

Integrating transcriptome-wide association studies with gene co-expression patterns

Draft manuscript

Text in red/red are internal comments

This manuscript ([permalink](#)) was automatically generated from [greenelab/phenoplier manuscript@70d2b48](#) on May 24, 2021.

Authors

- **Milton Pividori**

 [0000-0002-3035-4403](#) ·  [miltondp](#) ·  [miltondp](#)

Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA ·

Funded by The Gordon and Betty Moore Foundation GBMF 4552; The National Human Genome Research Institute (R01 HG010067)

- **Marylyn D. Ritchie**

 [0000-0002-1208-1720](#) ·  [MarylynRitchie](#)

Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

- **Casey S. Greene**

 [0000-0001-8713-9213](#) ·  [cgreene](#) ·  [GreeneScientist](#)

Center for Health AI, University of Colorado School of Medicine, Aurora, CO, 80045, USA; Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO, 80045, USA · Funded by The Gordon and Betty Moore Foundation (GBMF 4552); The National Human Genome Research Institute (R01 HG010067); The

National Cancer Institute (R01 CA237170)

Abstract

Introduction

Tissue specificity is a key feature of human disease, and genes with tissue-specific expression are enriched for disease associations [1,2,3]. Identifying the function of genes involves understanding the regulatory mechanisms that affect their expression across different tissues and cell types [4,5,6]. Large compendia describing key elements of regulatory DNA have been recently released or updated, which comprise chromatin-state annotations, high-resolution enhancers [7], DNase I hypersensitive sites maps [5], and the characterization of genetic effects on gene expression across different tissues [4]. The integration with genome-wide association studies (GWAS) on thousands of common diseases could dramatically improve the identification of these transcriptional mechanisms that, when dysregulated, often result in tissue- and cell lineage-specific pathology.

Owing to the readily available gene expression data across several tissues [4,8,9,10], a popular approach to identify these biological processes is transcription-wide association studies (TWAS), which integrates expression quantitative trait loci (eQTL) data to provide a mechanistic interpretation for genome-wide association studies (GWAS). This is done by testing whether perturbations in gene regulatory mechanisms mediate the association between genetic variants and human diseases [11,12,13,14]. However, TWAS has not been useful to detect tissue-specific effects [15,16], since eQTLs are generally shared across tissues. Alternative statistical approaches that connect GWAS with gene expression data can infer disease-relevant tissues and cell types [16,17,18,19,20,21], but they generally apply enrichment analysis techniques that do not account for widespread gene correlations due to technical noise (i.e. “batch effects”) [22,23]. In addition, they generally rely on small sets of expression data compared with the total RNA-seq samples available today [8,9].

Here we propose a polygenic method that maps gene associations from TWAS on >4,000 traits [24] into a latent representation learned from public gene expression repositories on tens of thousands of RNA-seq human samples [8,25]. This low-dimensional space comprises latent variables representing gene modules with coordinated expression across different tissues and cell types. We used a computational approach that can reduce technical noise by learning patterns that align with prior knowledge. When mapping gene-trait associations to this reduced expression space, we found that traits and diseases are associated with gene modules expressed in relevant cell types. Several of these associations found in PhenomeXcan also replicated in 309 curated disease phenotypes from the Electronic Medical Records and Genomics (eMERGE) network phase III [26]. Our approach is robust in finding meaningful module-trait associations even when individual genes involved in lipids metabolism do not reach genome-wide significance in lipids-related traits. We also show that our module-based approach is more accurate in predicting known drug-disease medical indications than using single gene-trait associations, potentially representing a more potent drug repositioning framework. Finally, we performed cluster analysis on traits mapped to this latent representation, with clusters highly stable across different resolutions. We were able to identify highly disease-specific gene modules expressed in relevant tissues for a wide range of complex diseases, including autoimmune, cardiovascular, and mental disorders.

Results

Framework for the integration of TWAS with gene co-expression patterns

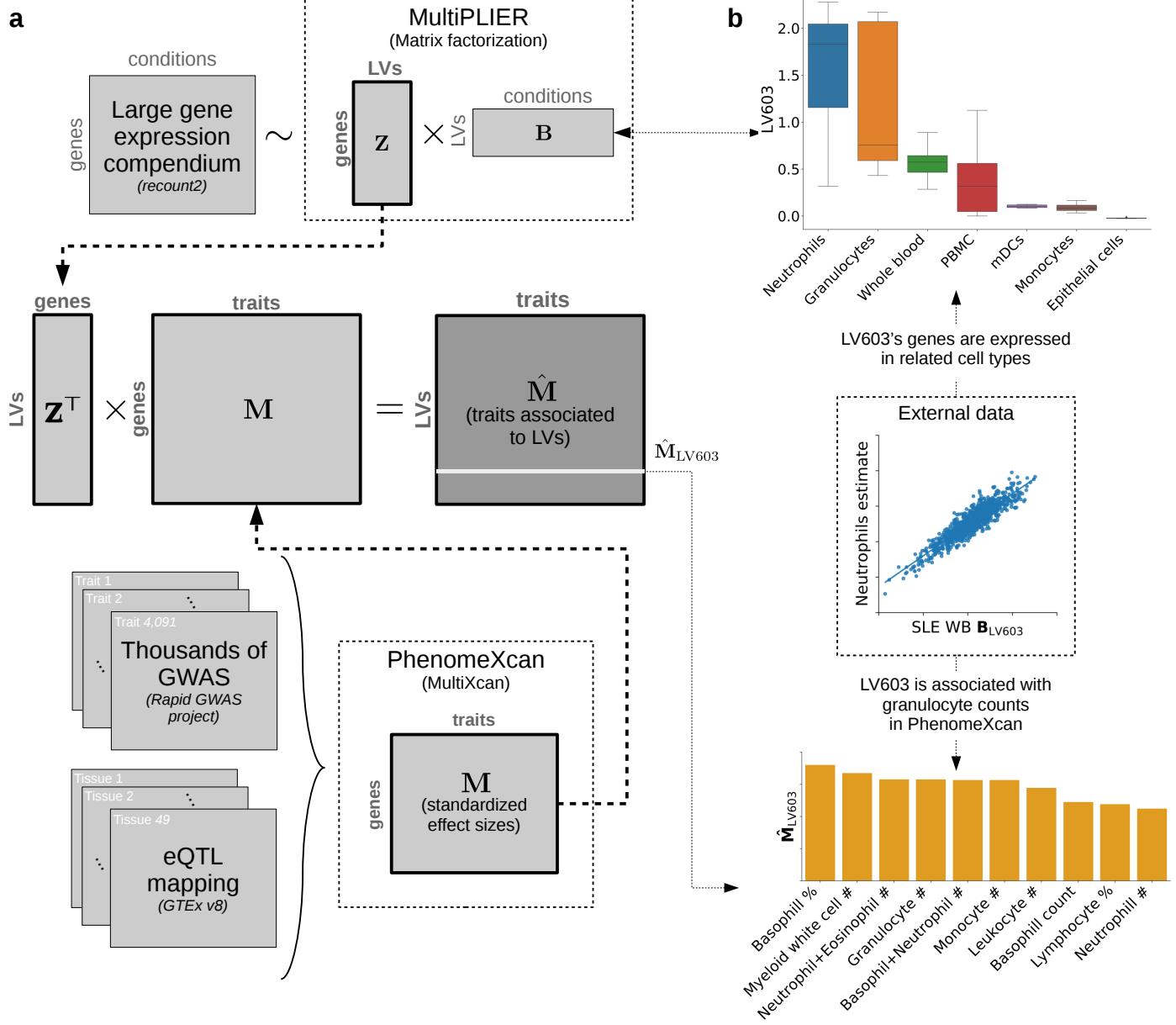


Figure 1: Schematic of the PhenoPLIER framework. **a)** The integration process between gene co-expression patterns from MultiPLIER (top) and TWAS results from PhenomeXcan. PhenoPLIER projects gene-based association results on thousands of traits to a latent space learned from large gene expression datasets. The process generates matrix \hat{M} , where each trait is now described by latent variables (LV) or gene modules. **b)** After the integration process, we found that neutrophil counts and other white blood cells (bottom) were ranked among the top 10 traits for an LV that was termed a neutrophil signature in the original MultiPLIER study. Genes in this LV are expressed in relevant cell types (top). PBMC: peripheral blood mononuclear cells; mDCs: myeloid dendritic cells.

In Figure 1, we show the main components of PhenoPLIER, our framework to integrate TWAS and gene co-expression patterns. The framework combines TWAS results with gene co-expression patterns by projecting gene-trait associations on a latent gene expression representation. We used PhenomeXcan [24], a massive TWAS resource on the UK Biobank [27] and other cohorts that provides results for 4,091 different diseases and traits. The latent gene expression representation was obtained by applying MultiPLIER [25], an unsupervised learning approach, on *recount2* [8]. Each of the 987 latent variables (LV) represents a gene module, essentially a group of genes with coordinated expression patterns (i.e., expressed together in the same tissues and cell types). Since modules might represent a functional set of genes regulated by the same transcriptional program [28,29], the projection of TWAS results into this latent space might provide context for their interpretation (see Methods). Our approach translates gene-trait associations to a gene module-trait score, allowing us to explore how different traits and diseases related to modules expressed in specific cell types and tissues, even under different developmental stages or stimuli. This is possible because the MultiPLIER

models also provide the experimental conditions (represented by matrix **B** in Figure 1 a) in which genes in a module are concurredly expressed.

In the original MultiPLIER study, the authors found an LV significantly associated with a known neutrophil pathway and highly correlated with neutrophil estimates from gene expression. We analyzed this LV using our approach (Figure 1 b), and found that 1) neutrophil counts and other white blood cell traits from PhenomeXcan were ranked among the top 10 traits for this LV, and 2) that the genes in this LV were expressed in neutrophil and other relevant cells. Moreover, using a generalized least squares approach similar to the gene-property analysis in MAGMA [30], we found that gene weights in this LV were predictive of gene associations for neutrophil count and percentage (FDR < 0.01) (see Methods). These initial results strongly suggested that shared patterns exist in the gene expression space (which has no GTEx samples) and the TWAS space (with gene models trained using GTEx v8); the approach also allows inferring the context-specific effects of gene modules on complex traits. We also show how the approach can aid translational efforts by mapping pharmacological perturbations to this latent space, enabling us to observe which compounds affect the transcriptional activity of gene modules.

Genes causally involved in lipids accumulation are associated with relevant traits and tissues

We found 271 genes associated with lipids accumulation by using a genome-wide lentiviral pooled CRISPR-Cas9 library targeting 19,114 genes in the human genome (see Methods). From these, we identified two gene-sets that either caused a decrease (96 genes in total, with eight high-confidence genes: *BLCAP*, *FBXW7*, *INSIG2*, *PCYT2*, *PTEN*, *SOX9*, *TCF7L2*, *UBE2J2*) or an increase of lipids (175 genes in total, with six high-confidence genes: *ACACA*, *DGAT2*, *HILPDA*, *MBTPS1*, *SCAP*, *SRPR*) (Supplementary File 1). Four gene modules were significantly enriched for these two gene-sets (FDR<0.05) (Supplementary Table 1). Next, we used these two lipids gene-sets to assess whether single gene-trait associations in PhenomeXcan recapitulated lipids-related traits. We show that our gene module-based approach can 1) robustly identify genetic associations even when known single-gene associations are not significant and 2) contextualize these results by identifying tissue and cell type-specific associations.

First, we assessed the genes' effects on all phenotypes by adding their *p*-values (as *z*-scores) and obtaining a ranked list of traits. The top associated traits for genes in the decreasing-lipids gene-set were highly relevant to lipid levels, such as hypertension, diastolic and systolic blood pressure, and vascular diseases, including also asthma and lung function (Supplementary Table 2). We performed the same operation for our gene module-based approach by considering 24 modules nominally enriched (unadjusted *p*-value < 0.05) with the decreasing-lipids gene-set (Gene-set enrichment analysis). In this case, we also found highly lipids-relevant traits among the top 25, including hypertension, blood pressure, cardiometabolic diseases like atherosclerosis, and celiac disease (Supplementary Table 3). This is particularly relevant because each of the 24 modules aggregated a specific weighted combination of almost 3,000 genes' *z*-scores across all traits. Thus, aggregating the effects of this number of genes and obtaining top-ranked lipids-relevant traits is unlikely to happen by chance (*p*-value < 0.01), suggesting that gene co-expression patterns are concordant with gene-trait associations.

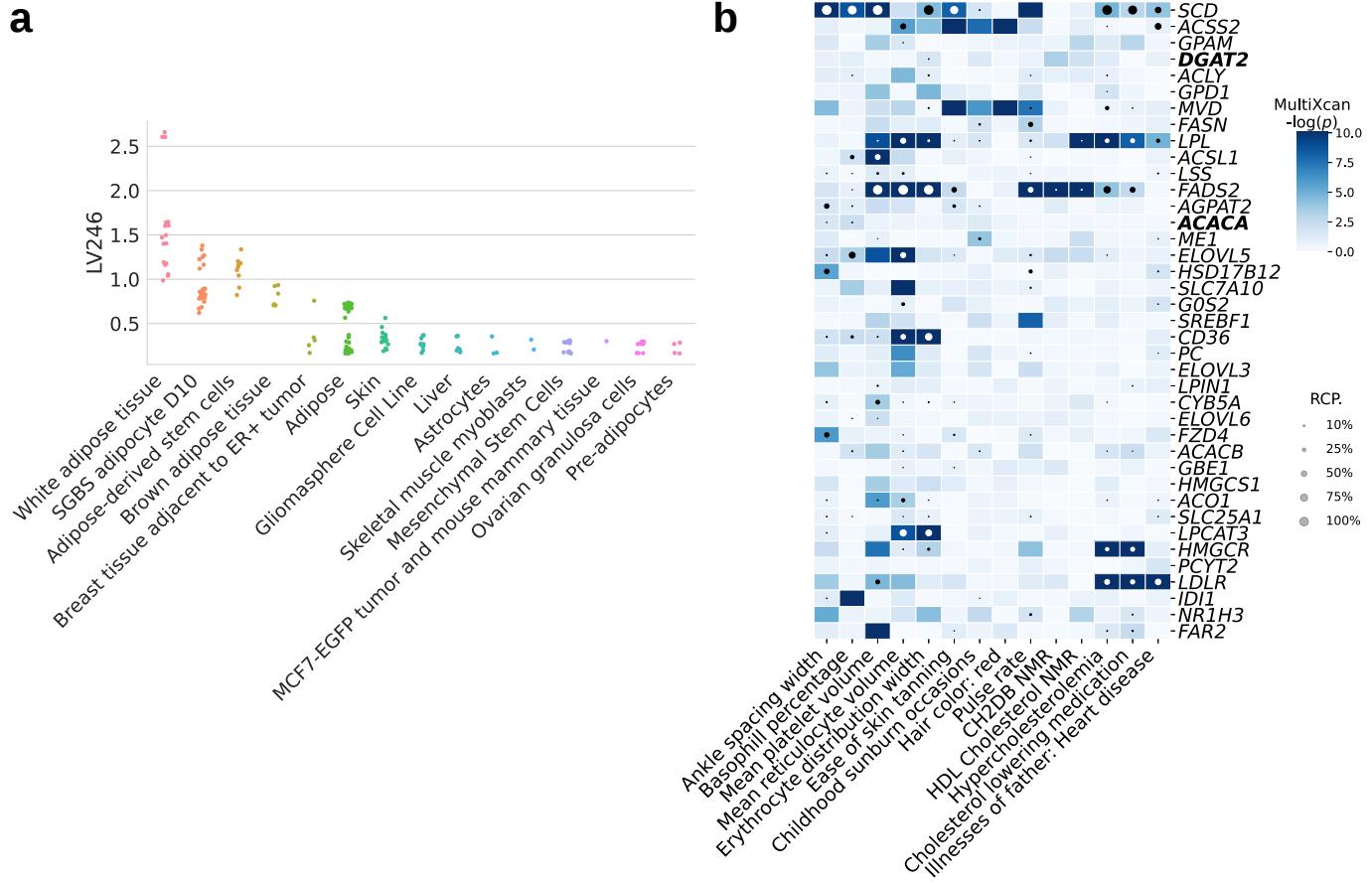


Figure 2: Tissues and traits associated with a gene module related to lipids metabolism (LV246). a) Top cell types/tissues where genes in LV246 are expressed on. Values in the *y*-axis come from matrix **B** in the MultiPLIER models (Figure 1 a). In the *x*-axis, cell types/tissues are sorted by the maximum value. **b)** Gene-trait associations (*S*-MultiXcan; thresholded at $-\log(p)=10$) and colocalization probability (fastENLOC) for the top traits in LV246. The top 40 genes in LV246 are shown, sorted by their module weight, from largest (top gene *SCD*) to smallest (gene *FAR2*); *DGAT2* and *ACACA*, in bold, are two of the six high-confidence genes in the increasing-lipids gene set from our CRISPR analyses. SGBS: Simpson Golabi Behmel Syndrome; CH2DB: CH₂ groups to double bonds ratio; NMR: nuclear magnetic resonance; HDL: high-density lipoprotein; RCP: locus regional colocalization probability.

When we considered the increasing-lipids set, genes and modules were associated with a more diverse set of traits, such as blood count tests, impedance measures, and bone-densitometry (Supplementary Tables 4 and 5). Additionally, gene modules were also associated with lung function, arterial stiffness, intraocular pressure, handgrip strength, rheumatoid arthritis, and celiac disease. One gene module (LV246) was also associated with lipids metabolism and triglyceride biosynthesis pathways in MultiPLIER (Supplementary Table 6). Using our approach, we found that its genes were mainly co-expressed in adipose tissue (Figure 2 a), which plays a key role in coordinating and regulating lipids metabolism. Gene-trait associations and colocalization for the top genes in this LV are shown in Figure 2 b). Additionally, gene weights for this LV significantly predicted gene associations for blood lipids and hypercholesterolemia (Supplementary Table 7). Two high-confidence genes from our CRISPR screening, *DGAT2* and *ACACA*, are responsible for encoding enzymes for triglycerides and fatty acid synthesis, and were among the top genes of LV246. However, as it can be seen in Figure 2 b, these two genes are not strongly associated with any of the top traits for this LV; other members of this module, such as *SCD*, *LPL*, *FADS2*, *HMGCR*, and *LDLR* were instead significantly associated and colocalized with lipid-relevant traits. This suggests that a module-based perspective can contextualize and reprioritize TWAS hits using modules of functionally related genes.

Our gene module-based approach more accurately predicts known disease therapeutics

We systematically evaluated our gene module-based approach from a translational perspective by assessing whether it could more accurately predict known treatments for disease. For this, we used the transcriptional responses to small molecule perturbations profiled in LINCS L1000 [31], which were further processed and mapped to DrugBank IDs [32,33,34]. Based on the established drug repurposing strategy that looks for reversed transcriptome patterns between genes and drug-induced perturbations [35,36], we used a framework for prioritizing drug candidates that uses imputed transcriptomes from GWAS [37]. For this, we computed a drug-disease score by anti-correlating the z -scores for a disease (from TWAS) and the z -scores for a drug (from LINCS) across sets of genes of different size (see Methods). Therefore, a large score for a drug-disease pair indicates that a higher (lower) predicted expression of disease-associated genes are down (up)-regulated by the drug, thus predicting a potential treatment. Similarly for the gene module approach, we estimated how pharmacological perturbations affected the gene module activity by projecting expression profiles of drugs into our latent representation (Methods). We used a manually-curated gold standard of drug-disease medical indications [33,38] across 53 diseases and 322 compounds to evaluate and compare the prediction performance.

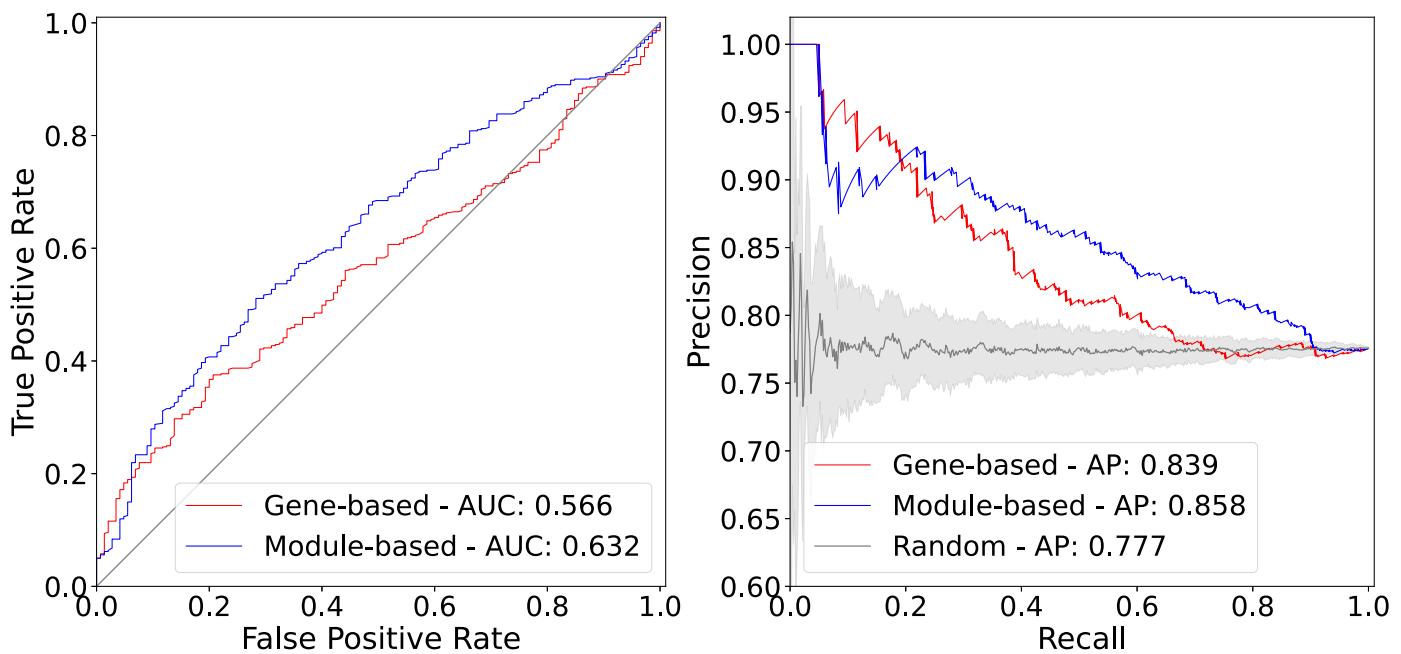


Figure 3: Drug-disease prediction performance for gene-based and module-based approaches. The receiver operating characteristic (ROC) (left) and the precision-recall curves (right) for a gene-based and our module-based approach. AUC: area under the curve; AP: average precision.

The ROC and precision-recall (PR) curves comparing both approaches are shown in Figure 3. Our proposed gene module-based method outperformed the gene-based one with an area under the curve of 0.632 and an average precision of 0.858. It is important to note that the gene-trait associations and drug-induced expression profiles projected into the latent space represent a compressed version of the entire set of results. The prediction results show that this low-dimensional space captures biologically meaningful patterns that can better represent and link pathophysiological processes with the mechanisms of action of drugs. In the following, with the aim to understand these results, we examined specific a drug-disease pair where both methods disagreed.

