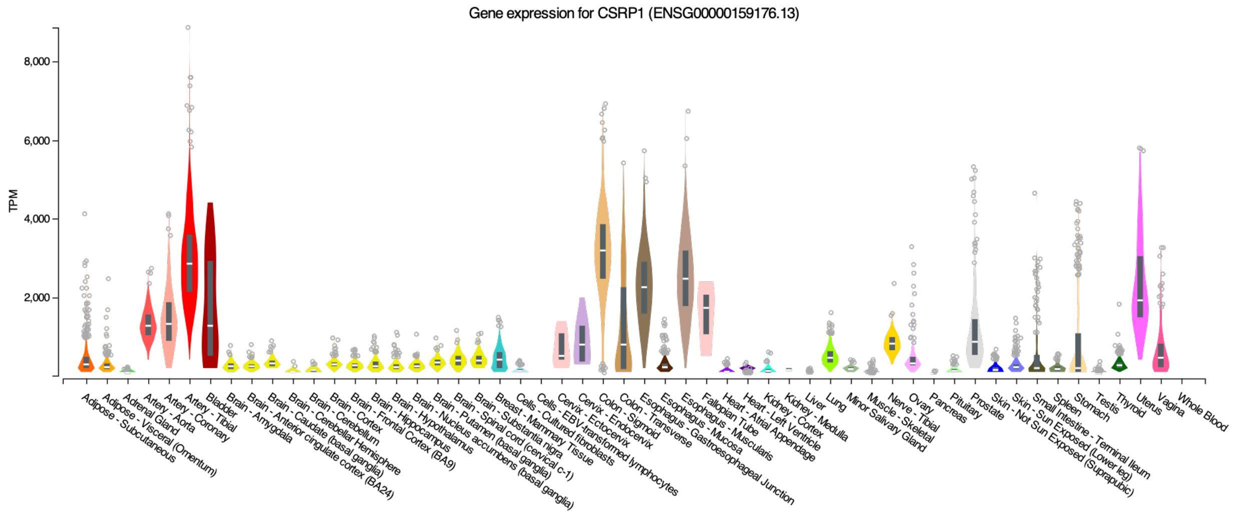

Figure S4: **CSRP1 expression.** TPM expression level of CSRP1 (ENSG00000159176) across all tissues available in GTEx. Plot obtained from the GTEx portal [15].

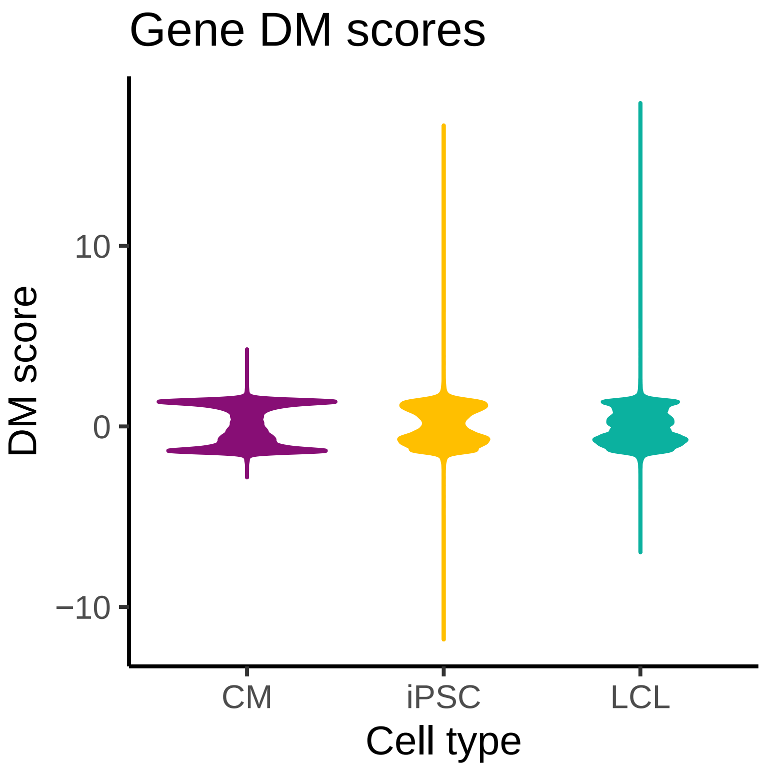

Figure S5: **Differential modularity scores.** Distributions of scaled differential modularity (DM) scores for 119 Yoruba individuals' EGRET networks in each cell type.

