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@60b94a0 on May 20, 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], but they generally apply enrichment analysis techniques that do not account for widespread gene correlations due to technical noise (i.e. “batch effects”) [21,22]. 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 [23] into a latent representation learned from public gene expression repositories on tens of thousands of RNA-seq human samples [8,24]. 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 to 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. Our approach is also more 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 association than using single gene-traits associations, and that our approach could be also useful to study mechanisms of action of drugs. Finally, we performed cluster analysis on traits mapped to this latent representation, with clusters highly stable across different resolutions. We found common and specific transcriptional processes associated with autoimmune and cardiovascular diseases.

`Notes:`

- We need to say more about the clustering of traits, and select a good example to summarize our results.
- I’m not including eMERGE replication here because I still have to work on that part.

Minor:

- I don’t like much the idea of mentioning “module-trait associations”, since that conveys the idea of a p-value, which we are not giving.

- Reviewers might ask for the sample sizes of GWAS. Some of them, in PhenoPLIER, are really underpowered, with very few cases; others are from large studies. We should be careful when picking our examples in the Results section.

Results

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

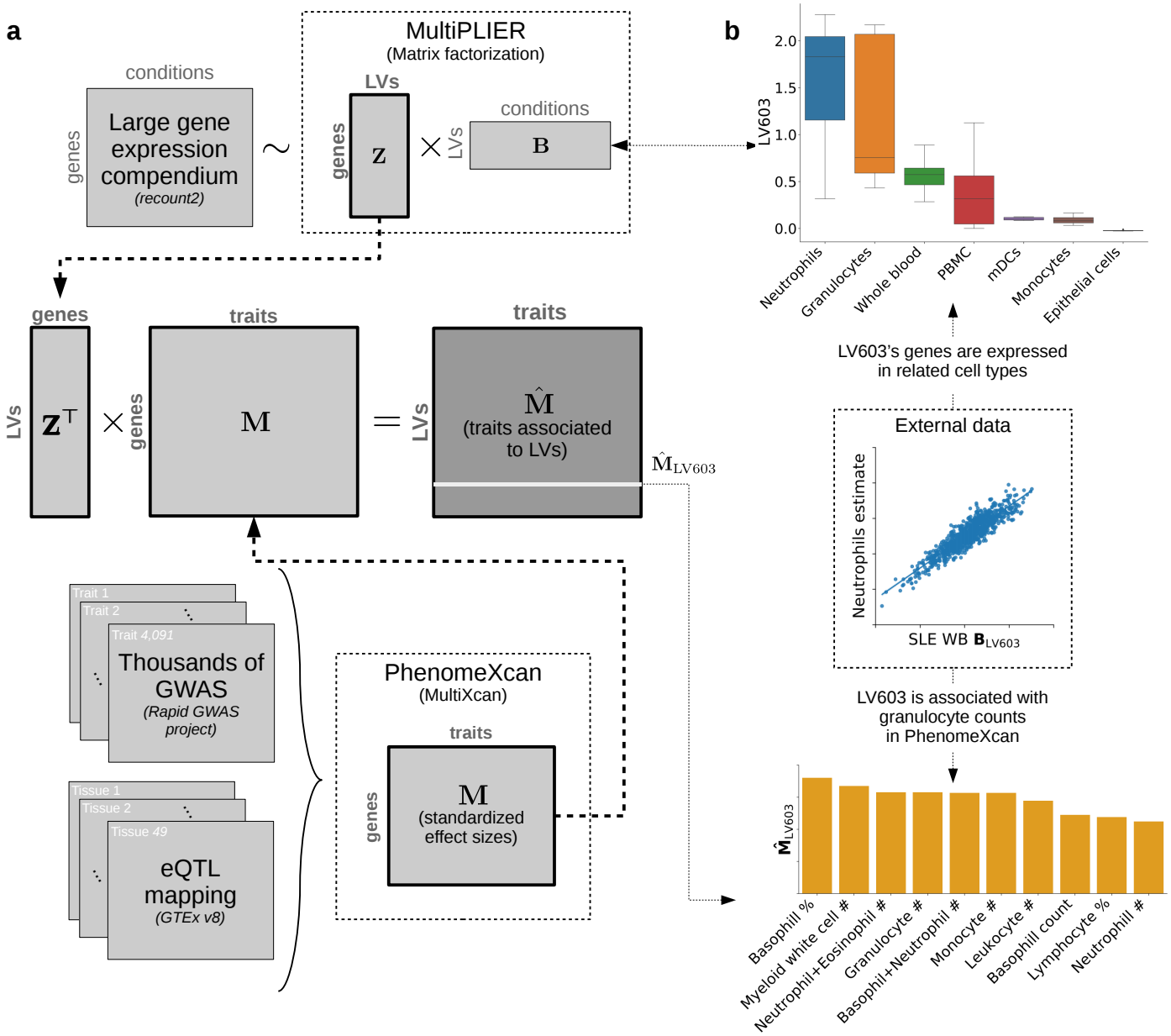


Figure 1: Schematic of the PhenoPLIER framework. a) The integration process between gene co-expression patterns from MultiPLIER (top) and TWAS results from PhenomeXcan. PhenoPLIER projects gene-based association results on thousands of traits to a latent space learned from large gene expression datasets. The process generates matrix $\hat{\mathbf{M}}$, 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 [23], a massive TWAS resource on the UK Biobank [25] and other cohorts that provides results for 4,091 different diseases and traits. The latent gene expression representation was obtained by applying MultiPLIER [24], 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 [26,27], 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 $\mathbf{B}$ in Figure 1 a) in which genes in a module are concurringly 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 [28], 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 492 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 high-confidence gene-sets that either caused a decrease (96 genes) or an increase (175 genes) of lipids. Next, we used these two gene-sets to assess whether singe gene-trait associations in PhenomeXcan recapitulated lipids-related traits. We show that our gene module-based approach is more robust to identify genetic associations with lipids-relevant traits, and that it can be used to contextualize these results by identifing tissue and cell type-specific gene-trait associations.

First, using these two gene-sets, we assessed the genes' effects on all 3,752 phenotypes in PhenomeXcan by adding their standardized effect sizes 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. Other associated traits included asthma and lung function. We also performed the same operation for our gene module-based approach by considering 24 modules significantly enriched with the decreasing-lipids gene-set (Gene-set enrichment analysis, FDR < 0.05). In this case, we also found highly lipids-relevant traits among the top 25, including hypertension, blood pressure, specific cardiometabolic diseases like atherosclerosis, and celiac disease. This is particularly relevant because each of the 24 modules aggregated a specific weighted combination of almost 3,000 genes' effect sizes across all 3,752 traits. Thus, aggregating the effects of this number of genes and obtaining top-ranked lipids-relevant traits is highly unlikely to happen by chance ($P < 0.001$, see Methods), suggesting that gene modules (discovered with an unsupervised method) represent functionally meaningful units.

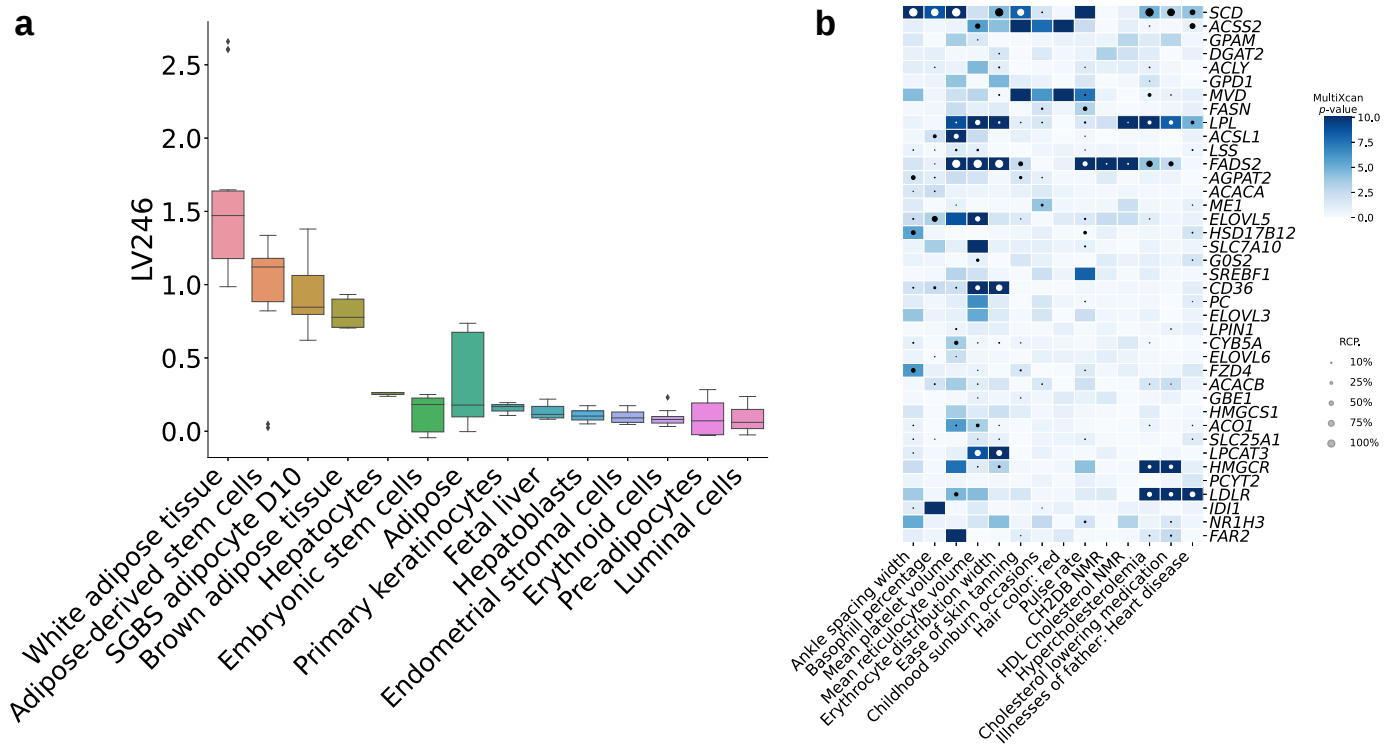


Figure 2: Tissues and traits associated with a gene module related to lipids metabolism (LV246). a) Top cell types/tissues where genes in LV246 are expressed on. Values in the y-axis come from matrix $\mathbf{B}$ in the MultiPLIER models (Figure 1 a). In the x-axis, cell types/tissues are sorted by the median value. **b)** Gene-trait associations (S-MultiXcan) and colocalization (fastENLOC) for the top traits in LV246. The top 40 genes in LV246 are shown, sorted by their module weigth, from largest (top gene *SCD*) to smallest (gene *FAR2*). SGBS: Simpson Golabi Behmel Syndrome; CH2DB: CH_2 groups to double bonds ratio; NMR: nuclear magnetic resonance; HDL: high-density lipoprotein.

When we considered the increasing-lipids set, genes and modules were associated with a more diverse set of traits, such as blood count tests, whole-body bioelectrical impedance measures, severe asthma, lung function, and rheumatoid arthritis. Additionally, gene modules were also associated with blood lipids, arterial stiffness, intraocular pressure, handgrip strength, and celiac disease. One gene module (LV246), significantly enriched with the increasing-lipids gene-set, was also associated with lipids metabolism and triglyceride biosynthesis pathways. In Figure 2 a, we used our module-based approach to show that LV246 genes are mainly co-expressed in adipose tissue, and to a less extent, liver cells (hepatocytes), which play key roles in coordinating and regulating lipids metabolism. This LV was associated with blood lipids, hypercholesterolemia, cholesterol lowering medication, and family history of heart disease, among others (Figure 2 b). Two high-confidence genes from our CRISPR screening, *DGAT2* and *ACACA* (responsible for encoding important enzymes for triglycerides and fatty acid synthesis), accounted for most of the gene-set enrichment signal for 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.

Notes:

- Improve description of CRISPR analysis.
- Genes *DGAT2* and *ACACA* are part of the high-confidence set, not the merged one (combining high and medium confidence). We might want to distinguish between them in Methods.
- It would be good at some point to have an LV that does not match a pathway. Otherwise, a reviewer could say “but this is similar to a method computing an association between pathways and traits, where is the novelty here?”. A potential candidate could be LV504, significantly enriched with the increasing-lipids gene-set, associated with medication for blood pressure, asthma, celiac

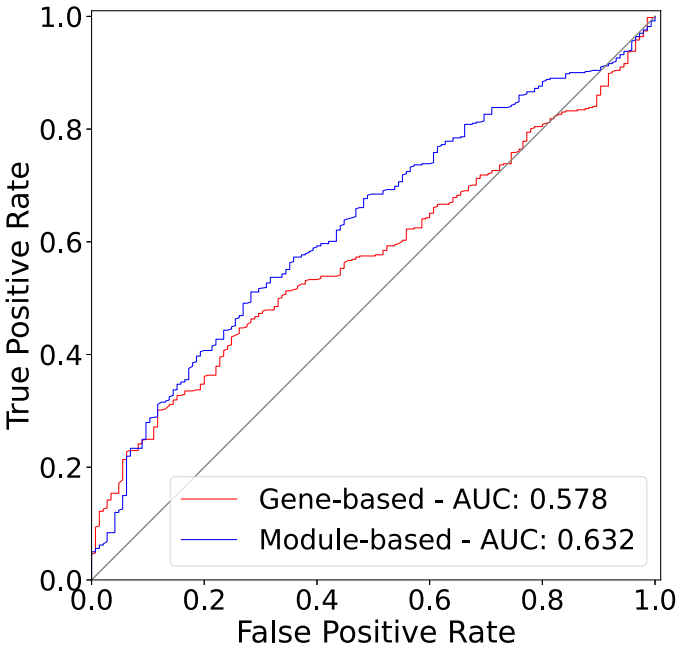
disease, and rheumatoid arthritis. Genes in this LV are expressed in skeletal muscle cells, intestinal subepithelial myofibroblasts, embryonic kidney cells, lung fibroblast cells, etc.
- We need to standardize the way we refer to our method (gene module-based approach, PhenoPLIER, etc).