Nicotinic acid (niacin) is a B vitamin widely used clinically to treat lipid disorders. Niacin exerts its effects on a number tissues, although not all its mechanisms have been documented [39,40]. This compound can increase high-density lipoprotein (HDL) by inhibiting an HDL catabolism receptor in liver. Niacin also inhibits diacylglycerol acyltransferase-2 (DGAT2), which decrease production of low-density lipoproteins (LDL) by modulating triglyceride synthesis in hepatocytes, or by inhibiting adipocyte triglyceride lipolysis [39]. Niacin was categorized in our gold standard as a disease-modifying indication for atherosclerosis (AT) and coronary artery disease (CAD), and not for

pancreatitis. For pancreatitis, both the gene-based and module-based methods assigned a negative score (below their averages), which agrees with the gold standard in that niacin does not therapeutically change the biology of this disease. For AT, the module-based approach predicted niacin as a therapeutic drug by scoring them with 0.52 (above the mean), whereas the gene-based method assigned negative scores of -0.01 (below the mean), respectively. To understand why the predictions by the module-based method were different, we obtained the LVs that positively contributed to the score by looking at large positive (negative) LV values for the disease and large negative (positive) LV values for the drug of interest (Supplementary Table). Notably, LV246 (analyzed previously) was among the top 20 modules contributing to the prediction of niacin as a therapeutic drug for AT. As shown in Figure 2, this module is mainly expressed in adipose cells and liver cells, and its top genes encode important enzymes involved in lipid biosynthesis. The module as a whole (considering the gene weights) is significantly associated with cardiovascular traits (Supplementary Table 7), and several of its top genes are significantly associated and colocalized with cardiovascular-related traits: *SCD* (10q24.31) is associated with hypercholesterolemia ($P=1.9e-5$) and its GWAS and eQTL signals are fully colocalized ($RCP=1.0$); *LPL* (8p21.3), which is known to be linked to different disorders of lipoprotein metabolism, is strongly associated with hypercholesterolemia ($P=7.5e-17$, $RCP=0.26$), and family history of heart disease ($P=1.7e-5$, $RCP=0.22$); other genes associated with hypercholesterolemia in this module are *FADS2* (11q12.2) ($P=9.42e-5$, $RCP=0.623$), *HMGCR* (5q13.3) ($P=1.3e-42$, $RCP=0.23$), and *LDLR* (19p13.2) ($P=9.9e-136$, $RCP=0.41$).

The analysis of other niacin-AT-contributing modules revealed additional known mechanisms of action of niacin. For example, GPR109A/HCAR2 is a G protein-coupled high-affinity niacin receptor in adipocytes and immune cells, including monocytes, macrophages, neutrophils and dendritic cells [41,42]. It was initially thought that the antiatherogenic effects of niacin were solely due to inhibition of lipolysis in adipose tissue. However, it has been shown that nicotinic acid can reduce atherosclerosis progression independently of its antidiabetic activity through the activation of GPR109A in immune cells [43], thus boosting anti-inflammatory processes and reversing cholesterol transport [44]. This alternative mechanism for niacin could have been hypothesized by examining the cell types where top modules positively contributing to the niacin-AT prediction are expressed: LV116 and LV931 (Supplementary Figures 11 and 12). Among these, we also found LV678 positively contributing to this prediction, which was significantly enriched with the lipids-decreasing genes from our CRISPR screening (Supplementary Table 1). This module was expressed in the heart and muscle cells (Supplementary Figure 10).

The projection of these two types of data into a common latent gene module-based representation could provide a more powerful framework for drug repositioning using data from genetic studies. Additionally, our approach could also be helpful to understand better the mechanism of pharmacological effect of known or experimental drugs. For example, one of the top modules affected by niacin (LV66, Supplementary Figure 13) is mainly expressed in ovarian granulosa cells. This compound has been very recently considered as a potential therapeutic for ovarian diseases [45,46], as it was found to promote follicle growth and inhibit granulosa cell apoptosis in animal models. Our proposed approach could be helpful to generate novel hypotheses to evaluate potential mechanisms of action of different drugs.

Clusters of traits in the gene module space are associated with relevant transcriptional processes

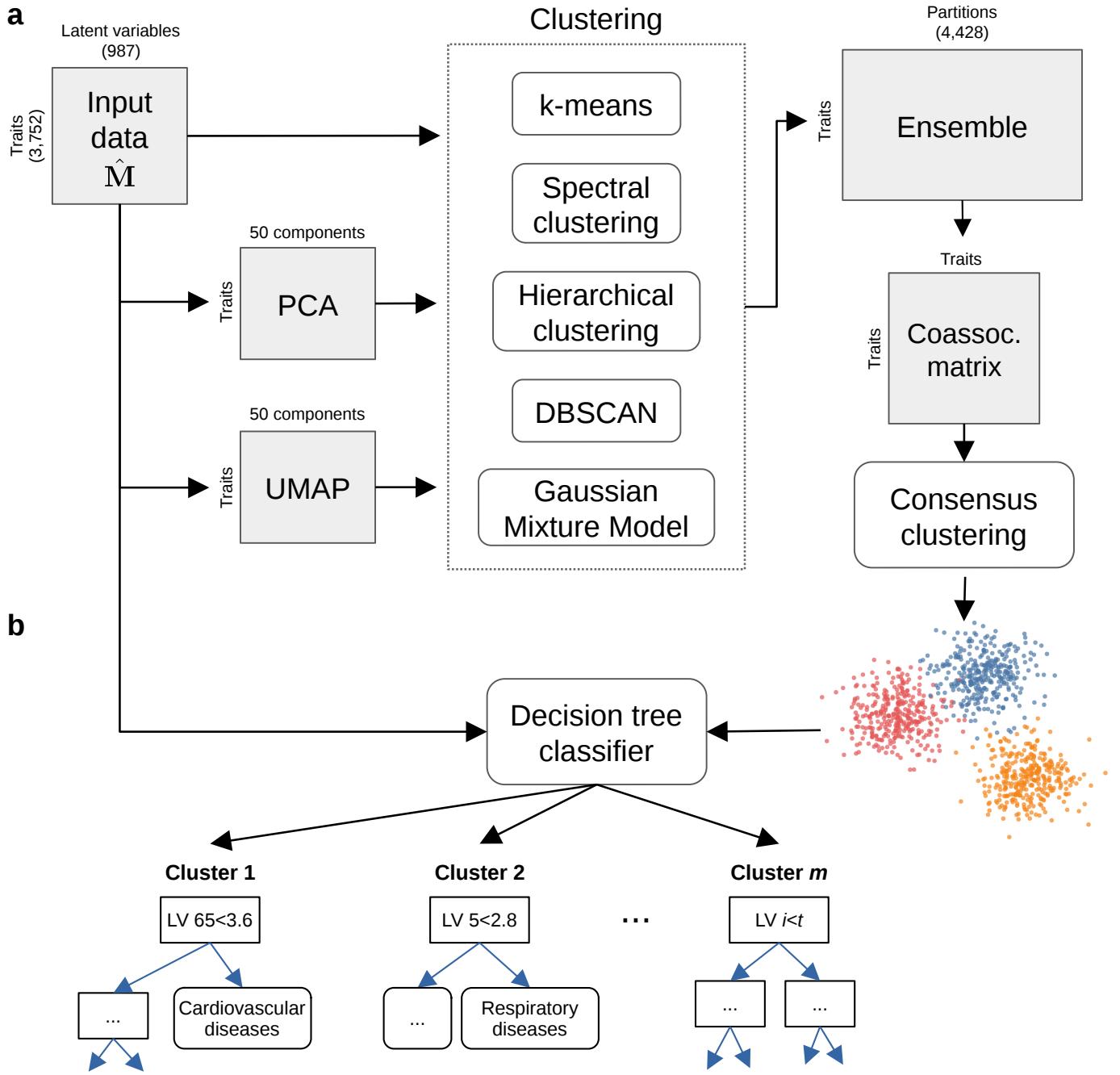


Figure 4: Cluster analysis on traits using the latent gene expression representation. **a)** The projection of TWAS results on $n=3,752$ traits into the latent gene expression representation is the input data to the clustering process. A linear (PCA) and non-linear (UMAP) dimensionality reduction techniques are applied to the input data, and the three data versions are processed by five different clustering algorithms. These algorithms derive partitions from the data using different sets of parameters (such as the number of clusters), leading to an ensemble of 4,428 partitions. Then, a distance matrix is derived by counting how many times a pair of traits were grouped in different clusters across the ensemble. Finally, a consensus function is applied to the distance matrix to generate consolidated partitions with different number of clusters (from 2 to $\sqrt{n} \approx 60$). These final solutions are represented in the clustering tree (Figure 5). **b)** The clusters found by the consensus function are used as labels to train a decision tree classifier on the original input data, which detects the LVs that better differentiate groups of traits.

The previous results suggest that \hat{M} represents a less noisy low-dimensional version of the data. Thus, we conducted cluster analysis on \hat{M} to find groups of traits that are similarly affected by the same transcriptional processes. To avoid using a single clustering algorithm (which implies using a single assumption about the structure of the data), we employed a consensus clustering approach where different methods with varying sets of parameters were applied on the data, and later combined into a consolidated solution [47,48,49] (Figure 4). Our clustering pipeline generated 15 final consensus clustering solutions with 5 to 29 clusters (Supplementary Figure 30). Instead of selecting a

specific number of clusters, we used a clustering tree [50] (Figure 5) to examine stable groups of traits across multiple resolutions. Finally, for the interpretation of the clusters, we trained a decision tree classifier (a highly interpretable machine learning model) on the input data $\hat{\mathbf{M}}$ using the clusters found as labels. This allowed us to quickly identify the latent variables/gene modules that better differentiated the groups of traits found (see Methods).

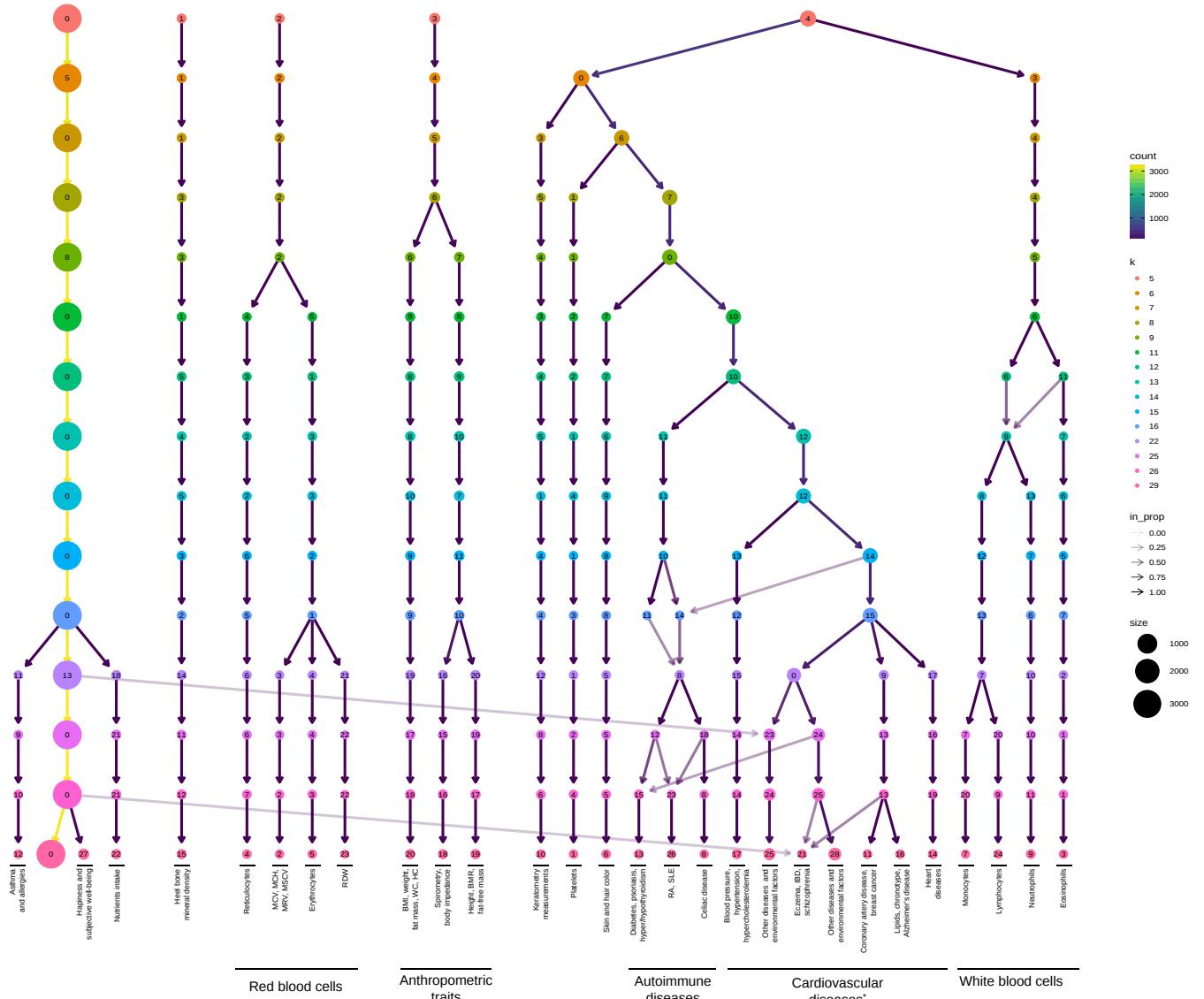


Figure 5: Clustering tree using multiple resolutions for clusters of traits. Each row represents a partition/grouping of the traits, and each circle is a cluster from that partition, and the number of clusters go from 5 to 29. Arrows indicate how traits in one cluster move across clusters from different partitions. Most of the clusters are preserved across different resolutions, showing highly stable solutions even with independent runs of the clustering algorithm. MCV: mean corpuscular volume; MCH: mean corpuscular hemoglobin; MRV: mean reticulocyte volume; MSCV: mean spheroid cell volume; RDW: red cell (erythrocyte) distribution width; BMI: body mass index; WC: waist circumference; HC: hip circumference; BMR: basal metabolic rate; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus; IBD: inflammatory bowel disease; *Descriptions of traits by cluster IDs (from left to right):* 12: also includes lymphocyte count and allergies such as allergic rhinitis or eczema; 4: includes reticulocyte count and percentage, immature reticulocyte fraction, and high light scatter reticulocytes count and percentage; 5: includes erythrocyte count, hemoglobin concentration, and hematocrit percentage; 18: also includes ankle spacing width; 1: includes platelet count, crit, mean volume, and distribution width; 13: diabetes refers to age when the diabetes was first diagnosed; 25: includes vascular problems such as angina, deep vein thrombosis (DVT), intraocular pressure, eye and mouth problems, pulse rate, hand-grip strength, several measurements of physical activity, jobs involving heavy physical work, types of transport used, intake of vitamin/mineral supplements, and various types of body pain and medications for pain relief; 21: also includes attention deficit hyperactivity disorder (ADHD), number of years of schooling completed, bone density, and intracranial volume measurement; 28: includes diabetes, gout, arthrosis, and respiratory diseases (and related medications such as ramipril, allopurinol, and lisinopril), urine assays, female-specific factors (age at menarche, menopause, first/last live birth), and several environmental/behavioral factors such as intake of a range of food/drink items including alcohol,

time spent outdoors and watching TV, smoking and sleeping habits, early-life factors (breastfed as a baby, maternal smoking around birth), education attainment, psychological and mental health, and health satisfaction; 11: also includes fasting blood glucose and insulin measurement; 16: lipids include high and low-density lipoprotein cholesterol (HDL and LDL), triglycerides, and average number of methylene groups per a double bond; 14: includes myocardial infarction, coronary atherosclerosis, ischaemic heart disease (wide definition). 9: includes neutrophil count, neutrophil+basophil count, neutrophil+eosinophil count, granulocyte count, leukocyte count, and myeloid cell count.

The clustering tree in Figure 5 shows five clear branches that are shown at the top with different numerical labels (from left to right): 0) a “large” branch that includes most of the traits that start to be subdivided only at $k=16$ (with asthma, subjective well-being traits, and nutrient intake clusters), 1) heel bone-densitometry measurements, 2) hematological assays on red blood cells, 3) physical measures, including spirometry and body impedance, and anthropometric traits with fat-free and fat mass measures in separate sub-branches, and 4) a “complex” branch including keratometry measurements, assays on white blood cells and platelets, skin and hair color traits, autoimmune disorders (type 1 diabetes, psoriasis, hyper/hypothyroidism, rheumatoid arthritis, systemic lupus erythematosus, celiac disease), and cardiovascular diseases (hypertension, coronary artery disease, myocardial infarction, hypercholesterolemia, and other cardiovascular-related traits such hand-grip strength [51], and environmental/behavioral factors such as physical activity and diet) (See Supplementary Files 1-5 for clustering results). All branches show relatively highly stable clusters, where the same traits are clustered together across different resolutions even with the consensus algorithm using random seeds at each level. The arrows between different clusters show how traits move from one group to another across different resolutions. This mainly happens between clusters within the “complex” branch, and between clusters from the “large” branch to the “complex” branch. This is expected, since the “complex” branch contains traits related to a wide range of different factors and thus are hard to categorize into a single cluster.

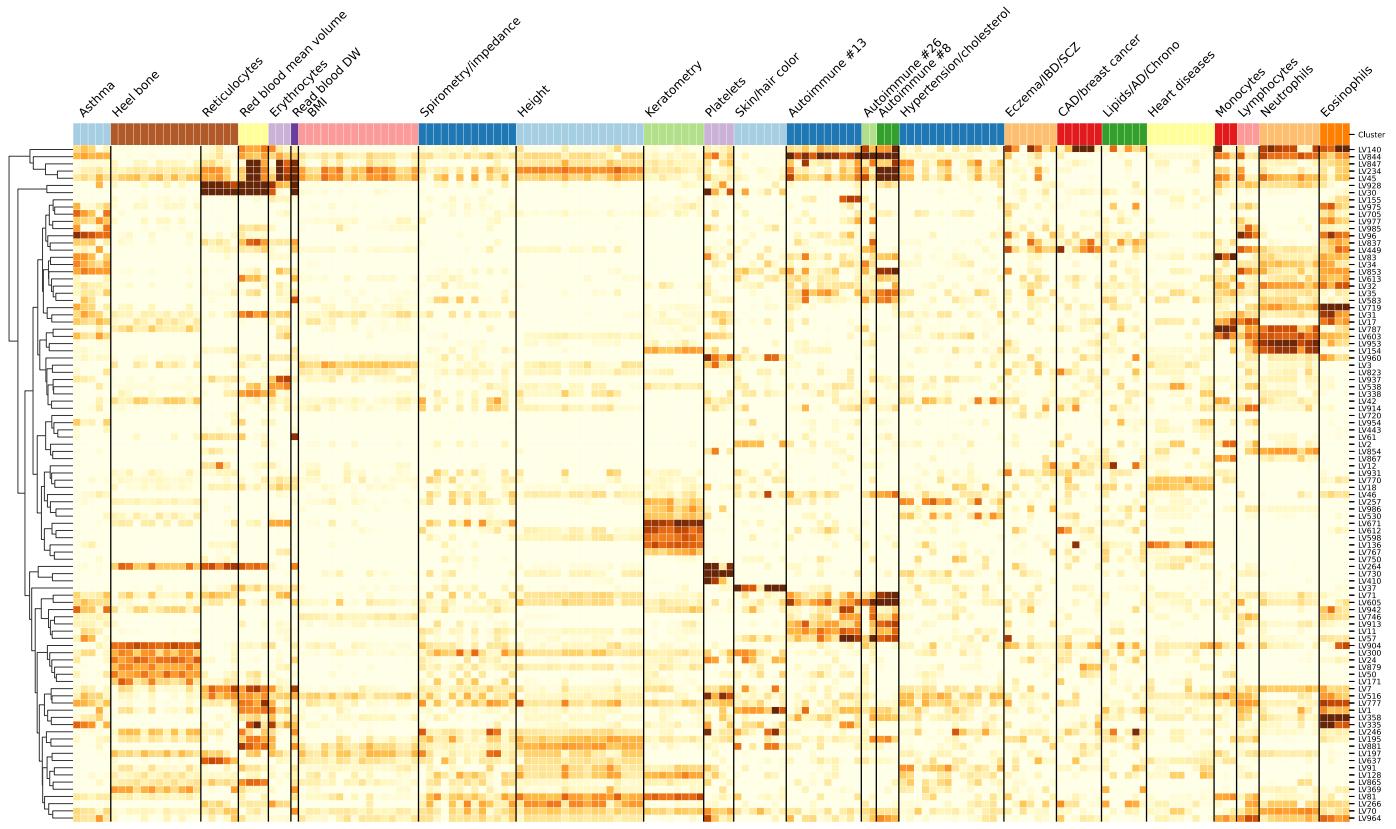


Figure 6: Cluster-specific and general transcriptional processes. The plot shows a submatrix of \hat{M} for the main trait clusters at $k=29$, considering only gene modules (rows) that align well with at least one known pathway. Values are standardized from -5 (lighter color) to 16 (darker color).

Next, we analyzed which gene modules are driving these clusters of traits. For that, we trained decision tree classifiers on the input data (Figure 4) using each cluster at $k=29$ (bottom of Figure 5) as

labels (see Methods). This yielded for each cluster the top gene modules, where several of them were well-aligned to existing pathways, and others were “novel” and expressed in relevant tissues. We summarized this in Figure 6, where it can be seen that some modules are highly specific to certain types of traits, and others seem to be associated with a wide range of different traits and diseases, thus potentially involved in more general biological functions. For example, modules such as LV928 and LV30 (Supplementary Figures 14 and 15), which are known to be related to early progenitors of the erythrocytes lineage [52], are predominantly expressed in early differentiation stages of erythropoiesis, and strongly associated with different assays on red blood cells (erythrocytes and reticulocytes). On the other side, others are highly specific, such as LV730, expressed in thrombocytes from different cancer samples (Supplementary Figures 16), and strongly associated with hematological assays on platelets; or LV598, whose genes are expressed in corneal endothelial cells (Supplementary Figures 17) and associated to keratometry measurements (FDR < 0.05; Supplementary Table 12).

The autoimmune diseases sub-branch also has significant gene modules associations expressed in relevant cell types. LV155 was strongly expressed in thyroid (Supplementary Figures 18), and significantly associated with hypothyroidism both in PhenomeXcan and eMERGE (FDR < 0.05; Supplementary Tables 13 and 14). LV844 was the most strongly associated gene module with autoimmune disorders (FDR < 1e-15; Supplementary Tables 15 and 16), and was expressed in a wide range of cell types, including blood, breast organoids, myeloma cells, lung fibroblasts, and different cell types from the brain (Supplementary Figures 19). Other important gene modules associated with autoimmunity in both PhenomeXcan and eMERGE are LV57 expressed in T cells (Supplementary Figure 20, and Supplementary Tables 17 and 18), and LV54 expressed in different soft tissue tumors, breast, lung, pterygia and epithelial cells (Supplementary Figure 21, and Supplementary Tables 19 and 20).

The cardiovascular sub-branch also exhibited significant associations, such as LV847 (Supplementary Figure 22) with blood pressure traits and hypertension (Supplementary Tables 21 and 22), which was expressed in CD19 (B cells) (which are related to preeclampsia [53]), Jurkat cells (T lymphocyte cells), and cervical carcinoma cell lines (the uterus was previously reported to be linked to blood pressure through a potential hormonal pathway [54,55]). LV136 was aligned with known collagen formation and muscle contraction pathways, and it was associated to coronary artery disease, myocardial infarction and keratometry measurements (Supplementary Tables 23 and 24), and expressed in a wide range of cell types, including fibroblasts, mesenchymal stem cells, osteoblasts, pancreatic stellate cells, cardiomyocytes, and adipocytes (Supplementary Figure 23). Lipids were clustered with chronotype and Alzheimer’s disease, and were significantly associated with several modules expressed mainly in brain cell types, including LV93 (Supplementary Figure 24, and Supplementary Tables 25 and 26), LV206 (Supplementary Figure 25, and Supplementary Tables 27 and 28), and LV260 (Supplementary Figure 26, and Supplementary Tables 29 and 30). These modules were associated mainly with cardiovascular traits in eMERGE.

