Representation learning based on proteomic profiles uncovers key cell types and biological processes contributing to the plasma proteome

Jiali Zhuang[1], Erin N. Smith[1,2] and Dorothée Diogo [1]

1 Takeda Development Center Americas, Inc., Cambridge, MA 02139

2 Current address: Genetic Epidemiology, Mirador Therapeutics, Inc., San Diego, CA 92121

Corresponding Author:

Jiali Zhuang

35 Lansdowne street, Cambridge, MA 02139, USA

jiali.zhuang@takeda.com

Abstract

The plasma proteome is a potential source of information on health status and physiological condition and holds great potential as candidate biomarkers for diagnosis, prognosis, intervention response monitoring, and patient stratification. As proteins in the plasma can be derived from numerous cellular and tissue sources, and their levels influenced by diverse mechanisms, a comprehensive assessment of patterns of protein variation could provide insight into mechanisms driving health and disease. By applying neural network-based representation learning and unsupervised clustering to the plasma proteomic profiles of 51,180 participants in the UK biobank, we identified 36 protein modules representing major cell types and biological processes present in the plasma proteome. We discovered that the overall abundances of proteins belonging to certain modules are associated with disease status and genetic variants. Those associations reflect complex and multi-faceted mechanisms that regulate protein abundances in circulation. An investigation into the protein modules associated with disease variants uncovered both known disease biology and novel findings that may translate into testable hypotheses. Our approach generates biologically relevant groupings of plasma proteins that can be deployed to inform the design of more predictive biomarker panels and shed new light on the effects of disease-associated genetic variants.

Introduction

Plasma protein concentrations have long been used as biomarkers for diagnosis and patient monitoring. Recent technological advances now permit accurate measurement of thousands of proteins simultaneously and offer a unique opportunity to comprehensively study the plasma proteome. UK Biobank Pharma Proteomics Project (PPP) employed the Olink proximity extension assay to profile the plasma proteomes of 54,219 UKB participants and identified over 12,000 significant associations between genetic variants and plasma protein levels [1]. The extensive genetic, phenotypic and clinical information available for the same participants allows for the discovery of genetic factors affecting plasma protein abundances as well as biomarkers for diagnosis and disease progression monitoring [2-4]. The relative contribution of a given cell type or tissue to the plasma proteome depends on multiple factors including cell counts, secretory activity level, metabolic rate, vasculature density and local endothelial permeability. The identification of the cell types and biological processes well represented in the circulating proteome informs the selection of disease areas that are most amenable to non-invasive monitoring using blood protein biomarkers. Ascertaining the tissue/cell type origin of a plasma protein based solely on expression enrichment can be challenging, considering the limited number of highly cell type specific proteins and the wide range of relative contribution levels across cell types. Proteins that originate from the same cell type or participate in the same biological process are more likely to share similar expression patterns across participants in the cohort. A data driven approach that starts with identifying co-expressed protein modules could therefore better unravel the cell types and biological processes represented in the plasma proteome and assign individual proteins to their source. Linear decomposition algorithms such as principal component analysis (PCA) and non-negative matrix factorization (NMF) have been applied to high-dimensional omics data for deconvolution purposes [5, 6]. However, the underlying biological factors do not always follow linear relationships. A variational autoencoder (VAE) model includes an encoder and a decoder and learns a lower dimensional representation

of the input data while minimizing information loss incurred by the dimension reduction. The encoder and decoder of VAEs are implemented with neural network architecture and are therefore able to model non-linear relationships. Here we present a computational framework that learns a low dimensional representation of proteins based on the plasma proteomic profiles of 51,180 UK Biobank participants with VAE and derives biologically relevant protein modules from unsupervised clustering in the latent space. We showed that most of those modules represent distinct cell types or biological processes and discovered associations with disease status and genetic variants. We demonstrated that dysregulation of protein modules reflects known biology of diseases and could provide additional insight into key aspects of complex diseases.

Results

Dimension reduction followed by unsupervised clustering reveals protein modules that represent key cell types and biological processes in the plasma proteome

We identified major cell types and biological processes represented in the plasma proteome using a computational framework that consists of two sequential steps: dimension reduction and clustering (**Figure 1**). In this framework, a vector of 51,180 elements is assigned to each protein, representing the normalized expression levels of the plasma protein measured in 51,180 participants. To derive an unentangled representation that captures the main sources of variation of protein abundances in the cohort, we built a variational autoencoder (VAE) that projected the protein vector into a lower dimensional latent space while minimizing the reconstruction loss relative to the original vector (for details see Methods). Next, we performed unsupervised clustering of proteins with Leiden algorithm [7] based on their proximity within the derived latent space. Proteins that are close to each other in the latent space embedding share similar expression profiles and are therefore more likely to originate from the same cell type/tissue or participate in the same biological process. The advantage of carrying out clustering on the latent space instead of the original input space is twofold: 1) recently developed non-linear clustering algorithms such as Leiden and Louvain are computationally expensive, therefore reducing the number of data dimension alleviates the computation burden; and 2) it mitigates the biases introduced by the presence of extensive correlations among the original features as VAE is designed to minimize correlation among the derived latent dimensions. With this computational framework, we assigned each plasma protein to one of the 36 distinct protein modules (**Supplementary Table S1**).

To test whether proteins in those modules are enriched in any cell type or biological process, we leveraged information from three orthogonal sources (for detail see Methods): 1) transcriptomic profiles of 54 major human tissues from GTEx and 116 cell types from Tabular Sapiens consortium [8, 9] (**Supplementary Figure S1 & S2**); 2) Biological processes and pathways annotated in Gene Ontology, KEGG and REACTOME (**Supplementary Table S2**); and 3) Blood test results including major blood cell counts and percentages generated in the same UK Biobank participants (**Supplementary Figure S3**). Information regarding preferential expression in tissues or cell types, enrichment in annotated biological processes and pathways, and correlation with blood cell/metabolite measurements were comprehensively evaluated before we manually assigned the protein modules to the cell type or biological process that they most likely represent. With this approach, we were able to annotate 32 of the 36 modules (**Figure 2a**).

Protein modules dysregulated in disease

We examined how those protein module levels vary by disease conditions. We used 3-digit ICD-10 codes in the UKB clinical record to determine the participants' disease status and identified plasma proteins that are associated with diseases by fitting a logistic regression model (Methods). For each module – disease pair, we tested if the proteins up- or down-regulated in the disease significantly overlap with those belonging to the module (for details see Methods). We uncovered many associations that match the tissue or known biology of the diseases (**Supplementary Table S3**). For instance, proteins up regulated in lung diseases including emphysema, interstitial pulmonary disease and chronic obstructive pulmonary disease are significantly enriched in the lung module (**Figure 2b**). Similarly, among the top diseases whose up-regulated proteins are overrepresented in the muscle module are spinal muscular atrophy, myopathies and primary disorders of muscles, possibly reflecting muscle tissue damage accompanying disease process (**Figure 2c**). Lichen planus and alopecia areata, both of which are autoimmune diseases of the skin, are associated with down regulation of keratinocyte module proteins (**Supplementary Table S3**), consistent with a loss of skin or follicle cells under immune system attacks.

Discovery of common genetic variants associated with changes in protein module levels

The UKB PPP consortium reported a comprehensive protein quantitative trait loci (pQTL) mapping that included 2,103 pleiotropic trans-pQTLs [1], where a locus is associated with the plasma abundances of multiple distal proteins. We hypothesized that one of the possible mechanisms underlying pleiotropic trans-pQTLs is when a genetic variant alters the overall contribution of a tissue/cell type or biological process, therefore affecting multiple proteins originating from the cell type or belonging to the biological process. We tested this hypothesis by testing whether proteins in the VAE-defined protein modules were systematically associated with selected genetic variants. In our analysis, we included 116,902 common genetic variants that are either significant quantitative trait loci (eQTL and pQTL) or coding variants with moderate to high impact predicted by Ensembl Variant Effect Predictor (VEP) [10] (for details see Methods). For each variant of interest, we collected the summary statistics from associations with 2,935 assays encompassing 2,917 unique plasma proteins and then for each protein module tested if the normalized effect sizes in the module significantly deviate from zero by fitting a Bayesian Regression model (for details see Methods). We identified 14,932 significant variant-module associations with false discovery rate < 1e-30 and estimated effect size > 1 (**Supplementary Table S4 & Figure S4**). Platelet and erythrocyte modules were associated with the largest number of loci, suggesting that the plasma proteome is highly sensitive to changes in these abundant blood cell types. One mechanism by which a locus associates with a protein module is via association with the cell counts. For instance, intronic SNP rs3917932 (1_36478315_C_G) is a splicing quantitative trait locus (sQTL) for the *CSF3R* gene, which encodes a receptor for granulocyte colony-stimulating factor (encoded by the *CSF3* gene). It significantly associates with elevated level of CSF3R and decreased level of CSF3 proteins in plasma, as well as reduced abundance of neutrophil module proteins and lower neutrophil counts in the blood (**Figure 3b & c**). A plausible explanation for the observed associations is that instead of localized to the plasma membrane, CSF3R proteins encoded by

the splicing isoform are predominantly released into circulation, bind to and sequester CSF hormones in the blood, thus leading to suppressed neutrophil maturation, low neutrophil counts and low concentrations of neutrophil proteins. Another possible mechanism underlying the association between a locus and the abundances of a protein module is through dysregulation of protein secretion. SNP rs10745925 (12_101825121_T_C) is an eQTL for the *GNPTAB* gene, which encodes subunits of the N-acetylglucosamine-1-phosphotransferase that catalyzes the formation of mannose 6-phosphate (M6P) markers on high mannose type oligosaccharides in the Golgi apparatus [11]. M6P modification is essential for M6P-receptor-mediated transport of lysosomal enzymes into lysosomes, therefore explaining the association between the genetic variant and the dramatic change in lysosomal enzyme abundances in plasma (**Figure 3d**). Gene-level collapsing analysis based on rare coding variants had previously revealed significant associations between loss of *GNPTAB* gene function and elevated lysosomal enzyme concentrations in plasma [12]. It is consistent with our observation that rs10745925, an eQTL associated with increased expression of *GNPTAB*, is also associated with decreased plasma abundances of lysosomal enzymes.