Minor:

- Add -log10(p-value) in the legend of figure.
- Maybe make *DGAT2* and *ACACA* gene names bold in figure.
- It would be great to be able to say "this LV is *significantly associated* with this trait". Some reviewers might want that. Maybe we could use the Summary-MultiXcan approach to estimate the multivariate regression coefficients from individual genes associations, and get a p-value for the module-trait association. This could be a future small project, maybe an application note. One way to quickly compute a p-value is to use MAGMA gene-set analysis.

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 [29], which were further processed and mapped to DrubBank IDs [30,31,32]. Based on the established drug repurposing strategy that looks for reversed transcriptome patterns between genes and drug-induced perturbations [33,34], we used a framework for prioritizing drug candidates that uses imputed transcriptomes from GWAS [35]. 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 (Equation ??). We used a manually-curated gold standard of drug-disease medical indications [31,36] across 53 diseases and 322 compounds to evaluate and compare the prediction performance.



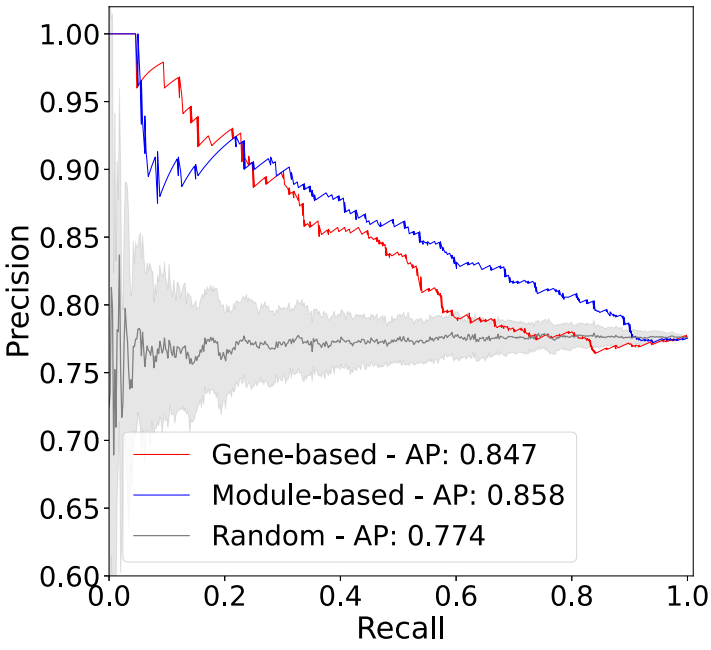


Figure 3: Drug-disease prediction performance for gene-based and module-based approaches. The receiver operating characteristic (ROC) (left) and the precision-recall curves (right) for a gene-based and our module-based approach. AUC: area under the curve; AP: average precision.

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 drug-disease pairs where both methods disagreed.

Nicotinic acid (niacin), a B vitamin widely used clinically to treat lipid disorders by exerting its effects on a number tissues, although not all its mechanisms have been documented [37,38]. 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 [37]. 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 and CAD, the module-based approach predicted niacin as a therapeutic drug by scoring them with 0.52 and 0.96 (above the mean), whereas the gene-based method assigned negative scores of -0.09 and -0.16 (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. 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 hepatocytes, its top genes encode important enzymes involved in lipid biosynthesis, and several of them 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).

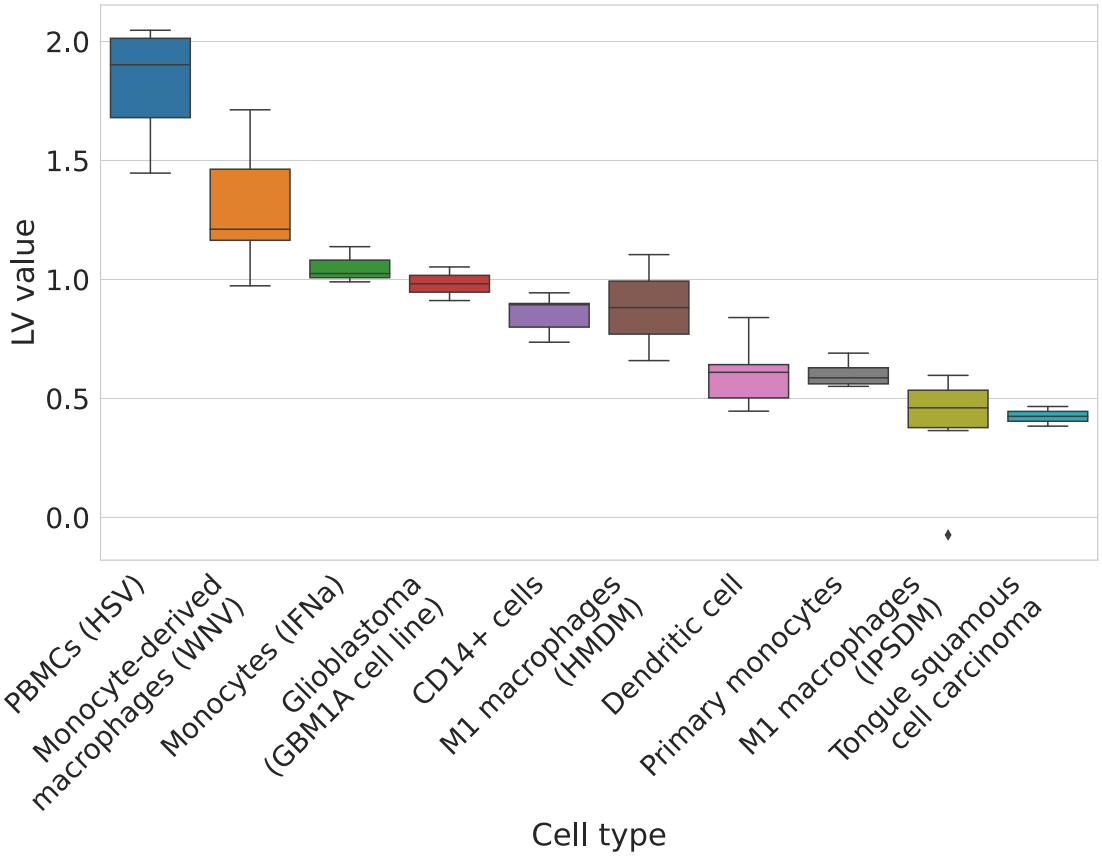


Figure 4: Cell types where the top 10 modules contributing for niacin-atherosclerosis prediction are expressed. Average module expression (y-axis) of different cell types (x-axis) across the top 10 latent variables/modules with a positive contribution for the niacin-AT prediction. The figure shows a clear immune cells signature, driven mainly by the top 2 modules: LV116 and LV931 (see Supplementary Figures 11 and 12). PBMCs: peripheral blood mononuclear cells; HSV: treated with herpes simplex virus; WNV: Infected with West Nile virus; IFNa: interferon-alpha treatment; HMDM: human peripheral blood mononuclear cell-derived macrophages; IPSDM: human induced pluripotent stem cell-derived macrophages;

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 [39,40]. 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 antidyslipidemic activity through the activation of GPR109A in immune cells [41], thus boosting anti-inflammatory processes and reversing cholesterol transport [42]. As shown in Figure 4, this alternative mechanism for niacin could have been hypothesized by examining the cell types where modules positively contributing to the niacin-AT prediction are expressed. Among these, we also found other potentially interesting modules that could represent mechanisms to explore, such as LV536 expressed in the bladder (Supplementary Figure 13) and LV885/LV840 expressed in kidneys (Supplementary Figures 14 and 15)

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 be also helpful to better understand the mechanism of pharmacological effect of known or experimental drugs. For example, one of the top modules affected by niacin (LV66, Supplementary Figure 16) is mainly expressed in ovarian granulosa cells. This compound has been very recently considered as a potential therapeutic for ovarian diseases [43,44], 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 hypothesis to evaluate potential mechanisms of action of different drugs.

`Notes/questions:`

- The main problem with this current section is the quality of LINCS L1000 data. It might be necessary to use Cmap build 02 also here, since there are some concerns about the LINCS imputation pipeline (https://think-lab.github.io/d/185/).
- It could be good to discuss cases where the gene-based approach performed better (there are several ones, even with cardiovascular diseases). This could potentially show that for some drug-disease pairs, maybe the compound targets a few genes instead of being a broad-spectrum/multi-tissue/multi-cell-type one like niacin.
- It would be great to have an expert in cardiovascular diseases and lipid disorder to review this part.
- Is it clear the message in Figure 4 ? An alternative is to just show the most interesting LVs instead of averaging all and showing the top cell types as it is now.

Ideas/minor:

- Maybe as part of the manuscript, we can provide the drug-predictions for all traits in PhenomeXcan for download.
- An interesting analysis could consist in keeping LVs aligned with pathways only; what happens with prediction performance? If it goes down, it means that among not-aligned LVs we have useful information to link diseases and drugs. It would be nice to be able to claim that.

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

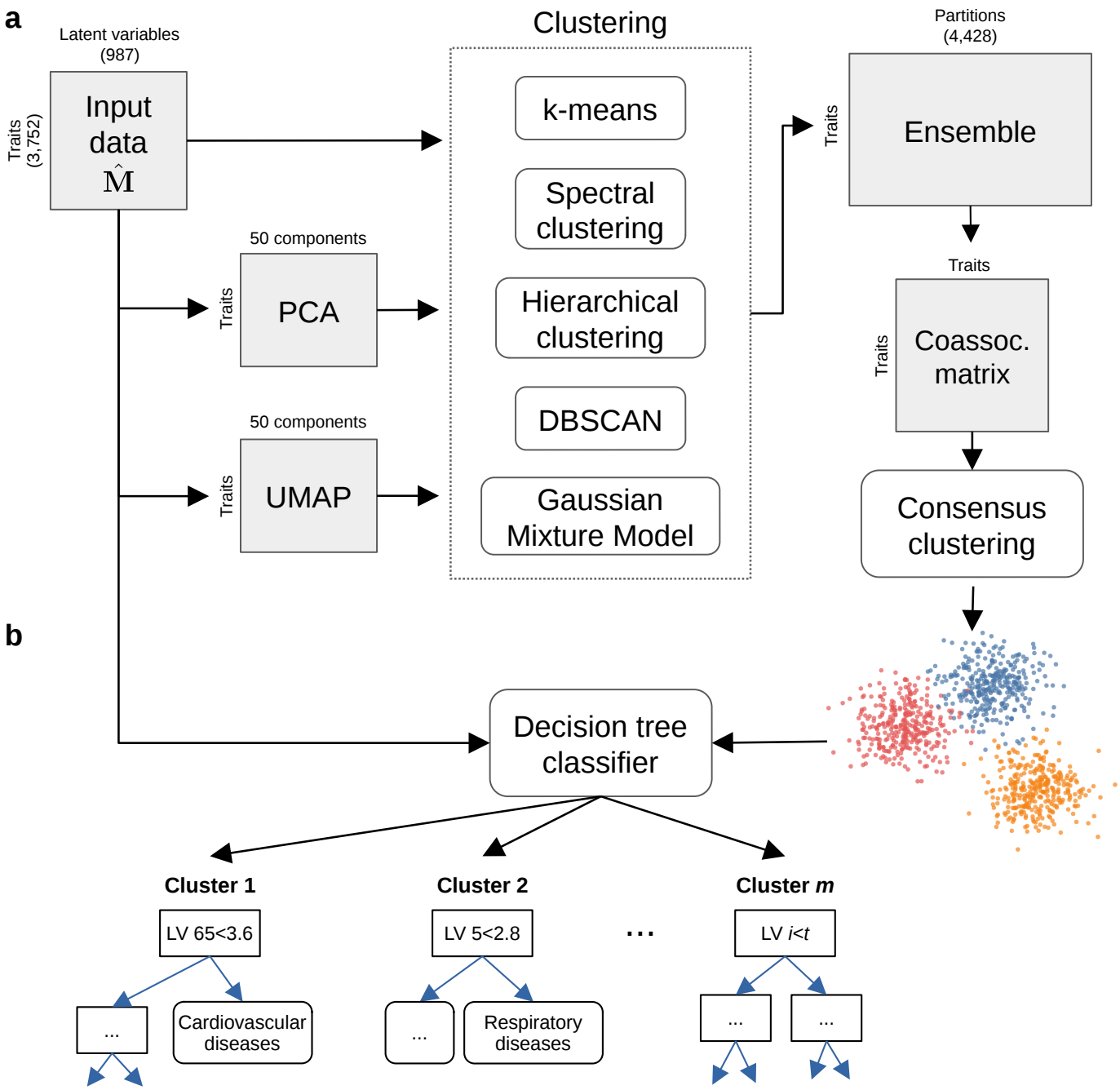


Figure 5: Cluster analysis on traits using the latent gene expression representation. a) The projection of TWAS results on n=3,752 traits into the latent gene expression representation is the input data to the clustering process. A linear (PCA) and non-linear (UMAP) dimensionality reduction techniques 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 6). **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 $\hat{\mathbf{M}}$ represents a less noisy low-dimentional version of the data. Thus, we conducted cluster analysis on $\hat{\mathbf{M}}$ 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 parameters were applied on the data, and later combined into a consolidated solution [45,46,47] (Figure 5). Our clustering pipeline generated 15 final consensus clustering solutions with 5 to 29 clusters (Supplementary Figure 33). Instead of selecting a

specific number of clusters, we used a clustering tree [48] (Figure 6) 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).