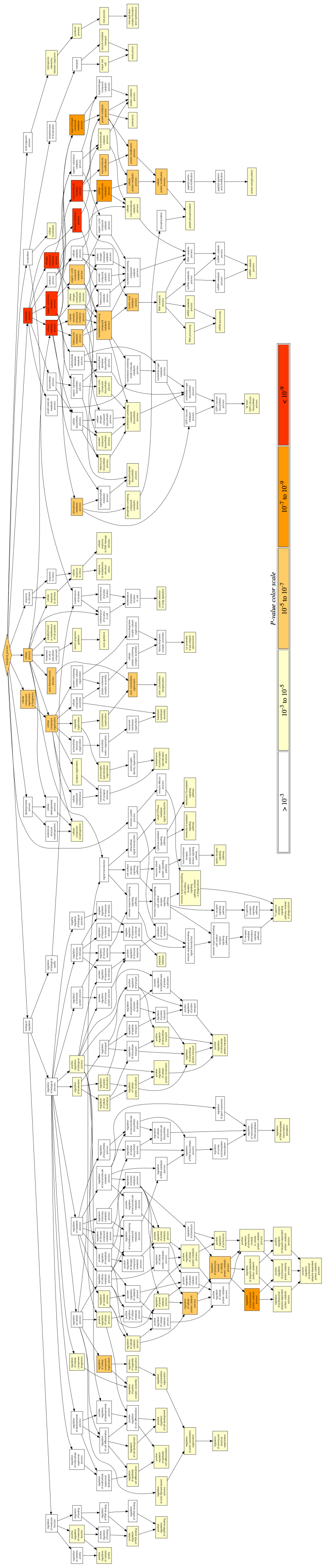

Figure S6: **Hierarchy of GO terms enriched in the genes with highest DM scores in individual 18 (genotype NA18523) in CMs, based on the mHG score test [31]. Enrichment performed using GORILLA [31].** GO terms are colored according to the significance of the p-value.