Protein modules help interpret effects of genetic variants associated with complex diseases

Plasma protein modules representing key cell types and biological processes may also be employed to analyze and interpret disease-associated genetic variants and to potentially shed new light on disease mechanisms. We searched published GWAS results hosted on Open Targets Genetics [13] and discovered that 354 of the identified module-modifying loci were associated with diseases (168 unique diseases) across multiple organ systems (for details, see Methods; **Figure 4a**). Examining the protein modules (and the cell types or pathways they represent) dysregulated by the disease-associated variants offers a unique opportunity to better understand their downstream effects. For instance, SNP rs738409, a missense variant in the *PNPLA3* gene that encodes a triacylglycerol lipase and regulates liver fat metabolism, has been found to strongly associate with liver cirrhosis and non-alcoholic fatty liver disease [14]. Consistent with previously published findings [15], the alternative allele of rs738409 (G) was associated with increased abundances of liver enzymes in our analysis, possibly reflecting hepatocyte damage incurred by the accumulation of lipid droplets (**Figure 4b**). Additionally, proteins in the lysosome, adaptive immunity, extracellular matrix and inflammation modules were also positively associated with rs738409. On the other hand, liver secretory proteins abundances were negatively associated with rs738409. These additional associations represent either secondary effects downstream of the primary effect (in this case the accumulation of lipid droplets in hepatocytes) or novel, independent events occurring in other cell types/tissues. SNP rs7256200 is a known risk variant for late onset Alzheimer's Disease (AD) located near the *APOC1* and *APOE* genes [16]. It is strongly associated with decreased levels of lysosomal enzymes in addition to lower APOE protein level in plasma (**Figure 4c**). Interestingly, significantly higher plasma abundance of myeloid and erythroid nuclear termination stage-specific protein (MENT), a protease inhibitor that inhibits several lysosomal enzymes including cathepsin K, cathepsin L, and cathepsin V [17], is also associated with the SNP. Dysregulation of lysosomal function and related autophagy process has been thought to play a key role in AD and age-related dementia [18]. Although by itself not a proof of causal relationship, the fact that an AD risk variant is simultaneously associated with lower plasma abundance of lysosomal

proteins lends more weight to the hypothesis that lysosomal dysfunction could be one of the potential causes of AD and dementia.

Many complex diseases are associated with multiple genetic loci, which may constitute different components of the same pathway or reflect the presence of multiple molecular mechanisms underlying disease risks, possibly with distinct etiologies or progression trajectories. Disease variants' impacts on the plasma proteome may help reveal potential disease subtypes. Type 1 diabetes is an autoimmune disease in which the immune system attacks the secretory cells in pancreas and affects insulin secretion. We found 34 module-modifying loci that were also associated with type 1 diabetes and by comparing their impacts on the protein modules we identified several variant clusters (**Figure 5a**). We also examined other diseases associated with these loci, hoping that additional phenotypic data may aid the interpretation of derived variant clusters. We found 6 loci significantly associated with pancreatic enzyme levels (labels colored in blue) and none of them is associated with additional autoimmune diseases besides type 1 diabetes. Whereas for loci that do not affect pancreas protein abundances, 60.7% (17 out of 28) of them are associated with additional autoimmune diseases besides type 1 diabetes (**Figure 5a**). The depletion (p-value = 0.024, chi-square test) of associations with additional autoimmune diseases for the pancreatic protein perturbing loci suggests that their association with type 1 diabetes may be mediated by a dysregulation in the pancreas that engenders autoimmune attacks against it. On the other hand, other variants likely act through pancreas-independent mechanisms that modulate the risk for not only type 1 diabetes but potentially other autoimmune diseases as well. Moreover, we found 14 genetic loci associated with dysregulation of proteins in the lipid metabolism module (**Figure 5b**). We observed that most (12 out of 14) of those loci were significantly associated with cardiovascular and metabolic diseases, which is consistent with the role that dysregulated lipid metabolism and obesity play in cardiovascular diseases and metabolic diseases such as type 2 diabetes [19, 20]. The direction of the association between a module-modifying variant and plasma protein levels may also be predictive of the direction of the variant's effect on disease risks. We analyzed 175 genetic associations from a GWAS on coronary artery disease (CAD) [21] and found 21 loci are significantly associated with an extracellular matrix organization module (FDR < 1e-20). We noticed that variants associated with decreased levels of ECM organization module proteins are predominantly (9 out of 10) risk variants for CAD while most of the variants associated with increased levels of ECM organization module (10 out of 11) are protective against CAD (**Figure 5c**, p-value = 0.00076, chi-square test).

Discussion

In this report, we applied a data-driven approach to deconvolute 2,935 plasma proteins into modules that likely represent major cell types and biological pathways present in the plasma proteome. The protein modules we found cover a wide range of cell types and tissues, supporting the notion that plasma proteome could become a useful tool for monitoring the health status of multiple organ systems and tissues. It comes as no surprise that several protein modules are assigned to blood cell types including platelets, erythrocytes, neutrophils and immune cells, given their direct access to the circulation. The platelet module was the largest module encompassing more than a fifth of all proteins in the Olink Explore 3072 panel. It is worth noting that many of the proteins belonging to the platelet module are not platelet specific

and some of them even have enriched expression in other tissues/cell types (**Supplementary Figure S1 & S2**). They were assigned to the platelet module because they share similar expression profiles with platelet-specific proteins such as Glycoprotein V Platelet (GP5) and Glycoprotein VI Platelet (GP6) and their abundance correlates with platelet counts. A likely explanation is that although those proteins are only moderately expressed in platelets, the proportion of the platelet contribution to the plasma proteome is so high that it constitutes a major source for those proteins in blood. This illustrates the potential pitfalls of relying solely on the expression levels measured in independent tissues/cell types and highlights the merit of taking a more data-driven approach to assign tissue/cell type origin for plasma proteins. A more reliable inference of tissue/cell type origin for plasma proteins will better inform panel design and protein selection for future biomarkers. With more proteins being added to the high-throughput assay panels, our approach may reveal more cell types or biological processes reflected in the plasma proteome in future studies.

Modulation of gene expression at the transcription level is merely one of the many possible mechanisms by which a genetic variant may alter the abundance of a protein in circulation. The discovery of numerous trans-pQTL loci from the original UKB-PPP report indicates extensive downstream effects beyond transcriptional modulation of the nearby *cis* gene that may include changes in cell counts, protein subcellular location (e.g., intracellular/membrane-bound versus secreted), protein leakage due to cell necrosis or apoptosis, and effects on transcription factor target gene expression. Those secondary effects often impact multiple proteins with relatively moderate effect sizes. Conventional pQTL analysis treats each protein as independent and adopts stringent *p*-value threshold to account for potential false discoveries due to multiple testing. The identification of protein modules allows us to consider a group of functionally related proteins as a whole and increases the statistical power of detecting associations with smaller effect sizes that would have been missed in conventional single protein pQTL analysis. For instance, rs1126313 (1_39174072_C_A) is significantly associated with decreased lipid metabolism module level (False Discovery Rate = 1.72e-39) while none of the individual proteins belonging to this module passed the significance threshold (nominal p-value < 1.7e-11) used in the pQTL analysis [1]. We highlighted examples of variant – protein module associations with known mechanisms and reported novel ones discovered in our analysis. Our findings suggested that genetic factors play a role in determining the relative contributions of many cell types and biological processes to the plasma proteome, and it is therefore crucial to incorporate genotypic information as a covariate in the training and classification of plasma protein-based models to adjust for variations in baseline levels across genetically diverse cohorts.

Genome-wide association studies have uncovered numerous genetic loci that are associated with predisposition to complex diseases. However, it is extremely challenging to unravel the molecular mechanisms by which genetic variation modulates the risk for a disease. Changes in plasma protein abundances associated with a disease variant may offer additional insight into the cell types and/or biological processes affected by the variant and lead to new hypothesis about the disease mechanisms. Data-driven grouping of plasma proteins into functional modules enables intuitive interpretation of disease variant effects and facilitates the identification of key cell types and biological processes for a given disease. We demonstrated examples where we use this approach to pinpoint the known biological process crucial to certain diseases (lipid metabolism in cardiovascular diseases and type 2 diabetes) and to reveal multiple potential etiologies (pancreas specific versus general autoimmune activity in the case of

type 1 diabetes). We note that not all proteomic changes associated with a disease variant are necessarily relevant to the disease and a holistic evaluation of additional evidence and known biology of the disease is essential. Nonetheless, the ability to link a disease variant to alterations in certain cell types and biological processes will provide valuable clues for a better understanding of disease mechanisms and empower target discovery for future therapeutics.

Methods

UKB data

UK Biobank is a longitudinal cohort study of >500,000 participants aged 40-69 years enrolled in study centers across England, Wales and Scotland between 2006 and 2019 [22]. In addition to extensive baseline data about medical histories, physical measurements and health behavior of participants, biosamples were collected from the participants (e.g., saliva, urine and blood). Ethical approval was granted for the UK Biobank by the North West Multi-Centre Research Ethics Committee and the National Health Service (NHS) National Research Ethics Service (ref: 11/NW/0382), all participants provided written informed consent, and all experiments were conducted in accordance to relevant regulations and guidelines. Additional information, including details on study protocol, are available online at https://biobank.ctsu.ox.ac.uk/. For a subset of 54,219 UKB participants, proteomic measurements were conducted by the UKB PPP consortium as described elsewhere [1]. Plasma proteomics data, genetic data and health data from the UKB participants were accessed under UK Biobank applications 65851 and 26041.

Representation learning with variational autoencoder and unsupervised clustering of proteins

Inverse rank normalized protein abundance measurements were used for representation learning. Proteins with missing values in >= 10% of the subjects were excluded, with 2,916 proteins (2,935 assays) retained for subsequent analysis. Subjects with missing values in >= 20% of the remaining proteins were excluded, with 51,180 subjects retained in the subsequent analysis. Both the encoder and decoder in the VAE are multilayer perceptron (MLP) with 2 hidden layers. The latent space has 10 dimensions. The layer sizes of the encoder network were 51,180 (input layer) – 512 (hidden layer) – 64 (hidden layer) – 20 (output layer, 10 for mean and 10 for variance). The layer sizes of the decoder were 10 (input layer) – 64 (hidden layer) – 512 (hidden layer) – 51,180 (output layer). Since the inverse rank normalized values approximately followed normal distribution, we defined the reconstruction loss as the difference of cumulative distribution function between the input and decoder output. The regularization term added to the loss function is the *KL* divergence as in standard VAE. The model was trained in 150 epochs, with batch size of 32 and learning rate 0.0001. The training of the VAE model was carried out with the PyTorch library (version 1.13.1). Leiden algorithm implemented in the Scanpy (version 1.10.1) package was used to perform protein clustering based on their coordinates in the latent embedding, with resolution parameter being set to 2.5. For each of the 45 initial modules identified, we calculated the coefficient on the 1st principal component as an estimate of the average level of that module across the cohort and computed pairwise correlations among the 45 modules. We then merged modules that are highly correlated across the subjects and reduced the number of modules to 36.

Annotation of protein modules and inference of tissue/cell type origin

For each of the 36 protein modules, we tested expression enrichment for the transcripts encoding member proteins across human tissues (GTEx RNA-seq) [8] and cell types (Tabular Sapiens scRNA-seq) [9] (**Supplementary Figure S1 & S2**). We also performed pathway enrichment analysis against annotated pathways/biological processes in Gene Ontology, KEGG and Reactome databases for each protein module using the enrichR package [23] (**Supplementary Table S2**). Furthermore, we correlated the 1st principal component coefficients of each protein module with demographic information and blood assay measurements across the cohort (**Supplementary Figure S3**) using the Pearson correlation function (scipy.stats.pearsonr) in the SciPy library (version 1.7.3). The cell type/biological process assignment was made after manual review of the results derived from analyses described above.