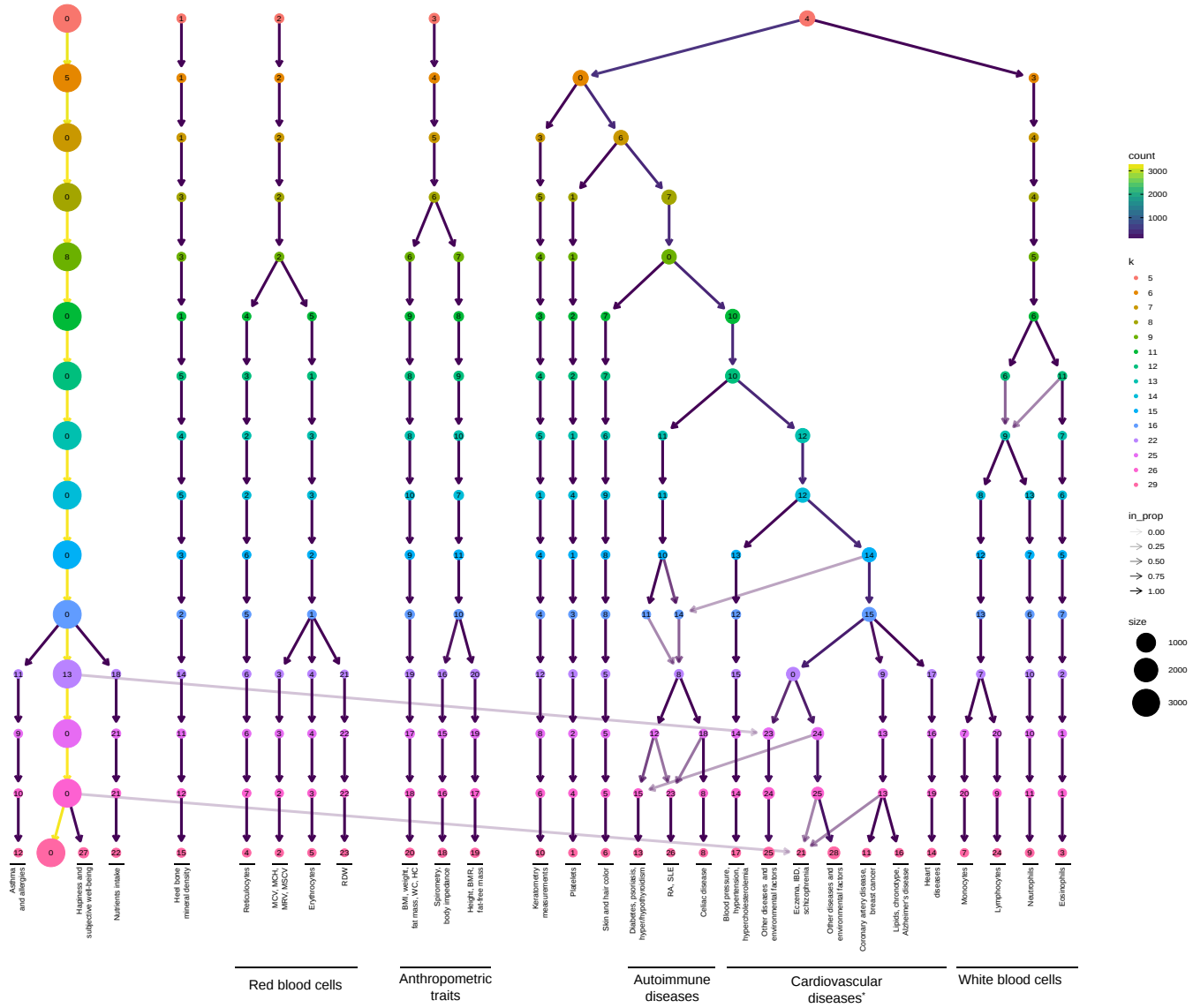


Figure 6: Clustering tree using multiple resolutions for clusters of traits. Each row represents a partition/grouping of the traits, and each circle is a cluster from that partition, and the number of clusters go from 5 to 29. Arrows indicate how traits in one cluster move across clusters from different 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 sphered 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 6 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 infraction, hypercholesterolemia, and other cardiovascular-related traits such hand-grip strength [49], 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.

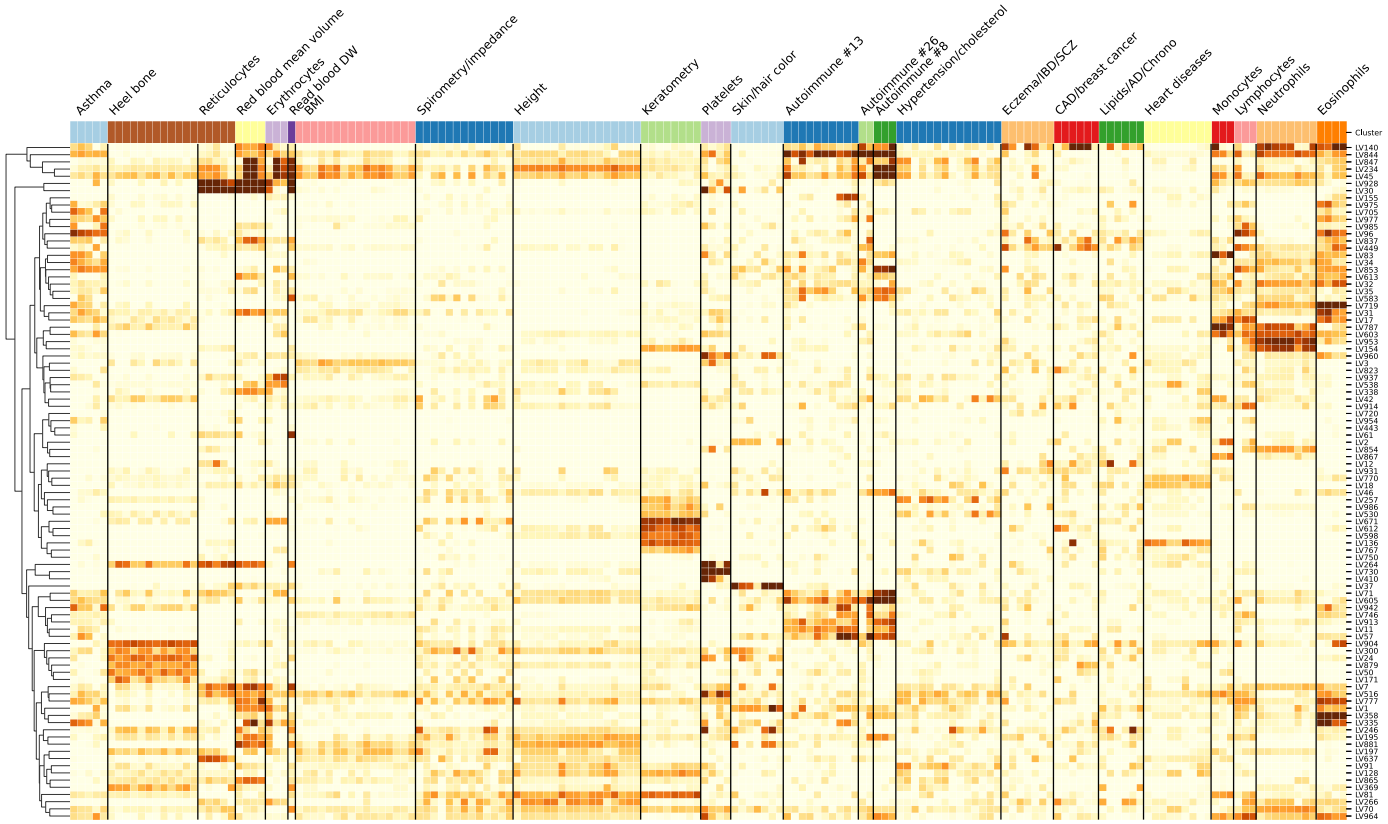


Figure 7: Cluster-specific and general transcriptional processes. The plot shows a submatrix of $\hat{\mathbf{M}}$ for the main trait clusters at k=29, considering only gene modules (rows) that align well with at least one known pathway. Values are standardized from -5 (lighter color) to 16 (darker color).

Next, we analyzed which gene modules are driving these clusters of traits. For that, we trained decision tree classifiers on the input data (Figure 5) using each cluster at k=29 (bottom of Figure 6) 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 7, 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 17 and 18), which are known to be related to early progenitors of the erythrocytes lineage [50], 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 19), and strongly associated with hemathological assays on platelets; or LV598, whose genes are expressed in corneal endothelial cells (Supplementary Figures 20) and associated to keratometry measurements (FDR < 0.05; Supplementary Table 1).

The autoimmune diseases sub-branch also has significant gene modules associations expressed in relevant cell types. LV155 was strongly expressed in thyroid (Supplementary Figures 21), and significantly associated with hypothyroidism both in PhenomeXcan and eMERGE (FDR < 0.05; Supplementary Tables 2 and 3). LV844 was the most strongly associated gene module with autoimmune disorders (FDR < 1e-15; Supplementary Tables 4 and 5), 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 22). Other important gene modules associated with autoimmunity in both PhenomeXcan and eMERGE are LV57 expressed in T cells (Supplementary Figure 23, and Supplementary Tables 6 and 7), and LV54 expressed in different soft tissue tumors, breast, lung, pterygia and epithelial cells (Supplementary Figure 24, and Supplementary Tables 8 and 9).

The cardiovascular sub-branch also exhibited significant associations, such as LV847 (Supplementary Figure 25) with blood pressure traits and hypertension (Supplementary Tables 10 and 11), which was expressed in CD19 (B cells) (which are related to preeclampsia [51]), Jurkcat 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 [52,53]). 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 12 and 13), and expressed in a wide range of cell types, including fibroblasts, mesenchymal stem cells, osteoblasts, pancreatic stellate cells, cardiomyocytes, and adipocytes (Supplementary Figure 26). 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 27, and Supplementary Tables 14 and 15), LV206 (Supplementary Figure 28, and Supplementary Tables 16 and 17), and LV260 (Supplementary Figure 29, and Supplementary Tables 18 and 19). 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 [54,55,56,57], and share several risk factors, including hypertension, high cholesterol, obesity, smoking, among others [58,59]. 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 20 and 22). LV21 was strongly expressed in a variety of soft tissue sarcomas, monocytes/macrophages (including microglia from cortex samples), and aortic valves (Supplementary Figure 30); as discussed previously, macrophages play a key role in the reverse cholesterol transport and thus atherogenesis [60]. LV5 was expressed in breast cancer and brain glioma samples, microglia (cortex), liver, and kidney, among other cell types (Supplementary Figure 31). 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 24), 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 32). 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

Conclusions

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 · PMID: 20624743 · PMCID: 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 · PMID: 14707169 · PMCID: 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 · PMID: 19104045 · PMCID: 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 · PMID: 32913098 · PMCID: 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 · PMID: 32728217 · PMCID: PMC7422677

6. **Mechanisms of tissue and cell-type specificity in heritable traits and diseases**
   Idan Hekselman, Esti Yeger-Lotem
   Nature Reviews Genetics (2020-01-08) https://doi.org/ggkx9v
   DOI: 10.1038/s41576-019-0200-9 · PMID: 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 · PMID: 33536621 · PMCID: 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 · PMID: 28398307 · PMCID: 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 · 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 · 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 · 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 · 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 · 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/f7p9zv
DOI: 10.1038/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 · PMID: 28238358 · PMCID: PMC5339290

16. **Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types**
Hilary K. Finucane, Yakir A. Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, … The Brainstorm Consortium
Nature Genetics (2018-04-09) https://doi.org/gdfjqt
DOI: 10.1038/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 · 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 · PMID: 24813542 · PMCID: PMC4147889

19. **Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine**
Padhraig Gormley, Verneri Anttila, Bendik S Winsvold, Priit Palta, Tonu 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 · PMID: 27322543 · PMCID: PMC5331903

20. **Biological interpretation of genome-wide association studies using predicted gene functions**
Tune H. Pers, Juha M. Karjalainen, Yingleong Chan, Harm-Jan Westra, Andrew R. Wood, Jian Yang, Julian C. Lui, Sailaja Vedantam, Stefan Gustafsson, Tonu Esko, … Genetic Investigation of ANthropometric Traits (GIANT) Consortium
Nature Communications (2015-01-19) https://doi.org/f3mwhd
DOI: 10.1038/ncomms6890 · PMID: 25597830 · PMCID: PMC4420238

21. **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 · PMID: 20838408 · PMCID: PMC3880143

22. **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 · PMID: 31249421 · PMCID: PMC7262669

23. **PhenomeXcan: Mapping the genome to the phenome 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 · PMID: 32917697

24. **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 · PMID: 31121115 · PMCID: PMC6538307

25. **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 · PMID: 30305743 · PMCID: PMC6786975

26. **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 · PMID: 16869964 · PMCID: PMC1560386

27. **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 · PMID: 31744546 · PMCID: PMC6864930

28. **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 · PMID: 25885710 · PMCID: PMC4401657

29. **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 · PMID: 29195078 · PMCID: PMC5990023

30. **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 · PMID: 24203711 · PMCID: PMC3965102

31. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, 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 · PMID: 28936969 · PMCID: PMC5640425

32. **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

33. **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 · PMID: 21849664 · PMCID: PMC3479650

34. **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 · PMID: 21849665 · PMCID: PMC3502016

35. **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 · PMID: 28805813

36. **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

37. **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 · PMID: 18375237

38. **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 · PMID: 30782960 · PMCID: PMC6446705

39. **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 · PMID: 16766048

40. **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/dgxg75
DOI: 10.1038/sj.jid.5700586 · PMID: 17008871

41. **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 · PMID: 21317532 · PMCID: PMC3048854

42. **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/atvbaha.108.179283 · PMID: 19797709 · PMCID: PMC2812788

43. **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 · PMID: 30415247

44. **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 · PMID: 33751926

45. **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 · PMID: 16355656

46. **Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions**
Alexander Strehl, Ghosh Joydeep
Journal of Machine Learning Research https://www.jmlr.org/papers/v3/strehl02a.html

47. **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

48. **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 · PMID: 30010766 · PMCID: PMC6057528

49. **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 · PMID: 25982160

50. **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 · PMID: 21241896 · PMCID: PMC3049864

51. **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 · PMID: 22353610

52. **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 · PMID: 30218068 · PMCID: PMC6138636

53. **Estrogen and hypertension**
Muhammad S. Ashraf, Wanpen Vongpatanasin
Current Hypertension Reports (2006-09) https://doi.org/d638rf
DOI: 10.1007/s11906-006-0080-1 · PMID: 16965722

54. **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 · PMID: 12093424

55. **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 · PMID: 24190959 · PMCID: PMC4107269

56. **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 · PMID: 28684403 · PMCID: PMC5579328

57. **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/gjzjgq
DOI: 10.1161/circresaha.118.313563 · PMID: 30605407 · PMCID: PMC6319653

58. **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/f6pjd4
DOI: 10.1186/s12916-014-0206-2 · PMID: 25385447 · PMCID: PMC4226890

59. **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 · PMID: 17483665

60. **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 · PMID: 16258026

61. **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 · PMID: 26771021 · PMCID: PMC4707969

62. **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 · PMID: 26392547 · PMCID: PMC4603492

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

Methods

MultiPLIER and Pathway-level information extractor (PLIER)

MultiPLIER [24] extracts patterns of co-expressed genes from recount2 [8], a large gene expression dataset. The approach applies the pathway-level information extractor method (PLIER) [22], which performs unsupervised learning using prior knowledge (cannonical 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 (i.e. a set of genes with coordinated expression patterns). This reduced the data dimensionality into 987 latent variables.