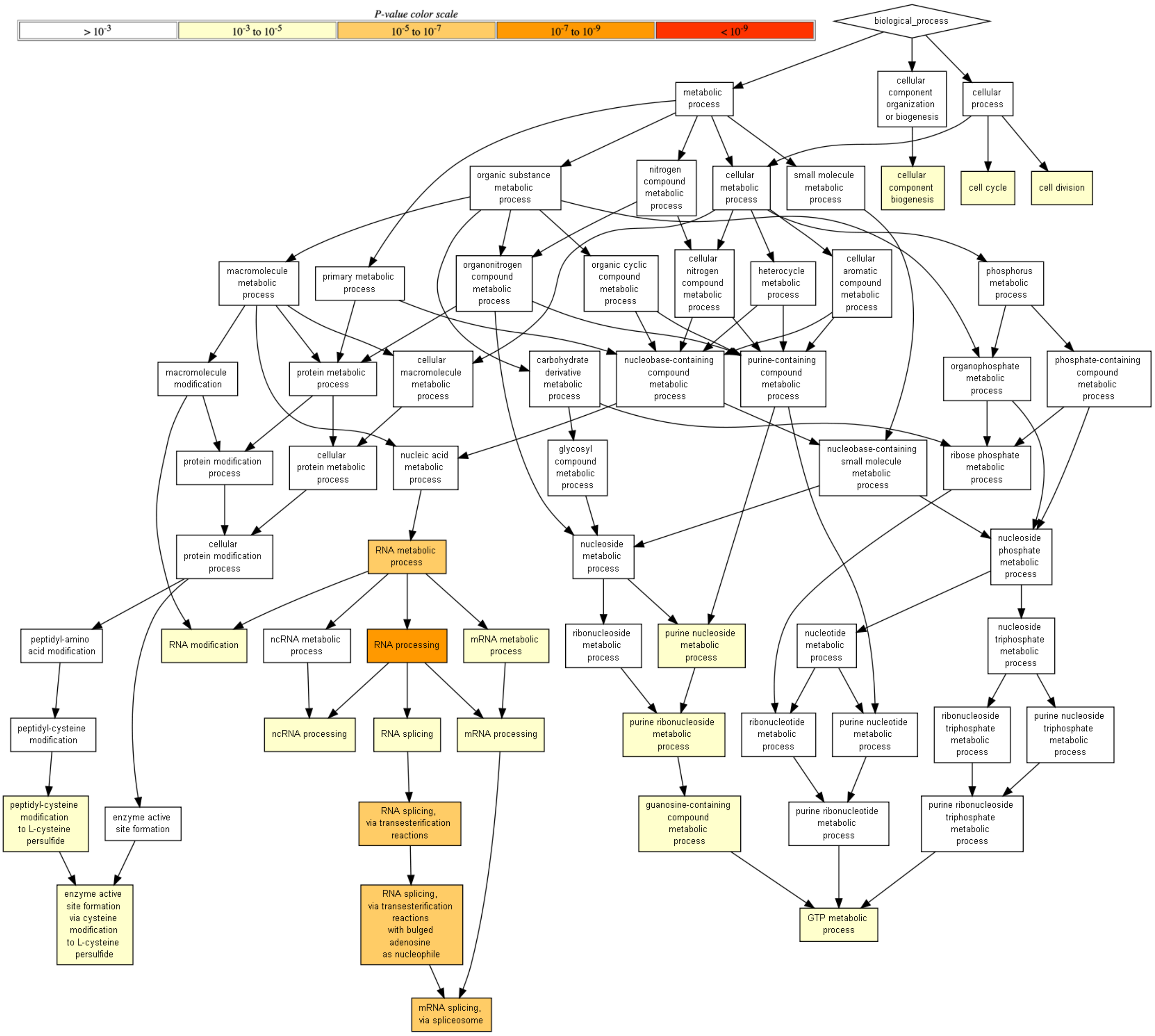
P-value color scale
> 10^-3
10^-3 to 10^-5
10^-5 to 10^-7
10^-7 to 10^-9
< 10^-9
biological_process
metabolic process
cellular component organization or biogenesis
cellular process
organic substance metabolic process
nitrogen compound metabolic process
cellular metabolic process
small molecule metabolic process
cellular component biogenesis
cell cycle
cell division
macromolecule metabolic process
primary metabolic process
organonitrogen compound metabolic process
organic cyclic compound metabolic process
cellular nitrogen compound metabolic process
heterocycle metabolic process
cellular aromatic compound metabolic process
phosphorus metabolic process
macromolecule modification
protein metabolic process
cellular macromolecule metabolic process
carbohydrate derivative metabolic process
nucleobase-containing compound metabolic process
purine-containing compound metabolic process
organophosphate metabolic process
phosphate-containing compound metabolic process
protein modification process
cellular protein metabolic process
nucleic acid metabolic process
glycosyl compound metabolic process
nucleobase-containing small molecule metabolic process
ribose phosphate metabolic process
cellular protein modification process
RNA metabolic process
nucleoside metabolic process
nucleoside phosphate metabolic process
peptidyl-amino acid modification
RNA modification
ncRNA metabolic process
RNA processing
mRNA metabolic process
ribonucleoside metabolic process
purine nucleoside metabolic process
nucleotide metabolic process
nucleoside triphosphate metabolic process
peptidyl-cysteine modification
ncRNA processing
RNA splicing
mRNA processing
purine ribonucleoside metabolic process
ribonucleotide metabolic process
purine nucleotide metabolic process
ribonucleoside triphosphate metabolic process
purine nucleoside triphosphate metabolic process
peptidyl-cysteine modification to L-cysteine persulfide
enzyme active site formation
RNA splicing, via transesterification reactions
guanosine-containing compound metabolic process
purine ribonucleotide metabolic process
purine ribonucleoside triphosphate metabolic process
enzyme active site formation via cysteine modification to L-cysteine persulfide
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile
GTP metabolic process
mRNA splicing, via spliceosome

Figure S7: **Hierarchy of GO terms enriched in the genes with highest DM scores in individual 18 (genotype NA18523) in iPSCs, based on the mHG score test [31].** Enrichment performed using GORILLA [31]. GO terms are colored according to the significance of the p-value.