Within the cardiovascular sub-branch, we also found mental and neurodevelopmental disorders such as Alzheimer’s disease, schizophrenia, and attention deficit hyperactivity disorder (ADHD). These disorders were previously linked to the cardiovascular system [56,57,58,59], and share several risk factors, including hypertension, high cholesterol, obesity, smoking, among others [60,61]. In our results, however, these diseases were grouped by potentially shared transcriptional processes expressed in specific tissues/cell types. Alzheimer’s disease, for example, was significantly associated with LV21 (FDR < 1e-18) and with LV5 (FDR < 0.01) (Supplementary Tables 31 and 33). LV21 was strongly expressed in a variety of soft tissue sarcomas, monocytes/macrophages (including microglia from cortex samples), and aortic valves (Supplementary Figure 27); as discussed previously, macrophages play a key role in the reverse cholesterol transport and thus atherogenesis [62]. LV5 was expressed in breast cancer and brain glioma samples, microglia (cortex), liver, and kidney, among other cell types (Supplementary Figure 28). LV21 and LV5 were also strongly associated with lipids:

LDL cholesterol (FDR < 0.001) and triglycerides (FDR < 0.05 and FDR < 0.001, respectively). Additionally, LV5 was associated with depression traits from the UK Biobank. ADHD was the only significantly associated trait for LV434 (FDR < 0.01) (Supplementary Table 35), which was expressed in breast cancer and glioma cells, cerebral organoids, and several different cell populations from the brain: fetal neurons (replicating and quiescence), microglia, and astrocytes (Supplementary Figure 29). Schizophrenia was not significantly associated with any gene module tested in our analysis. None of these LVs were aligned to prior pathways, which might represent potentially novel transcriptional processes affecting the cardiovascular and central nervous system.

Discussion

We have introduced a novel computational approach that can map existing TWAS results into a latent gene expression representation to infer cell type-specific effects for complex diseases. Leveraging one of the largest gene expression compendia, we integrated gene co-expression patterns on tens of thousands of RNA-seq samples covering different tissues and cell types under a variety of stimuli and developmental stages. We showed that this latent representation is robust to capture relevant associations even when known single gene-trait effects are not detected. Moreover, projecting gene-trait and gene-drug associations into this common representation can better link pathophysiological processes with the mechanisms of action of drugs, which led to a more accurate prediction of known disease treatments. Finally, we showed how the approach could be used to group diseases and traits affected by the same transcriptional processes, which showed disease-specific modules expressed in highly relevant tissues.

The key difference of our approach with other alternatives is that we can incorporate patterns learned in RNA-seq datasets with tissues and cell types measured in a large variety of conditions. As we showed, gene modules are expressed in very specific cell types, and this has the advantage of identifying potentially causal cell types with more precision. We provide examples of modules clearly expressed in one type of cell (such as adipose in LV246, thyroid in LV155, or ovary in LV66), whereas others seem to be activated across different contexts and thus capture a more realistic multi-tissue scenario for complex diseases (for example, LV136 is associated with coronary artery disease and expressed in fibroblasts, osteoblasts, pancreas, liver, and cardiomyocytes). To our knowledge, this is a novel approach for the identification of tissue and cell type-specific effects that is computationally simple to implement.

Our approach assumes that gene modules with coordinated expression controlled by a perturbed transcriptional process will also manifest coordinated pathological effects on relevant diseases. This leads to several limitations related to 1) the unsupervised learning process that detects modules of co-expressed genes (MultiPLIER), and 2) the statistical methods that link SNP-trait to gene-trait associations through eQTL (TWAS/PhenomeXcan). First, MultiPLIER applies a matrix factorization algorithm to extract latent variables/modules across a large expression dataset. These modules could represent real transcriptional processes or technical factors ("batch effects"). Although the deconvolution process successfully uses prior knowledge (pathways) to segregate noise from authentic biological patterns, several detected modules could be either novel, in the best scenario, or just noise. On the other hand, the underlying factorization method is restricted to linear combinations, which could miss important and more complex co-expression patterns. Second, TWAS approaches have several limitations that can lead to false positives [63,64]. Like GWAS, which generally detects groups of associated variants in LD (linkage disequilibrium), TWAS usually identifies several genes within the same locus [14,65]. This is due to sharing of GWAS variants in gene expression models, to correlated expression of nearby genes, or even correlation of their predicted expression due to eQTLs in LD, among others [63]. These two factors can combine to produce spurious module-trait associations, but our results suggest that the first aspect (module detection) could be the most problematic and probably more challenging to address.

Our approach is also in line with previous studies showing that drugs with genetic support are more likely to succeed through the drug development pipeline [37,66]. We show that a gene module perspective can more accurately predict known therapeutics and potentially serve as a more effective drug repositioning framework than considering single-gene effects. This outcome is explained by the biologically meaningful patterns captured by the latent gene expression space. The projection of single gene-trait associations and gene-compound profiles into this common representation translated into a stronger link between pathophysiological processes with the mechanisms of action of drugs.

Our consensus clustering pipeline combined different algorithms and assumptions on data structure to deliver highly stable groups of traits differentiated by relevant transcriptional processes. Considering groups of related diseases was previously shown to be more powerful to detect shared genetic etiology [67,68]; Our low-dimensional expression representation was used to detect groups of diseases previously suspected to share common pathophysiological mechanisms, such as cardiovascular, autoimmune, and mental diseases. The approach, however, also has several limitations related to 1) the data we used for grouping traits and 2) the fit of the clustering algorithms used for highly multi-factorial traits. First, we used the S-MultiXcan associations, which only provide the association strength between a gene and a trait, but with no direction of effect. This makes groups of traits harder to interpret since we can only infer that transcriptional processes are associated with traits in the same cluster, but they could actually have opposite effects for some traits than others. Second, we employed hard-partitioning algorithms (one trait belongs exclusively to one cluster) where the distance between two traits takes into account all gene modules. However, it is likely for two complex diseases to share just a few biological processes instead of being similar across all of them. These more complex scenarios cannot be captured by our clustering pipeline, as it can be seen for traits in the “complex” branch, which are hard to categorize into a single cluster or branch. This limitation was circumvented to some extent by using the hierarchical structure provided by clustering trees.

The proposed approach will benefit from larger gene expression datasets on different tissues and cell types measured on highly heterogeneous conditions. High-quality attributes for RNA-seq samples are critical for a successful application of our method, and we welcome current efforts to improve this situation [69]. This will enable the identification of more precise tissue- and cell lineage-specific effects on complex diseases.

Methods

PhenomeXcan: gene-based associations on 4,091 traits

We used TWAS results from PhenomeXcan [24] on 4091 traits for 22515 genes. PhenomeXcan was built using publicly available GWAS summary statistics to compute 1) gene-based associations with the PrediXcan family of methods [13,14,70], and 2) a posterior probability of colocalization between GWAS loci and *cis*-eQTL with fastENLOC [24,71]. The PrediXcan family of methods first build prediction models using data from the Genotype-Tissue Expression project (GTEx v8) [4] for gene expression imputation and then correlate this predicted expression with the phenotype of interest. This family is comprised of S-PrediXcan [70] (which computes a gene-tissue-trait association using GWAS as input) and S-MultiXcan [13] (which computes a gene-trait association by aggregating evidence of associations across all tissues).

We refer to the standardized effect sizes (*z*-scores) of S-PrediXcan across n traits and m genes in tissue t as $\mathbf{M}^t \in \mathbb{R}^{n \times m}$. For S-MultiXcan we do not have the direction of effect, and we used the p -values converted to *z*-scores $\mathbf{M} = \Phi^{-1}(1 - p/2)$, where Φ^{-1} is the probit function. Higher *z*-scores correspond to stronger associations.

MultiPLIER and Pathway-level information extractor (PLIER)

MultiPLIER [25] extracts patterns of co-expressed genes from recount2 [8], a large gene expression dataset. The approach applies the pathway-level information extractor method (PLIER) [23], which performs unsupervised learning using prior knowledge (canonical pathways) to reduce technical noise. Via a matrix factorization approach, PLIER deconvolutes the gene expression data into a set of latent variables (LV), where each represents a gene module. This reduced the data dimensionality into 987 latent variables or gene modules.

Given a gene expression dataset $\mathbf{Y}^{m \times c}$ with m genes and c experimental conditions and a prior knowledge matrix $\mathbf{C} \in \{0, 1\}^{m \times p}$ for p MSigDB pathways [72] (so that $\mathbf{C}_{ij} = 1$ if gene i belongs to pathway j), PLIER finds \mathbf{U} , \mathbf{Z} , and \mathbf{B} minimizing

$$\|\mathbf{Y} - \mathbf{Z}\mathbf{B}\|_F^2 + \lambda_1 \|\mathbf{Z} - \mathbf{C}\mathbf{U}\|_F^2 + \lambda_2 \|\mathbf{B}\|_F^2 + \lambda_3 \|\mathbf{U}\|_{L^1} \quad (1)$$

subject to $\mathbf{U} > 0$, $\mathbf{Z} > 0$; $\mathbf{Z}^{m \times l}$ are the gene loadings with l latent variables, $\mathbf{B}^{l \times c}$ is the latent space for c conditions, $\mathbf{U}^{p \times l}$ specifies which of the p prior-information pathways in \mathbf{C} are represented for each LV, and λ_i are different regularization parameters used in the training step. \mathbf{Z} is a low-dimensional representation of the gene space where each LV aligns as much as possible to prior knowledge, and it might represent either a known or novel gene module (i.e., a meaningful biological pattern) or noise.

We projected \mathbf{M} (either from S-PrediXcan across each tissue, or S-MultiXcan) into the low-dimensional gene module space learned by MultiPLIER using:

$$\hat{\mathbf{M}} = (\mathbf{Z}^\top \mathbf{Z} + \lambda_2 \mathbf{I})^{-1} \mathbf{Z}^\top \mathbf{M}. \quad (2)$$

where in $\hat{\mathbf{M}}^{l \times n}$ all traits in PhenomeXcan are now described by gene modules.

CRISPR-Cas9 screening

Add details

Gene module-trait associations

To compute an association between a gene module and a trait, we used an approach similar to the gene-property analysis in MAGMA [30], which is essentially a competitive test using gene weights from \mathbf{Z} to predict gene z -scores from \mathbf{M} . Thus, the regression model uses genes as data points by fitting $\mathbf{m} = \beta_0 + \mathbf{z}\beta_z + \epsilon$, where $\epsilon \sim \text{MVN}(0, \hat{\Sigma})$, \mathbf{m} are gene p -values (for a trait) from S-MultiXcan that we transformed to z -scores as mentioned before. Since we are only interested in whether genes with a stronger membership to a module (highest weights) are more associated with the phenotype, we performed a one-sided test on the coefficient β_z with the null hypothesis of $\beta_z = 0$ against the alternative $\beta_z > 0$. Since the error terms ϵ could be correlated due to correlation between predicted expression, we used a generalized least squares approach instead of standard linear regression. To calculate $\hat{\Sigma}$, we first estimated the correlation of predicted expression for each gene pair $(\mathbf{t}_i, \mathbf{t}_j)$ in tissue t using equations from [13, 70]:

$$\begin{aligned}
\hat{\Sigma}_{ij}^t &= \text{Cor}(\mathbf{t}_i, \mathbf{t}_j) \\
&= \frac{\text{Cov}(\mathbf{t}_i, \mathbf{t}_j)}{\sqrt{\widehat{\text{var}}(\mathbf{t}_i)\widehat{\text{var}}(\mathbf{t}_j)}} \\
&= \frac{\text{Cov}(\sum_{a \in \text{model}_i} w_a^i X_a, \sum_{b \in \text{model}_j} w_b^j X_b)}{\sqrt{\widehat{\text{var}}(\mathbf{t}_i)\widehat{\text{var}}(\mathbf{t}_j)}} \\
&= \frac{\sum_{\substack{a \in \text{model}_i \\ b \in \text{model}_j}} w_a^i w_b^j \text{Cov}(X_a, X_b)}{\sqrt{\widehat{\text{var}}(\mathbf{t}_i)\widehat{\text{var}}(\mathbf{t}_j)}} \\
&= \frac{\sum_{\substack{a \in \text{model}_i \\ b \in \text{model}_j}} w_a^i w_b^j \Gamma_{ab}}{\sqrt{\widehat{\text{var}}(\mathbf{t}_i)\widehat{\text{var}}(\mathbf{t}_j)}},
\end{aligned} \tag{3}$$

where $\Gamma = \widehat{\text{var}}(\mathbf{X}) = (\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{X} - \bar{\mathbf{X}})/(m - 1)$ is the genotype covariance matrix using 1000 Genomes Project data [73,74]. The variances for predicted gene expression of gene i is estimated as:

$$\begin{aligned}
\widehat{\text{var}}(\mathbf{t}_i) &= (\mathbf{W}^i)^\top \Gamma^i \mathbf{W}^i \\
&= \sum_{\substack{a \in \text{model}_i \\ b \in \text{model}_i}} w_a^i w_b^i \Gamma_{ab}^i.
\end{aligned} \tag{4}$$

Finally, $\hat{\Sigma} = \sum_t \hat{\Sigma}^t / |t|$ where $|t|=49$ is the number of tissues.

Drug-disease prediction

For the drug-disease prediction, we used a method based on a drug repositioning framework previously used for psychiatry traits [37] where gene-trait associations are anticorrelated with expression profiles for drugs. For the single-gene approach, we computed a drug-disease score by multiplying each S-PrediXcan set of results in tissue t , \mathbf{M}^t , with the transcriptional responses profiled in LINCS L1000 [31], $\mathbf{L}^{c \times m}$ (for c compounds): $\mathbf{D}^{t,k} = -1 \cdot \mathbf{M}^{t,k} \mathbf{L}^\top$, where k refers to the number of most significant gene associations in \mathbf{M}^t for each trait. As suggested in [37], k could be either all genes or the top 50, 100, 250, and 500; then we average score ranks across all k and obtain \mathbf{D}^t . Finally, for each drug-disease pair, we took the maximum prediction score across all tissues:

$$\mathbf{D}_{ij} = \max\{\mathbf{D}_{ij}^t \mid \forall t\}.$$

The same procedure was used for the gene module-based approach, where we projected S-PrediXcan results into our latent representation, leading to $\hat{\mathbf{M}}^t$; and also \mathbf{L} , leading to $\hat{\mathbf{L}}^{l \times c}$. Finally, $\mathbf{D}^{t,k} = -1 \cdot \hat{\mathbf{M}}^{t,k} \hat{\mathbf{L}}^\top$, where in this case k could be all LVs or the top 5, 10, 25 and 50 (since have an order of magnitude less LVs than genes).

Since the gold standard of drug-disease medical indications used contained Disease Ontology IDs (DOID) [75], we mapped PhenomeXcan traits to the Experimental Factor Ontology [76] using [77], and then to DOID.

Consensus clustering of traits

We performed two preprocessing steps on the S-MultiXcan results before the cluster analysis procedure. First, we combined results in \mathbf{M} (S-MultiXcan) for traits that mapped to the same Experimental Factor Ontology (EFO) [76] term using the Stouffer's method: $\sum w_i M_{ij} / \sqrt{\sum w_i^2}$, where w_i is a weight based on the GWAS sample size for trait i , and M_{ij} is the z -score for gene j . Second, we standardized all z -scores for each trait i by their sum to reduce the effect of highly polygenic traits: $M_{ij} / \sum M_{ij}$. Finally, we projected this data matrix using Equation 2, obtaining $\hat{\mathbf{M}}$ with $n=3752$ traits and $l=987$ LVs as the input of our clustering pipeline.

A partitioning of $\hat{\mathbf{M}}$ with n traits into k clusters is represented as a label vector $\pi \in \mathbb{N}^n$. Consensus clustering approaches consist of two steps: 1) the generation of an ensemble Π with r partitions of the dataset: $\Pi = \{\pi_1, \pi_2, \dots, \pi_r\}$, and 2) the combination of the ensemble into a consolidated solution defined as:

$$\pi^* = \arg \max_{\hat{\pi}} Q(\{|\mathcal{L}^i| \phi(\hat{\pi}_{\mathcal{L}^i}, \pi_{i\mathcal{L}^i}) \mid i \in \{1, \dots, r\}\}), \quad (5)$$

where \mathcal{L}^i is a set of data indices with known cluster labels for partition i , $\phi: \mathbb{N}^n \times \mathbb{N}^n \rightarrow \mathbb{R}$ is a function that measures the similarity between two partitions, and Q is a measure of central tendency, such as the mean or median. We used the adjusted Rand index (ARI) [78] for ϕ , and the median for Q . To obtain π^* , we define a consensus function $\Gamma: \mathbb{N}^{n \times r} \rightarrow \mathbb{N}^n$ with Π as the input. We used consensus functions based on the evidence accumulation clustering (EAC) paradigm [79], where Π is first transformed into a distance matrix $\mathbf{D}_{ij} = d_{ij}/r$, where d_{ij} is the number of times traits i and j were grouped in different clusters across all r partitions in Π . Then, Γ can be any similarity-based clustering algorithm, which is applied on \mathbf{D} to derive the final partition π^* .

For the ensemble generation step, we used different algorithms to create a highly diverse set of partitions (see Figure 4), since diversity is an important property for ensembles [49, 80, 81]. We used three data representations: the raw dataset, its projection into the top 50 principal components, and the embedding learned by UMAP [82] using 50 components. For each of these, we applied five clustering algorithms, covering a wide range of different assumptions on the data structure: k -means [83], spectral clustering [84], a Gaussian mixture model (GMM), hierarchical clustering, and DBSCAN [85]. For k -means, spectral clustering and GMM, we specified a range of k between 2 and $\sqrt{n} \approx 60$, and for each k we generated five partitions using random seeds. For hierarchical clustering, for each k we generated four partitions using four common linkage criteria: ward, complete, average and single. For DBSCAN, we combined different ranges for parameters ϵ (the maximum distance between two data points to be considered part of the same neighborhood) and $minPts$ (the minimum number of data points in a neighborhood for a data point to be considered a core point). Specifically, we used $minPts$ values from 2 to 125, and for each data version, we determined a plausible range of ϵ values by observing the distribution of the mean distance of the $minPts$ -nearest neighbors across all data points. Since some combinations of $minPts$ and ϵ might not produce a meaningful partition (for instance, when all points are detected as noisy or only one cluster is found), we resampled partitions generated by DBSCAN to ensure an equal representation in the ensemble. This procedure generated a final ensemble of 4428 partitions.

Finally, we used spectral clustering on \mathbf{D} to derive the final consensus partitions. \mathbf{D} was first transformed into a similarity matrix by applying an RBF kernel $\exp(-\gamma \mathbf{D}^2)$ using four different values for γ that we empirically determined to work best. Thus for each k between 2 and 60, we derived four consensus partitions and selected the one that maximized Equation 5. We further filtered this set of

59 solutions to keep only those with an ensemble agreement larger than the 75th percentile, leaving a total of 15 final consensus partitions shown in Figure 5.

Clusters interpretation

We used a supervised learning approach to interpret clustering results by detecting which gene modules are the most important for clusters of traits. For this, we used the highest resolution partition ($k=29$, although any could be used) to train a decision tree model using each of the clusters as labels and the projected data $\hat{\mathbf{M}}$ as the training samples. For each k , we built a set of binary labels with the current cluster's traits as the positive class and the rest of the traits as the negative class. Then, we selected the LV in the root node of the trained model only if its threshold was positive and larger than one standard deviation. Next, we removed this LV from $\hat{\mathbf{M}}$ (regardless of being previously selected or not) and trained the model again. We repeated this procedure 20 times to extract the top 20 LVs that better discriminate traits in a cluster from the rest.

References

1. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes.

James J Cai, Dmitri A Petrov

Genome biology and evolution (2010-07-12) <https://www.ncbi.nlm.nih.gov/pubmed/20624743>

DOI: [10.1093/gbe/evq019](https://doi.org/10.1093/gbe/evq019) · PMID: [20624743](https://pubmed.ncbi.nlm.nih.gov/20624743/) · PMCID: [PMC2997544](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC2997544/)

2. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes.