Given a gene expression dataset $\mathbf{Y}^{n \times p}$ with n genes and p conditions and a prior knowledge matrix $\mathbf{C} \in \{0,1\}^{n \times m}$ for m gene sets (so that $\mathbf{C}_{ij} = 1$ if gene i belongs to gene set j), (e.g., gene sets from MSigDB [61]), PLIER finds $\mathbf{U}$, $\mathbf{Z}$, and $\mathbf{B}$ minimizing

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

subject to $\mathbf{U} > 0, \mathbf{Z} > 0$; $\mathbf{Z}^{n \times l}$ are the gene loadings with l latent variables, $\mathbf{B}^{l \times p}$ is the latent space for p conditions, $\mathbf{U}^{m \times l}$ specifies which of the m prior-information gene sets 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 a known or novel gene module (i.e., a meaningful biological pattern) or noise.

CRISPR-Cas9 screening

`Add details`

Consensus clustering of traits in PhenomeXcan

Dimensionality reduction

Ensemble creation: clustering algorithms and parameters

- list of methods and its parameters

Ensemble combination and consensus functions

- evidence accumulation approach
- we also used a spectral clustering approach
- for each k, we picked the partition that maximized the agreement with the ensemble
- show figure where we select the ks that are greater than the median

Clusters interpretation

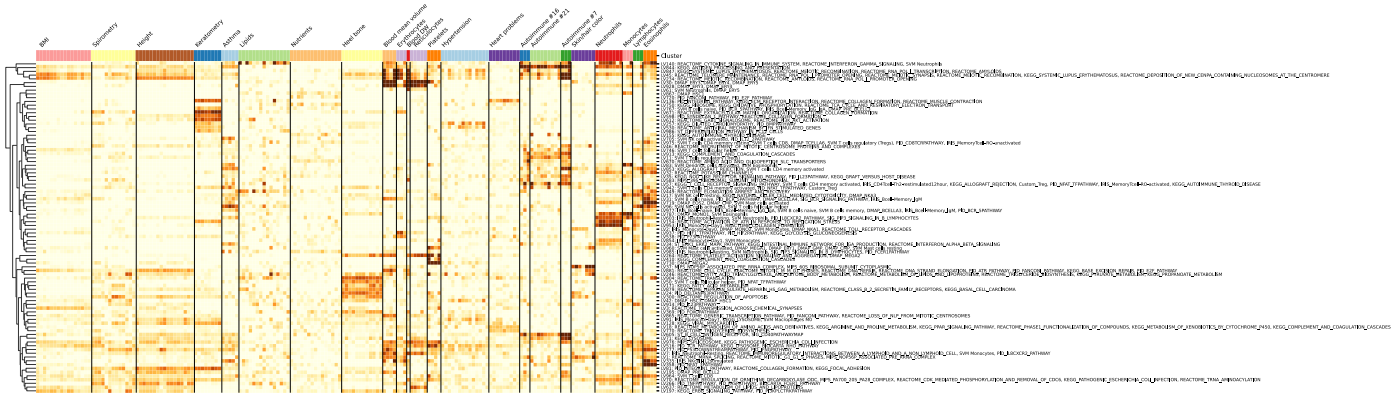


Figure 8: Pathways associated with gene modules in Figure 7. `(Early draft version; instead of this figure, we might want to just add a table with information about LVs)` This figure is equivalent to Figure 7 but, instead of cell types, labels on the right show all the associated pathways for each latent variable.

Drug-disease predictions

NOT FINISHED

We used the dot product of the S-PrediXcan z-score for each gene-disease pair, and the z-score for each gene-drug pair in LINCS L1000, multiplied by -1.

To obtain a drug-disease association for the gene module-mapped TWAS results, we first projected LINCS L1000 data into this latent representation using Equation (??), thus leading to a matrix with the expression profiles of drugs mapped to latent variables. This can be interpreted as the effects of compounds on gene modules activity. Then, similarly as before, we anti-correlated gene module-traits scores and module expression profiles of drugs.

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 esimates than expected by chance. Probably I just need to add a proper caption for each figure, and reference them from the main text.

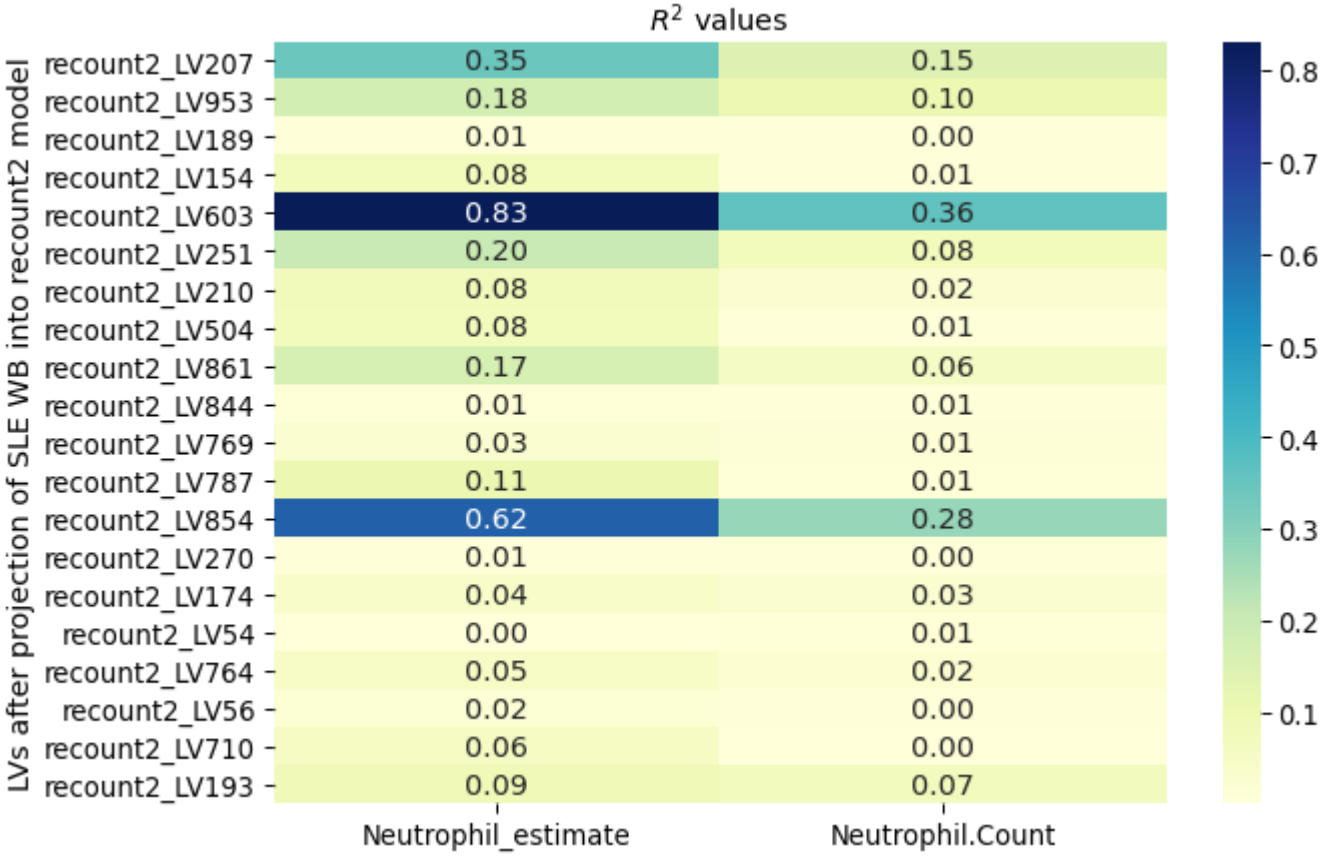


Figure 9: Correlation of neutrophil counts with top LVs associated with neutrophils traits.

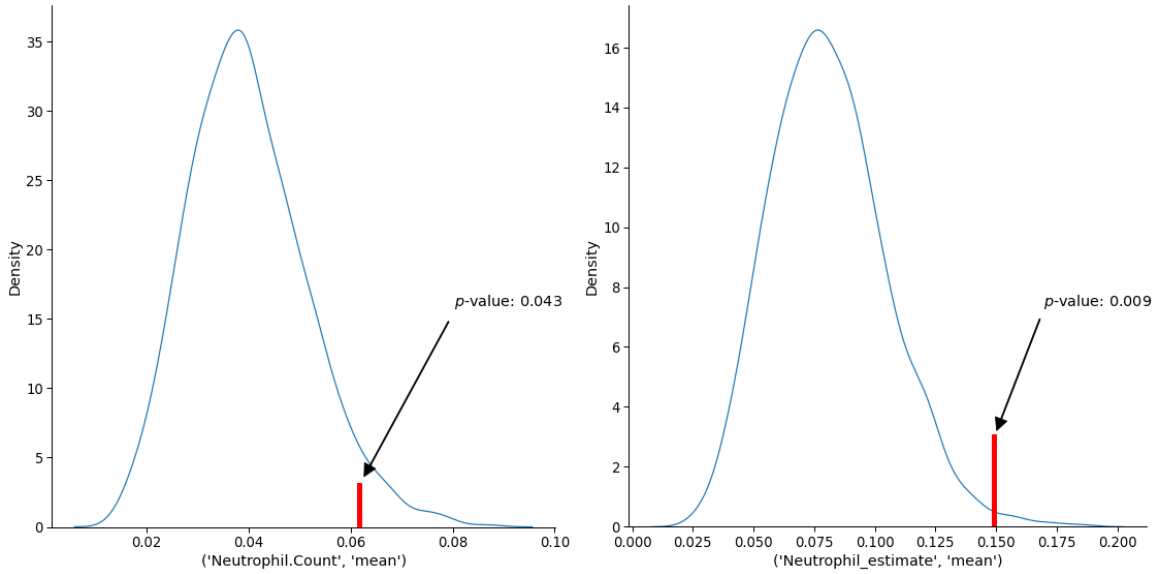


Figure 10: Significance of neutrophil counts correlation.

LV116 cell types

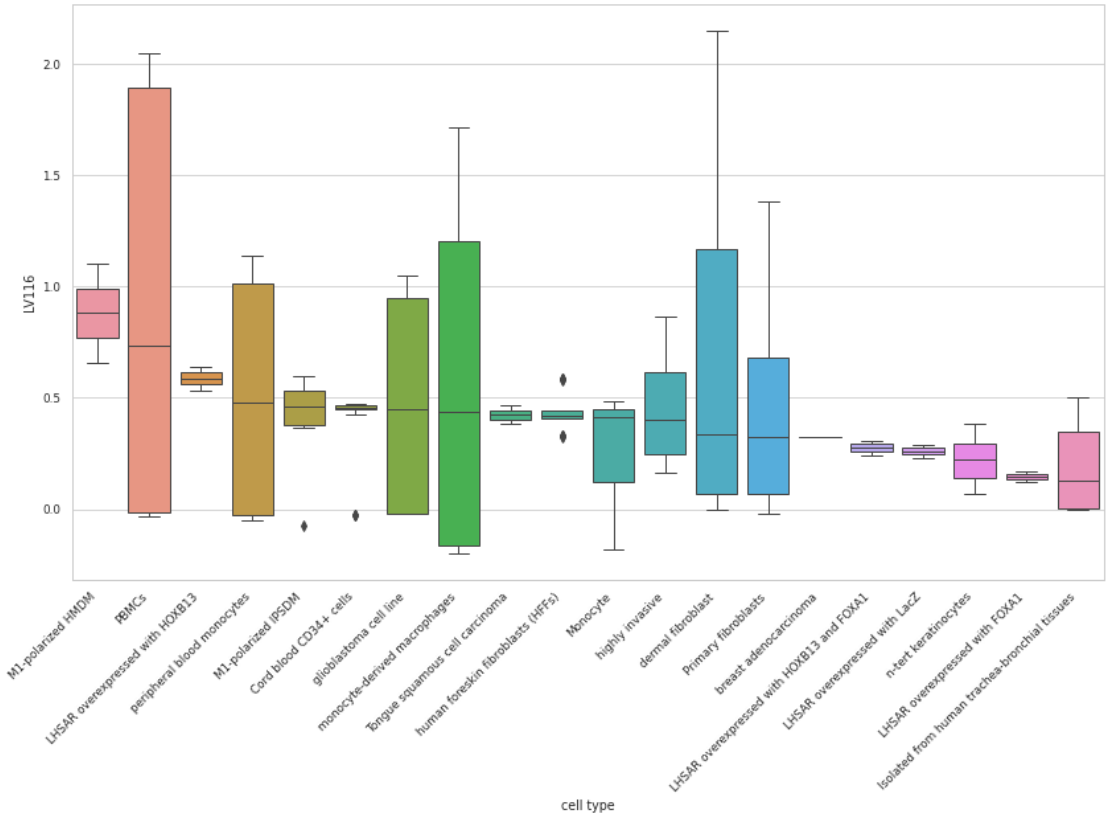


Figure 11: Cell types for LV116.