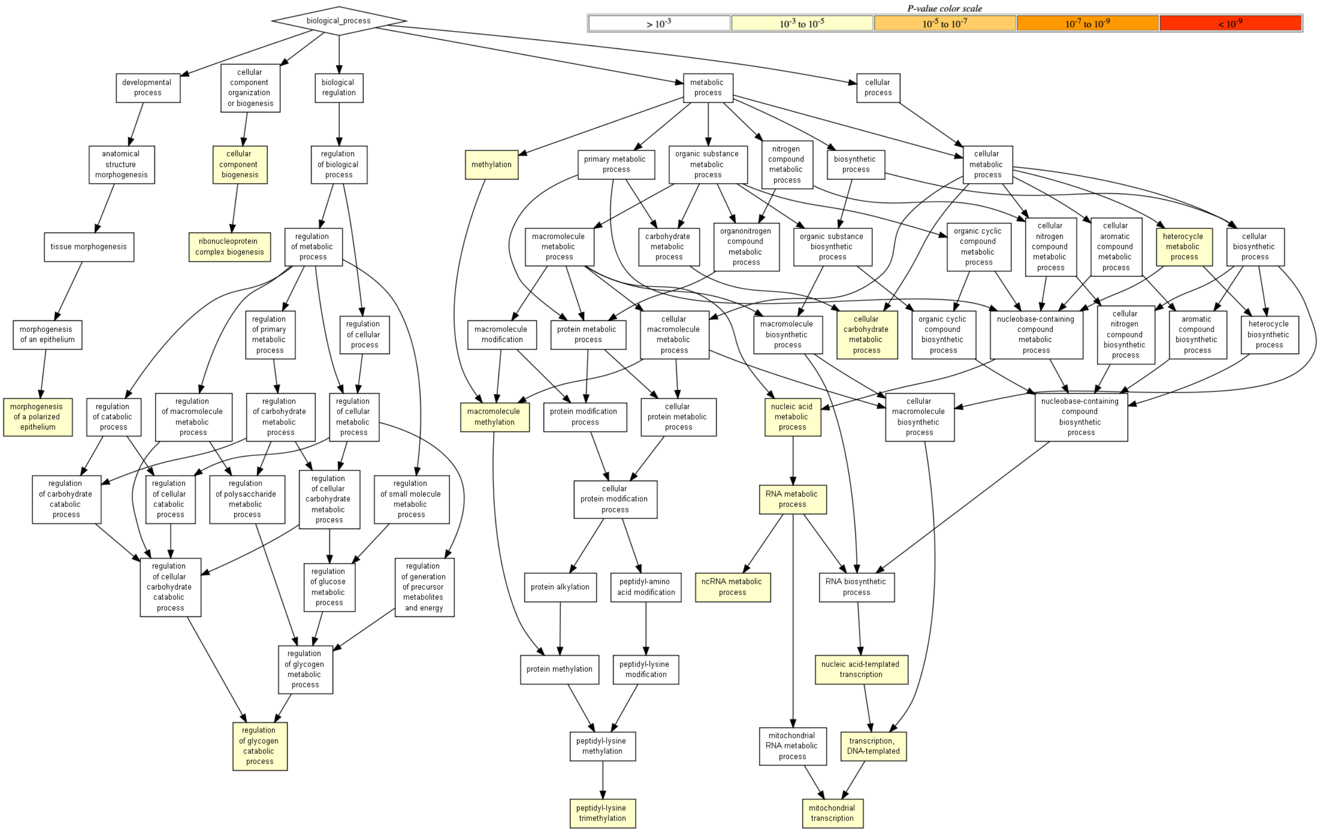
P-value color scale
> 10^{-3}
10^{-3} to 10^{-5}
10^{-5} to 10^{-7}
10^{-7} to 10^{-9}
< 10^{-9}
biological_process
developmental process
cellular component organization or biogenesis
biological regulation
metabolic process
cellular process
anatomical structure morphogenesis
cellular component biogenesis
regulation of biological process
methylation
primary metabolic process
organic substance metabolic process
nitrogen compound metabolic process
biosynthetic process
cellular metabolic process
tissue morphogenesis
ribonucleoprotein complex biogenesis
regulation of metabolic process
macromolecule metabolic process
carbohydrate metabolic process
organonitrogen compound metabolic process
organic substance biosynthetic process
organic cyclic compound metabolic process
cellular nitrogen compound metabolic process
cellular aromatic compound metabolic process
heterocycle metabolic process
cellular biosynthetic process
morphogenesis of an epithelium
regulation of primary metabolic process
regulation of cellular process
macromolecule modification
protein metabolic process
cellular macromolecule metabolic process
macromolecule biosynthetic process
cellular carbohydrate metabolic process
organic cyclic compound biosynthetic process
nucleobase-containing compound metabolic process
cellular nitrogen compound biosynthetic process
aromatic compound biosynthetic process
heterocycle biosynthetic process
morphogenesis of a polarized epithelium
regulation of catabolic process
regulation of macromolecule metabolic process
regulation of carbohydrate metabolic process
regulation of cellular metabolic process
macromolecule methylation
protein modification process
cellular protein metabolic process
nucleic acid metabolic process
cellular macromolecule biosynthetic process
nucleobase-containing compound biosynthetic process
regulation of carbohydrate catabolic process
regulation of cellular catabolic process
regulation of polysaccharide metabolic process
regulation of cellular carbohydrate metabolic process
regulation of small molecule metabolic process
cellular protein modification process
RNA metabolic process
regulation of cellular carbohydrate catabolic process
regulation of glucose metabolic process
regulation of generation of precursor metabolites and energy
protein alkylation
peptidyl-amino acid modification
ncRNA metabolic process
RNA biosynthetic process
regulation of glycogen metabolic process
protein methylation
peptidyl-lysine modification
nucleic acid-templated transcription
regulation of glycogen catabolic process
peptidyl-lysine methylation
mitochondrial RNA metabolic process
transcription, DNA-templated
peptidyl-lysine trimethylation
mitochondrial transcription