Assessing protein modules dysregulation in diseases

Case-control logistic regression analysis was performed on first occurrences data (Category 1712, which included hospital inpatient data, self-reports, primary care data and death register records) defined by UK Biobank for all 3-digit ICD-10 codes. Covariates included in the regression are: time between plasma sample and assay (tbms_covar) + age + sex + age_sex + age2 + age2_sex + factor(Proteomics Batch) + factor(ukb_centre_fct) + factor(array_fct) + 20 genetic principal components. We defined up- and down-regulated proteins for each ICD-10 code with a loose threshold of false discovery rate (FDR) < 0.1. Next, for each disease (3-digit ICD-10 code)-module pair we tested if the proteins up-/down-regulated in the disease are significantly enriched in the module with hypergeometric test, which was implemented using “scipy.stats.hypergeom.cdf” function in the Scipy library (version 1.7.3). False discovery rate (“Benjamini-hochberg” method) based adjustment for multiple testing correction was carried out using the “statsmodels.stats.multitest” function in the statsmodels library (version 0.13.5). Significant associations with FDR < 1e-5 and number of overlapped proteins >= 25 and fraction of overlapped proteins > 0.05 are included in Supplementary Table S3.

Identify associations between selected genomic loci and protein module levels

Full summary statistics from genome-wide pQTL analysis generated by the UKB PPP consortium was downloaded from Synapse (https://doi.org/10.7303/syn51364943). We selected genetic variants from 3 sources for the subsequent analysis: 1) exon variants reported in gnomAD v4.1 with minor allele frequency >= 1% in at least one ancestry group and predicted impact >= MODERATE by the Variant Effect Predictor (VEP) [10, 24]; 2) significant eQTL loci (nominal *p*-value < 1e-15) from eQTL Catalogue [25] (with linkage disequilibrium clumping for variants within 500kb window and r^2 >= 0.8); 3) significant pQTL loci reported by the UKB PPP consortium [1]. The combination of the above three sources led to a total of 116,902 variants with summary statistics. For each variant, we calculated the normalized effect size on each protein by dividing the estimated coefficient (β) by the standard error. To minimize the impact of outliers, normalized effect sizes that are more than 5 standard deviations (s.d.) away from the mean are fixed at mean ± 5 s.d. We then fitted a Bayesian linear regression model with a centered elliptic Gaussian distribution as the prior on the normalized effect sizes of proteins belonging to a protein module. It is also known as Automatic Relevance Determination (ARD) model which automatically prunes features that are estimated to have no effects and hence leads to sparser solutions. The fitting was carried out using the “sklearn.linear_model.ARDRegression” function in the scikit-learn library (version 1.0.2). The maximum a posteriori (MAP) values of the coefficients were taken as the estimates for the effect

sizes of the modules. P-values were calculated from the posterior (normal) distributions of the coefficients. False discovery rate ("Benjamini-hochberg" method) based adjustment for multiple testing was carried out using the "statsmodels.stats.multitest" function. LD clumping was performed within 500kb windows to retain only the most significant variant to represent the locus. Significant associations with FDR < 1e-30 and absolute MAP > 1 are included in Supplementary Table S4.

Disease associations for module-modifying variants

Disease associations for module-modifying variants were obtained from Open Targets Genetics platform via the "pheWAS" API endpoint. Only associations with *p*-value < 5e-8 were retained in subsequent analysis. Diseases belonging to the "cell proliferation disorder" category were not shown in Figure 5b & c. The summary statistics for the coronary artery disease study (GCST010866) were downloaded from the GWAS Catalog.

Acknowledgements

This research has been conducted using the UK Biobank Resource under application numbers 65851 and 26041. We thank the UK Biobank team and participants for making this valuable resource possible. We thank the research and development teams at the 13 participating UKB-PPP companies (Alnylam Pharmaceuticals, Amgen, AstraZeneca, Biogen, Calico, Bristol-Myers Squibb, Genentech, GlaxoSmithKlein (GSK), Janssen Pharmaceuticals, Novo Nordisk, Pfizer, Regeneron and Takeda) for funding the proteomics measurements.

Data Availability

Data from UK biobank cannot be shared openly and are subject to UK Biobank ethical approval. Further information about applying for UK Biobank data access can be obtained from the UK Biobank website (https://www.ukbiobank.ac.uk) or by emailing UK Biobank (ukbiobank@ukbiobank.ac.uk).

Author contributions

J.Z. conceived and designed the analysis. J.Z. and E.N.S. conducted the analysis. J.Z., E.N.S. and D.D wrote the manuscript.

Funding

Takeda Development Center Americas, Inc. provided funding. They had no role in study design, data analysis or preparation of the manuscript.

Declaration of interests

J.Z. and D.D. are employees of Takeda Development Center Americas, Inc. At the time of the study, E.N.S was an employee of Takeda Development Center Americas, Inc. J.Z., E.N.S and D.D. are stockholders of Takeda Pharmaceuticals Company Limited. All authors declare no non-financial competing interests.

Figure legend

Figure 1.

A schematic representation of the workflow that derives and annotates biologically relevant protein modules based on Olink proteomics measurements from the UKB cohort.

Figure 2.

a) Number of member proteins in each of the identified protein modules. b) & c) Diseases in which up-regulated plasma proteins are enriched in the lung (b) and muscle (c) modules. Shades of green represents the logarithm of false discovery rate. Sizes of the dots are proportional to the fraction of up-regulated proteins that belong to the module.

Figure 3.

a) Number of loci significantly associated with the abundance of each of the protein modules. Only modules associated with at least one locus are shown. b) Distribution of normalized effect sizes of genetic variant rs3917932 (1_36478315_C_G) across protein modules. c) Forest plot showing phenotypes associated with variant rs3917932 (1_36478315_C_G). Sizes of the boxes are proportional to the study cohort sizes. Numbers in the parathesis on y-axis labels are *p*-values. All associations with *p*-value < 1e-30 are shown. d) Distribution of normalized effect sizes of genetic variant rs7964859 (12_101825121_T_C) across protein modules.

Figure 4.

a) Number of module-modifying loci associated with trait categories. b) & c) Distribution of estimated effect sizes (MAP) of genetic variants rs738409 (22_43928847_C_G) (b) and rs7256200 (19_44912678_G_T) (c) on inferred protein modules. Shades of red are proportional to the logarithm of false discovery rate. Error bars represent posterior standard deviation.

Figure 5.

a) The heatmap on the left depicts the significance of associations between variants (rows) and protein modules (columns) for module-modifying variants that are also associated with type 1 diabetes. Shades of red are proportional to the logarithm of false discovery rate of the association. Variants are clustered using hierarchical clustering. Variants with blue labels belong to a cluster that show strong association with the pancreas module. The heatmap on the right depicts association between the same variants (in the same order as the left heatmap) and diseases. Green grids indicate the presence of significant association while white grids indicate the absence of association. b) Same as a) except for variants that are associated with changes in the lipid metabolism module. c) Fraction of variants that confer protection versus risk for coronary artery disease grouped by the direction of their effects on the extracellular matrix organization module.

1. Sun, B.B., et al., *Plasma proteomic associations with genetics and health in the UK Biobank.* Nature, 2023. **622**(7982): p. 329-338.
2. Carrasco-Zanini, J., et al., *Proteomic signatures improve risk prediction for common and rare diseases.* Nat Med, 2024. **30**(9): p. 2489-2498.
3. Gadd, D.A., et al., *Blood protein assessment of leading incident diseases and mortality in the UK Biobank.* Nat Aging, 2024. **4**(7): p. 939-948.
4. Garg, M., et al., *Disease prediction with multi-omics and biomarkers empowers case-control genetic discoveries in the UK Biobank.* Nat Genet, 2024. **56**(9): p. 1821-1831.
5. Nyamundanda, G., L. Brennan, and I.C. Gormley, *Probabilistic principal component analysis for metabolomic data.* Bmc Bioinformatics, 2010. **11**.
6. Toden, S., et al., *Noninvasive characterization of Alzheimer's disease by circulating, cell-free messenger RNA next-generation sequencing.* Sci Adv, 2020. **6**(50).
7. Traag, V.A., L. Waltman, and N.J. van Eck, *From Louvain to Leiden: guaranteeing well-connected communities.* Sci Rep, 2019. **9**(1): p. 5233.
8. Consortium, G.T., *The GTEx Consortium atlas of genetic regulatory effects across human tissues.* Science, 2020. **369**(6509): p. 1318-1330.
9. Tabula Sapiens, C., et al., *The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans.* Science, 2022. **376**(6594): p. eabl4896.
10. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.
11. Ludwig, N.F., et al., *GNPTAB missense mutations cause loss of GlcNAc-1-phosphotransferase activity in mucolipidosis type II through distinct mechanisms.* Int J Biochem Cell Biol, 2017. **92**: p. 90-94.
12. Dhindsa, R.S., et al., *Rare variant associations with plasma protein levels in the UK Biobank.* Nature, 2023. **622**(7982): p. 339-347.
13. Ghoussaini, M., et al., *Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics.* Nucleic Acids Res, 2021. **49**(D1): p. D1311-D1320.
14. Trepo, E., et al., *PNPLA3 gene in liver diseases.* J Hepatol, 2016. **65**(2): p. 399-412.
15. Barton, A.R., et al., *Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses.* Nat Genet, 2021. **53**(8): p. 1260-1269.
16. Eissman, J.M., et al., *Sex-specific genetic architecture of late-life memory performance.* Alzheimers Dement, 2024. **20**(2): p. 1250-1267.
17. Irving, J.A., et al., *Inhibitory activity of a heterochromatin-associated serpin (MENT) against papain-like cysteine proteinases affects chromatin structure and blocks cell proliferation.* J Biol Chem, 2002. **277**(15): p. 13192-201.
18. Zhang, W., et al., *Impairment of the autophagy-lysosomal pathway in Alzheimer's diseases: Pathogenic mechanisms and therapeutic potential.* Acta Pharm Sin B, 2022. **12**(3): p. 1019-1040.
19. Klein, S., et al., *Why does obesity cause diabetes?* Cell Metab, 2022. **34**(1): p. 11-20.
20. Powell-Wiley, T.M., et al., *Obesity and Cardiovascular Disease: A Scientific Statement From the American Heart Association.* Circulation, 2021. **143**(21): p. e984-e1010.
21. Koyama, S., et al., *Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease.* Nat Genet, 2020. **52**(11): p. 1169-1177.
22. Bycroft, C., et al., *The UK Biobank resource with deep phenotyping and genomic data.* Nature, 2018. **562**(7726): p. 203-209.

23. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
24. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
25. Kerimov, N., et al., *A compendium of uniformly processed human gene expression and splicing quantitative trait loci.* Nat Genet, 2021. **53**(9): p. 1290-1299.

Figure 1

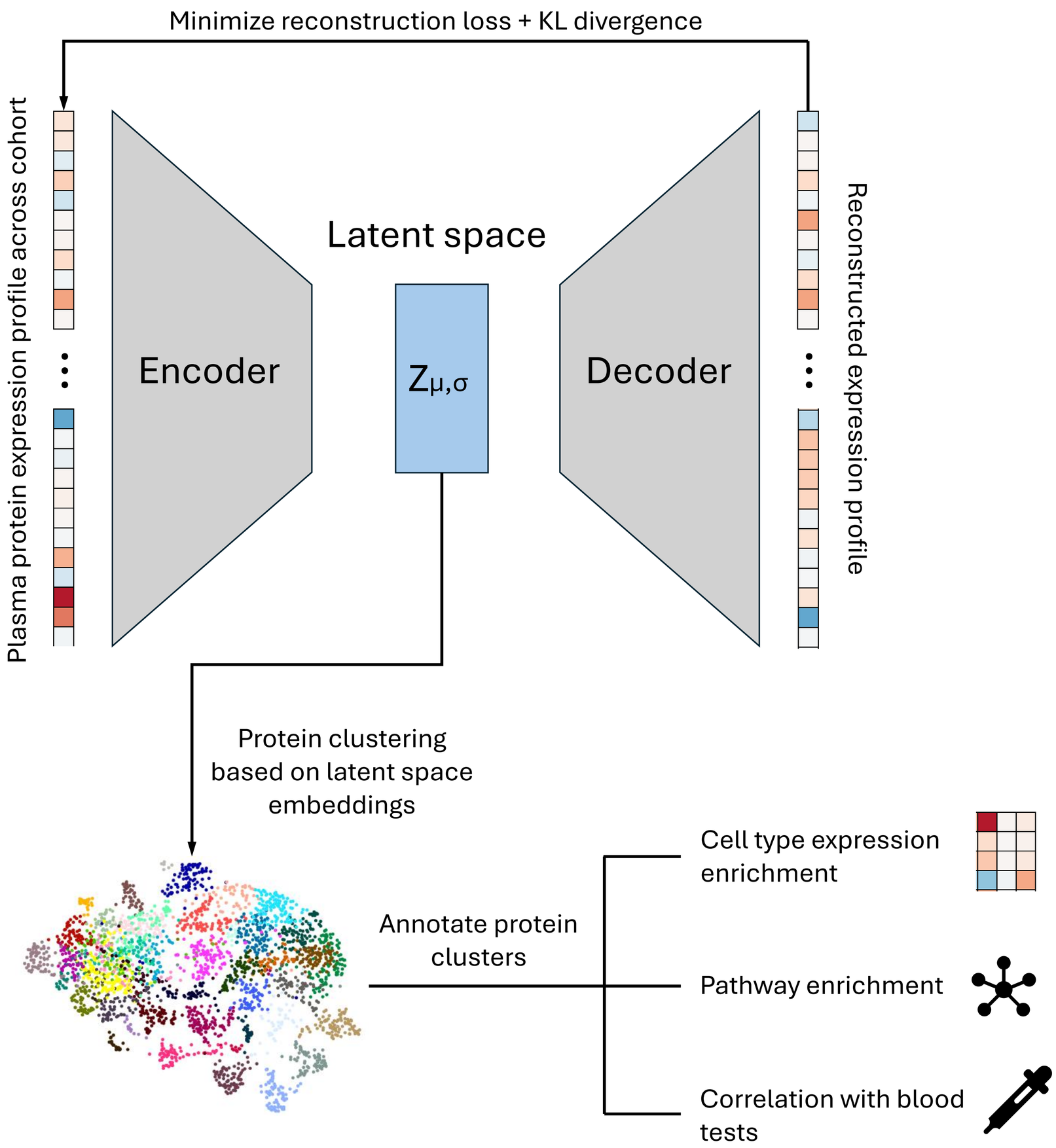
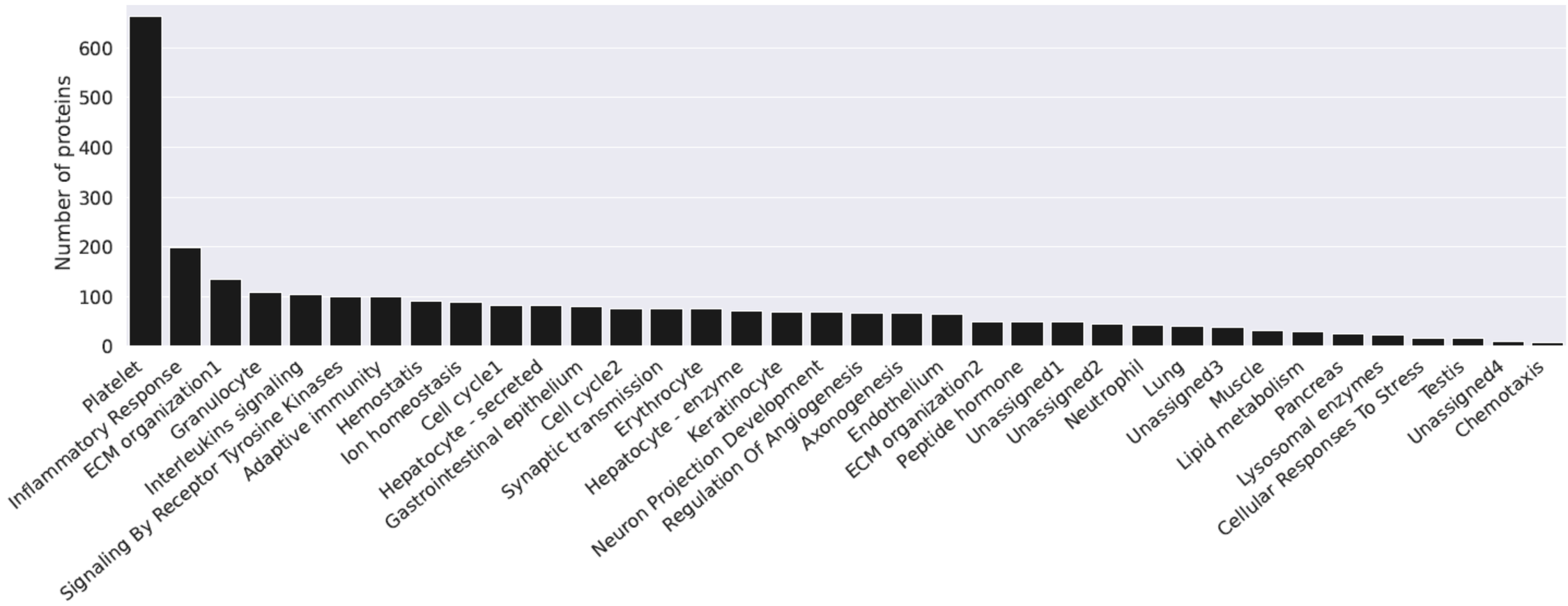
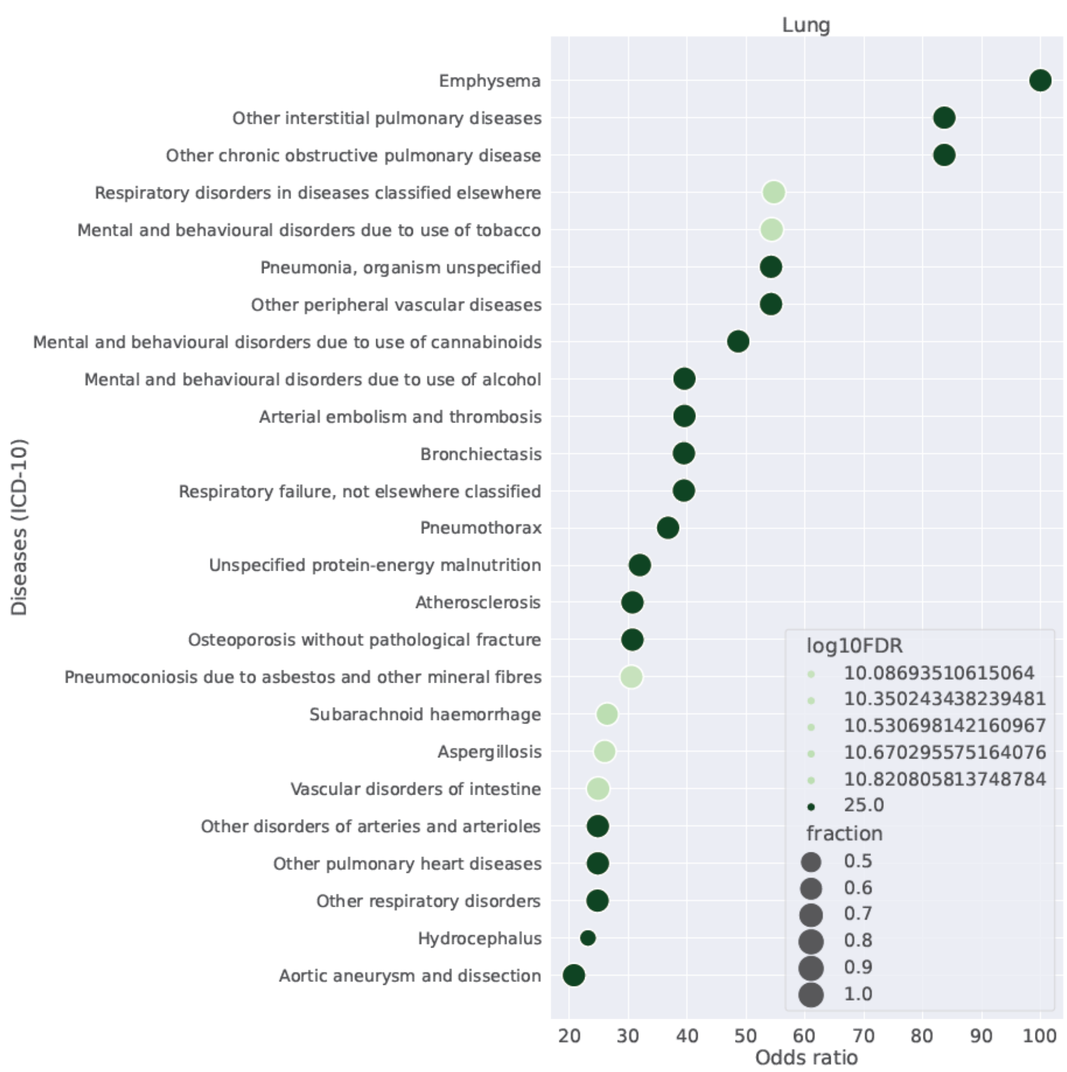