Eitan E Winter, Leo Goodstadt, Chris P Ponting

Genome research (2004-01) <https://www.ncbi.nlm.nih.gov/pubmed/14707169>

DOI: [10.1101/gr.1924004](https://doi.org/10.1101/gr.1924004) · PMID: [14707169](https://pubmed.ncbi.nlm.nih.gov/14707169/) · PMCID: [PMC314278](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC314278/)

3. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes

K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, S. Brunak

Proceedings of the National Academy of Sciences (2008-12-22) <https://doi.org/d5qcv9>

DOI: [10.1073/pnas.0810772105](https://doi.org/10.1073/pnas.0810772105) · PMID: [19104045](https://pubmed.ncbi.nlm.nih.gov/19104045/) · PMCID: [PMC2606902](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC2606902/)

4. The GTEx Consortium atlas of genetic regulatory effects across human tissues

The GTEx Consortium

Science (2020-09-11) <https://doi.org/ghbnhr>

DOI: [10.1126/science.aaz1776](https://doi.org/10.1126/science.aaz1776) · PMID: [32913098](https://pubmed.ncbi.nlm.nih.gov/32913098/) · PMCID: [PMC7737656](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC7737656/)

5. Index and biological spectrum of human DNase I hypersensitive sites

Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, ... John Stamatoyannopoulos
Nature (2020-07-29) <https://doi.org/gg6dhp>

DOI: [10.1038/s41586-020-2559-3](https://doi.org/10.1038/s41586-020-2559-3) · PMID: [32728217](https://pubmed.ncbi.nlm.nih.gov/32728217/) · PMCID: [PMC7422677](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC7422677/)

6. Mechanisms of tissue and cell-type specificity in heritable traits and diseases

Idan Heckselman, Esti Yeger-Lotem

Nature Reviews Genetics (2020-01-08) <https://doi.org/ggkx9v>

DOI: [10.1038/s41576-019-0200-9](https://doi.org/10.1038/s41576-019-0200-9) · PMID: [31913361](https://pubmed.ncbi.nlm.nih.gov/31913361/)

7. Regulatory genomic circuitry of human disease loci by integrative epigenomics

Carles A. Boix, Benjamin T. James, Yongjin P. Park, Wouter Meuleman, Manolis Kellis
Nature (2021-02-03) <https://doi.org/ghzkhr>

DOI: [10.1038/s41586-020-03145-z](https://doi.org/10.1038/s41586-020-03145-z) · PMID: [33536621](https://pubmed.ncbi.nlm.nih.gov/33536621/) · PMCID: [PMC7875769](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC7875769/)

8. Reproducible RNA-seq analysis using recount2

Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, Jeffrey T Leek

Nature Biotechnology (2017-04-11) <https://doi.org/gf75hp>

DOI: [10.1038/nbt.3838](https://doi.org/10.1038/nbt.3838) · PMID: [28398307](https://pubmed.ncbi.nlm.nih.gov/28398307/) · PMCID: [PMC6742427](https://pmc.ncbi.nlm.nih.gov/pmc/articles/PMC6742427/)

9. Massive mining of publicly available RNA-seq data from human and mouse

Alexander Lachmann, Denis Torre, Alexandra B. Keenan, Kathleen M. Jagodnik, Hoyjin J. Lee, Lily Wang, Moshe C. Silverstein, Avi Ma'ayan

Nature Communications (2018-04-10) <https://doi.org/gc92dr>
DOI: [10.1038/s41467-018-03751-6](https://doi.org/s41467-018-03751-6) · PMID: [29636450](#) · PMCID: [PMC5893633](#)

10. Identification of therapeutic targets from genetic association studies using hierarchical component analysis

Hao-Chih Lee, Osamu Ichikawa, Benjamin S. Glicksberg, Aparna A. Divaraniya, Christine E. Becker, Pankaj Agarwal, Joel T. Dudley

BioData Mining (2020-06-17) <https://doi.org/gjp5pf>

DOI: [10.1186/s13040-020-00216-9](https://doi.org/s13040-020-00216-9) · PMID: [32565911](#) · PMCID: [PMC7301559](#)

11. Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia

Shizhen Tang, Aron S. Buchman, Philip L. De Jager, David A. Bennett, Michael P. Epstein, Jingjing Yang

PLOS Genetics (2021-04-02) <https://doi.org/gjpr3j>

DOI: [10.1371/journal.pgen.1009482](https://doi.org/10.1371/journal.pgen.1009482) · PMID: [33798195](#) · PMCID: [PMC8046351](#)

12. Integrative approaches for large-scale transcriptome-wide association studies

Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC de Geus, Dorret I Boomsma, Fred A Wright, ... Bogdan Pasaniuc

Nature Genetics (2016-02-08) <https://doi.org/f3vf4p>

DOI: [10.1038/ng.3506](https://doi.org/ng.3506) · PMID: [26854917](#) · PMCID: [PMC4767558](#)

13. Integrating predicted transcriptome from multiple tissues improves association detection

Alvaro N. Barbeira, Milton Pividori, Jiamao Zheng, Heather E. Wheeler, Dan L. Nicolae, Hae Kyung Im

PLOS Genetics (2019-01-22) <https://doi.org/ghs8vx>

DOI: [10.1371/journal.pgen.1007889](https://doi.org/10.1371/journal.pgen.1007889) · PMID: [30668570](#) · PMCID: [PMC6358100](#)

14. A gene-based association method for mapping traits using reference transcriptome data

Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, ... GTEx Consortium

Nature Genetics (2015-08-10) <https://doi.org/f7p9zy>

DOI: [10.1038/ng.3367](https://doi.org/ng.3367) · PMID: [26258848](#) · PMCID: [PMC4552594](#)

15. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits

Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, Bogdan Pasaniuc

The American Journal of Human Genetics (2017-03) <https://doi.org/f9wvsg>

DOI: [10.1016/j.ajhg.2017.01.031](https://doi.org/10.1016/j.ajhg.2017.01.031) · PMID: [28238358](#) · PMCID: [PMC5339290](#)

16. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types

Hilary K. Finucane, Yakir A. Reshef, Verner Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shores, ... The Brainstorm Consortium

Nature Genetics (2018-04-09) <https://doi.org/gdfjqt>

DOI: [10.1038/s41588-018-0081-4](https://doi.org/s41588-018-0081-4) · PMID: [29632380](#) · PMCID: [PMC5896795](#)

17. Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets

Xinli Hu, Hyun Kim, Eli Stahl, Robert Plenge, Mark Daly, Soumya Raychaudhuri

The American Journal of Human Genetics (2011-10) <https://doi.org/fpghp4>

DOI: [10.1016/j.ajhg.2011.09.002](https://doi.org/10.1016/j.ajhg.2011.09.002) · PMID: [21963258](#) · PMCID: [PMC3188838](#)

18. **SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci**

Kamil Slowikowski, Xinli Hu, Soumya Raychaudhuri

Bioinformatics (2014-09-01) <https://doi.org/f6j6v3>

DOI: [10.1093/bioinformatics/btu326](https://doi.org/btu326) · PMID: [24813542](https://pubmed.ncbi.nlm.nih.gov/24813542/) · PMCID: [PMC4147889](https://pubmed.ncbi.nlm.nih.gov/PMC4147889/)

19. **Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine**

Padhraig Gormley, Verner Anttila, Bendik S Winsvold, Priit Palta, Tõnu Esko, Tune H Pers, Kai-How Farh, Ester Cuenca-Leon, Mikko Muona, Nicholas A Furlotte, ... International Headache Genetics Consortium

Nature Genetics (2016-06-20) <https://doi.org/bmzx>

DOI: [10.1038/ng.3598](https://doi.org/ng.3598) · PMID: [27322543](https://pubmed.ncbi.nlm.nih.gov/27322543/) · PMCID: [PMC5331903](https://pubmed.ncbi.nlm.nih.gov/PMC5331903/)

20. **Biological interpretation of genome-wide association studies using predicted gene functions**

Tune H. Pers, Juha M. Karjalainen, Yinglong Chan, Harm-Jan Westra, Andrew R. Wood, Jian Yang, Julian C. Lui, Sailaja Vedantam, Stefan Gustafsson, Tõnu Esko, ... Genetic Investigation of ANthropometric Traits (GIANT) Consortium

Nature Communications (2015-01-19) <https://doi.org/f3mwhd>

DOI: [10.1038/ncomms6890](https://doi.org/ncomms6890) · PMID: [25597830](https://pubmed.ncbi.nlm.nih.gov/25597830/) · PMCID: [PMC4420238](https://pubmed.ncbi.nlm.nih.gov/PMC4420238/)

21. **Estimating the causal tissues for complex traits and diseases**

Halit Onken, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, Emmanouil T Dermitzakis, GTEx Consortium

Nature Genetics (2017-10-23) <https://doi.org/ggrr72>

DOI: [10.1038/ng.3981](https://doi.org/ng.3981) · PMID: [29058715](https://pubmed.ncbi.nlm.nih.gov/29058715/)

22. **Tackling the widespread and critical impact of batch effects in high-throughput data**

Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, Rafael A. Irizarry

Nature Reviews Genetics (2010-09-14) <https://doi.org/cfr324>

DOI: [10.1038/nrg2825](https://doi.org/nrg2825) · PMID: [20838408](https://pubmed.ncbi.nlm.nih.gov/20838408/) · PMCID: [PMC3880143](https://pubmed.ncbi.nlm.nih.gov/PMC3880143/)

23. **Pathway-level information extractor (PLIER) for gene expression data**

Weiguang Mao, Elena Zaslavsky, Boris M. Hartmann, Stuart C. Sealfon, Maria Chikina

Nature Methods (2019-06-27) <https://doi.org/gf75g6>

DOI: [10.1038/s41592-019-0456-1](https://doi.org/s41592-019-0456-1) · PMID: [31249421](https://pubmed.ncbi.nlm.nih.gov/31249421/) · PMCID: [PMC7262669](https://pubmed.ncbi.nlm.nih.gov/PMC7262669/)

24. **PhenomeXcan: Mapping the genome to the phenotype through the transcriptome**

Milton Pividori, Padma S. Rajagopal, Alvaro Barbeira, Yanyu Liang, Owen Melia, Lisa Bastarache, YoSon Park, GTEx Consortium, Xiaoquan Wen, Hae K. Im

Science Advances (2020-09) <https://doi.org/ghbvbf>

DOI: [10.1126/sciadv.aba2083](https://doi.org/sciadv.aba2083) · PMID: [32917697](https://pubmed.ncbi.nlm.nih.gov/32917697/)

25. **MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease**

Jaclyn N. Taroni, Peter C. Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A. Merkel, Casey S. Greene

Cell Systems (2019-05) <https://doi.org/gf75g5>

DOI: [10.1016/j.cels.2019.04.003](https://doi.org/10.1016/j.cels.2019.04.003) · PMID: [31121115](https://pubmed.ncbi.nlm.nih.gov/31121115/) · PMCID: [PMC6538307](https://pubmed.ncbi.nlm.nih.gov/PMC6538307/)

26. **The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future**

Omri Gottesman, Helena Kuivaniemi, Gerard Tromp, W. Andrew Faustett, Rongling Li, Teri A. Manolio, Saskia C. Sanderson, Joseph Kannry, Randi Zinberg, Melissa A. Basford, ... and The

eMERGE Network

Genetics in Medicine (2013-06-06) <https://doi.org/f5dwbt>

DOI: [10.1038/gim.2013.72](https://doi.org/10.1038/gim.2013.72) · PMID: [23743551](https://pubmed.ncbi.nlm.nih.gov/23743551/) · PMCID: [PMC3795928](https://pubmed.ncbi.nlm.nih.gov/PMC3795928/)

27. The UK Biobank resource with deep phenotyping and genomic data

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, ... Jonathan Marchini

Nature (2018-10-10) <https://doi.org/gfb7h2>

DOI: [10.1038/s41586-018-0579-z](https://doi.org/10.1038/s41586-018-0579-z) · PMID: [30305743](https://pubmed.ncbi.nlm.nih.gov/30305743/) · PMCID: [PMC6786975](https://pubmed.ncbi.nlm.nih.gov/PMC6786975/)

28. Finding function: evaluation methods for functional genomic data

Chad L Myers, Daniel R Barrett, Matthew A Hibbs, Curtis Huttenhower, Olga G Troyanskaya

BMC Genomics (2006-07-25) <https://doi.org/fg6wnk>

DOI: [10.1186/1471-2164-7-187](https://doi.org/10.1186/1471-2164-7-187) · PMID: [16869964](https://pubmed.ncbi.nlm.nih.gov/16869964/) · PMCID: [PMC1560386](https://pubmed.ncbi.nlm.nih.gov/PMC1560386/)

29. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens

Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz Hamid, ... Iddo Friedberg

Genome Biology (2019-11-19) <https://doi.org/ggnxpz>

DOI: [10.1186/s13059-019-1835-8](https://doi.org/10.1186/s13059-019-1835-8) · PMID: [31744546](https://pubmed.ncbi.nlm.nih.gov/31744546/) · PMCID: [PMC6864930](https://pubmed.ncbi.nlm.nih.gov/PMC6864930/)

30. MAGMA: Generalized Gene-Set Analysis of GWAS Data

Christiaan A. de Leeuw, Joris M. Mooij, Tom Heskes, Danielle Posthuma

PLOS Computational Biology (2015-04-17) <https://doi.org/gf92gp>

DOI: [10.1371/journal.pcbi.1004219](https://doi.org/10.1371/journal.pcbi.1004219) · PMID: [25885710](https://pubmed.ncbi.nlm.nih.gov/25885710/) · PMCID: [PMC4401657](https://pubmed.ncbi.nlm.nih.gov/PMC4401657/)

31. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, ... Todd R. Golub

Cell (2017-11) <https://doi.org/cgwt>

DOI: [10.1016/j.cell.2017.10.049](https://doi.org/10.1016/j.cell.2017.10.049) · PMID: [29195078](https://pubmed.ncbi.nlm.nih.gov/29195078/) · PMCID: [PMC5990023](https://pubmed.ncbi.nlm.nih.gov/PMC5990023/)

32. DrugBank 4.0: shedding new light on drug metabolism

Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, ... David S. Wishart

Nucleic Acids Research (2014-01) <https://doi.org/f3mn6d>

DOI: [10.1093/nar/gkt1068](https://doi.org/10.1093/nar/gkt1068) · PMID: [24203711](https://pubmed.ncbi.nlm.nih.gov/24203711/) · PMCID: [PMC3965102](https://pubmed.ncbi.nlm.nih.gov/PMC3965102/)

33. Systematic integration of biomedical knowledge prioritizes drugs for repurposing

Daniel Scott Himmelstein, Antoine Lizée, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini

eLife (2017-09-22) <https://doi.org/cdfk>

DOI: [10.7554/elife.26726](https://doi.org/10.7554/elife.26726) · PMID: [28936969](https://pubmed.ncbi.nlm.nih.gov/28936969/) · PMCID: [PMC5640425](https://pubmed.ncbi.nlm.nih.gov/PMC5640425/)

34. Dhimmel/Lincs V2.0: Refined Consensus Signatures From Lincs L1000

Daniel Himmelstein, Leo Brueggeman, Sergio Baranzini

Zenodo (2016-03-08) <https://doi.org/f3mqvr>

DOI: [10.5281/zenodo.47223](https://doi.org/10.5281/zenodo.47223)

35. Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease

J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J.

Pasricha, A. J. Butte

Science Translational Medicine (2011-08-17) <https://doi.org/bmh5ts>

DOI: [10.1126/scitranslmed.3002648](https://doi.org/10.1126/scitranslmed.3002648) · PMID: [21849664](https://pubmed.ncbi.nlm.nih.gov/21849664/) · PMCID: [PMC3479650](https://pubmed.ncbi.nlm.nih.gov/PMC3479650/)

36. Discovery and Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data

M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, A. J. Butte

Science Translational Medicine (2011-08-17) <https://doi.org/c3fwxv>

DOI: [10.1126/scitranslmed.3001318](https://doi.org/10.1126/scitranslmed.3001318) · PMID: [21849665](https://pubmed.ncbi.nlm.nih.gov/21849665/) · PMCID: [PMC3502016](https://pubmed.ncbi.nlm.nih.gov/PMC3502016/)

37. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry

Hon-Cheong So, Carlos Kwan-Long Chau, Wan-To Chiu, Kin-Sang Ho, Cho-Pong Lo, Stephanie Ho-Yue Yim, Pak-Chung Sham

Nature Neuroscience (2017-08-14) <https://doi.org/gbrssh>

DOI: [10.1038/nn.4618](https://doi.org/10.1038/nn.4618) · PMID: [28805813](https://pubmed.ncbi.nlm.nih.gov/28805813/)

38. Dhimmel/Indications V1.0. Pharmacotherapydb: The Open Catalog Of Drug Therapies For Disease

Daniel S. Himmelstein, Pouya Khankhanian, Christine S. Hessler, Ari J. Green, Sergio E. Baranzini

Zenodo (2016-03-15) <https://doi.org/f3mqwb>

DOI: [10.5281/zenodo.47664](https://doi.org/10.5281/zenodo.47664)

39. Mechanism of Action of Niacin

Vaijinath S. Kamanna, Moti L. Kashyap

The American Journal of Cardiology (2008-04) <https://doi.org/c8zwdt>

DOI: [10.1016/j.amjcard.2008.02.029](https://doi.org/10.1016/j.amjcard.2008.02.029) · PMID: [18375237](https://pubmed.ncbi.nlm.nih.gov/18375237/)

40. Niacin: an old lipid drug in a new NAD⁺ dress

Mario Romani, Dina Carina Hofer, Elena Katsyuba, Johan Auwerx

Journal of Lipid Research (2019-04) <https://doi.org/gjpjft>

DOI: [10.1194/jlr.s092007](https://doi.org/10.1194/jlr.s092007) · PMID: [30782960](https://pubmed.ncbi.nlm.nih.gov/30782960/) · PMCID: [PMC6446705](https://pubmed.ncbi.nlm.nih.gov/PMC6446705/)

41. The nicotinic acid receptor GPR109A (HM74A or PUMA-G) as a new therapeutic target

S OFFERMANNS

Trends in Pharmacological Sciences (2006-07) <https://doi.org/fgb4tr>

DOI: [10.1016/j.tips.2006.05.008](https://doi.org/10.1016/j.tips.2006.05.008) · PMID: [16766048](https://pubmed.ncbi.nlm.nih.gov/16766048/)

42. Langerhans Cells Release Prostaglandin D2 in Response to Nicotinic Acid

Dominique Maciejewski-Lenoir, Jeremy G. Richman, Yaron Hakak, Ibragim Gaidarov, Dominic P. Behan, Daniel T. Connolly

Journal of Investigative Dermatology (2006-12) <https://doi.org/dgwg75>

DOI: [10.1038/sj.jid.5700586](https://doi.org/10.1038/sj.jid.5700586) · PMID: [17008871](https://pubmed.ncbi.nlm.nih.gov/17008871/)

43. Nicotinic acid inhibits progression of atherosclerosis in mice through its receptor GPR109A expressed by immune cells

Martina Lukasova, Camille Malaval, Andreas Gille, Jukka Kero, Stefan Offermanns

Journal of Clinical Investigation (2011-03-01) <https://doi.org/cqftcq>

DOI: [10.1172/jci41651](https://doi.org/10.1172/jci41651) · PMID: [21317532](https://pubmed.ncbi.nlm.nih.gov/21317532/) · PMCID: [PMC3048854](https://pubmed.ncbi.nlm.nih.gov/PMC3048854/)

44. Role of HDL, ABCA1, and ABCG1 Transporters in Cholesterol Efflux and Immune Responses

Laurent Yvan-Charvet, Nan Wang, Alan R. Tall

Arteriosclerosis, Thrombosis, and Vascular Biology (2010-02) <https://doi.org/ds23w6>
DOI: [10.1161/atvaha.108.179283](https://doi.org/10.1161/atvaha.108.179283) · PMID: [19797709](https://pubmed.ncbi.nlm.nih.gov/19797709/) · PMCID: [PMC2812788](https://pubmed.ncbi.nlm.nih.gov/PMC2812788/)

45. Niacin Inhibits Apoptosis and Rescues Premature Ovarian Failure

Shufang Wang, Min Sun, Ling Yu, Yixuan Wang, Yuanqing Yao, Deqing Wang
Cellular Physiology and Biochemistry (2018) <https://doi.org/gfqvcq>
DOI: [10.1159/000495051](https://doi.org/10.1159/000495051) · PMID: [30415247](https://pubmed.ncbi.nlm.nih.gov/30415247/)

46. Chronic niacin administration ameliorates ovulation, histological changes in the ovary and adiponectin concentrations in a rat model of polycystic ovary syndrome

Negin Asadi, Mahin Izadi, Ali Aflatounian, Mansour Esmaeili-Dehaj, Mohammad Ebrahim Rezvani, Zeinab Hafizi
Reproduction, Fertility and Development (2021) <https://doi.org/gjpjkt>
DOI: [10.1071/rd20306](https://doi.org/10.1071/rd20306) · PMID: [33751926](https://pubmed.ncbi.nlm.nih.gov/33751926/)

47. Clustering ensembles: models of consensus and weak partitions

A. Topchy, A. K. Jain, W. Punch
IEEE Transactions on Pattern Analysis and Machine Intelligence (2005-12) <https://doi.org/c8z32x>
DOI: [10.1109/tpami.2005.237](https://doi.org/10.1109/tpami.2005.237) · PMID: [16355656](https://pubmed.ncbi.nlm.nih.gov/16355656/)

48. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions

Alexander Strehl, Ghosh Joydeep
Journal of Machine Learning Research (2002) <https://www.jmlr.org/papers/v3/strehl02a.html>

49. Diversity control for improving the analysis of consensus clustering

Milton Pividori, Georgina Stegmayer, Diego H. Milone
Information Sciences (2016-09) <https://doi.org/ghtqbk>
DOI: [10.1016/j.ins.2016.04.027](https://doi.org/10.1016/j.ins.2016.04.027)

50. Clustering trees: a visualization for evaluating clusterings at multiple resolutions

Luke Zappia, Alicia Oshlack
GigaScience (2018-07) <https://doi.org/gfzqf5>
DOI: [10.1093/gigascience/giy083](https://doi.org/10.1093/gigascience/giy083) · PMID: [30010766](https://pubmed.ncbi.nlm.nih.gov/30010766/) · PMCID: [PMC6057528](https://pubmed.ncbi.nlm.nih.gov/PMC6057528/)

51. Prognostic value of grip strength: findings from the Prospective Urban Rural Epidemiology (PURE) study.

Darryl P Leong, Koon K Teo, Sumathy Rangarajan, Patricio Lopez-Jaramillo, Alvaro Avezum, Andres Orlandini, Pamela Seron, Suad H Ahmed, Annika Rosengren, Roya Kelishadi, ...
Lancet (London, England) (2015-05-13) <https://www.ncbi.nlm.nih.gov/pubmed/25982160>
DOI: [10.1016/s0140-6736\(14\)62000-6](https://doi.org/10.1016/s0140-6736(14)62000-6) · PMID: [25982160](https://pubmed.ncbi.nlm.nih.gov/25982160/)

52. Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis

Noa Novershtern, Aravind Subramanian, Lee N. Lawton, Raymond H. Mak, W. Nicholas Haining, Marie E. McConkey, Naomi Habib, Nir Yosef, Cindy Y. Chang, Tal Shay, ... Benjamin L. Ebert
Cell (2011-01) <https://doi.org/cf5k92>
DOI: [10.1016/j.cell.2011.01.004](https://doi.org/10.1016/j.cell.2011.01.004) · PMID: [21241896](https://pubmed.ncbi.nlm.nih.gov/21241896/) · PMCID: [PMC3049864](https://pubmed.ncbi.nlm.nih.gov/PMC3049864/)

53. CD19⁺ CD5⁺ Cells as Indicators of Preeclampsia

Federico Jensen, Gerd Wallukat, Florian Herse, Oliver Budner, Tarek El-Mousleh, Serban-Dan Costa, Ralf Dechend, Ana Claudia Zenclussen
Hypertension (2012-04) <https://doi.org/gj36rs>
DOI: [10.1161/hypertensionaha.111.188276](https://doi.org/10.1161/hypertensionaha.111.188276) · PMID: [22353610](https://pubmed.ncbi.nlm.nih.gov/22353610/)

54. Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure

Christiaan A. de Leeuw, Sven Stringer, Ilona A. Dekkers, Tom Heskes, Danielle Posthuma
Nature Communications (2018-09-14) <https://doi.org/gd6d85>
DOI: [10.1038/s41467-018-06022-6](https://doi.org/10.1038/s41467-018-06022-6) · PMID: [30218068](#) · PMCID: [PMC6138636](#)

55. Estrogen and hypertension

Muhammad S. Ashraf, Wanpen Vongpatanasin
Current Hypertension Reports (2006-09) <https://doi.org/d638rf>
DOI: [10.1007/s11906-006-0080-1](https://doi.org/10.1007/s11906-006-0080-1) · PMID: [16965722](#)

56. Depression as a predictor for coronary heart disease. a review and meta-analysis.

Reiner Rugulies
American journal of preventive medicine (2002-07)
<https://www.ncbi.nlm.nih.gov/pubmed/12093424>
DOI: [10.1016/s0749-3797\(02\)00439-7](https://doi.org/10.1016/s0749-3797(02)00439-7) · PMID: [12093424](#)

57. Mental Disorders Across the Adult Life Course and Future Coronary Heart Disease

Catharine R. Gale, G. David Batty, David P. J. Osborn, Per Tynelius, Finn Rasmussen
Circulation (2014-01-14) <https://doi.org/qm4>
DOI: [10.1161/circulationaha.113.002065](https://doi.org/10.1161/circulationaha.113.002065) · PMID: [24190959](#) · PMCID: [PMC4107269](#)

58. Mortality gap for people with bipolar disorder and schizophrenia: UK-based cohort study 2000–2014

Joseph F. Hayes, Louise Marston, Kate Walters, Michael B. King, David P. J. Osborn
British Journal of Psychiatry (2018-01-02) <https://doi.org/gbwcjx>
DOI: [10.1192/bjp.bp.117.202606](https://doi.org/10.1192/bjp.bp.117.202606) · PMID: [28684403](#) · PMCID: [PMC5579328](#)

59. Getting to the Heart of Alzheimer Disease

Joshua M. Tublin, Jeremy M. Adelstein, Federica del Monte, Colin K. Combs, Loren E. Wold
Circulation Research (2019-01-04) <https://doi.org/gjzjgg>
DOI: [10.1161/circresaha.118.313563](https://doi.org/10.1161/circresaha.118.313563) · PMID: [30605407](#) · PMCID: [PMC6319653](#)

60. The overlap between vascular disease and Alzheimer's disease - lessons from pathology

Johannes Attems, Kurt A Jellinger
BMC Medicine (2014-11-11) <https://doi.org/f6pj4>
DOI: [10.1186/s12916-014-0206-2](https://doi.org/10.1186/s12916-014-0206-2) · PMID: [25385447](#) · PMCID: [PMC4226890](#)