Figure S8: **Hierarchy of GO terms enriched in the genes with highest DM scores in individual 18 (genotype NA18523) in LCLs, based on the mHG score test [31].** Enrichment performed using GORILLA [31]. GO terms are colored according to the significance of the p-value.

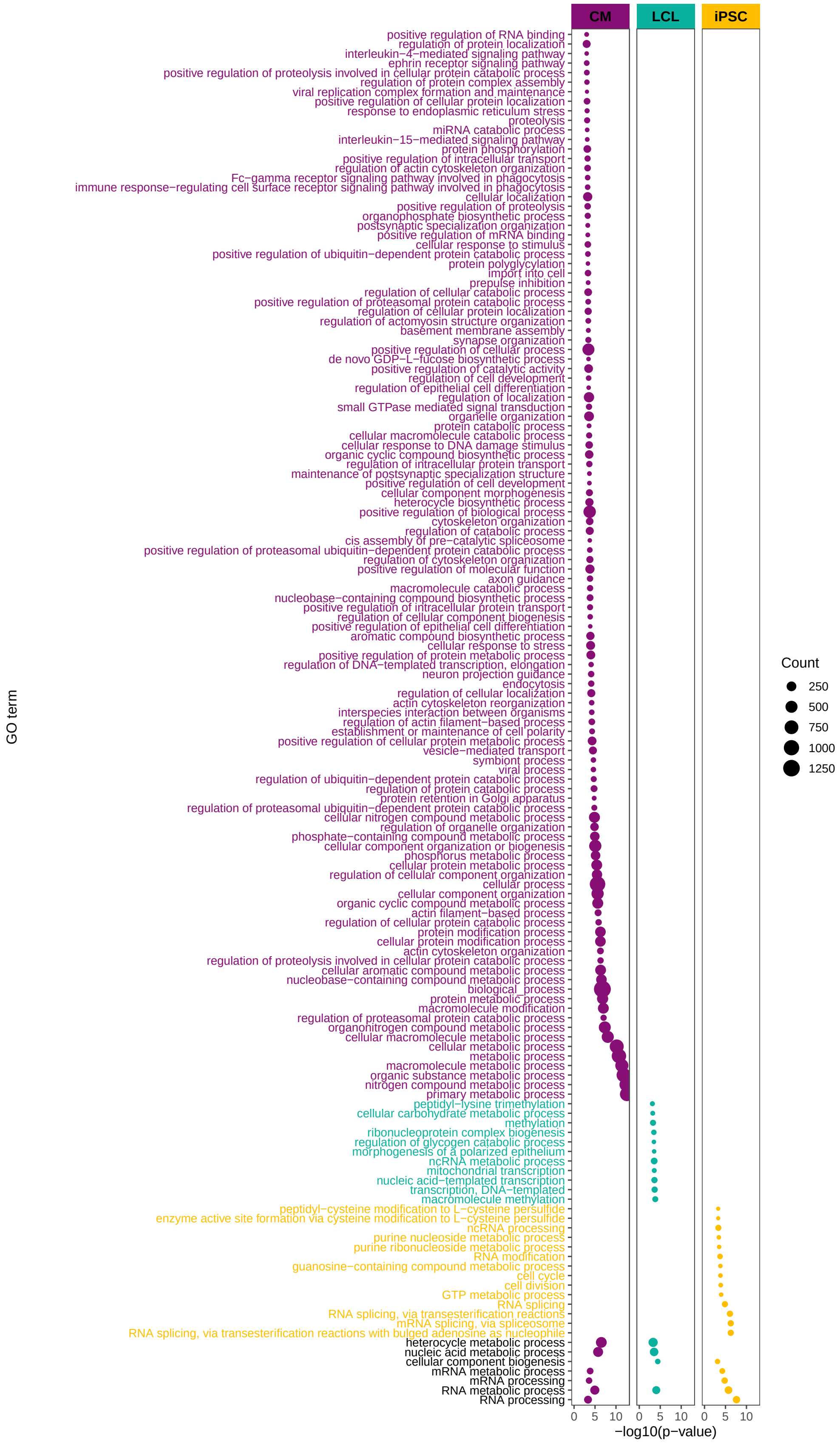

Figure S9: G**O terms enriched in genes with high DM scores of individual 18**, the individual with the highest TF disruption