Minimize reconstruction loss + KL divergence
Plasma protein expression profile across cohort
Encoder
Latent space
$Z_{\mu,\sigma}$
Decoder
Reconstructed expression profile
Protein clustering based on latent space embeddings
Annotate protein clusters
Cell type expression enrichment
Pathway enrichment
Correlation with blood tests

Figure 2

a


Number of proteins
0
100
200
300
400
500
600
Platelet
Inflammatory Response
ECM organization1
Granulocyte
Interleukins signaling
Signaling By Receptor Tyrosine Kinases
Adaptive immunity
Hemostatis
Ion homeostasis
Cell cycle1
Hepatocyte - secreted
Gastrointestinal epithelium
Cell cycle2
Synaptic transmission
Erythrocyte
Hepatocyte - enzyme
Keratinocyte
Neuron Projection Development
Regulation Of Angiogenesis
Axonogenesis
Endothelium
ECM organization2
Peptide hormone
Unassigned1
Unassigned2
Neutrophil
Lung
Unassigned3
Muscle
Lipid metabolism
Pancreas
Lysosomal enzymes
Cellular Responses To Stress
Testis
Unassigned4
Chemotaxis

b


Lung
Diseases (ICD-10)
Emphysema
Other interstitial pulmonary diseases
Other chronic obstructive pulmonary disease
Respiratory disorders in diseases classified elsewhere
Mental and behavioural disorders due to use of tobacco
Pneumonia, organism unspecified
Other peripheral vascular diseases
Mental and behavioural disorders due to use of cannabinoids
Mental and behavioural disorders due to use of alcohol
Arterial embolism and thrombosis
Bronchiectasis
Respiratory failure, not elsewhere classified
Pneumothorax
Unspecified protein-energy malnutrition
Atherosclerosis
Osteoporosis without pathological fracture
Pneumoconiosis due to asbestos and other mineral fibres
Subarachnoid haemorrhage
Aspergillosis
Vascular disorders of intestine
Other disorders of arteries and arterioles
Other pulmonary heart diseases
Other respiratory disorders
Hydrocephalus
Aortic aneurysm and dissection
20
30
40
50
60
70
80
90
100
Odds ratio
log10FDR
10.08693510615064
10.350243438239481
10.530698142160967
10.670295575164076
10.820805813748784
25.0
fraction
0.5
0.6
0.7
0.8
0.9
1.0

c

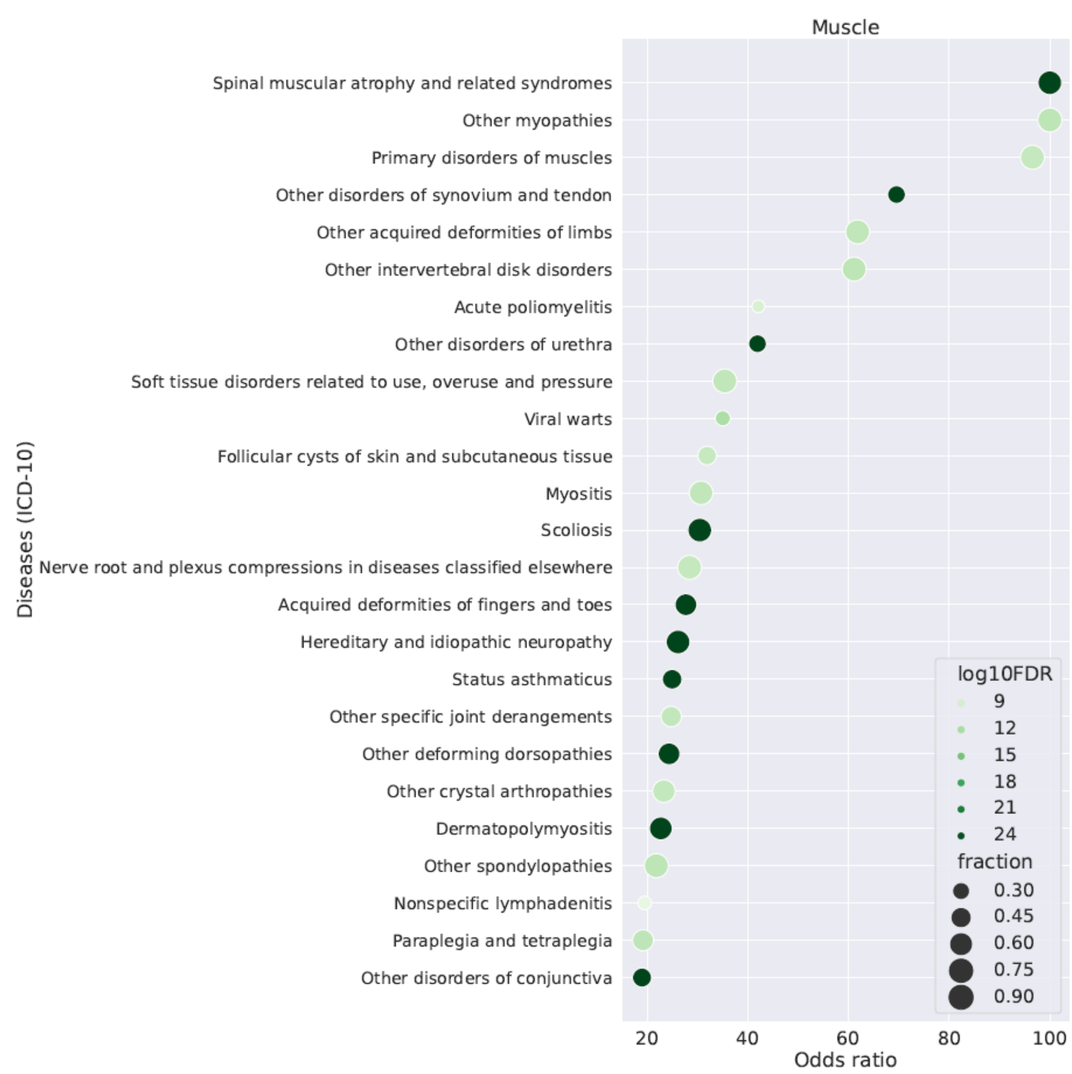
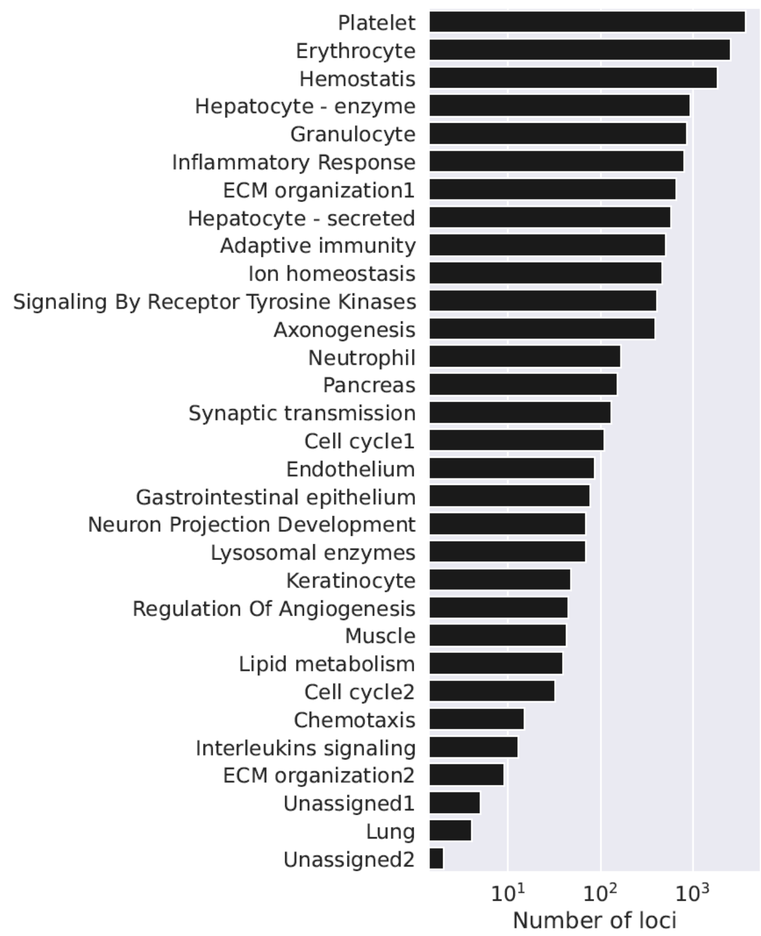
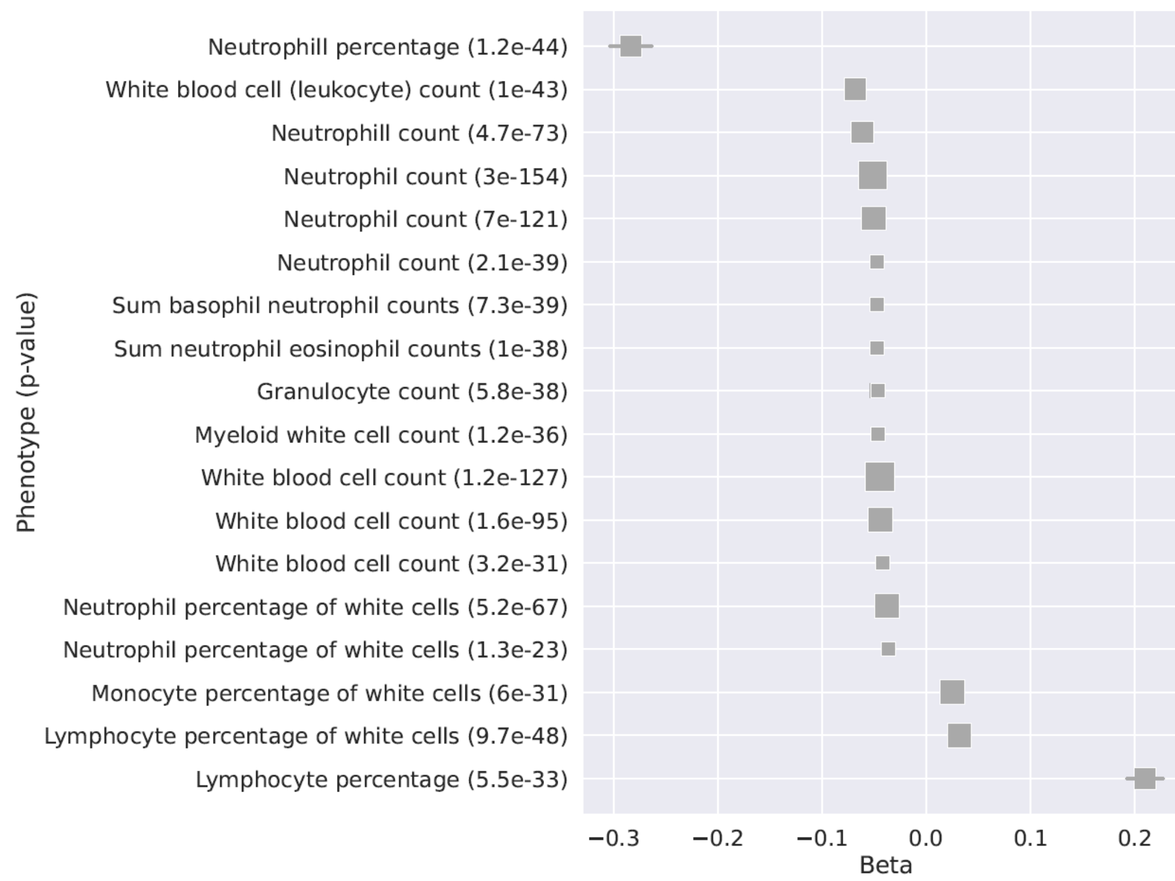