61. Cardiovascular Risk Factors for Alzheimer's Disease

Clive Rosendorff, Michal S. Beeri, Jeremy M. Silverman
The American Journal of Geriatric Cardiology (2007-03) <https://doi.org/bpfw5d>
DOI: [10.1111/j.1076-7460.2007.06696.x](https://doi.org/10.1111/j.1076-7460.2007.06696.x) · PMID: [17483665](#)

62. Reverse cholesterol transport and cholesterol efflux in atherosclerosis

R. Ohashi, H. Mu, X. Wang, Q. Yao, C. Chen
QJM: An International Journal of Medicine (2005-12) <https://doi.org/dn2fgt>
DOI: [10.1093/qjmed/hci136](https://doi.org/10.1093/qjmed/hci136) · PMID: [16258026](#)

63. Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N. Barbeira, David A. Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, ... Anshul Kundaje

64. Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations

Abhay Hukku, Milton Pividori, Francesca Luca, Roger Pique-Regi, Hae Kyung Im, Xiaoquan Wen
The American Journal of Human Genetics (2021-01) <https://doi.org/gj58gg>
DOI: [10.1016/j.ajhg.2020.11.012](https://doi.org/10.1016/j.ajhg.2020.11.012) · PMID: [33308443](https://pubmed.ncbi.nlm.nih.gov/33308443/) · PMCID: [PMC7820626](https://pubmed.ncbi.nlm.nih.gov/PMC7820626/)

65. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights

Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K. Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, Steven McCarroll, Benjamin M. Neale, ... Schizophrenia Working Group of the Psychiatric Genomics Consortium

Nature Genetics (2018-04-09) <https://doi.org/gdfdf2>
DOI: [10.1038/s41588-018-0092-1](https://doi.org/10.1038/s41588-018-0092-1) · PMID: [29632383](https://pubmed.ncbi.nlm.nih.gov/29632383/) · PMCID: [PMC5942893](https://pubmed.ncbi.nlm.nih.gov/PMC5942893/)

66. The support of human genetic evidence for approved drug indications

Matthew R Nelson, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, ... Philippe Sanseau

Nature Genetics (2015-06-29) <https://doi.org/f3mn52>
DOI: [10.1038/ng.3314](https://doi.org/10.1038/ng.3314) · PMID: [26121088](https://pubmed.ncbi.nlm.nih.gov/26121088/)

67. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology

Manuel A Ferreira, Judith M Vonk, Hansjörg Baurecht, Ingo Marenholz, Chao Tian, Joshua D Hoffman, Quinta Helmer, Annika Tillander, Vilhelmina Ullemar, Jenny van Dongen, ... LifeLines Cohort Study

Nature Genetics (2017-10-30) <https://doi.org/gchg62>
DOI: [10.1038/ng.3985](https://doi.org/10.1038/ng.3985) · PMID: [29083406](https://pubmed.ncbi.nlm.nih.gov/29083406/) · PMCID: [PMC5989923](https://pubmed.ncbi.nlm.nih.gov/PMC5989923/)

68. A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases

Zhaozhong Zhu, Phil H. Lee, Mark D. Chaffin, Wonil Chung, Po-Ru Loh, Quan Lu, David C. Christiani, Liming Liang

Nature Genetics (2018-05-21) <https://doi.org/gdpmtn>
DOI: [10.1038/s41588-018-0121-0](https://doi.org/10.1038/s41588-018-0121-0) · PMID: [29785011](https://pubmed.ncbi.nlm.nih.gov/29785011/) · PMCID: [PMC5980765](https://pubmed.ncbi.nlm.nih.gov/PMC5980765/)

69. Systematic tissue annotations of -omics samples by modeling unstructured metadata

Nathaniel T. Hawkins, Marc Maldaveri, Anna Yannakopoulos, Lindsay A. Guare, Arjun Krishnan
Cold Spring Harbor Laboratory (2021-05-20) <https://doi.org/gj2pkc>
DOI: [10.1101/2021.05.10.443525](https://doi.org/10.1101/2021.05.10.443525)

70. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics

Alvaro N. Barbeira, Scott P. Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E. Wheeler, Jason M. Torres, Eric S. Torstenson, Kaanan P. Shah, Tzintzuni Garcia, Todd L. Edwards, ... GTEx Consortium

Nature Communications (2018-05-08) <https://doi.org/gdjvp5>
DOI: [10.1038/s41467-018-03621-1](https://doi.org/10.1038/s41467-018-03621-1) · PMID: [29739930](https://pubmed.ncbi.nlm.nih.gov/29739930/) · PMCID: [PMC5940825](https://pubmed.ncbi.nlm.nih.gov/PMC5940825/)

71. Probabilistic Colocalization of Genetic Variants from Complex and Molecular Traits: Promise and Limitations

Abhay Hukku, Milton Pividori, Francesca Luca, Roger Pique-Regi, Hae Kyung Im, Xiaoquan Wen

72. The Molecular Signatures Database Hallmark Gene Set Collection

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, Pablo Tamayo

Cell Systems (2015-12) <https://doi.org/gf78hq>

DOI: [10.1016/j.cels.2015.12.004](https://doi.org/10.1016/j.cels.2015.12.004) · PMID: [26771021](https://pubmed.ncbi.nlm.nih.gov/26771021/) · PMCID: [PMC4707969](https://pmcid.ncbi.nlm.nih.gov/pmc/articles/PMC4707969/)

73. A global reference for human genetic variation

The 1000 Genomes Project Consortium

Nature (2015-09-30) <https://doi.org/73d>

DOI: [10.1038/nature15393](https://doi.org/10.1038/nature15393) · PMID: [26432245](https://pubmed.ncbi.nlm.nih.gov/26432245/) · PMCID: [PMC4750478](https://pmcid.ncbi.nlm.nih.gov/pmc/articles/PMC4750478/)

74. GWAS summary statistics imputation support data and integration with PrediXcan MASHR

Alvaro Numa Barbeira, Hae Kyung Im

Zenodo (2019-12-10) <https://doi.org/gj6jwq>

DOI: [10.5281/zenodo.3657902](https://doi.org/10.5281/zenodo.3657902)

75. Human Disease Ontology 2018 update: classification, content and workflow expansion

Lynn M Schriml, Elvira Mitraka, James Munro, Becky Tauber, Mike Schor, Lance Nickle, Victor Felix, Linda Jeng, Cynthia Bearer, Richard Lichenstein, ... Carol Greene

Nucleic Acids Research (2019-01-08) <https://doi.org/ggx9wp>

DOI: [10.1093/nar/gky1032](https://doi.org/10.1093/nar/gky1032) · PMID: [30407550](https://pubmed.ncbi.nlm.nih.gov/30407550/) · PMCID: [PMC6323977](https://pmcid.ncbi.nlm.nih.gov/pmc/articles/PMC6323977/)

76. Modeling sample variables with an Experimental Factor Ontology

James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, Helen Parkinson

Bioinformatics (2010-04-15) <https://doi.org/dsb6vt>

DOI: [10.1093/bioinformatics/btq099](https://doi.org/10.1093/bioinformatics/btq099) · PMID: [20200009](https://pubmed.ncbi.nlm.nih.gov/20200009/) · PMCID: [PMC2853691](https://pmcid.ncbi.nlm.nih.gov/pmc/articles/PMC2853691/)

77. EBISPORT/EFO-UKB-mappings

EBISPORT

(2021-03-15) <https://github.com/EBISPORT/EFO-UKB-mappings>

78. Comparing partitions

Lawrence Hubert, Phipps Arabie

Journal of Classification (1985-12) <https://doi.org/bphmzh>

DOI: [10.1007/bf01908075](https://doi.org/10.1007/bf01908075)

79. Combining multiple clusterings using evidence accumulation

Ana L. N. Fred, Anil K. Jain

IEEE Transactions on Pattern Analysis and Machine Intelligence (2005-06) <https://doi.org/bsknv6>

DOI: [10.1109/tpami.2005.113](https://doi.org/10.1109/tpami.2005.113) · PMID: [15943417](https://pubmed.ncbi.nlm.nih.gov/15943417/)

80. A Link-Based Approach to the Cluster Ensemble Problem

Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, Chris Price

IEEE Transactions on Pattern Analysis and Machine Intelligence (2011-12) <https://doi.org/cqgkh3>

DOI: [10.1109/tpami.2011.84](https://doi.org/10.1109/tpami.2011.84) · PMID: [21576752](https://pubmed.ncbi.nlm.nih.gov/21576752/)

81. Hybrid clustering solution selection strategy

Zhiwen Yu, Le Li, Yunjun Gao, Jane You, Jiming Liu, Hau-San Wong, Guoqiang Han

82. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Leland McInnes, John Healy, James Melville
arXiv (2020-09-21) <https://arxiv.org/abs/1802.03426>

83. k-means++: the advantages of careful seeding

David Arthur, Sergei Vassilvitskii
Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (2007)
<http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf>

84. On Spectral Clustering: Analysis and an algorithm

Andrew Ng, Michael Jordan, Yair Weiss
Advances in Neural Information Processing Systems (2001)
<https://ai.stanford.edu/~ang/papers/nips01-spectral.pdf>

85. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu
Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (1996) <https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf>

86. Human pluripotent stem cell-derived neural constructs for predicting neural toxicity

Michael P. Schwartz, Zhonggang Hou, Nicholas E. Propson, Jue Zhang, Collin J. Engstrom, Vitor Santos Costa, Peng Jiang, Bao Kim Nguyen, Jennifer M. Bolin, William Daly, ... James A. Thomson
Proceedings of the National Academy of Sciences (2015-10-06) <https://doi.org/f7vpzd>
DOI: [10.1073/pnas.1516645112](https://doi.org/10.1073/pnas.1516645112) · PMID: [26392547](https://pubmed.ncbi.nlm.nih.gov/26392547/) · PMCID: [PMC4603492](https://pubmed.ncbi.nlm.nih.gov/PMC4603492/)

87. Homo sapiens (ID 232177) - BioProject - NCBI
<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA232177>

Supplementary material

Top latent variables associated with neutrophils

NOT FINISHED

This section aims to show that the top LVs related to neutrophil counts are more correlated to neutrophil counts or estimates than expected by chance. Probably I just need to add a proper caption for each figure, and reference them from the main text.

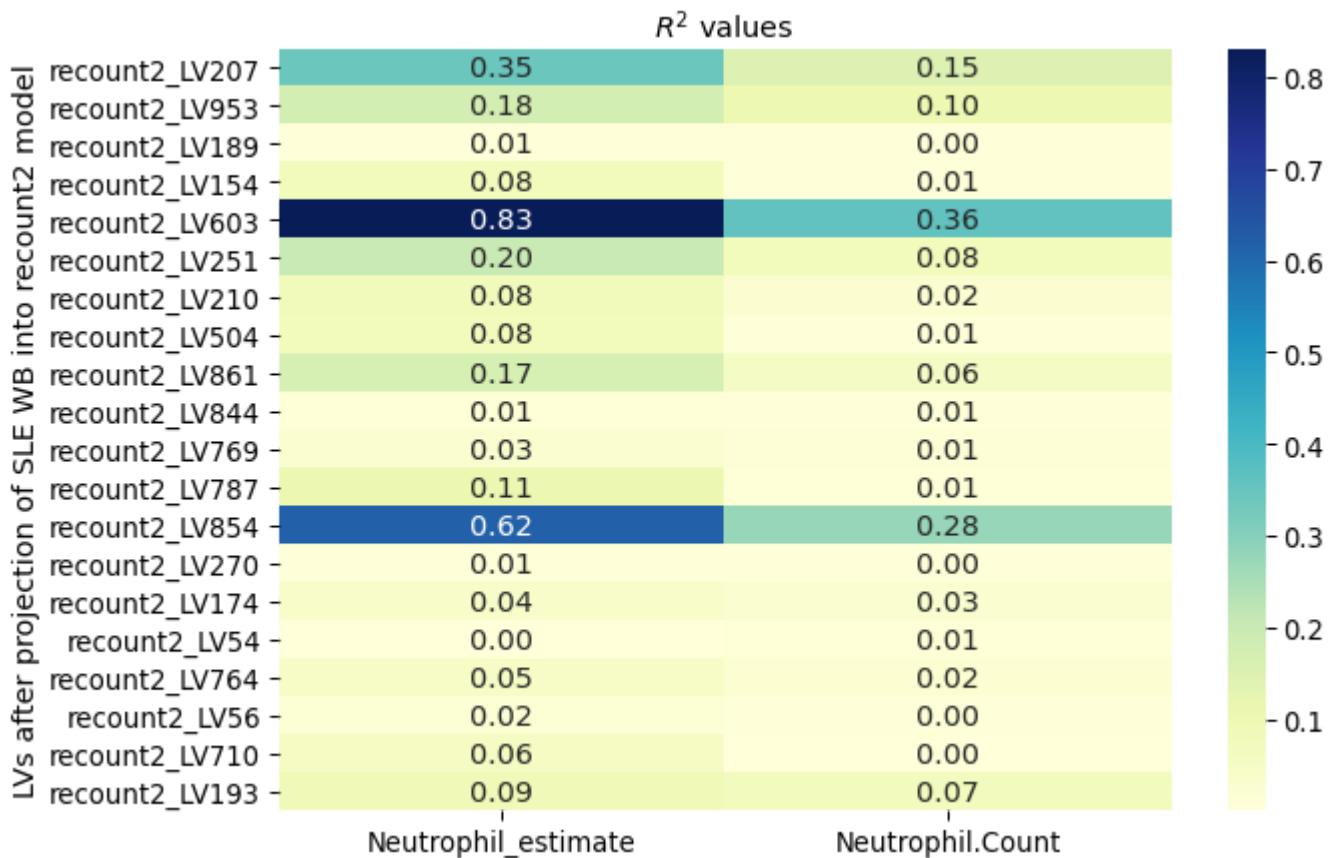


Figure 7: Correlation of neutrophil counts with top LVs associated with neutrophils traits.

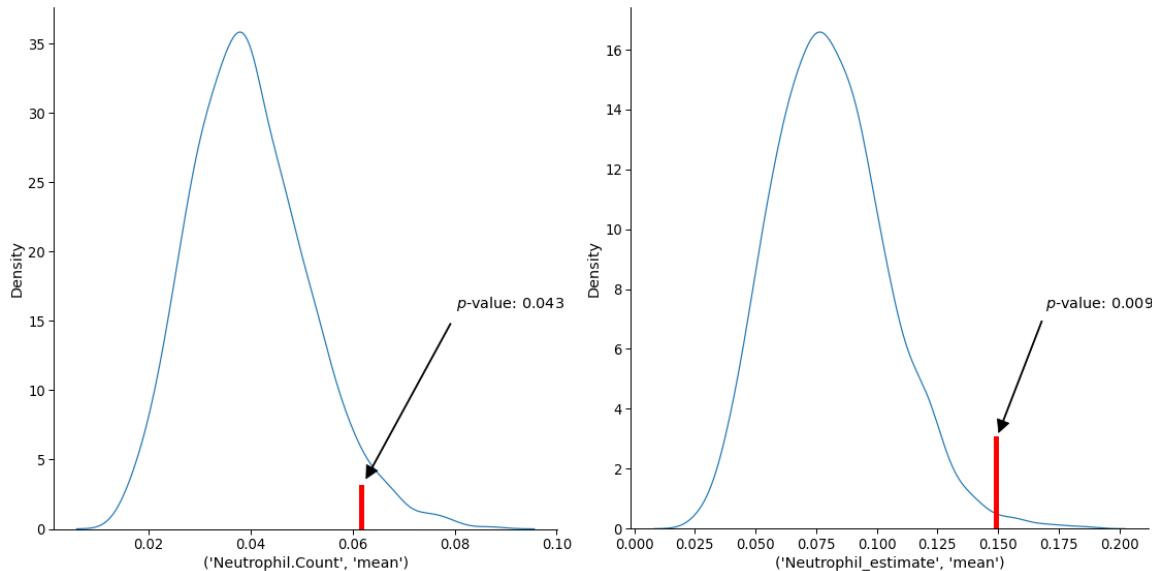


Figure 8: Significance of neutrophil counts correlation.

Gene modules enriched for lipids gene-sets

Table 1: Gene modules (LVs) enriched for lipids gene-sets found with CRISPR screening.

Gene module	Lipids gene-set	p-value	FDR
LV678	decrease	2.61e-07	2.57e-04
LV707	increase	1.74e-07	2.57e-04
LV905	increase	4.29e-05	2.82e-02
LV915	increase	6.37e-05	3.14e-02

Top traits across lipids-associated genes and modules

Table 2: Top 25 traits associated with genes from the lipids-decreasing gene-set found with CRISPR screening.

Order	Trait	Category
1	Vascular/heart problems diagnosed by doctor: High blood pressure	Diseases (cardiovascular)
2	Diastolic blood pressure, automated reading	Blood pressure
3	Non-cancer illness code, self-reported: hypertension	Diseases (cardiovascular)
4	Suggestive for eosinophilic asthma	Diseases (FinnGen)
5	Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Blood pressure medication	Medication
6	Forced expiratory volume in 1-second (FEV1), predicted	Spirometry
7	Vascular/heart problems diagnosed by doctor: None of the above	Diseases (cardiovascular)
8	Treatment/medication code: levothyroxine sodium	Medications
9	Haematocrit percentage	Blood count
10	Treatment/medication code: lisinopril	Medications
11	Haemoglobin concentration	Blood count
12	Job coding: counter clerk, bank clerk, cashier, post office clerk	Employment history
13	Acute alcohol intoxication	Diseases (FinnGen)
14	Systolic blood pressure, automated reading	Blood pressure
15	Platelet count	Blood count
16	Red Blood Cell Count	Blood
17	Peak expiratory flow (PEF)	Spirometry
18	Sitting height	Body size measures
19	Treatment/medication code: bendroflumethiazide	Medications
20	Age started wearing glasses or contact lenses	Eyesight
21	Comparative height size at age 10	Early life factors
22	Workplace very cold: Often	Employment history
23	Salt added to food	Diet
24	Difficulty concentrating during worst period of anxiety	Anxiety
25	Treatment/medication code: metformin	Medications

Table 3: Top 25 traits associated with gene modules (LVs) enriched for the lipids-decreasing gene-set found with CRISPR screening.

Order	Trait	Category
1	Non-cancer illness code, self-reported: malabsorption/coeliac disease	Diseases (gastrointestinal/abdominal)

Order	Trait	Category
2	Diastolic blood pressure, automated reading	Blood pressure
3	Immature reticulocyte fraction	Blood count
4	Treatment/medication code: ferrous salt product	Medications
5	Vascular/heart problems diagnosed by doctor: None of the above	Diseases (cardiovascular)
6	Platelet distribution width	Blood count
7	Unstable angina pectoris	Diseases (FinnGen)
8	Vascular/heart problems diagnosed by doctor: High blood pressure	Diseases (cardiovascular)
9	Nucleated red blood cell count	Blood count
10	Diagnoses - main ICD10: K90 Intestinal malabsorption	Diseases (ICD10 main)
11	Coeliac disease	Diseases (FinnGen)
12	Non-cancer illness code, self-reported: hypertension	Diseases (cardiovascular)
13	Nucleated red blood cell percentage	Blood count
14	Relative age of first facial hair	Male-specific factors
15	Treatment/medication code: thiamine preparation	Medications
16	Diagnoses - main ICD10: I70 Atherosclerosis	Diseases (ICD10 main)
17	White Blood Cell Count	Blood
18	Treatment/medication code: gtn 400micrograms spray	Medications
19	Treatment/medication code: singulair 10mg tablet	Medications
20	Difficulty not smoking for 1 day	Smoking
21	Mean reticulocyte volume	Blood count
22	Other malignant neoplasms of skin	Diseases (FinnGen)
23	Length of working week for main job	Employment
24	Pulse rate, automated reading	Blood pressure
25	Milk type used: Skimmed	Diet

Table 4: Top 25 traits associated with genes from the lipids-increasing gene-set found with CRISPR screening.

Order	Trait	Category
1	Lymphocyte percentage	Blood count
2	Neutrophill percentage	Blood count
3	Neutrophill count	Blood count
4	Red blood cell (erythrocyte) count	Blood count
5	Lymphocyte Count	Blood
6	Trunk predicted mass	Impedance measures
7	Trunk fat-free mass	Impedance measures
8	Mean corpuscular volume	Blood count
9	Mean sphered cell volume	Blood count
10	White blood cell (leukocyte) count	Blood count

Order	Trait	Category
11	Skin colour	Sun exposure
12	Arm fat-free mass (left)	Impedance measures
13	Impedance of arm (left)	Impedance measures
14	Mean reticulocyte volume	Blood count
15	Whole body water mass	Impedance measures
16	Impedance of arm (right)	Impedance measures
17	Mean corpuscular haemoglobin	Blood count
18	Whole body fat-free mass	Impedance measures
19	Arm predicted mass (left)	Impedance measures
20	Arm predicted mass (right)	Impedance measures
21	Arm fat-free mass (right)	Impedance measures
22	Hair colour (natural, before greying): Red	Sun exposure
23	Ease of skin tanning	Sun exposure
24	High light scatter reticulocyte count	Blood count
25	White Blood Cell Count	Blood

Table 5: Top 25 traits associated with gene modules (LVs) enriched for the lipids-increasing gene-set found with CRISPR screening.