score for ERG. Point size corresponds to the the number of high-DM genes annotated with the corresponding GO term.

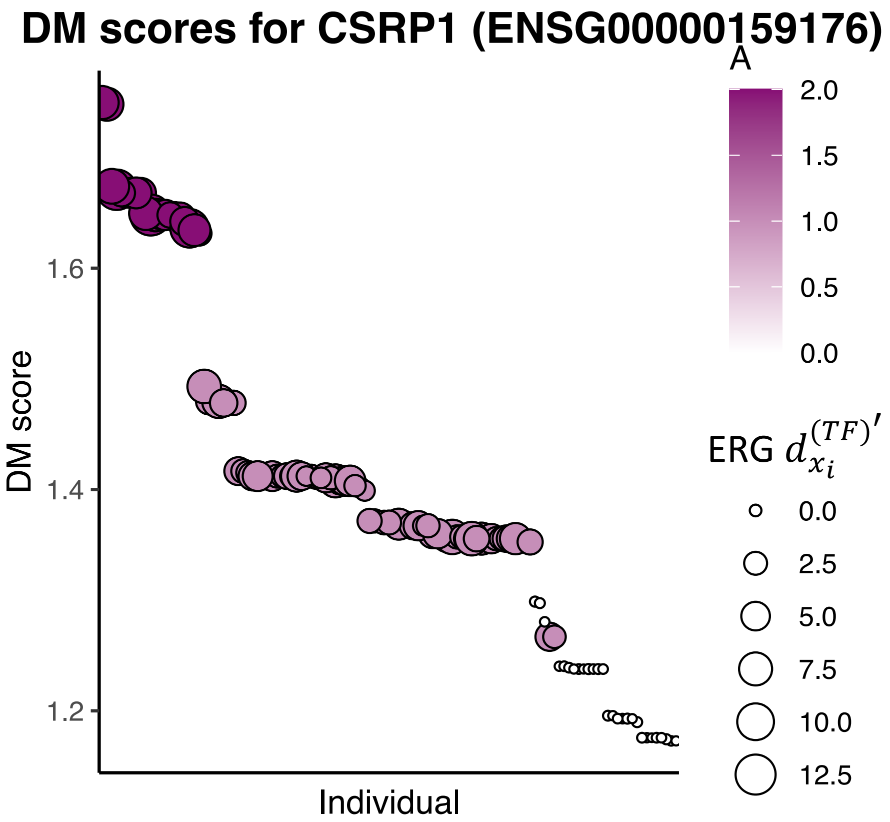

Figure S10: **DM scores of CSRP1 for 119 individuals in Yoruba population.** Point color intensity corresponds to the individual's alternate allele dosage A for the SNP within the ERG binding motif, and point size corresponds to the TF disruption $d_{x_i}^{(TF)'}$ score of ERG.