Muscle
Diseases (ICD-10)
Spinal muscular atrophy and related syndromes
Other myopathies
Primary disorders of muscles
Other disorders of synovium and tendon
Other acquired deformities of limbs
Other intervertebral disk disorders
Acute poliomyelitis
Other disorders of urethra
Soft tissue disorders related to use, overuse and pressure
Viral warts
Follicular cysts of skin and subcutaneous tissue
Myositis
Scoliosis
Nerve root and plexus compressions in diseases classified elsewhere
Acquired deformities of fingers and toes
Hereditary and idiopathic neuropathy
Status asthmaticus
Other specific joint derangements
Other deforming dorsopathies
Other crystal arthropathies
Dermatopolymyositis
Other spondylopathies
Nonspecific lymphadenitis
Paraplegia and tetraplegia
Other disorders of conjunctiva
20
40
60
80
100
Odds ratio
log10FDR
9
12
15
18
21
24
fraction
0.30
0.45
0.60
0.75
0.90

Figure 3

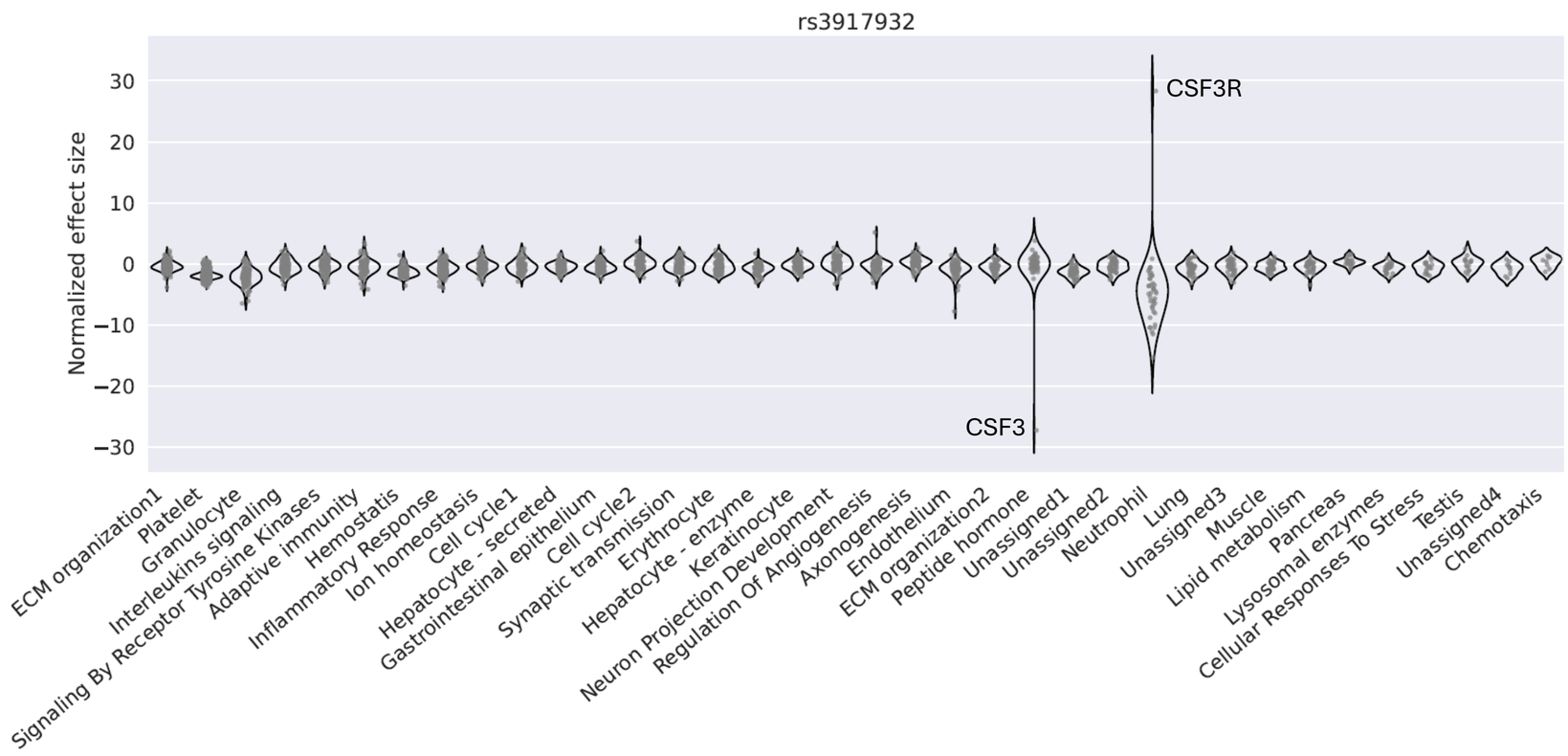
a



c



b



d

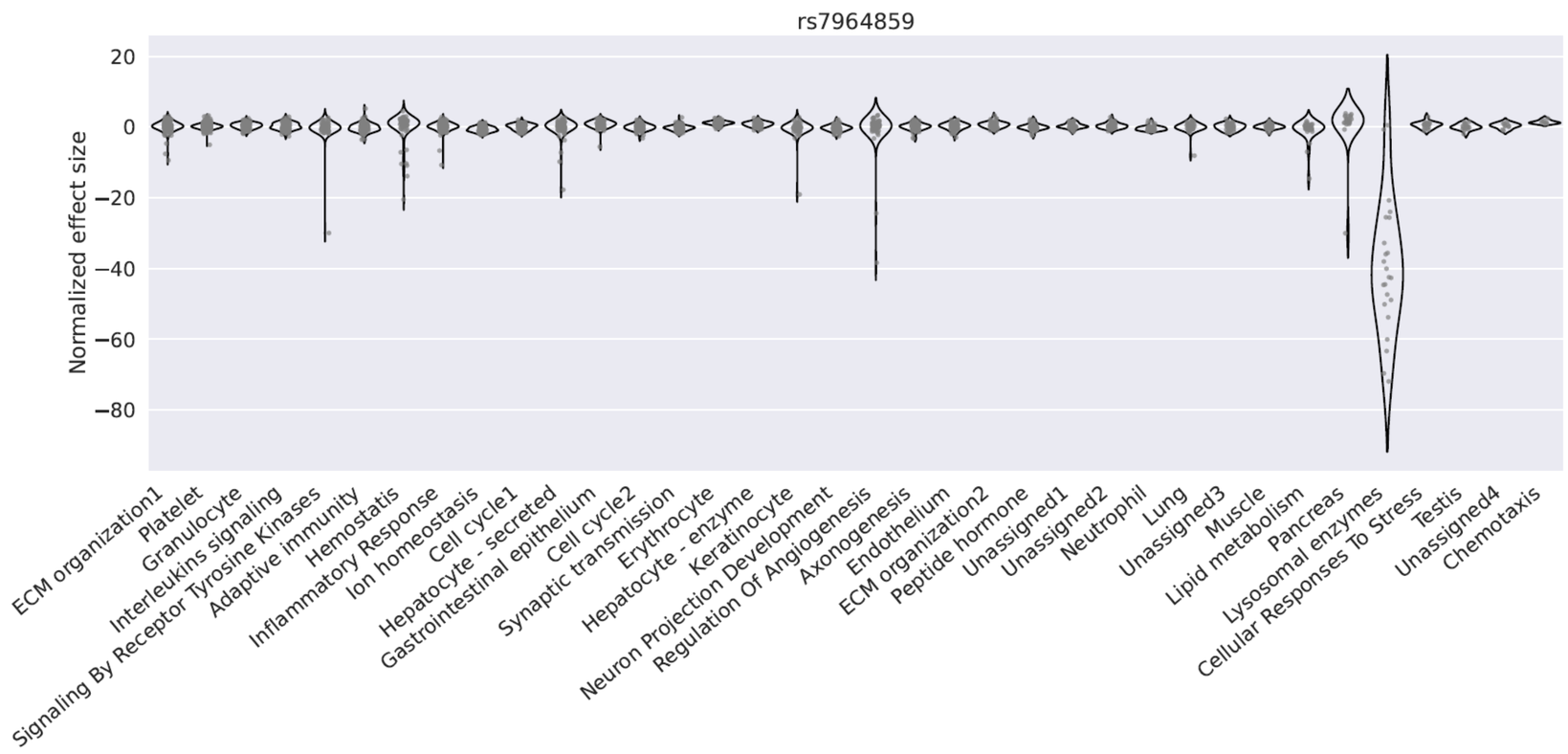
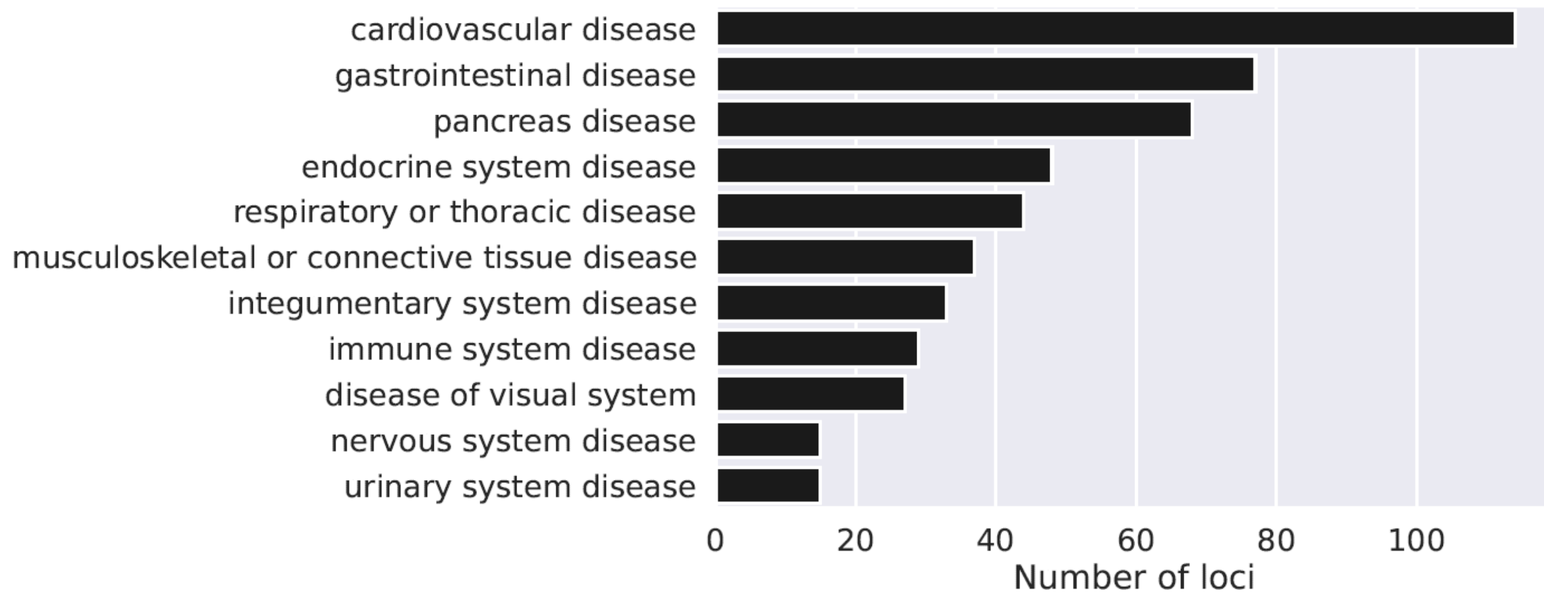
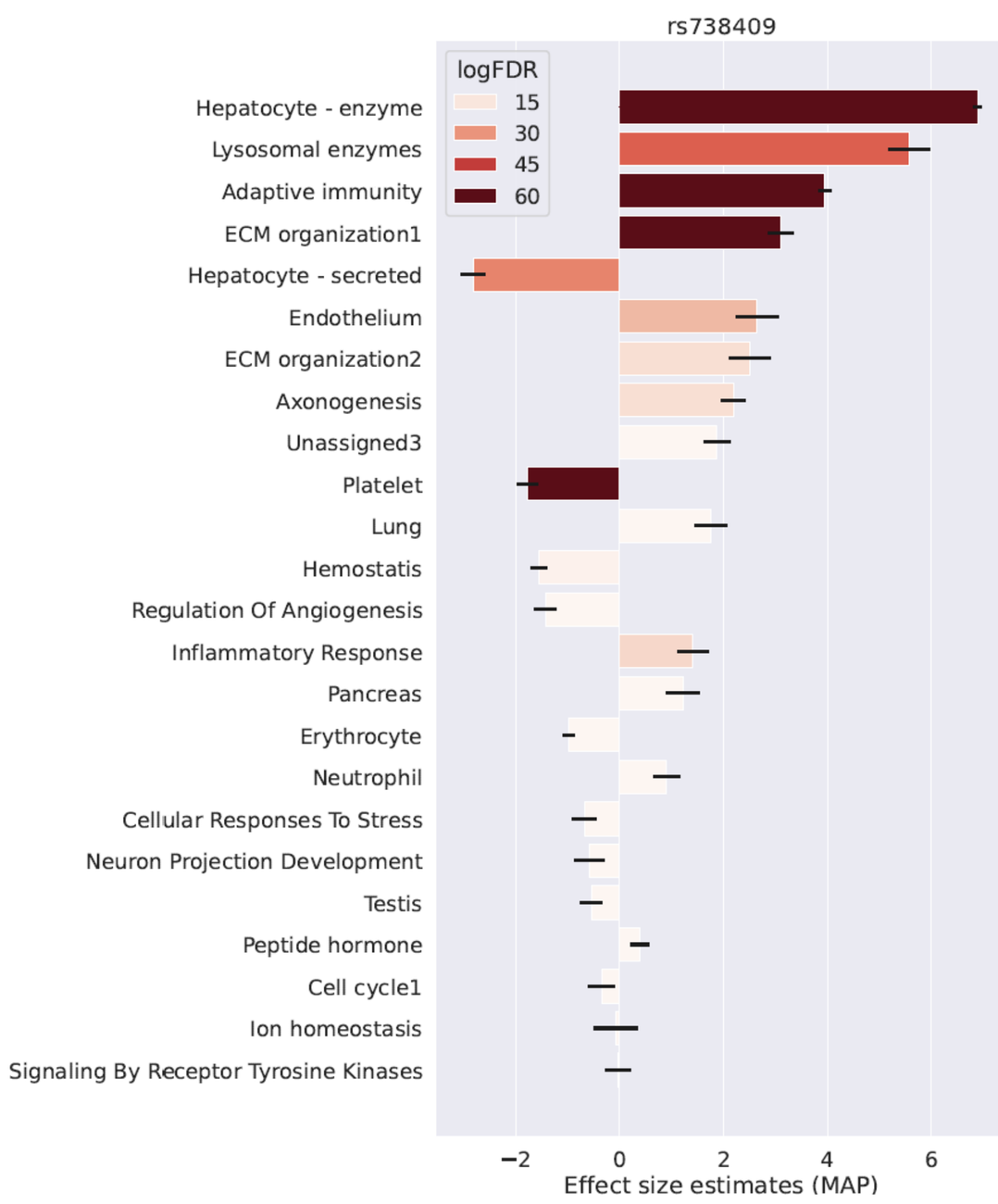