Order	Trait	Category
1	Ankle spacing width	Bone-densitometry of heel
2	Ankle spacing width (left)	Bone-densitometry of heel
3	Ankle spacing width (right)	Bone-densitometry of heel
4	Job SOC coding: Advertising and public relations managers	Employment history
5	Hair colour (natural, before greying): Red	Sun exposure
6	Sitting height	Body size measures
7	Platelet distribution width	Blood count
8	Non-cancer illness code, self-reported: malabsorption/coeliac disease	Diseases (gastrointestinal/abdominal)
9	Job coding: advertising or public relations manager, media/publicity manager, campaign/fundraising manager	Employment history
10	Forced expiratory volume in 1-second (FEV1), predicted	Spirometry
11	Heel Broadband ultrasound attenuation, direct entry	Bone-densitometry of heel
12	Intra-ocular pressure, Goldmann-correlated (right)	Intraocular pressure
13	Hearing test done: No, I am unable to do this	Hearing test
14	Rheumatoid Arthritis	Diseases (ICD10 main)
15	Red blood cell (erythrocyte) distribution width	Blood count
16	Job coding: childminder, au pair, children's nanny	Employment history

Order	Trait	Category
17	Heel bone mineral density (BMD)	Bone-densitometry of heel
18	Heel quantitative ultrasound index (QUI), direct entry	Bone-densitometry of heel
19	Heel bone mineral density (BMD) T-score, automated	Bone-densitometry of heel
20	Job SOC coding: Hand craft occupations n.e.c.	Employment history
21	Reason for glasses/contact lenses: For just reading/near work as you are getting older (called 'presbyopia')	Eyesight
22	Intra-ocular pressure, Goldmann-correlated (left)	Intraocular pressure
23	Pulse wave peak to peak time	Arterial stiffness
24	Hand grip strength (left)	Hand grip strength
25	Treatment/medication code: luteine	Medications

LV246

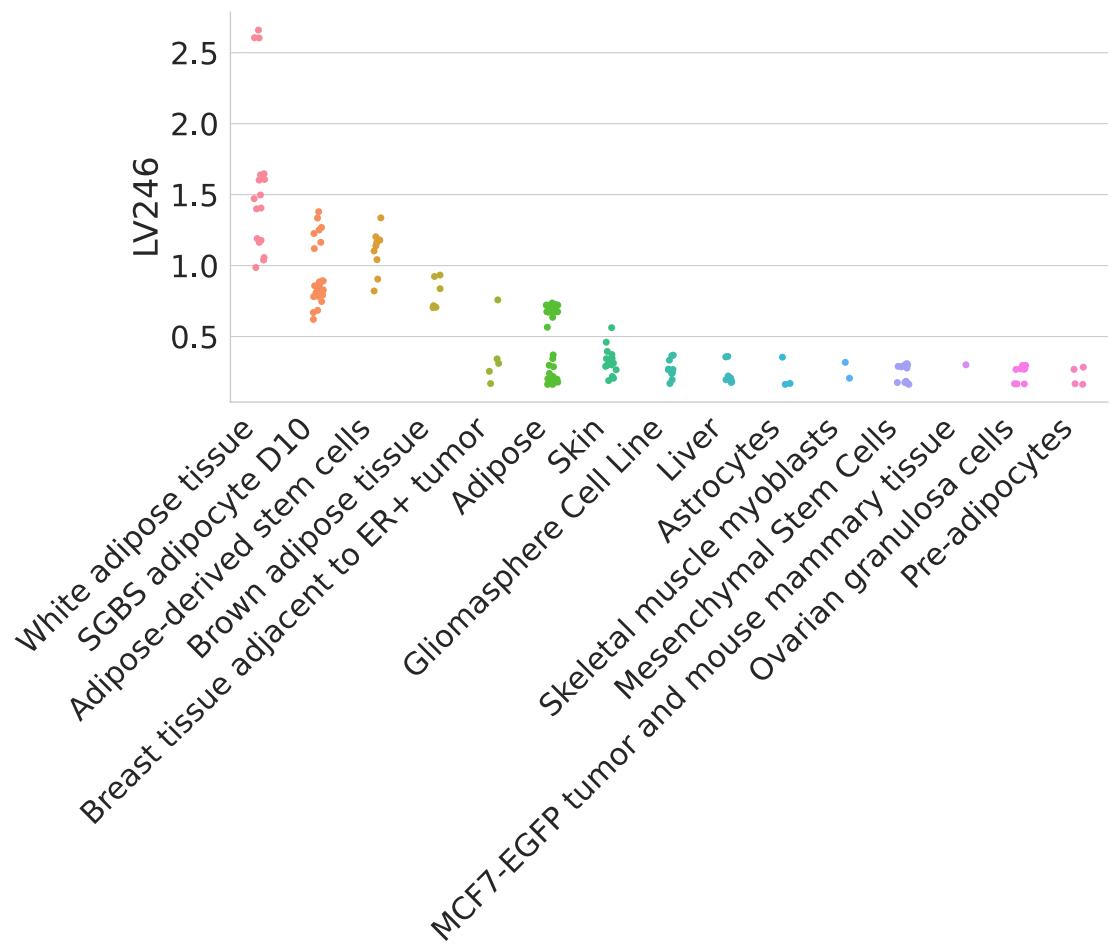


Figure 9: Cell types for LV246.

Table 6: Pathways aligned to LV246.

Pathway	AUC	p-value (adjusted)
REACTOME_FATTY_ACID_TRIACYLGLYCEROL_AND_KETONE_BODY_METABOLISM	0.89	3.97e-16
REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS	0.67	1.14e-08

Pathway	AUC	p-value (adjusted)
REACTOME_TRIGLYCERIDE BIOSYNTHESIS	0.86	6.52e-04
KEGG_PYRUVATE_METABOLISM	0.82	2.66e-03
KEGG_PROPANOATE_METABOLISM	0.83	4.27e-03

Table 7: Significant trait associations of LV246 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
CH2DB NMR	24,154		29 / 16	9.36e-11
Non-cancer illness code, self-reported: high cholesterol	361,141	43,957	29 / 17	5.24e-05
Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Cholesterol lowering medication	193,148	24,247	29 / 17	9.34e-03
HDL Cholesterol NMR	19,270		29 / 16	9.34e-03
Fasting Glucose	46,186		29 / 11	4.13e-02

Table 8: Significant trait associations of LV246 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
---------	-------------------	-------------	-------	--------------------

LV678

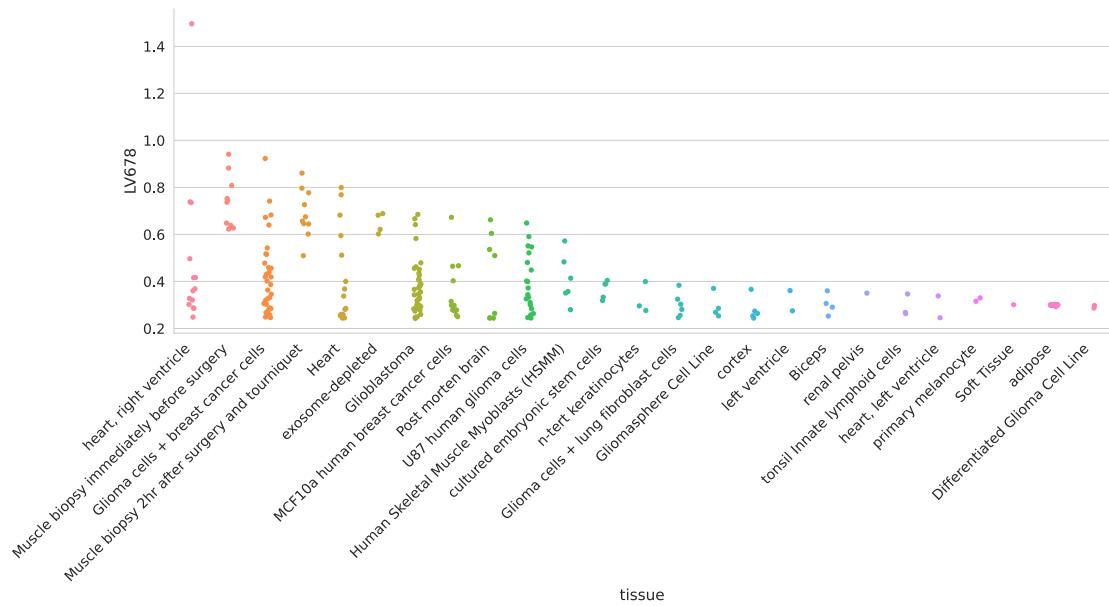


Figure 10: Cell types for LV678.

Table 9: Pathways aligned to LV678.

Pathway	AUC	p-value (adjusted)
---------	-----	--------------------

Pathway	AUC	p-value (adjusted)
KEGG_OXIDATIVE_PHOSPHORYLATION	0.98	5.75e-14
REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_SYNTHESIS_BY_CHEMIOSMOTIC_COUPLING_AND_HEAT_PRODUCTION_BY_UNCOUPLING_PROTEINS_	0.99	5.94e-11
REACTOME_RESPIRATORY_ELECTRON_TRANSPORT	1.00	3.10e-09
REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT	0.86	9.66e-09
MIPS_55S_RIBOSOME_MITOCHONDRIAL	0.81	8.20e-05
REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	0.69	6.03e-03
REACTOME_MITOCHONDRIAL_PROTEIN_IMPORT	0.74	1.99e-02

Table 10: Significant trait associations of LV678 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Vascular/heart problems diagnosed by doctor: Heart attack	360,420	8,288	29 / 14	1.08e-02
Inflammatory Bowel Disease	34,652	12,882	29 / 21	2.35e-02
Non-cancer illness code, self-reported: heart attack/myocardial infarction	361,141	8,239	29 / 14	2.35e-02

Table 11: Significant trait associations of LV678 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
---------	-------------------	-------------	-------	--------------------

LV116

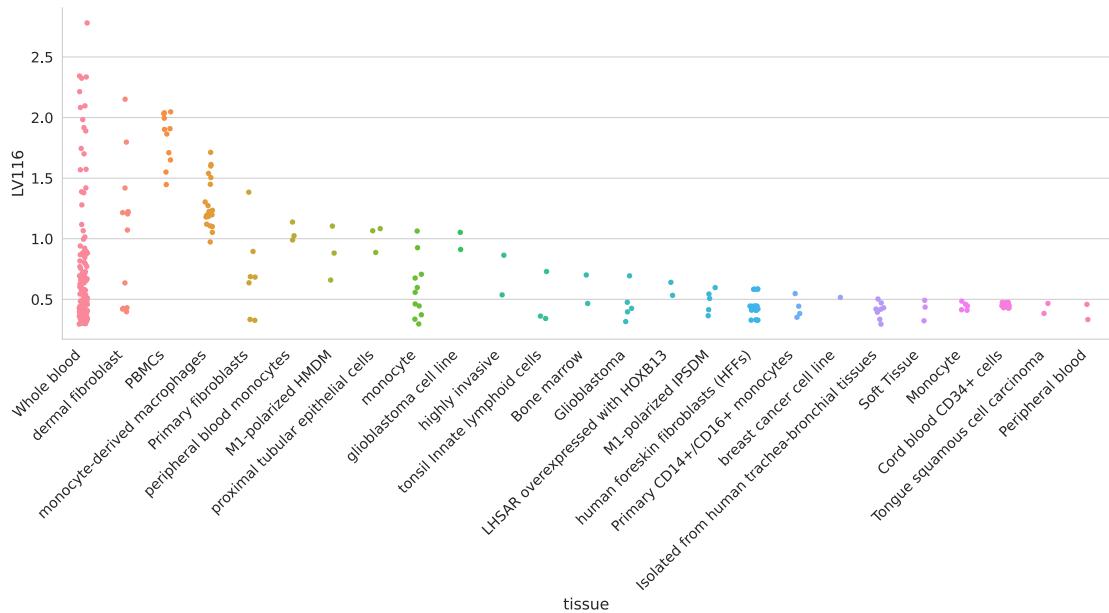


Figure 11: Cell types for LV116.

LV931

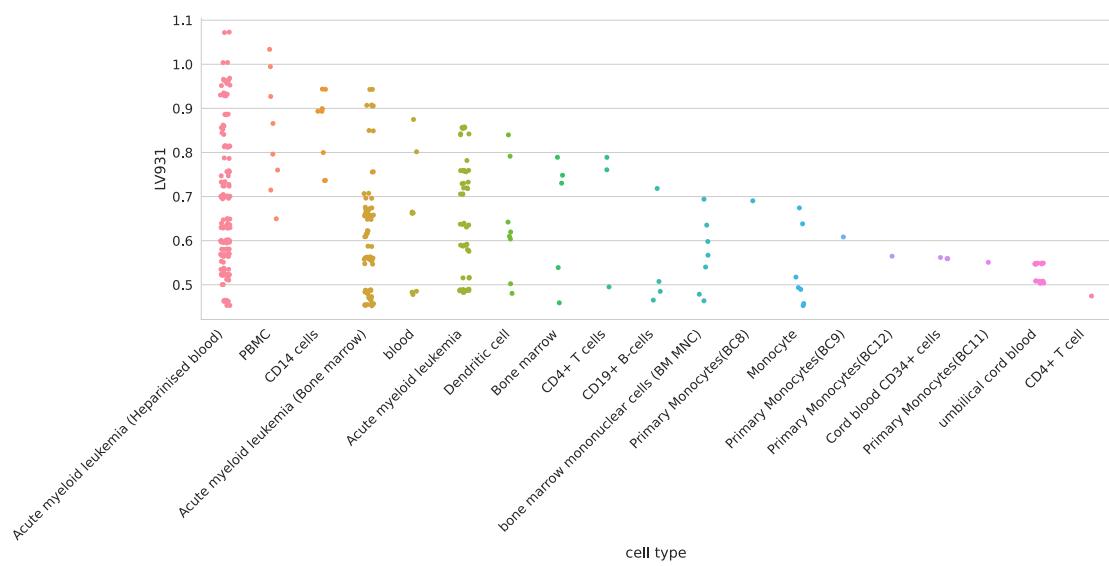


Figure 12: Cell types for LV931.

LV66

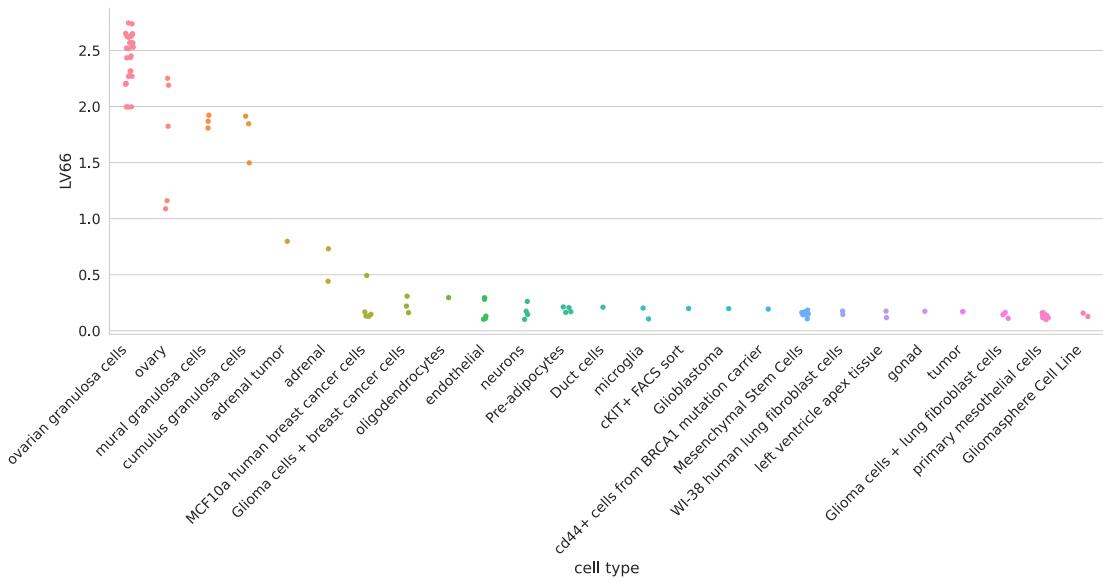


Figure 13: Cell types for LV66.

LV928

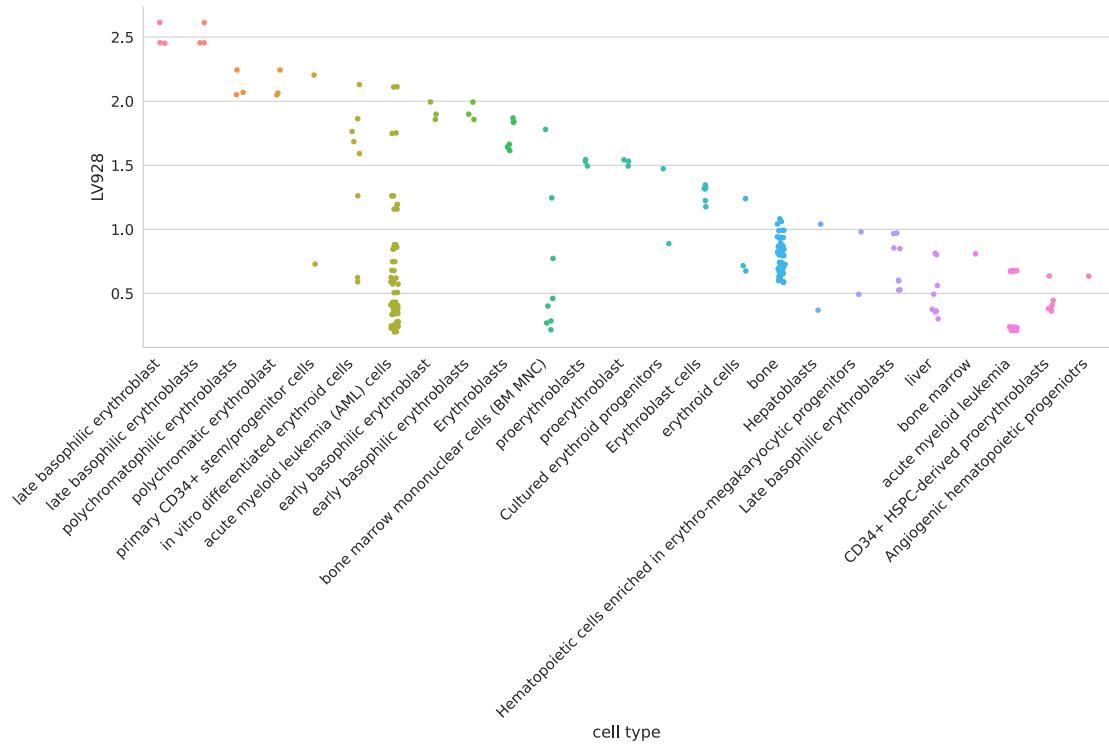


Figure 14: Cell types for LV928.

LV30

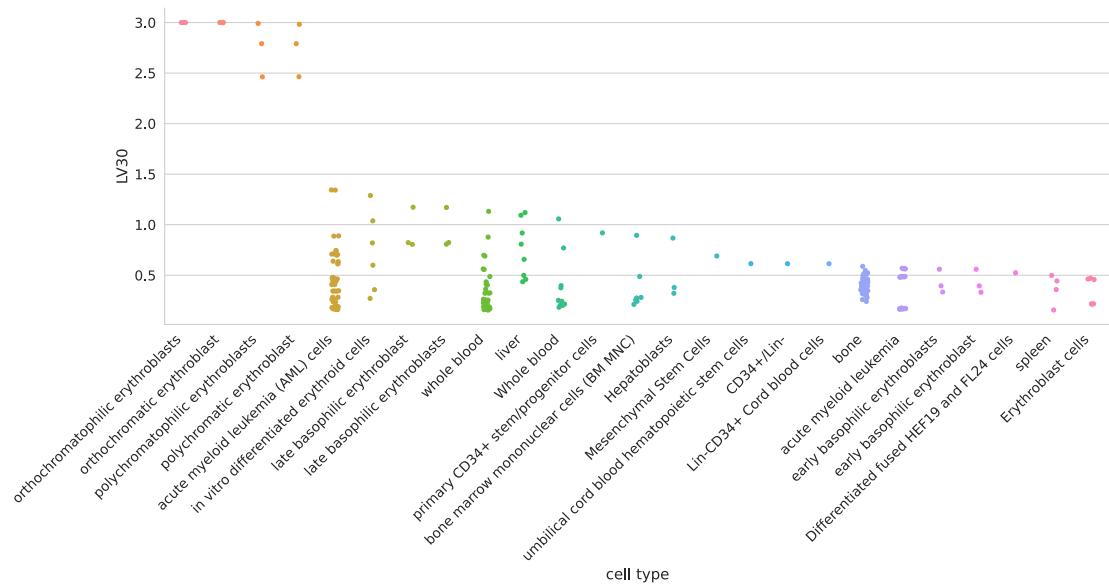


Figure 15: Cell types for LV30.

LV730

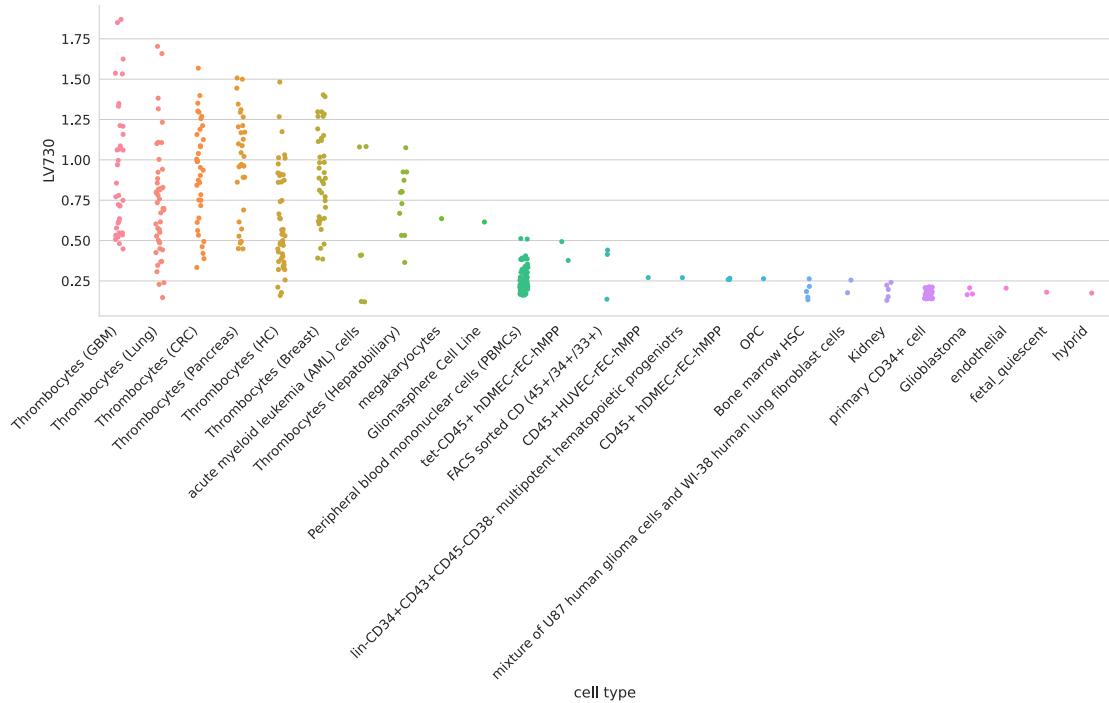


Figure 16: Cell types for LV730.

LV598

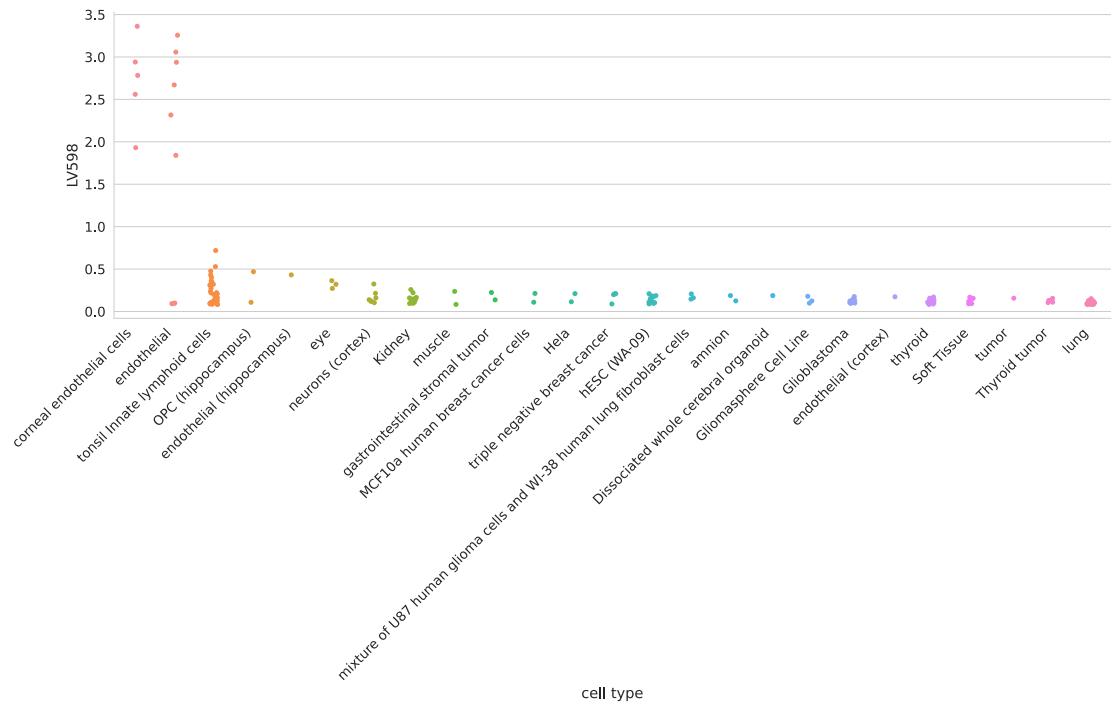


Figure 17: Cell types for LV598.