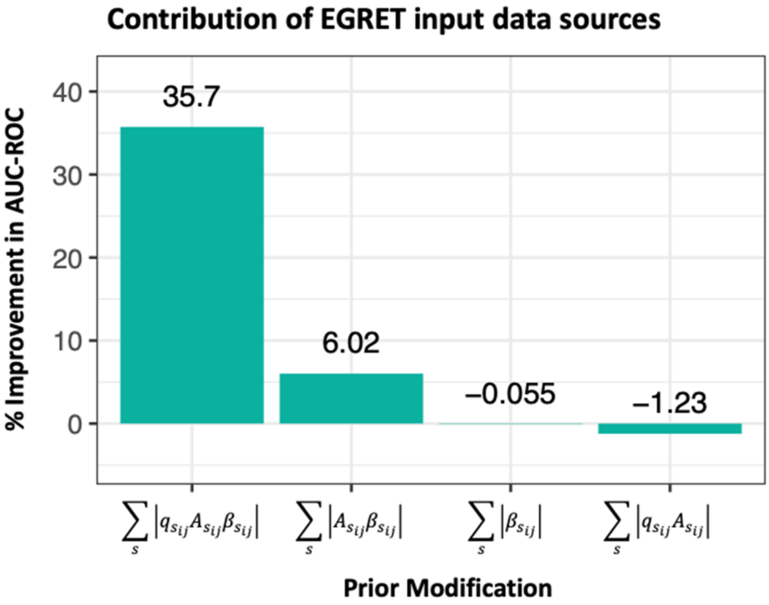

Figure S11: **Contribution of different data types to EGRET.** Percentage improvement in the prediction of ChIP-seq TF binding for EGRET in GM12878, compared to the ability of the baseline, genotype-agnostic network to predict ChIP-seq TF binding in GM12878. Each bar represents the AUC-ROC improvement when using a different combination of data types in the prior modification, for each SNP s with QBiC effect q, alternate allele count A and eQTL beta value β.

Supplementary Tables

Table S1: Improvement in AUC-ROC for the prediction of ChIP-seq binding in GM12878 when using EGRET edge weights, over using baseline genotype-agnostic edge-weights, for different cutoffs of $d_{x_{ij}}^{(E)}$. Total number of negatives (N), total number of positives (P), improvement in the AUC-ROC as well as the Delong p-value for the improvement are reported.

$d_{x_{ij}}^{(E)}$ cutoff	**N**	**P**	**AUC improvement**	**Delong p-value**
0.1	226	133	-0.05	0.96
0.15	132	81	-0.07	0.95
0.2	90	76	-0.01	0.55
0.25	72	75	0.08	0.07
0.3	70	72	0.09	0.05
0.35	**57**	**65**	**0.14**	**0.01**
0.4	57	64	0.13	0.02
0.45	57	64	0.13	0.02
0.5	57	64	0.13	0.02
0.55	57	62	0.11	0.04
0.6	57	62	0.11	0.04
0.65	57	61	0.10	0.05
0.7	57	61	0.10	0.05
0.75	57	58	0.07	0.12
0.8	57	58	0.07	0.12
0.85	57	58	0.07	0.12
0.9	57	57	0.06	0.16
1	56	56	0.06	0.16

Table S2: Improvement in AUC-ROC for the prediction of ChIP-seq binding in K562 when using EGRET edge weights, over using baseline genotype-agnostic edge-weights, for different cutoffs of $d_{x_{ij}}^{(E)}$. Total number of negatives (N), total number of positives (P), improvement in the AUC-ROC as well as the Delong p-value for the improvement are reported.

$d_{x_{ij}}^{(E)}$ **cutoff**	**N**	**P**	**AUC improvement**	**Delong p-value**
0.1	750	547	-0.01	0.88
0.15	408	283	-0.03	0.90
0.2	235	161	-0.05	0.93
0.25	149	127	-0.01	0.55
0.3	105	97	0.03	0.29
0.35	**75**	**78**	**0.11**	**0.03**
0.4	68	72	0.14	0.01
0.45	67	70	0.13	0.02
0.5	67	69	0.13	0.02
0.55	67	68	0.12	0.03
0.6	67	68	0.12	0.03
0.65	64	63	0.12	0.03
0.7	61	57	0.11	0.05
0.75	61	57	0.11	0.05
0.8	61	57	0.11	0.05
0.85	61	57	0.11	0.05
0.9	61	57	0.11	0.05
1	61	56	0.11	0.05

Table S3: GWAS catalog study references for CAD genes.