Figure 4

a


cardiovascular disease
gastrointestinal disease
pancreas disease
endocrine system disease
respiratory or thoracic disease
musculoskeletal or connective tissue disease
integumentary system disease
immune system disease
disease of visual system
nervous system disease
urinary system disease
0
20
40
60
80
100
Number of loci

b


rs738409
logFDR
15
30
45
60
Hepatocyte - enzyme
Lysosomal enzymes
Adaptive immunity
ECM organization1
Hepatocyte - secreted
Endothelium
ECM organization2
Axonogenesis
Unassigned3
Platelet
Lung
Hemostatis
Regulation Of Angiogenesis
Inflammatory Response
Pancreas
Erythrocyte
Neutrophil
Cellular Responses To Stress
Neuron Projection Development
Testis
Peptide hormone
Cell cycle1
Ion homeostasis
Signaling By Receptor Tyrosine Kinases
−2
0
2
4
6
Effect size estimates (MAP)

c

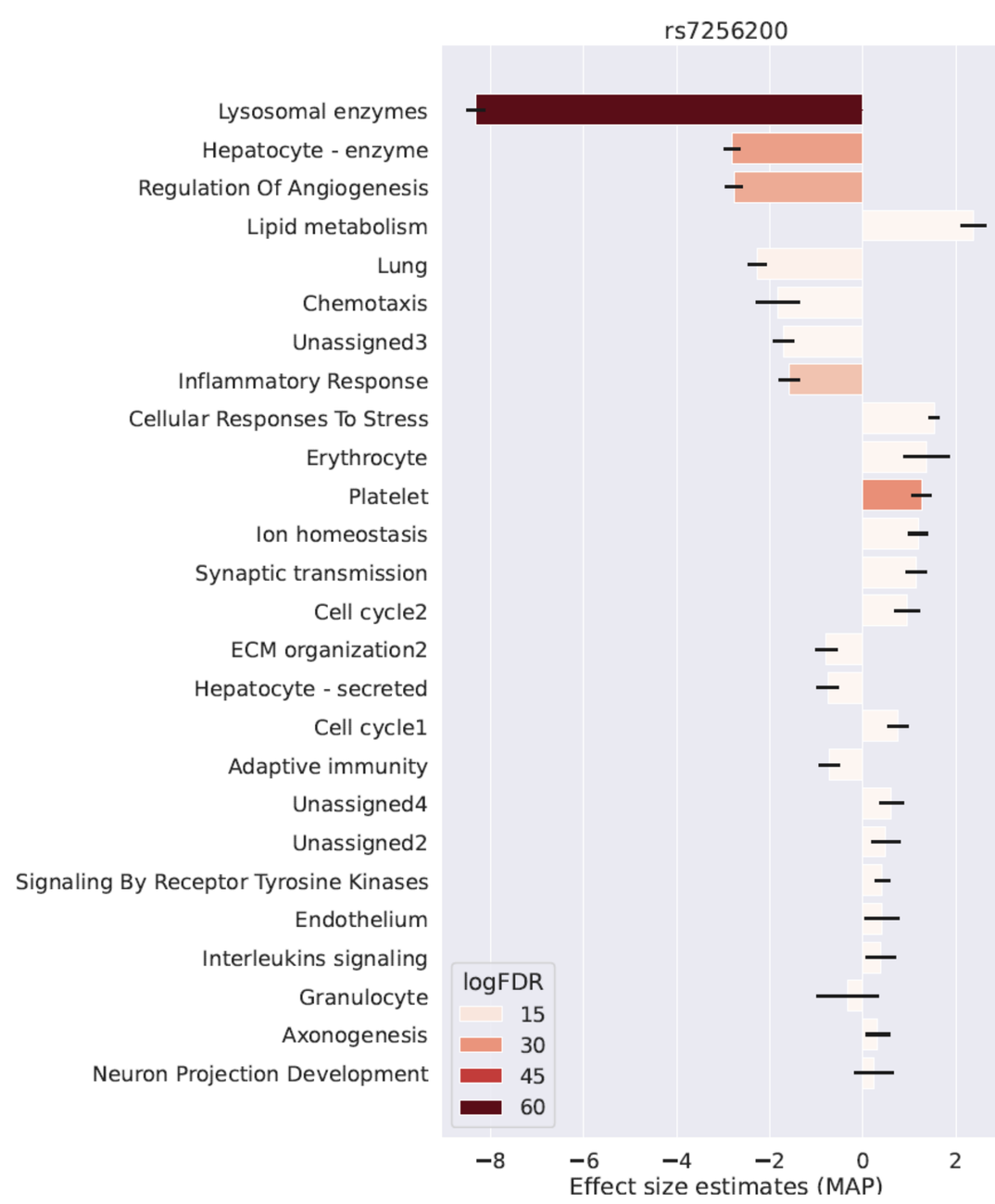
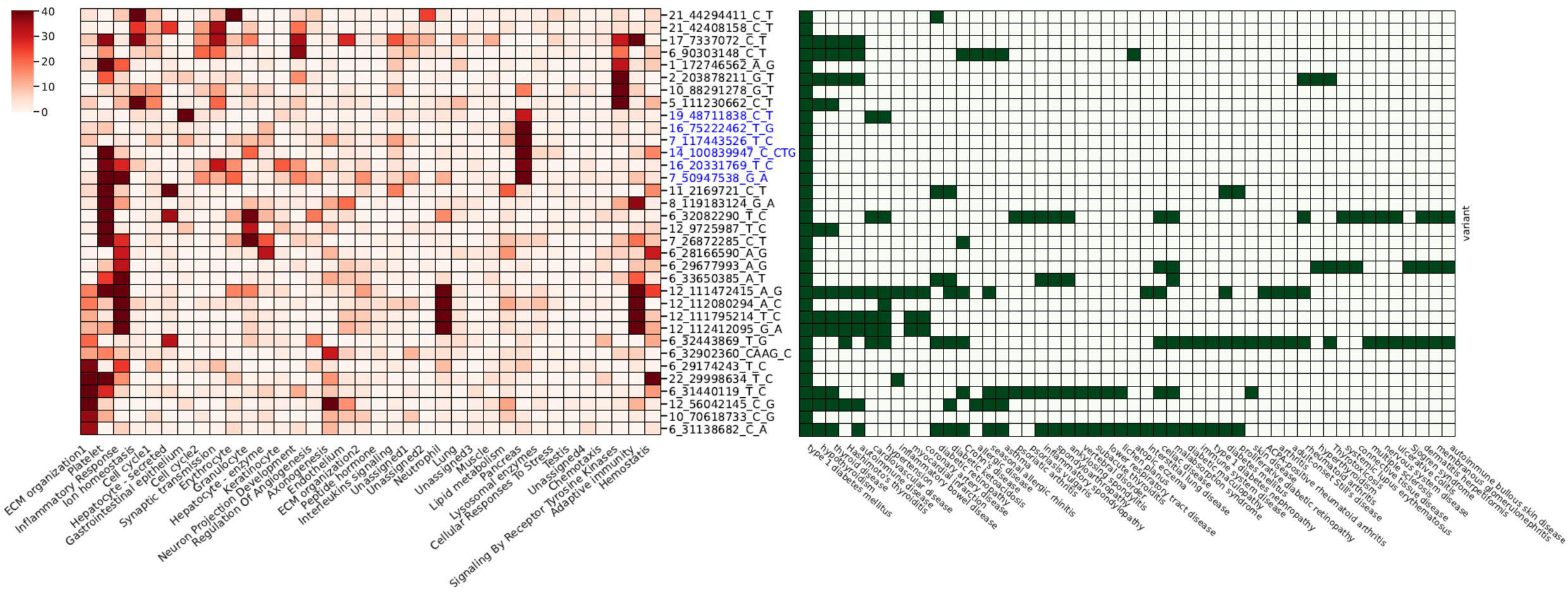
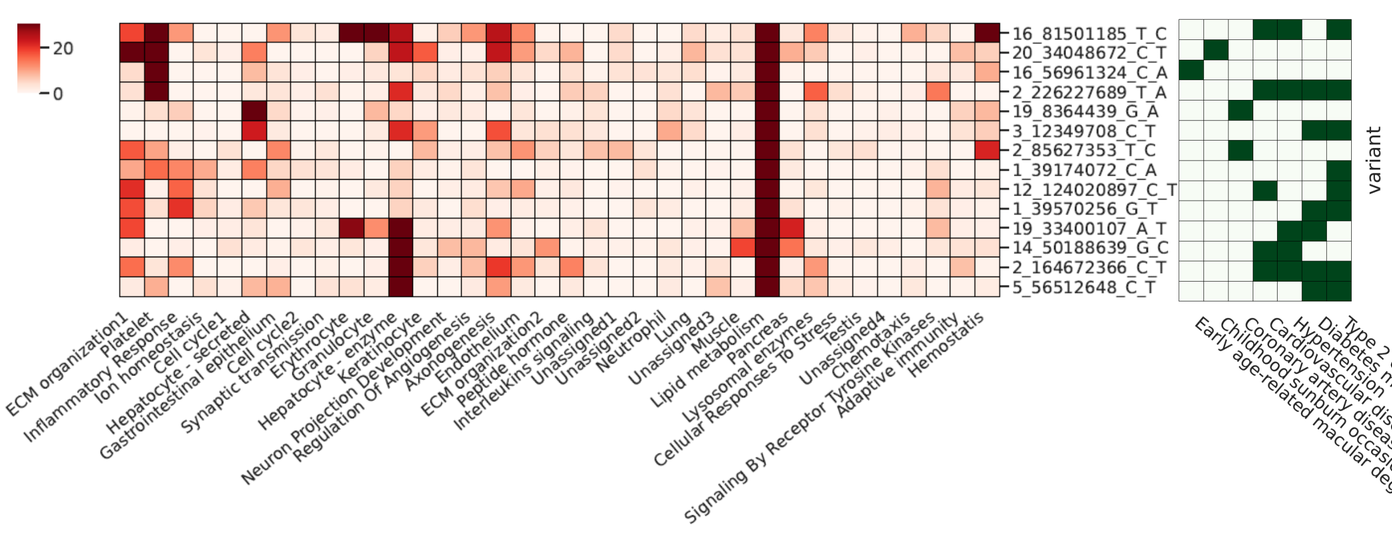