LV931 cell types

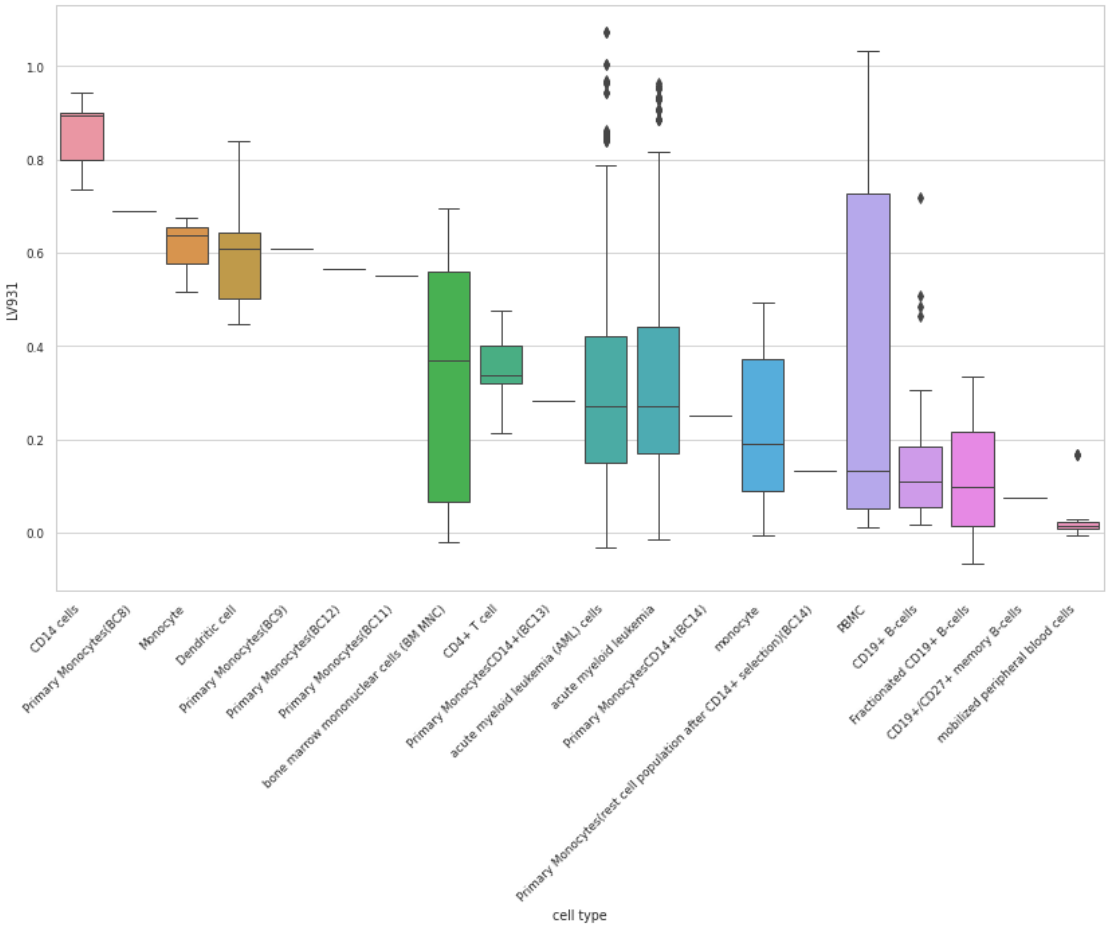


Figure 12: Cell types for LV931.

LV536 cell types

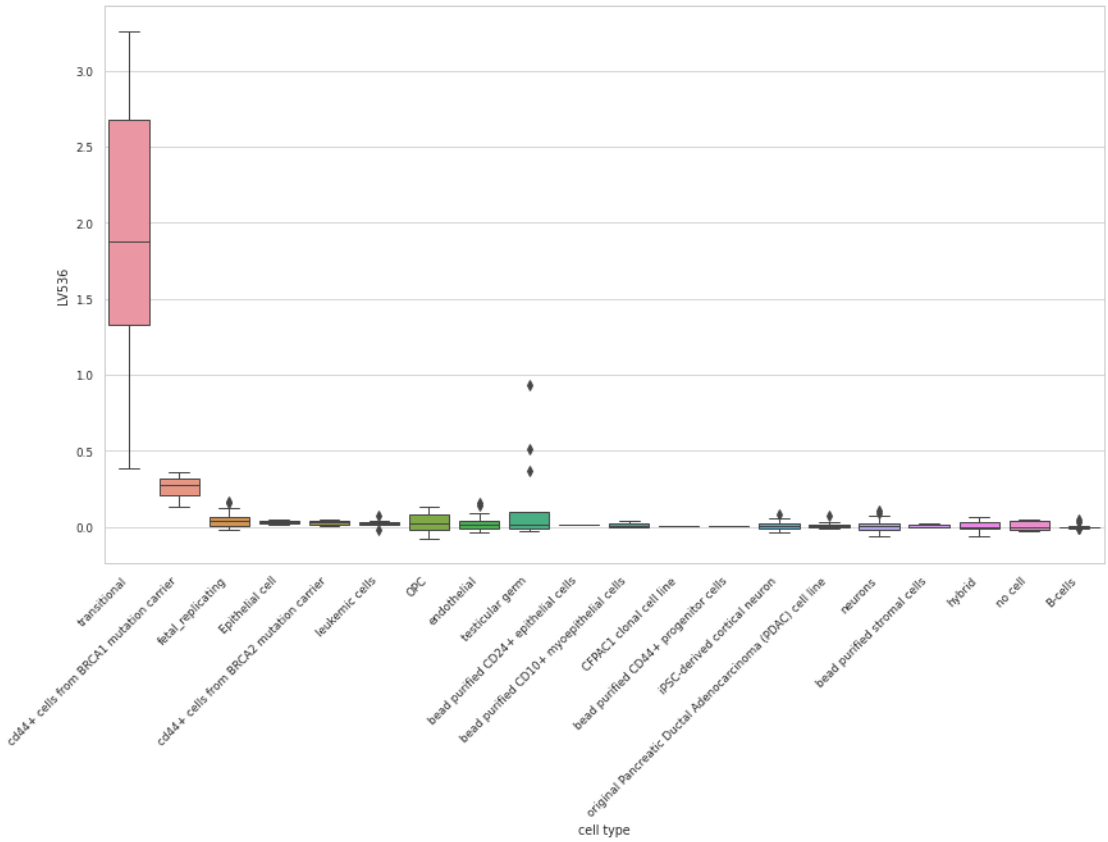


Figure 13: Cell types for LV536. FIXME: *transitional* here refers to bladder: https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP007947

LV885 cell types

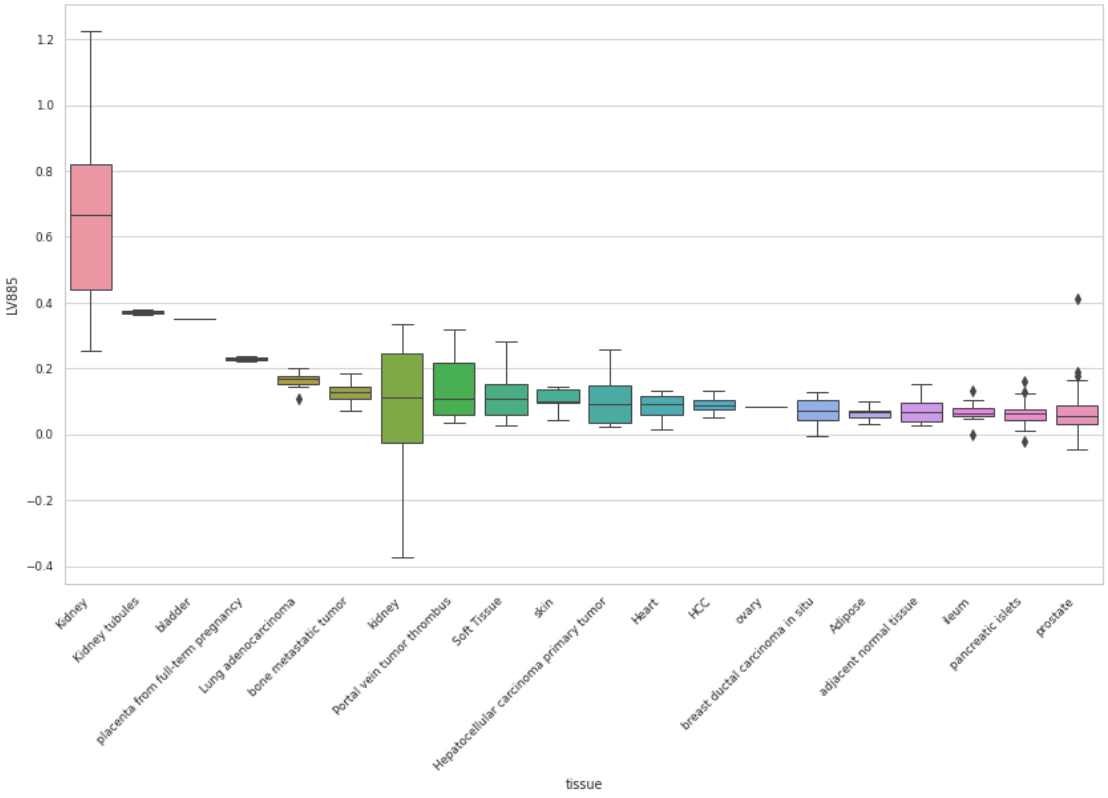


Figure 14: Cell types for LV885.

LV840 cell types

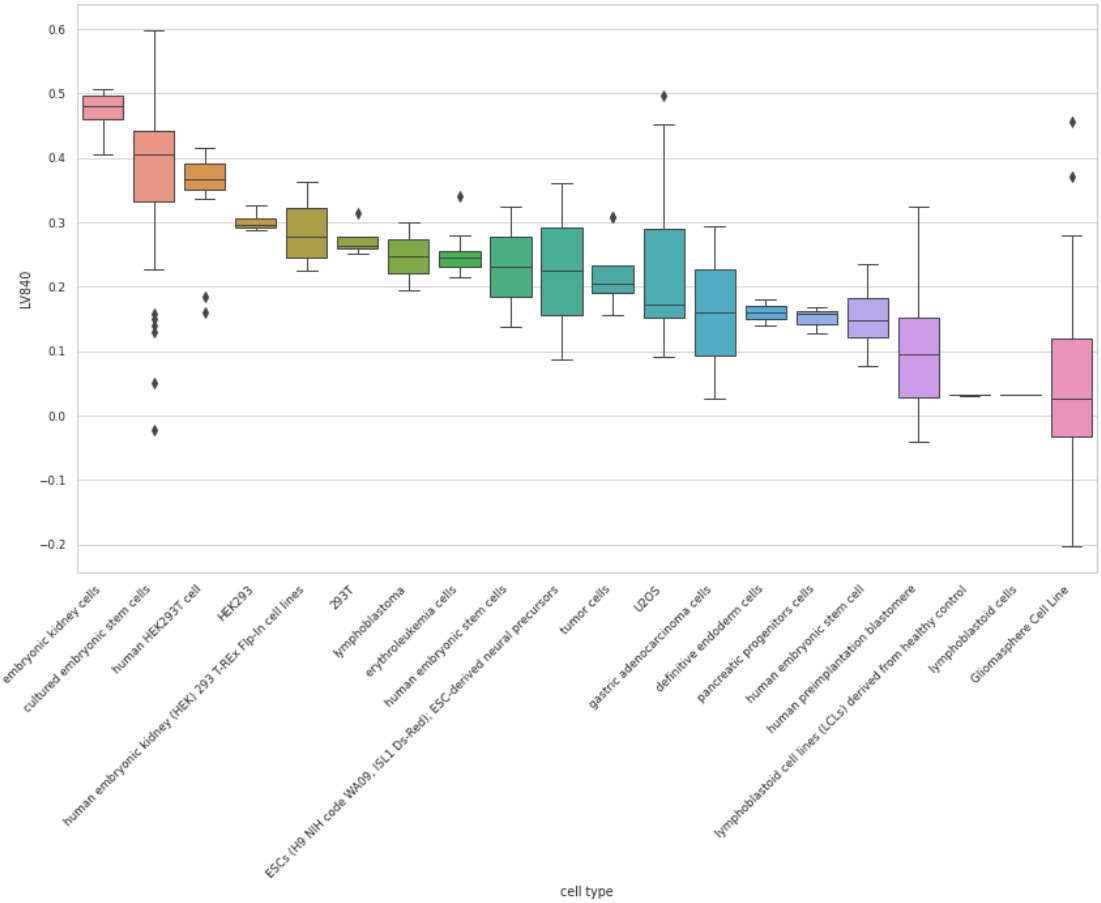


Figure 15: Cell types for LV840.