Table 12: Significant trait associations of LV598 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
6mm strong meridian (right)	66,256		29 / 10	4.13e-07
6mm weak meridian (right)	66,256		29 / 10	2.63e-06
6mm strong meridian (left)	65,551		29 / 10	3.13e-06
3mm strong meridian (left)	75,398		29 / 10	3.24e-06
6mm weak meridian (left)	65,551		29 / 10	1.53e-05
3mm weak meridian (left)	75,398		29 / 10	2.00e-05
3mm strong meridian (right)	75,410		29 / 10	3.70e-05
3mm weak meridian (right)	75,410		29 / 10	4.81e-05

LV155

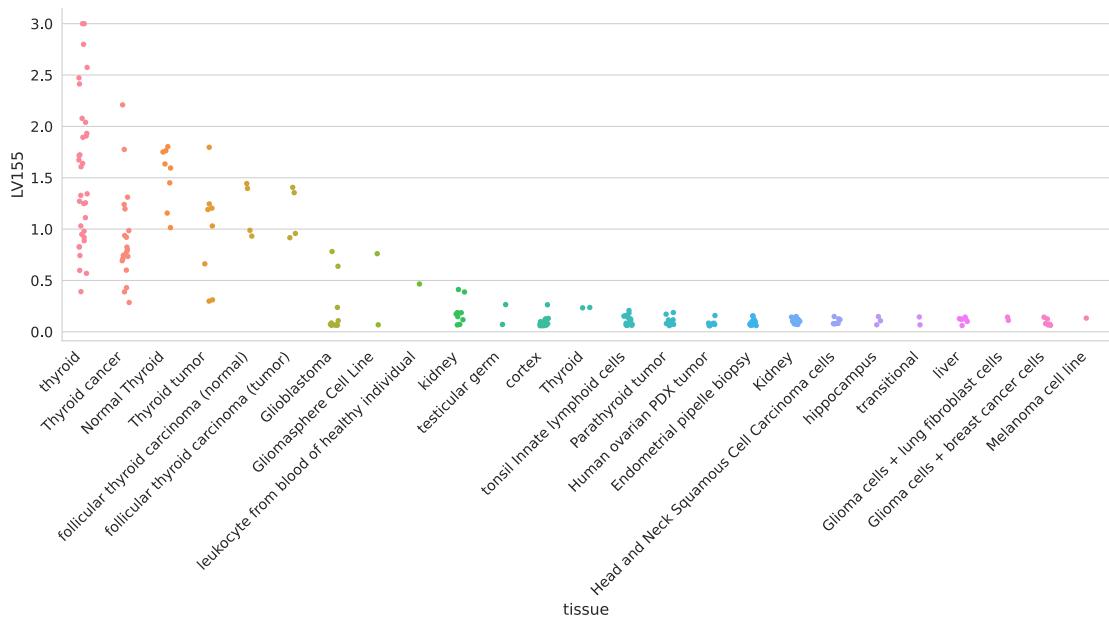


Figure 18: Cell types for LV155.

Table 13: Significant trait associations of LV155 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Non-cancer illness code, self-reported: hypothyroidism/myxoedema	361,141	17,5 74	29 / 13	2.01e-03
Non-cancer illness code, self-reported: hyperthyroidism/thyrotoxicosis	361,141	2,73 0	29 / 13	1.29e-02
Treatment/medication code: levothyroxine sodium	361,141	14,6 89	29 / 13	1.41e-02

Table 14: Significant trait associations of LV155 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
244.2	Acquired hypothyroidism	45,839	1,155	2.19e-02
427.9	Palpitations	35,214	6,092	4.43e-02

LV844

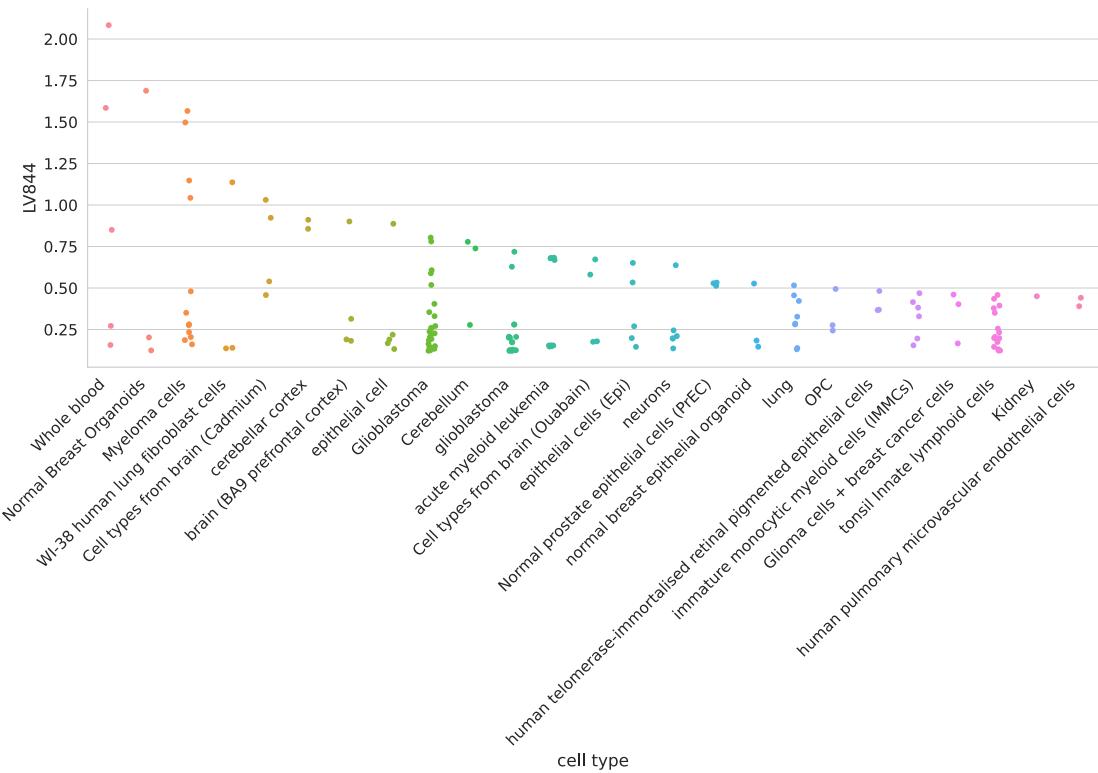


Figure 19: Cell types for LV844. **Note** “Cell types from brain” come from [86], treated with different chemicals; I don’t have enough information to separate cell types.

Table 15: Significant trait associations of LV844 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Rheumatoid Arthritis	80,799	19,234	29 / 26	4.27e-57
Non-cancer illness code, self-reported: malabsorption/coeliac disease	361,141	1,587	29 / 8	4.83e-43
Coeliac disease	361,194	842	29 / 8	4.76e-41
Diagnoses - main ICD10: K90 Intestinal malabsorption	361,194	922	29 / 8	1.41e-39
Started insulin within one year diagnosis of diabetes	16,415	1,999	29 / 13	1.78e-37
Systemic Lupus Erythematosus	23,210	7,219	29 / 26	1.41e-34
Age diabetes diagnosed	16,166		29 / 13	3.93e-34
Never eat eggs, dairy, wheat, sugar: Wheat products	359,777	9,573	29 / 13	2.78e-31
Non-cancer illness code, self-reported: hyperthyroidism/thyrotoxicosis	361,141	2,730	29 / 13	6.08e-30
Treatment/medication code: insulin product	361,141	3,545	29 / 13	3.05e-25
Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Insulin	193,148	1,476	29 / 13	4.63e-23
Medication for cholesterol, blood pressure or diabetes: Insulin	165,340	2,248	29 / 13	1.92e-20

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Non-cancer illness code, self-reported: hypothyroidism/myxoedema	361,141	17,574	29 / 13	4.96e-20
Treatment/medication code: levothyroxine sodium	361,141	14,689	29 / 13	4.01e-19
Non-cancer illness code, self-reported: psoriasis	361,141	4,192	29 / 13	9.28e-16

Table 16: Significant trait associations of LV844 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
714.1	Rheumatoid arthritis	49,453	2,541	8.22e-09
250.1	Type 1 diabetes	42,723	2,450	2.54e-08
714	Rheumatoid arthritis and other inflammatory polyarthropathies	50,215	3,303	5.06e-07
440	Atherosclerosis	47,471	4,993	3.15e-03
578.8	Hemorrhage of rectum and anus	47,545	1,991	3.15e-03
585.32	End stage renal disease	43,309	1,842	4.38e-03
440.2	Atherosclerosis of the extremities	45,524	3,046	5.00e-03
514.2	Solitary pulmonary nodule	50,389	2,270	6.16e-03
444	Arterial embolism and thrombosis	43,378	900	1.36e-02
558	Noninfectious gastroenteritis	40,177	3,191	2.94e-02
747.11	Cardiac shunt/ heart septal defect	58,364	1,037	3.60e-02
585	Renal failure	51,437	9,970	3.87e-02
443.9	Peripheral vascular disease, unspecified	46,926	4,448	4.43e-02

LV57

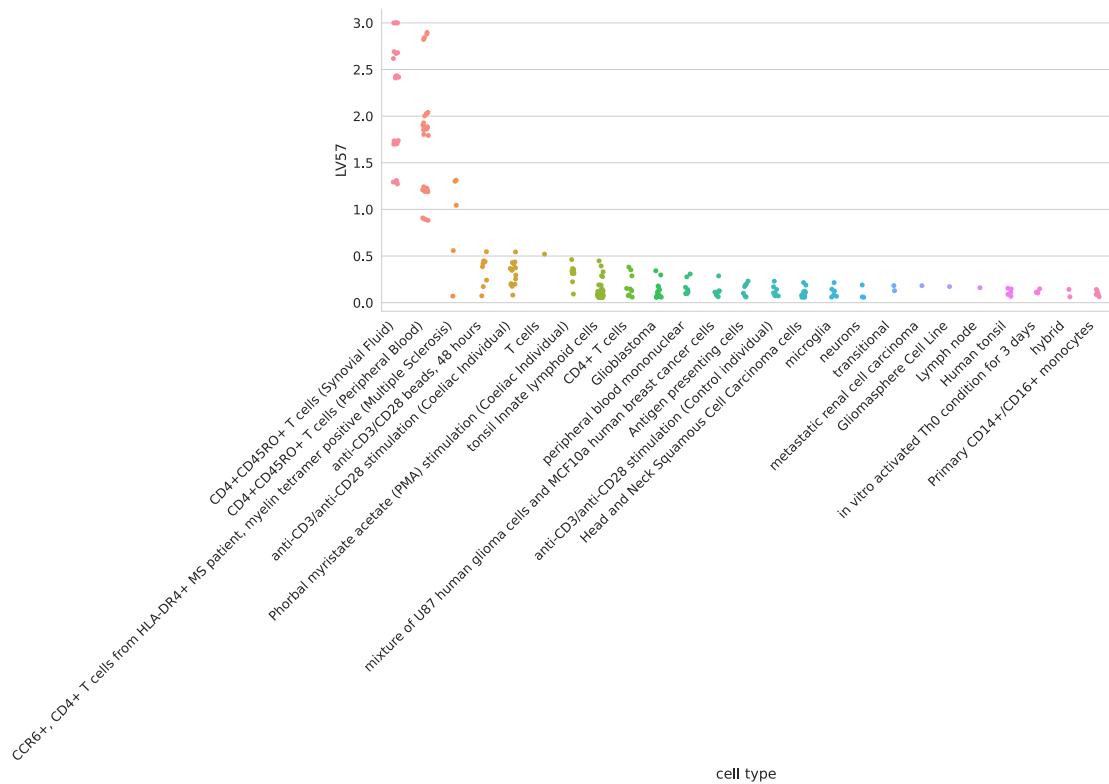


Figure 20: Cell types for LV57.

Table 17: Significant trait associations of LV57 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Non-cancer illness code, self-reported: hypothyroidism/myxoedema	361,141	17,574	29 / 13	1.17e-24
Treatment/medication code: levothyroxine sodium	361,141	14,689	29 / 13	6.07e-23
Non-cancer illness code, self-reported: hyperthyroidism/thyrotoxicosis	361,141	2,730	29 / 13	1.16e-06
Started insulin within one year diagnosis of diabetes	16,415	1,999	29 / 13	8.17e-05
Treatment/medication code: insulin product	361,141	3,545	29 / 13	6.33e-04
Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Insulin	193,148	1,476	29 / 13	1.13e-03
Medication for cholesterol, blood pressure or diabetes: Insulin	165,340	2,248	29 / 13	4.50e-03

Table 18: Significant trait associations of LV57 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
244	Hypothyroidism	54,404	9,720	3.97e-09
244.4	Hypothyroidism NOS	53,968	9,284	3.97e-09
279	Disorders involving the immune mechanism	56,771	3,309	4.93e-03
514.2	Solitary pulmonary nodule	50,389	2,270	1.19e-02

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
714	Rheumatoid arthritis and other inflammatory polyarthropathies	50,215	3,303	1.68e-02
452.2	Deep vein thrombosis [DVT]	38,791	2,131	4.37e-02

LV54

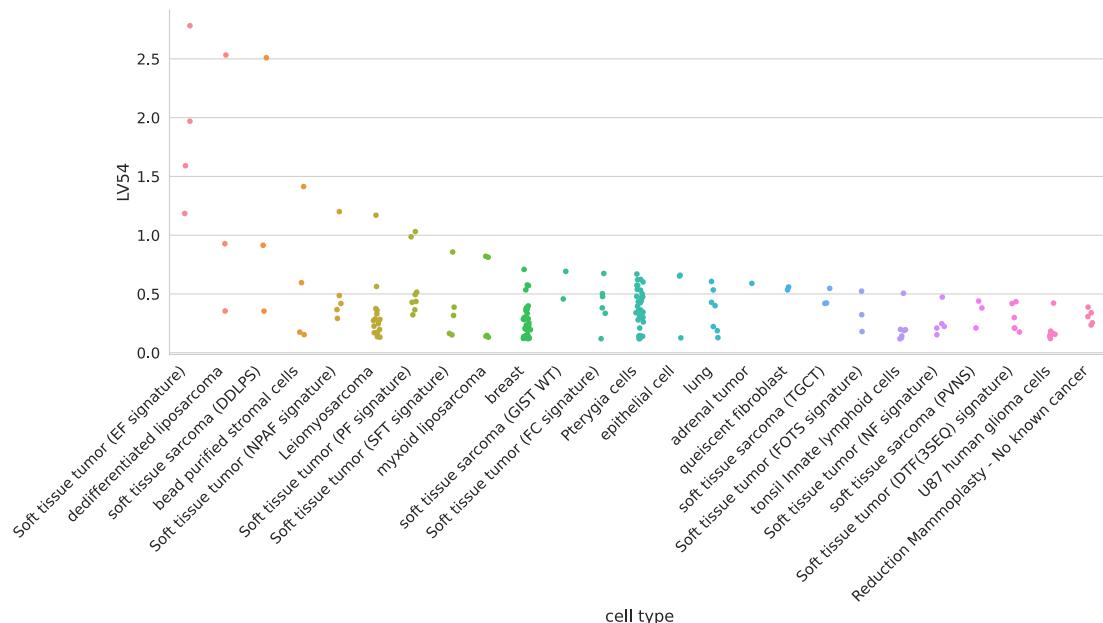


Figure 21: Cell types for LV54.

Table 19: Significant trait associations of LV54 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Diagnoses - main ICD10: K90 Intestinal malabsorption	361,194	922	29 / 8	5.09e-25
Coeliac disease	361,194	842	29 / 8	7.77e-25
Never eat eggs, dairy, wheat, sugar: Wheat products	359,777	9,5 73	29 / 13	6.33e-23
Systemic Lupus Erythematosus	23,210	7,2 19	29 / 26	1.32e-22
Started insulin within one year diagnosis of diabetes	16,415	1,9 99	29 / 13	3.84e-20
Non-cancer illness code, self-reported: hyperthyroidism/thyrotoxicosis	361,141	2,7 30	29 / 13	9.59e-19
Treatment/medication code: insulin product	361,141	3,5 45	29 / 13	5.07e-18
Age diabetes diagnosed	16,166		29 / 13	1.28e-17
Non-cancer illness code, self-reported: malabsorption/coeliac disease	361,141	1,5 87	29 / 8	1.36e-14
Medication for cholesterol, blood pressure or diabetes: Insulin	165,340	2,2 48	29 / 13	8.67e-14

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Non-cancer illness code, self-reported: psoriasis	361,141	4,192	29 / 13	2.61e-13
Rheumatoid Arthritis	80,799	19,234	29 / 26	3.11e-13
Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Insulin	193,148	1,476	29 / 13	3.89e-12
Treatment/medication code: levothyroxine sodium	361,141	14,689	29 / 13	5.92e-10
Non-cancer illness code, self-reported: hypothyroidism/myxoedema	361,141	17,574	29 / 13	3.31e-08

Table 20: Significant trait associations of LV54 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
250.1	Type 1 diabetes	42,723	2,450	2.04e-13
244	Hypothyroidism	54,404	9,720	5.10e-06
244.4	Hypothyroidism NOS	53,968	9,284	5.37e-06
695	Erythematous conditions	48,347	4,210	4.25e-05
714	Rheumatoid arthritis and other inflammatory polyarthropathies	50,215	3,303	3.06e-04
440	Atherosclerosis	47,471	4,993	8.88e-04
585	Renal failure	51,437	9,970	3.40e-03
585.32	End stage renal disease	43,309	1,842	3.64e-03
585.33	Chronic Kidney Disease, Stage III	46,279	4,812	3.64e-03
285.2	Anemia of chronic disease	39,673	2,606	7.62e-03
415.1	Acute pulmonary heart disease	49,887	1,857	8.67e-03
285.21	Anemia in chronic kidney disease	38,616	1,549	1.16e-02
743	Osteoporosis, osteopenia and pathological fracture	55,165	11,990	1.31e-02
415.11	Pulmonary embolism and infarction, acute	49,867	1,837	1.39e-02
577	Diseases of pancreas	60,538	1,795	1.42e-02
585.1	Acute renal failure	46,803	5,336	1.51e-02
195	Cancer, suspected or other	50,040	2,250	1.52e-02
440.2	Atherosclerosis of the extremities	45,524	3,046	1.89e-02
714.1	Rheumatoid arthritis	49,453	2,541	3.18e-02
458.9	Hypotension NOS	50,150	3,241	3.32e-02

LV847

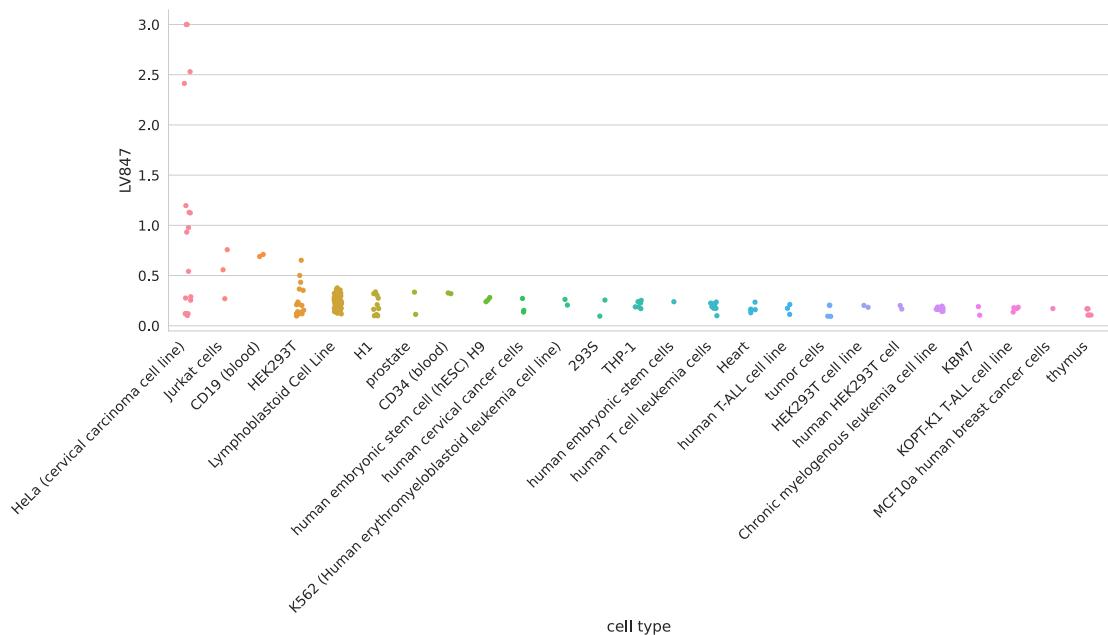


Figure 22: Cell types for LV847.