rs7256200
Lysosomal enzymes
Hepatocyte - enzyme
Regulation Of Angiogenesis
Lipid metabolism
Lung
Chemotaxis
Unassigned3
Inflammatory Response
Cellular Responses To Stress
Erythrocyte
Platelet
Ion homeostasis
Synaptic transmission
Cell cycle2
ECM organization2
Hepatocyte - secreted
Cell cycle1
Adaptive immunity
Unassigned4
Unassigned2
Signaling By Receptor Tyrosine Kinases
Endothelium
Interleukins signaling
Granulocyte
Axonogenesis
Neuron Projection Development
logFDR
15
30
45
60
−8
−6
−4
−2
0
2
Effect size estimates (MAP)

Figure 5

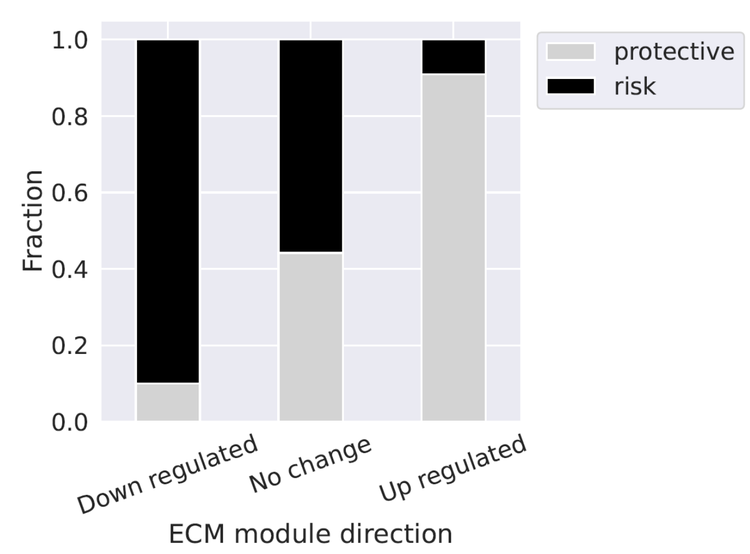
a


40
30
20
10
0
21_44294411_C_T
21_42408158_C_T
17_7337072_C_T
6_90303148_C_T
1_172746562_A_G
2_203878211_G_T
10_88291278_G_T
5_111230662_C_T
19_48711838_C_T
16_75222462_T_G
7_117443526_T_C
14_100839947_C_CTG
16_20331769_T_C
7_50947538_G_A
11_2169721_C_T
8_119183124_G_A
6_32082290_T_C
12_9725987_T_C
7_26872285_C_T
6_28166590_A_G
6_29677993_A_G
6_33650385_A_T
12_111472415_A_G
12_112080294_A_C
12_111795214_T_C
12_112412095_G_A
6_32443869_T_G
6_32902360_CAAG_C
6_29174243_T_C
22_29998634_T_C
6_31440119_T_C
12_56042145_C_G
10_70618733_C_G
6_31138682_C_A
variant
ECM organization1
Platelet
Inflammatory Response
Ion homeostasis
Hepatocyte - secreted
Cell cycle1
Gastrointestinal epithelium
Cell cycle2
Synaptic transmission
Erythrocyte
Granulocyte
Hepatocyte - enzyme
Keratinocyte
Neuron Projection Development
Regulation Of Angiogenesis
Axonogenesis
Endothelium
ECM organization2
Peptide hormone
Interleukins signaling
Unassigned1
Unassigned2
Neutrophil
Lung
Unassigned3
Muscle
Lipid metabolism
Pancreas
Lysosomal enzymes
Cellular Responses To Stress
Testis
Unassigned4
Chemotaxis
Signaling By Receptor Tyrosine Kinases
Adaptive immunity
Hemostatis
type 1 diabetes mellitus
hypothyroidism
thyroid disease
Hashimoto's thyroiditis
autoimmune disease
cardiovascular disease
hypertension
inflammatory disease
myocardial infarction
coronary artery disease
diabetic retinopathy
Crohn's disease
allergic disease
seasonal allergic rhinitis
asthma
psoriatic arthritis
psoriasis vulgaris
inflammatory bowel disease
spondyloarthropathy
ankylosing spondylitis
vertebral disorders
subacute thyroiditis
lower respiratory tract disease
lichen planus
atopic eczema
interstitial lung disease
celiac disease
malabsorption syndrome
diabetic macuopathy
immune system disease
type 1 diabetes mellitus
diabetes nephropathy
proliferative diabetic retinopathy
skin disease
ACPA-positive rheumatoid arthritis
arthritis
adult-onset Still's disease
rheumatoid arthritis
hyperthyroidism
Thyrotoxicosis
Systemic lupus erythematosus
connective tissue disease
multiple sclerosis
nervous system disease
ulcerative colitis
Sjogren syndrome
dermatitis herpetiformis
membranous glomerulonephritis
autoimmune bullous skin disease

b


20
0
16_81501185_T_C
20_34048672_C_T
16_56961324_C_A
2_226227689_T_A
19_8364439_G_A
3_12349708_C_T
2_85627353_T_C
1_39174072_C_A
12_124020897_C_T
1_39570256_G_T
19_33400107_A_T
14_50188639_G_C
2_164672366_C_T
5_56512648_C_T
variant
ECM organization1
Platelet
Inflammatory Response
Ion homeostasis
Hepatocyte - secreted
Cell cycle1
Gastrointestinal epithelium
Cell cycle2
Synaptic transmission
Erythrocyte
Granulocyte
Hepatocyte - enzyme
Keratinocyte
Neuron Projection Development
Regulation Of Angiogenesis
Axonogenesis
Endothelium
ECM organization2
Peptide hormone
Interleukins signaling
Unassigned1
Unassigned2
Neutrophil
Lung
Unassigned3
Muscle
Lipid metabolism
Pancreas
Lysosomal enzymes
Cellular Responses To Stress
Testis
Unassigned4
Chemotaxis
Signaling By Receptor Tyrosine Kinases
Adaptive immunity
Hemostatis
Type 2 diabetes
Diabetes mellitus
Hypertension
Cardiovascular disease
Coronary artery disease
Childhood sunburn occasions
Early age-related macular degeneration

c


1.0
0.8
0.6
0.4
0.2
0.0
Fraction
protective
risk
Down regulated
No change
Up regulated
ECM module direction