LV66 cell types

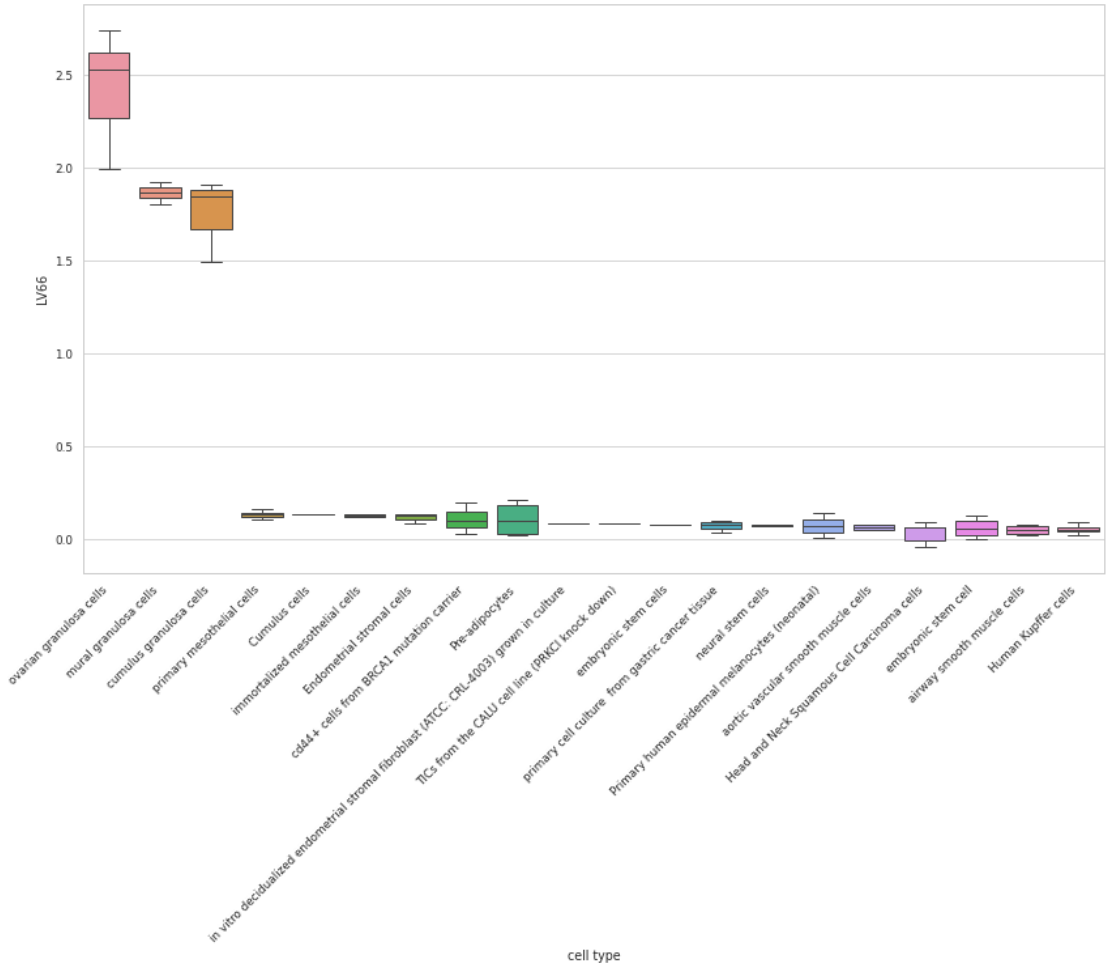


Figure 16: Cell types for LV66.

LV928

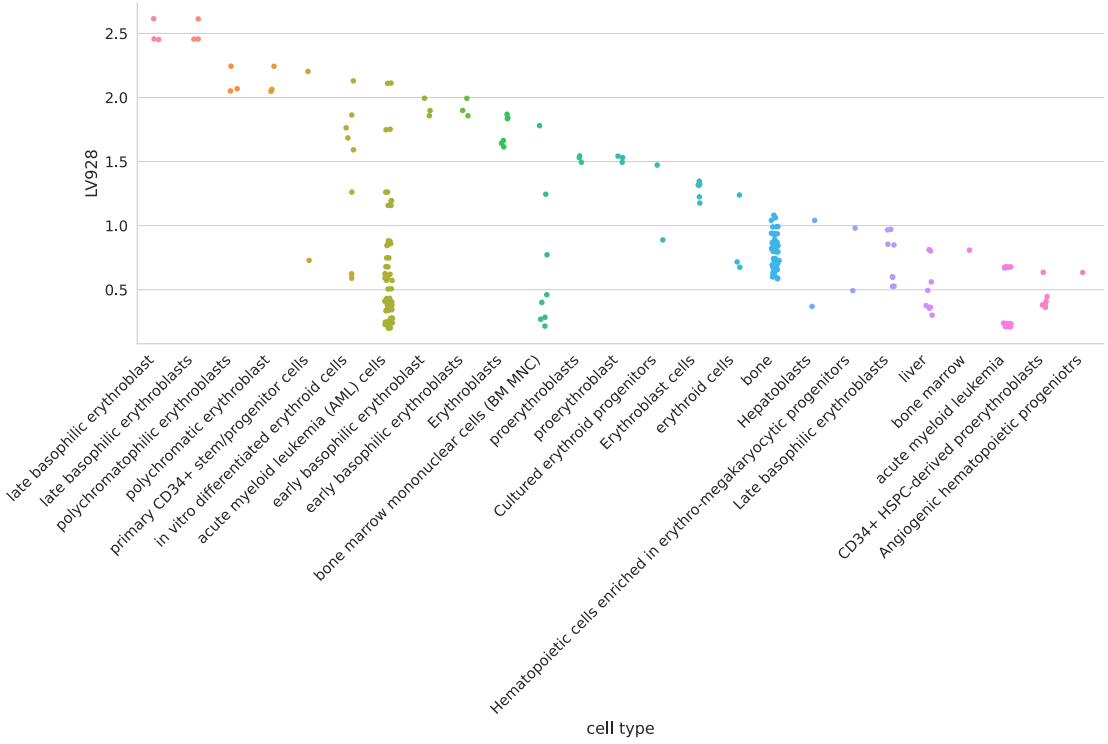


Figure 17: Cell types for LV928.

LV30

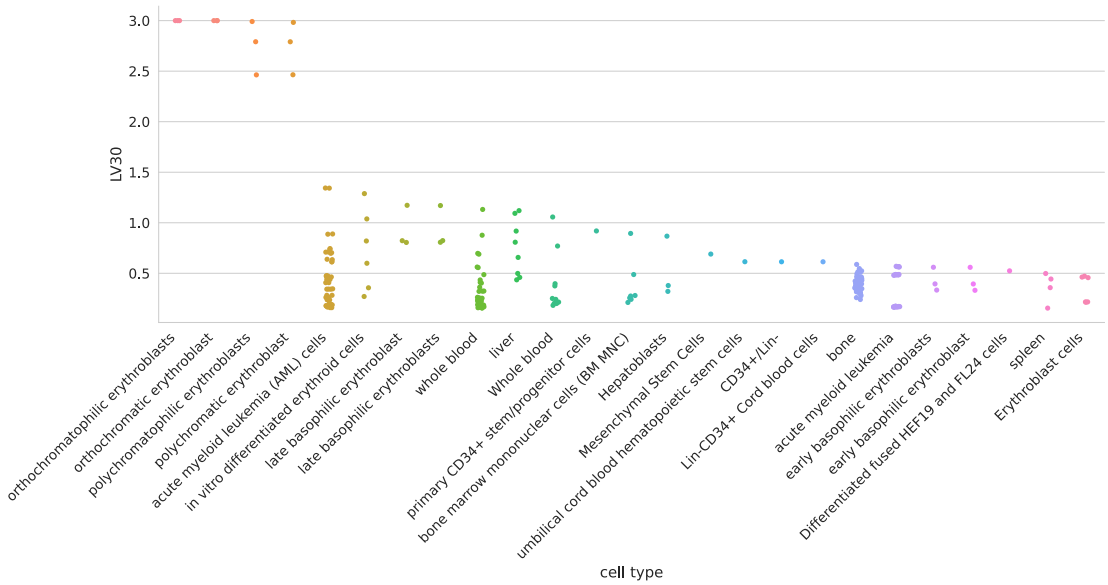


Figure 18: Cell types for LV30.

LV730

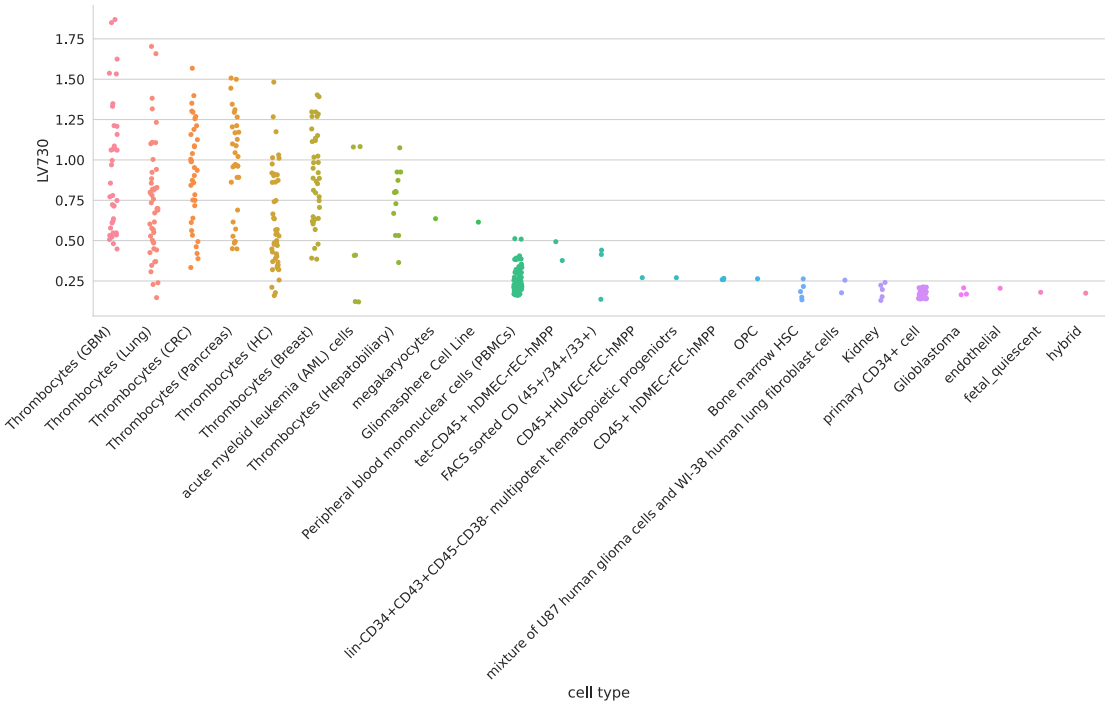


Figure 19: Cell types for LV730.

LV598

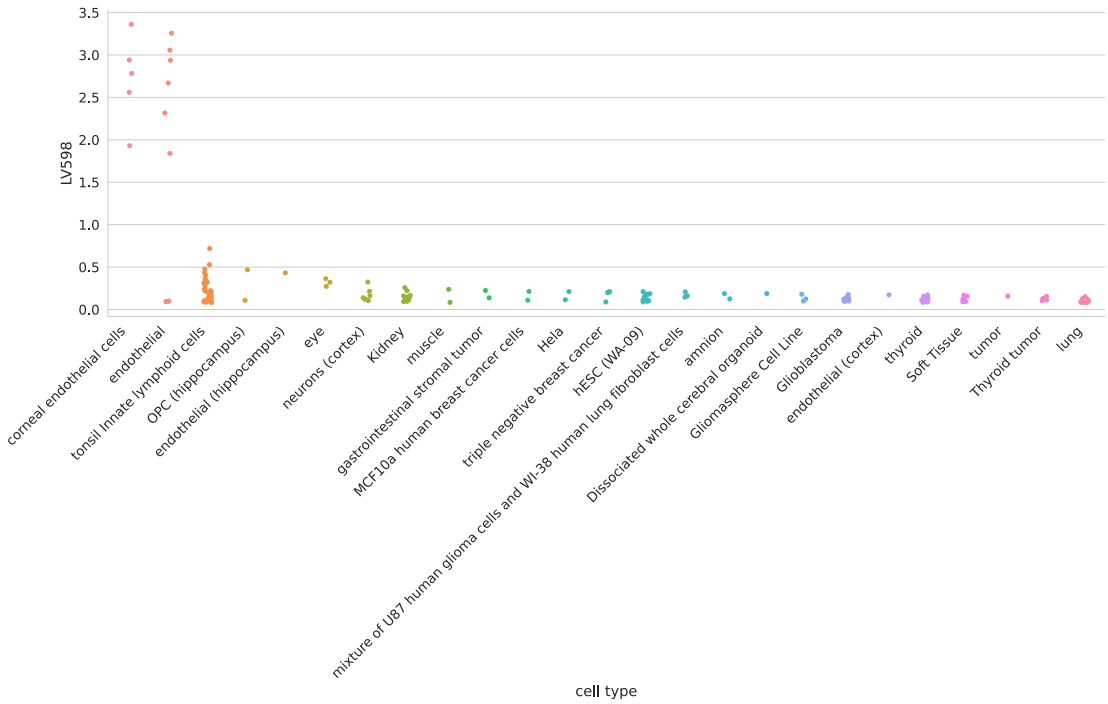


Figure 20: Cell types for LV598.