Table 21: Significant trait associations of LV847 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Blood pressure medication	193,148	33,519	29 / 17	1.95e-18
Vascular/heart problems diagnosed by doctor: None of the above	360,420	253,565	29 / 17	4.07e-15
Vascular/heart problems diagnosed by doctor: High blood pressure	360,420	97,139	29 / 17	6.99e-14
Non-cancer illness code, self-reported: hypertension	361,141	93,560	29 / 17	1.48e-13
Treatment/medication code: bendroflumethiazide	361,141	20,196	29 / 17	1.41e-08
Medication for cholesterol, blood pressure or diabetes: Blood pressure medication	165,340	40,987	29 / 17	1.47e-07
Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: None of the above	193,148	133,338	29 / 17	1.55e-06
Diastolic blood pressure, automated reading	340,162		29 / 17	3.76e-06
Medication for cholesterol, blood pressure or diabetes: None of the above	165,340	110,372	29 / 17	6.36e-06

Table 22: Significant trait associations of LV847 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
585.32	End stage renal disease	43,309	1,842	1.88e-08
442.1	Aortic aneurysm	45,589	3,111	5.23e-06
411.3	Angina pectoris	43,503	4,382	2.14e-05
415.11	Pulmonary embolism and infarction, acute	49,867	1,837	5.13e-05
416	Cardiomegaly	53,289	5,259	6.50e-05
415.1	Acute pulmonary heart disease	49,887	1,857	7.28e-05
411	Ischemic Heart Disease	54,275	15,154	5.49e-04
401.2	Hypertensive heart and/or renal disease	30,405	6,253	1.28e-03
519	Other diseases of respiratory system, not elsewhere classified	56,909	2,056	1.28e-03
411.8	Other chronic ischemic heart disease, unspecified	44,123	5,002	1.42e-03
427.6	Premature beats	31,575	2,453	5.65e-03
687.1	Rash and other nonspecific skin eruption	47,039	4,964	9.88e-03
185	Cancer of prostate	52,630	2,815	1.03e-02
591	Urinary tract infection	49,727	10,016	1.34e-02
442.11	Abdominal aortic aneurysm	44,531	2,053	2.08e-02
427.21	Atrial fibrillation	37,743	8,621	2.26e-02
389.1	Sensorineural hearing loss	53,672	4,318	2.73e-02
427.2	Atrial fibrillation and flutter	37,934	8,812	4.50e-02

LV136

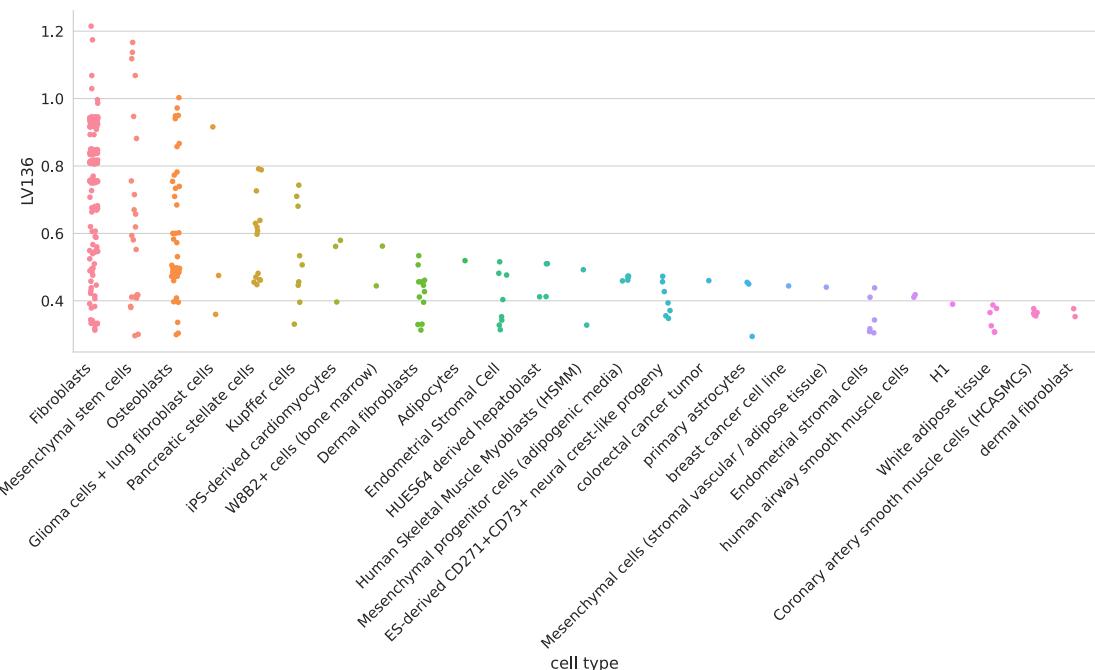


Figure 23: Cell types for LV136. Pulmonary microvascular endothelial cells were exposed to hypoxia for 24 hours or more [87];

Table 23: Significant trait associations of LV136 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
3mm strong meridian (right)	75,410		29 / 10	9.19e-11
6mm strong meridian (left)	65,551		29 / 10	2.06e-09
6mm strong meridian (right)	66,256		29 / 10	2.38e-09
3mm strong meridian (left)	75,398		29 / 10	1.34e-08
3mm weak meridian (right)	75,410		29 / 10	1.67e-08
Coronary Artery Disease	184,305	60,8 01	29 / 11	1.67e-08
6mm weak meridian (right)	66,256		29 / 10	3.21e-08
3mm weak meridian (left)	75,398		29 / 10	5.20e-08
6mm weak meridian (left)	65,551		29 / 10	1.21e-07
Coronary atherosclerosis	361,194	14,3 34	29 / 14	3.90e-06
Ischaemic heart disease, wide definition	361,194	20,8 57	29 / 14	7.22e-06
Vascular/heart problems diagnosed by doctor: Heart attack	360,420	8,28 8	29 / 14	2.93e-04
Myocardial infarction	361,194	7,01 8	29 / 14	6.33e-04
Myocardial infarction, strict	361,194	7,01 8	29 / 14	6.33e-04
Diagnoses - main ICD10: I21 Acute myocardial infarction	361,194	5,94 8	29 / 14	9.92e-04
Non-cancer illness code, self-reported: heart attack/myocardial infarction	361,141	8,23 9	29 / 14	1.40e-03
Major coronary heart disease event excluding revascularizations	361,194	10,1 57	29 / 14	1.85e-02
Major coronary heart disease event	361,194	10,1 57	29 / 14	1.85e-02
Fasting Insulin	38,238		29 / 11	3.85e-02

Table 24: Significant trait associations of LV136 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
747.1	Cardiac congenital anomalies	59,198	1,871	4.71e-02
411.4	Coronary atherosclerosis	52,836	13,715	4.80e-02

LV93

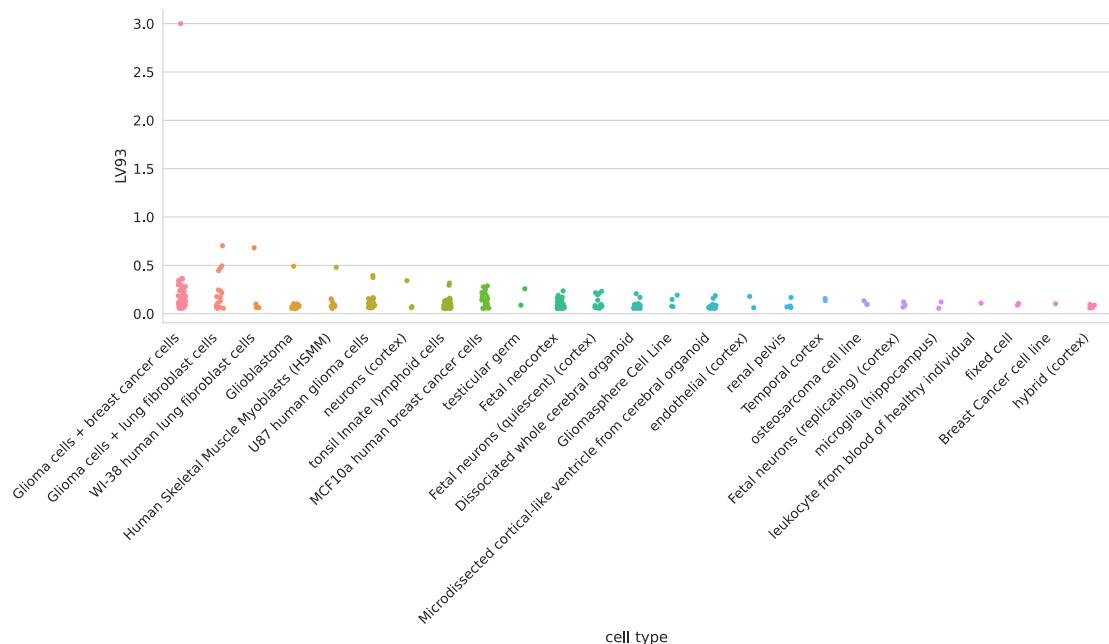


Figure 24: Cell types for LV93.

Table 25: Significant trait associations of LV93 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
CH2DB NMR	24,154		29 / 16	9.61e-24
Chronotype	128,266		29 / 16	1.17e-03
HDL Cholesterol NMR	19,270		29 / 16	2.99e-03

Table 26: Significant trait associations of LV93 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
208	Benign neoplasm of colon	55,694	8,597	6.21e-03
440.2	Atherosclerosis of the extremities	45,524	3,046	1.31e-02
444	Arterial embolism and thrombosis	43,378	900	4.06e-02

LV206

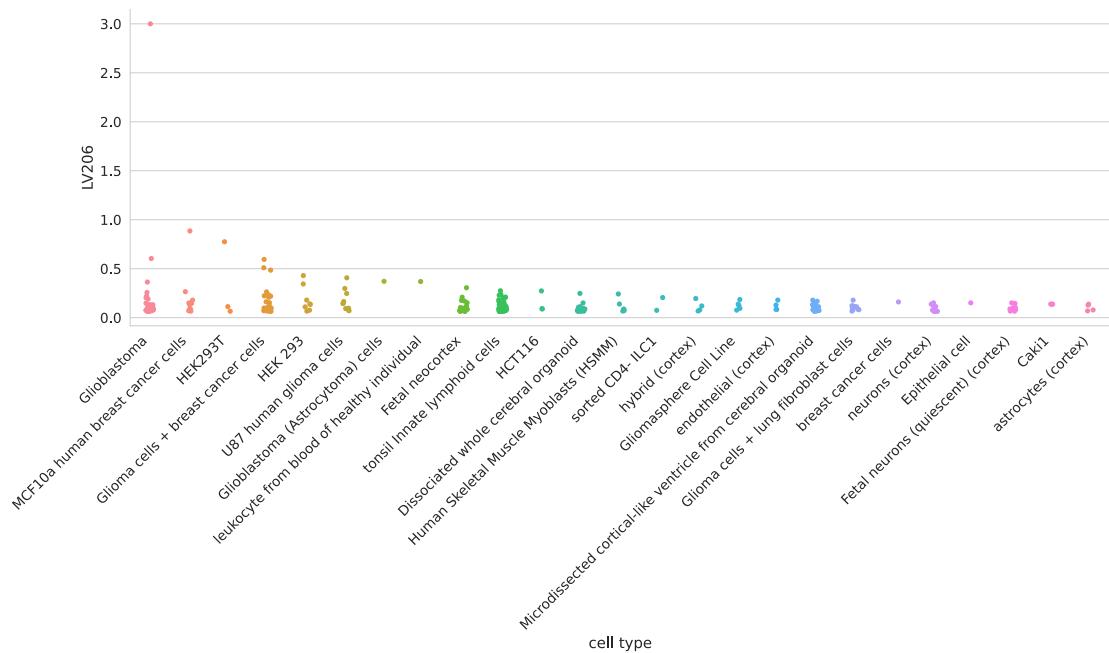


Figure 25: Cell types for LV206.

Table 27: Significant trait associations of LV206 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
CH2DB NMR	24,154		29 / 16	7.67e-21
HDL Cholesterol NMR	19,270		29 / 16	6.46e-03

Table 28: Significant trait associations of LV206 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
458	Hypotension	51,341	4,432	1.41e-02
286.9	Abnormal coagulation profile	48,006	800	1.54e-02
458.9	Hypotension NOS	50,150	3,241	1.58e-02
428.2	Heart failure NOS	48,178	3,584	1.65e-02

LV260

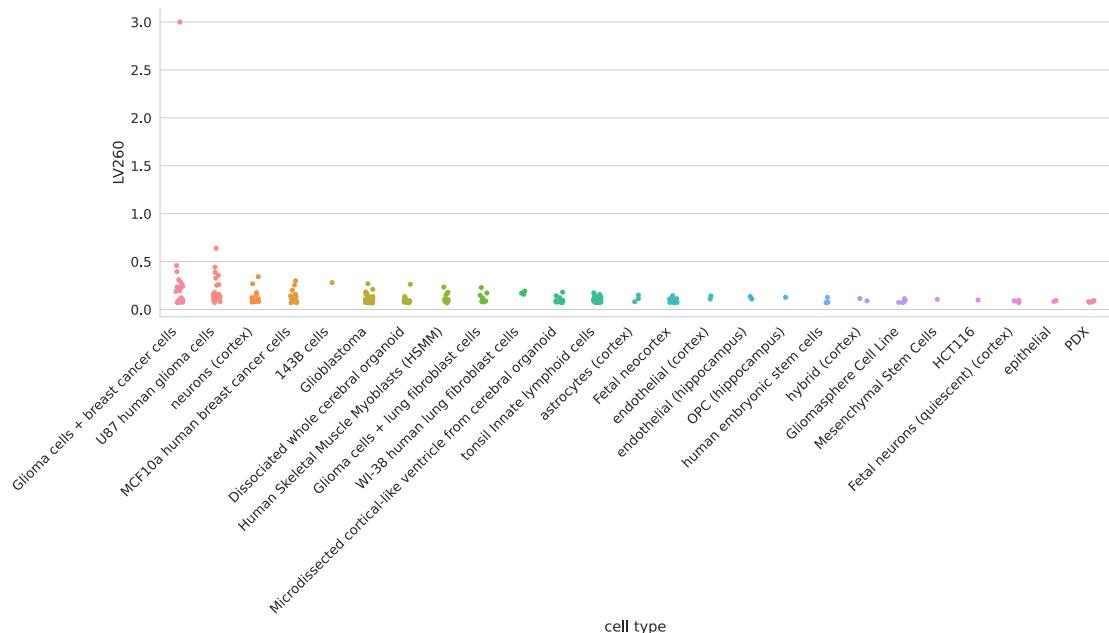


Figure 26: Cell types for LV260.

Table 29: Significant trait associations of LV260 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
CH2DB NMR	24,154		29 / 16	5.96e-17
HDL Cholesterol NMR	19,270		29 / 16	2.37e-02

Table 30: Significant trait associations of LV260 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
427.6	Premature beats	31,575	2,453	2.85e-02
426.3	Bundle branch block	31,827	2,705	4.80e-02

LV21

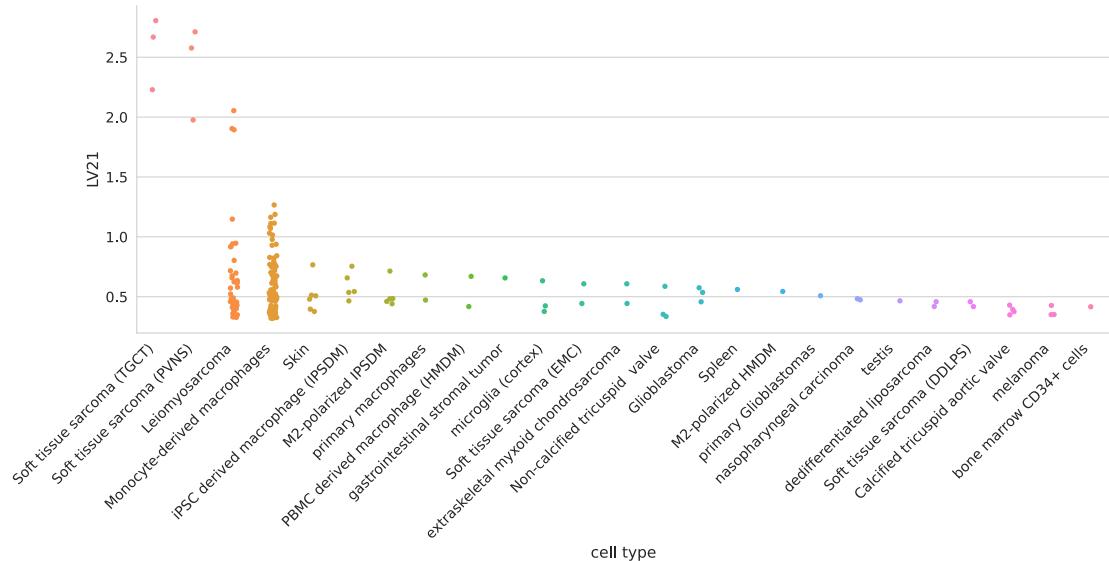


Figure 27: Cell types for LV21.

Table 31: Significant trait associations of LV21 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Alzheimers Disease	54,162	17,008	29 / 16	1.64e-19
LDL Cholesterol NMR	13,527		29 / 16	1.18e-04
Triglycerides NMR	21,559		29 / 16	2.19e-02

Table 32: Significant trait associations of LV21 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
573	Other disorders of liver	47,826	2,524	1.37e-02
577	Diseases of pancreas	60,538	1,795	2.15e-02

LV5

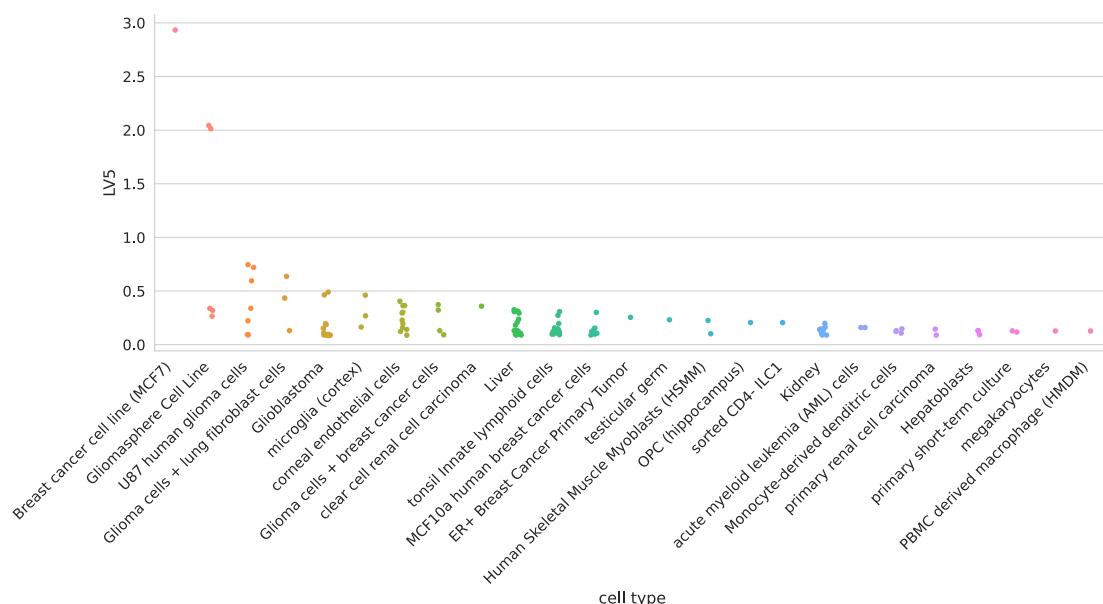


Figure 28: Cell types for LV5.

Table 33: Significant trait associations of LV5 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
LDL Cholesterol NMR	13,527		29 / 16	1.78e-04
Triglycerides NMR	21,559		29 / 16	5.00e-04
Alzheimers Disease	54,162	17,008	29 / 16	3.06e-03
Ever had prolonged feelings of sadness or depression	117,763	64,374	29 / 27	8.69e-03
Substances taken for depression: Medication prescribed to you (for at least two weeks)	117,763	28,351	29 / 27	1.03e-02
Recent feelings of depression	117,656		29 / 27	1.32e-02
Ever contemplated self-harm	117,610		29 / 27	1.89e-02
Recent lack of interest or pleasure in doing things	117,757		29 / 27	2.08e-02
Amount of alcohol drunk on a typical drinking day	108,256		29 / 27	3.50e-02

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Ever sought or received professional help for mental distress	117,677	46,020	29 / 27	3.92e-02
General happiness	117,442		29 / 27	4.74e-02

Table 34: Significant trait associations of LV5 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
241	Nontoxic nodular goiter	47,842	3,158	8.98e-03
241.1	Nontoxic uninodular goiter	47,125	2,441	2.57e-02
241.2	Nontoxic multinodular goiter	46,465	1,781	4.43e-02

LV434

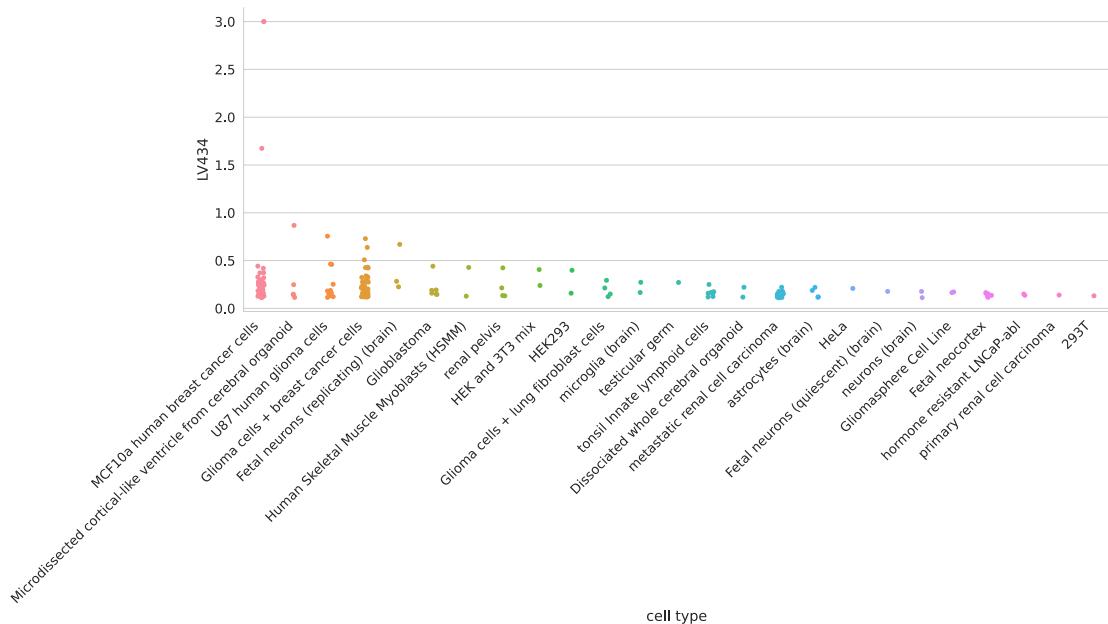


Figure 29: Cell types for LV434. FIXME: create a section for all LVs and clarify how figures were generated, like: The top samples in human breast cancer cells have larger LV values that were thresholded at 3.0 for visualization purposes. HEK293 is a cell line derived from human embryonic kidney cells; 3T3 is a cell line derived from mouse embryonic fibroblasts.

Table 35: Significant trait associations of LV434 in PhenomeXcan.

Trait description	Sample size	Cases	Partition/cluster number	p-value (adjusted)
Attention Deficit Hyperactivity Disorder	53,293	19,099	29 / 21	7.01e-03

Table 36: Significant trait associations of LV434 in eMERGE.

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
722	Intervertebral disc disorders	47,659	7,458	6.65e-03
721	Spondylosis and allied disorders	47,517	7,316	7.62e-03
250.4	Abnormal glucose	45,220	4,947	1.02e-02
721.1	Spondylosis without myelopathy	47,315	7,114	1.22e-02

Phecode	Trait description	Sample size	Cases	p-value (adjusted)
720	Spinal stenosis	44,807	4,606	1.74e-02
288	Diseases of white blood cells	47,288	2,802	2.10e-02
796	Elevated prostate specific antigen [PSA]	51,990	2,175	3.09e-02
288.2	Elevated white blood cell count	46,595	2,109	3.54e-02
079	Viral infection	46,991	1,934	4.19e-02

Top gene modules in the prediction of Niacin as a therapeutic drug

Table 37: Top gene modules/LVs responsible for predicting Niacin as a therapeutic drug for atherosclerosis.

LV	Value for Niacin	Value for atherosclerosis	Difference

Agreement of consensus clustering partitions with the ensemble by number of clusters

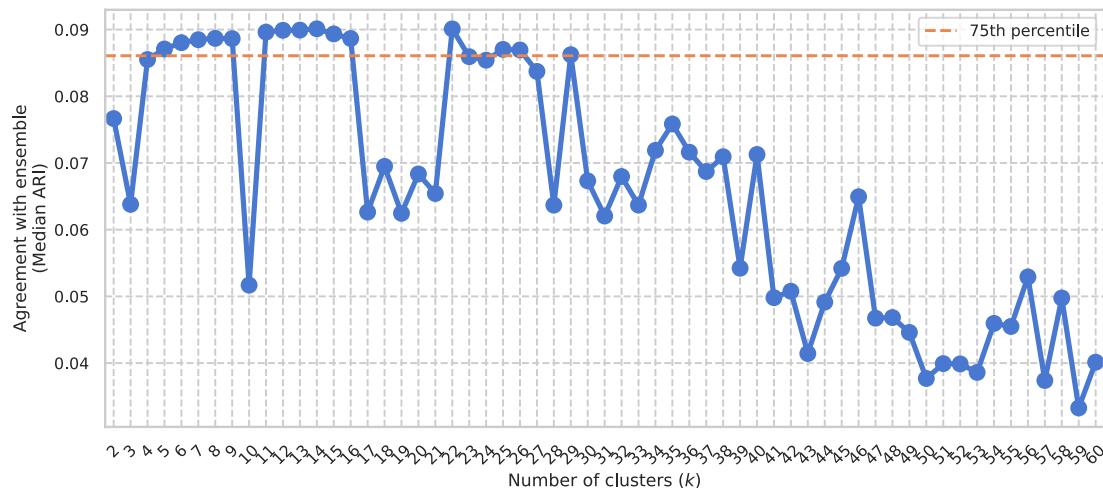


Figure 30: Final selected partitions for follow-up analysis. From all consensus clustering partitions generated with k from 2 to 60, we selected those with a median adjusted Rand index (ARI) with the ensemble members greater than the 75th percentile.