PMID	First author	Date	Journal	Study	Ref
21239051	Reilly MP	2011-01-14	Lancet	Identification of ADAMTS7 as a novel locus for coronary atherosclerosis and association of ABO with myocardial infarction in the presence of coronary atherosclerosis: two genome-wide association studies.	[37]
24262325	Dichgans M	2013-11-21	Stroke	Shared genetic susceptibility to ischemic stroke and coronary artery disease: a genome-wide analysis of common variants.	[38]
26343387	Nikpay M	2015-09-07	Nat Genet	A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease.	[39]
26708285	Wakil SM	2016-02-01	Atherosclerosis	A genome-wide association study reveals susceptibility loci for myocardial infarction/coronary artery disease in Saudi Arabs.	[40]
28714974	Klarin D	2017-07-17	Nat Genet	Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease.	[41]
29212778	van der Harst P	2017-12-06	Circ Res	Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease.	[42]
29263402	Han Y	2017-12-20	Sci Rep	Genome-wide association study identifies a missense variant at APOA5 for coronary artery disease in Multi-Ethnic Cohorts from Southeast Asia.	[43]
29472232	Li Y	2018-02-22	Arterioscler Thromb Vasc Biol	Genome-Wide Association and Functional Studies Identify SCML4 and THSD7A as Novel Susceptibility Genes for Coronary Artery Disease.	[44]
30104761	Zhou W	2018-08-13	Nat Genet	Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies.	[45]
30402224	Yamada Y	2018-09-17	Biomed Rep	Identification of 26 novel loci that confer susceptibility to early-onset coronary artery disease in a Japanese population.	[46]

Table S4: GWAS catalog study references for CD genes.

PMID	Journal	Study	Ref
17435756	Nat Genet	Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis.	[47]
17447842	PLoS Genet	Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4.	[48]
17554261	Nat Genet	Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility.	[49]
17554300	Nature	Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.	[50]
17684544	PLoS One	Systematic association mapping identifies NELL1 as a novel IBD disease gene.	[51]
17804789	Proc Natl Acad Sci U S A	Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci.	[52]
18587394	Nat Genet	Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.	[53]
20570966	Hum Mol Genet	Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease.	[54]
21102463	Nat Genet	Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.	[55]
22293688	Eur J Hum Genet	1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data.	[56]
22412388	PLoS Genet	A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci.	[57]
22936669	Gut	A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at RBX1-EP300.	[58]
23128233	Nature	Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.	[59]
23266558	Gastroenterology	A genome-wide association study identifies 2 susceptibility Loci for Crohn's disease in a Japanese population.	[60]
23850713	Gut	Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations.	[61]
25489960	Inflamm Bowel Dis	Immunochip analysis identification of 6 additional susceptibility loci for Crohn's disease in Koreans.	[62]
26192919	Nat Genet	Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.	[63]
26278503	Gastroenterology	Characterization of genetic loci that affect susceptibility to inflammatory bowel diseases in African Americans.	[64]
26891255	Inflamm Bowel Dis	HLA-C*01 is a Risk Factor for Crohn's Disease.	[65]
28008999	Sci Rep	Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the Polish population.	[66]
28067908	Nat Genet	Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease.	[67]
30500874	J Crohns Colitis	A genome-wide association study identifying RAP1A as a novel susceptibility gene for Crohn's disease in Japanese individuals.	[68]

Table S5: GO terms enriched in the genes with highest DM scores in individual 18 (genotype NA18523) in CMs. Enrichment performed using GORILLA [31]. N - total number of genes, B - total number of genes associated with a specific GO term, n - number of genes in the top of the user's input list, b - number of genes in the intersection

GO Term	Description	P-value FDR	q-value	Enrichment	N	B	n	b	Genes
See attached excel sheet									

Table S6: GO terms enriched in the genes with highest DM scores in individual 18 (genotype NA18523) in LCLs. Enrichment performed using GORILLA [31]. N - total number of genes, B - total number of genes associated with a specific GO term, n - number of genes in the top of the user's input list, b - number of genes in the intersection

GO Term	Description	P-value FDR	q-value	Enrichment	N	B	n	b	Genes
See attached excel sheet									

Table S7: GO terms enriched in the genes with highest DM scores in individual 18 (genotype NA18523) in iPSCs. Enrichment performed using GORILLA [31]. N - total number of genes, B - total number of genes associated with a specific GO term, n - number of genes in the top of the user's input list, b - number of genes in the intersection

GO Term	Description	P-value FDR	q-value	Enrichment	N	B	n	b	Genes
See attached excel sheet									