Table 1: 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

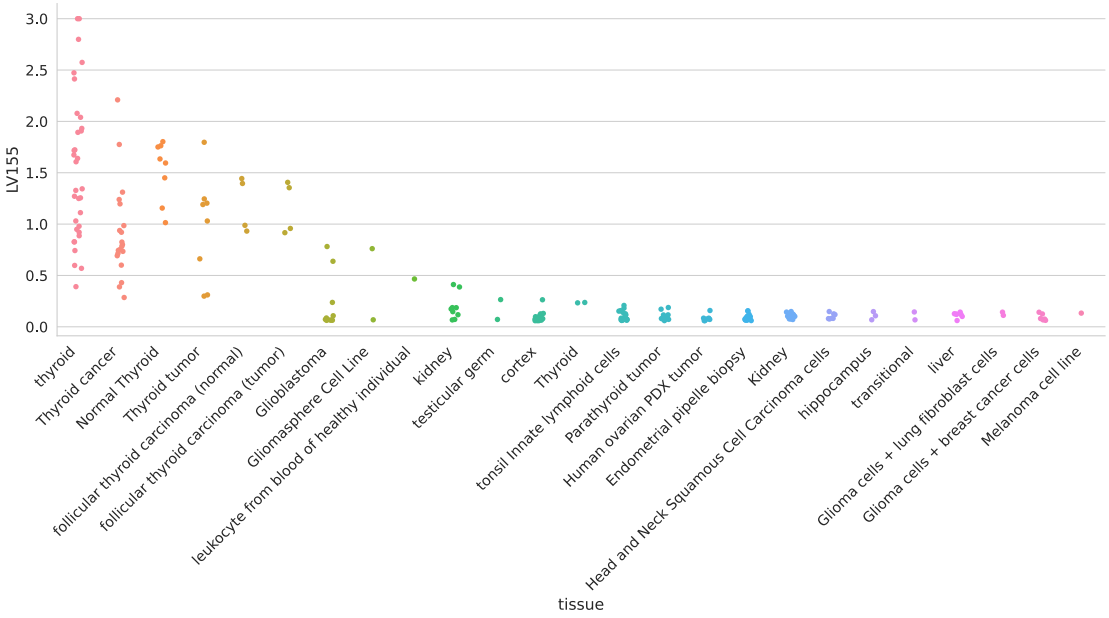


Figure 21: Cell types for LV155.

Table 2: 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	17,574	29 / 13	2.01e-03
Non-cancer illness code, self-reported: hyperthyroidism/thyrotoxicosis	361,141	2,730	29 / 13	1.29e-02
Treatment/medication code: levothyroxine sodium	361,141	14,689	29 / 13	1.41e-02

Table 3: 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

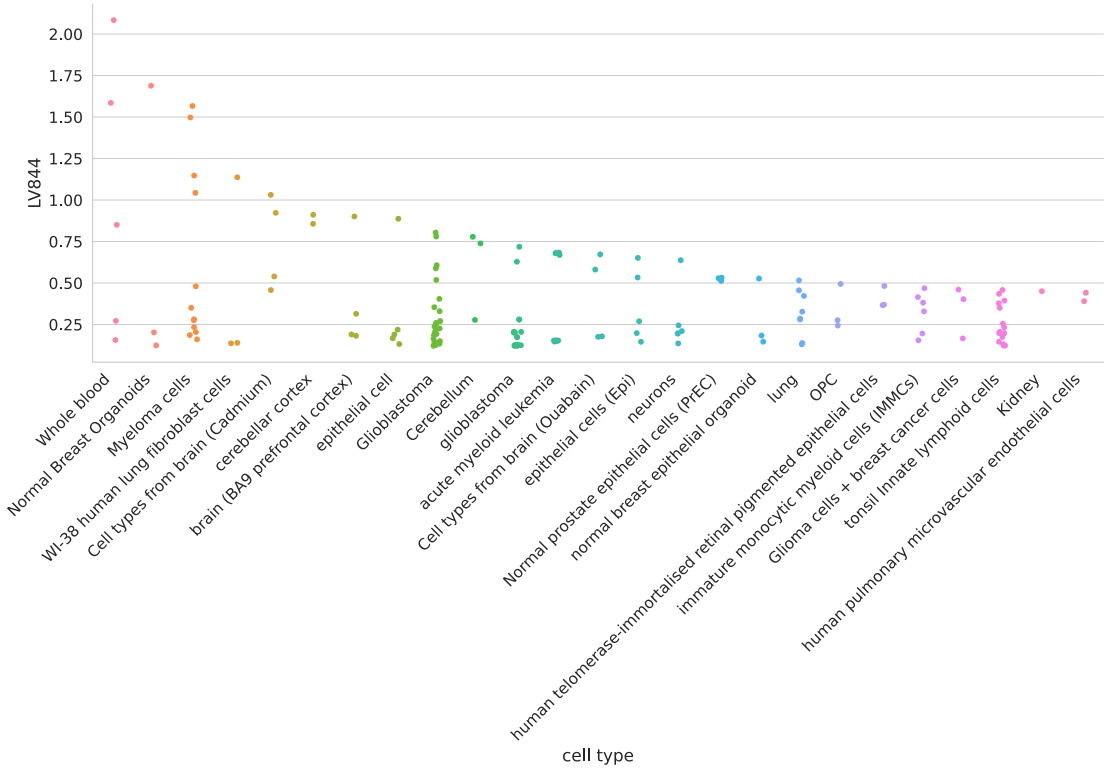


Figure 22: Cell types for LV844. Note "Cell types from brain" come from [62], treated with different chemichals; I don't have enough information to separate cell types.

Table 4: 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 5: 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

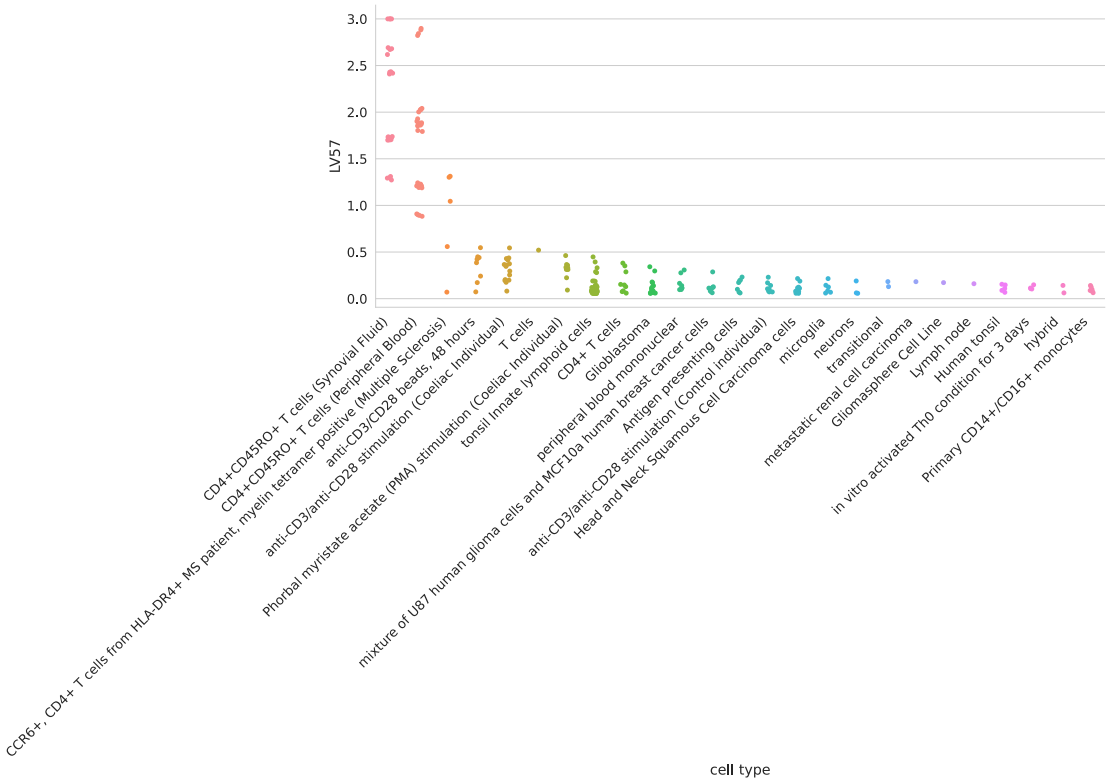


Figure 23: Cell types for LV57.

Table 6: 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 7: 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

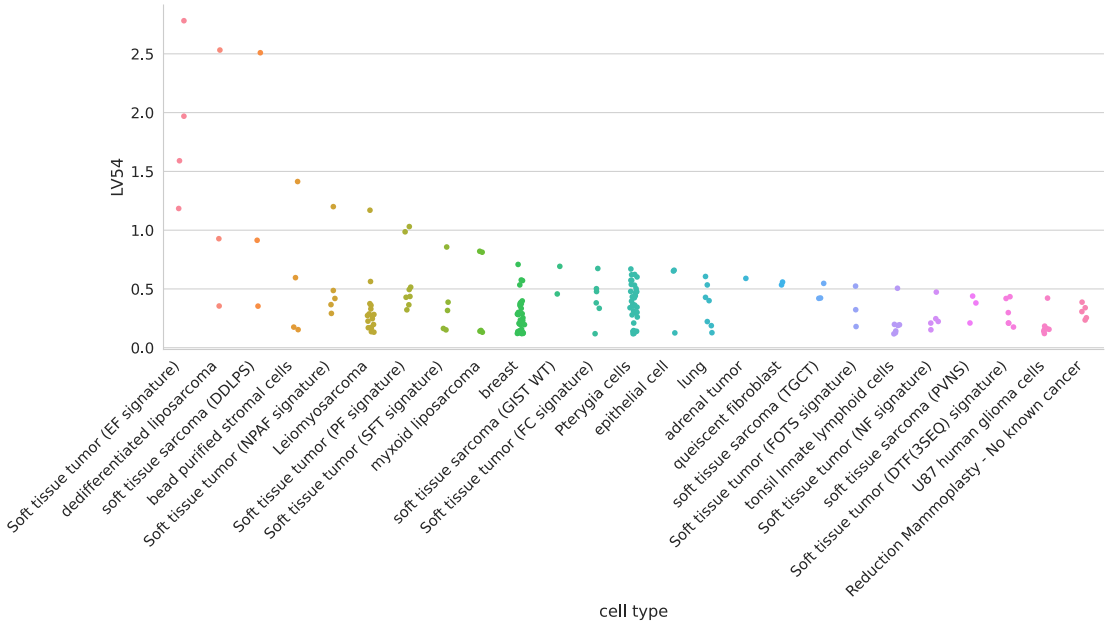


Figure 24: Cell types for LV54.

Table 8: 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	9,573	29 / 13	6.33e-23
Systemic Lupus Erythematosus	23,210	7,219	29 / 26	1.32e-22
Started insulin within one year diagnosis of diabetes	16,415	1,999	29 / 13	3.84e-20
Non-cancer illness code, self-reported: hyperthyroidism/thyrotoxicosis	361,141	2,730	29 / 13	9.59e-19
Treatment/medication code: insulin product	361,141	3,545	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	1,587	29 / 8	1.36e-14
Medication for cholesterol, blood pressure or diabetes: Insulin	165,340	2,248	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 9: 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

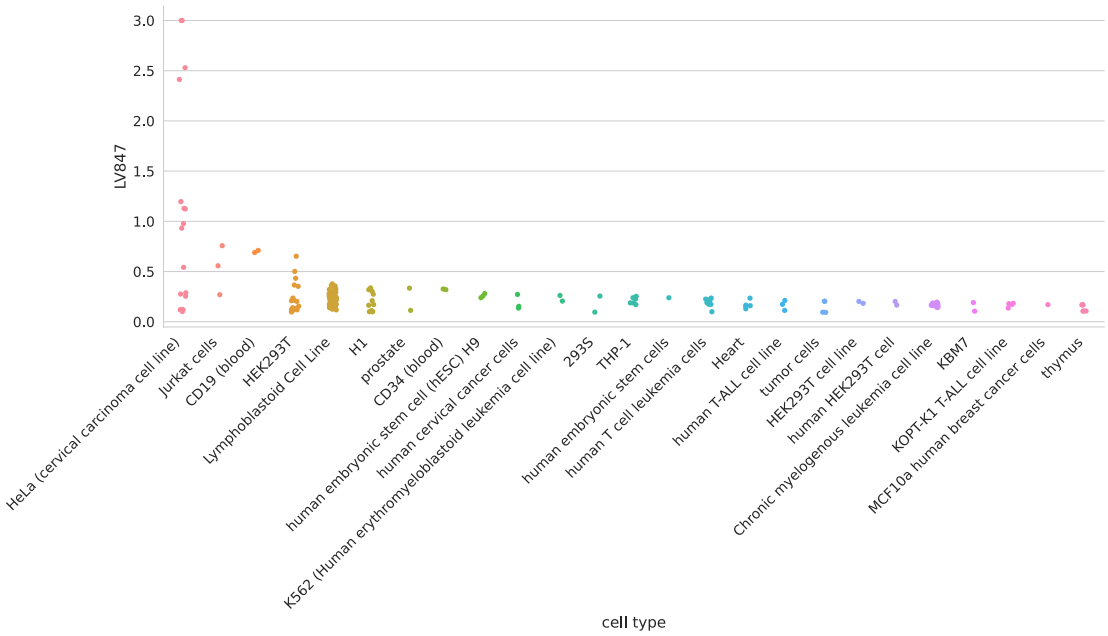


Figure 25: Cell types for LV847.

Table 10: 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 11: 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

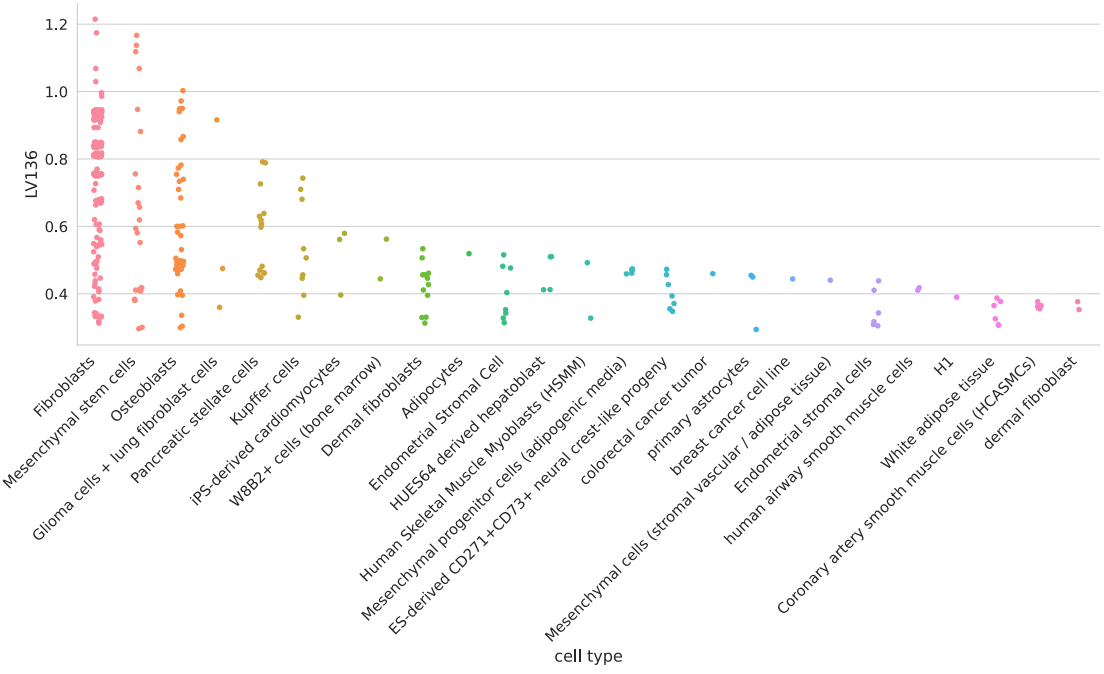


Figure 26: Cell types for LV136. Pulmonary microvascular endothelial cells were exposed to hypoxia for 24 hours or more [63];

Table 12: 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	60,801	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	14,334	29 / 14	3.90e-06
Ischaemic heart disease, wide definition	361,194	20,857	29 / 14	7.22e-06
Vascular/heart problems diagnosed by doctor: Heart attack	360,420	8,288	29 / 14	2.93e-04
Myocardial infarction	361,194	7,018	29 / 14	6.33e-04
Myocardial infarction, strict	361,194	7,018	29 / 14	6.33e-04
Diagnoses - main ICD10: I21 Acute myocardial infarction	361,194	5,948	29 / 14	9.92e-04
Non-cancer illness code, self-reported: heart attack/myocardial infarction	361,141	8,239	29 / 14	1.40e-03
Major coronary heart disease event excluding revascularizations	361,194	10,157	29 / 14	1.85e-02
Major coronary heart disease event	361,194	10,157	29 / 14	1.85e-02
Fasting Insulin	38,238		29 / 11	3.85e-02

Table 13: 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

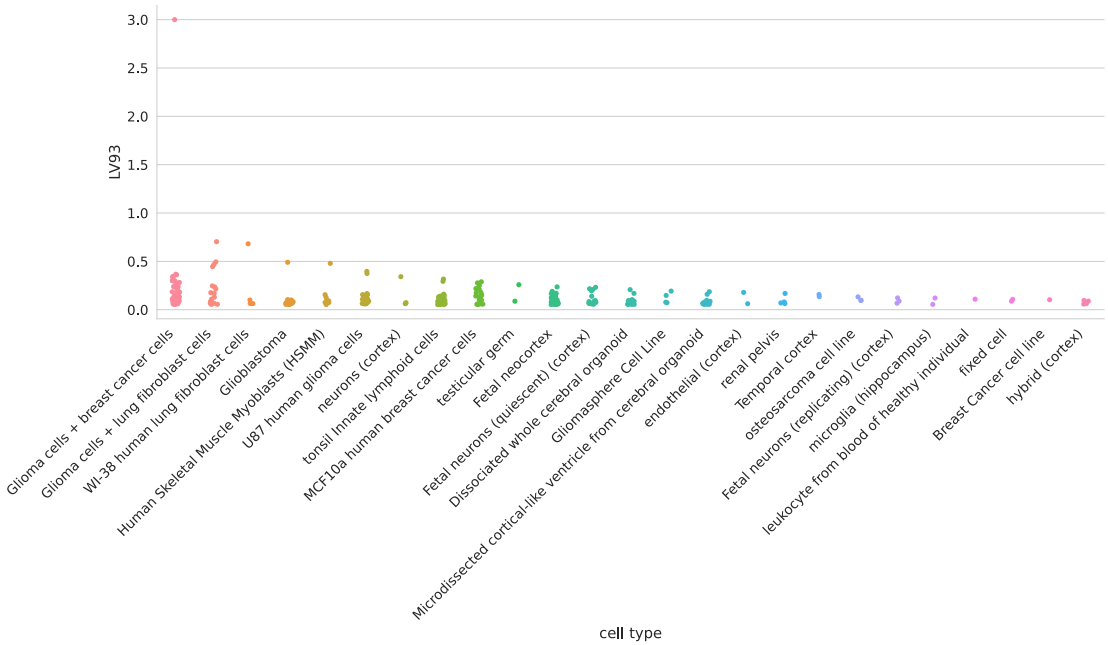


Figure 27: Cell types for LV93.

Table 14: 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 15: 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

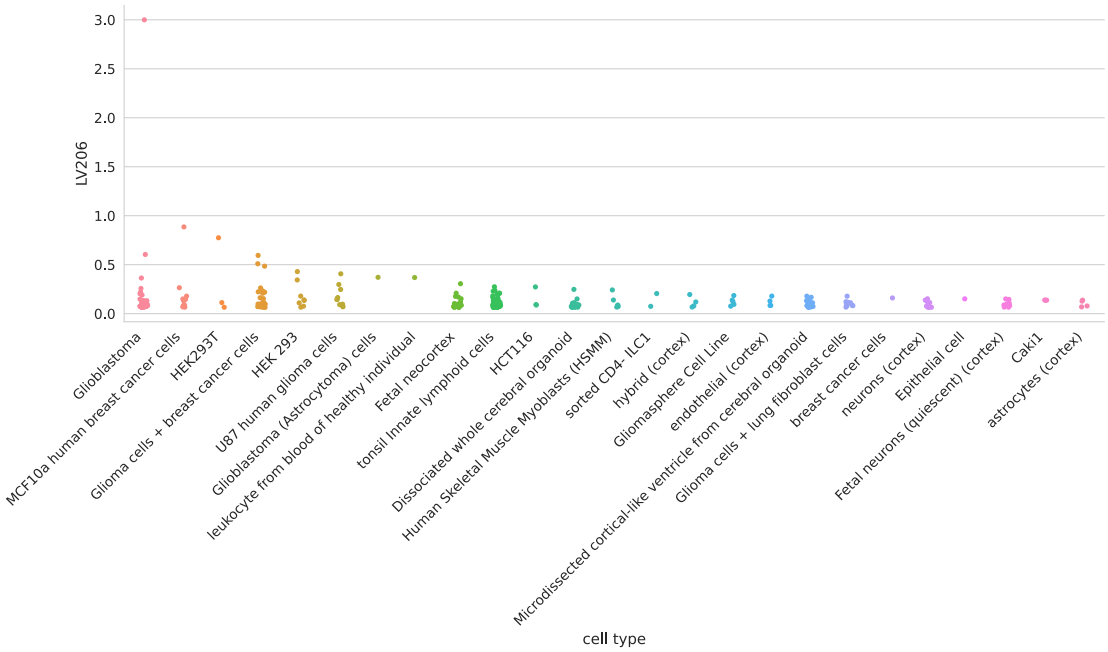


Figure 28: Cell types for LV206.

Table 16: 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 17: 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

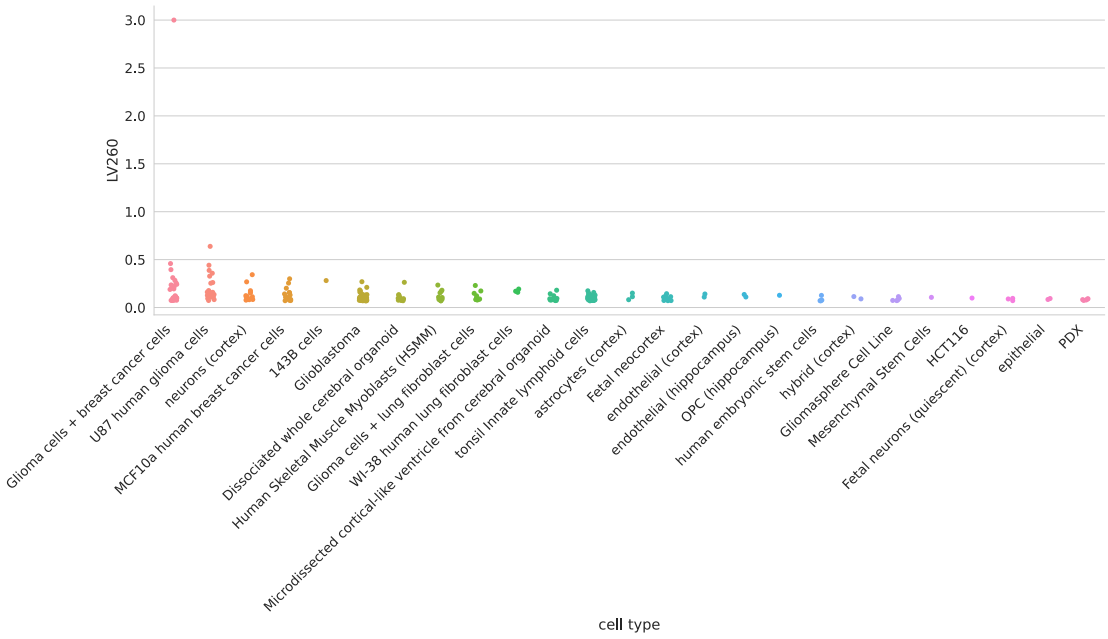


Figure 29: Cell types for LV260.

Table 18: 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 19: 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

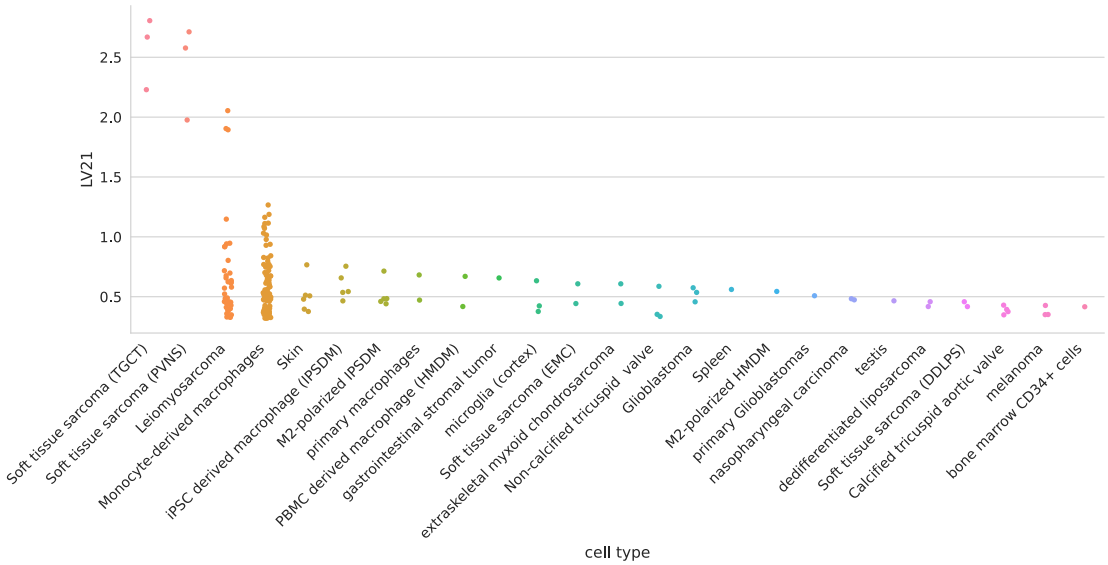


Figure 30: Cell types for LV21.

Table 20: 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 21: 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

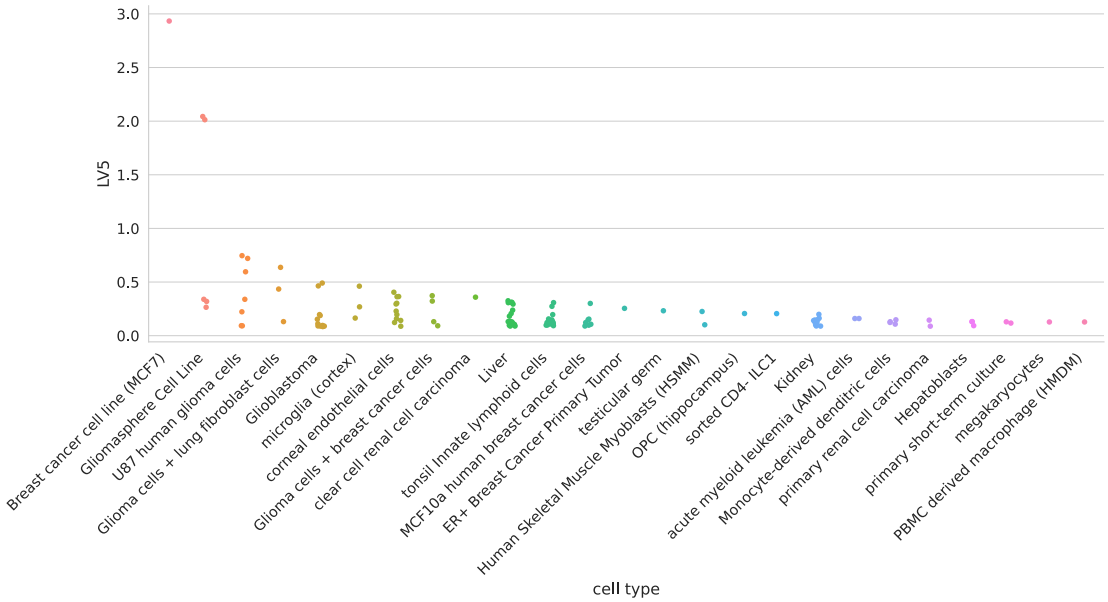


Figure 31: Cell types for LV5.

Table 22: 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 23: 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

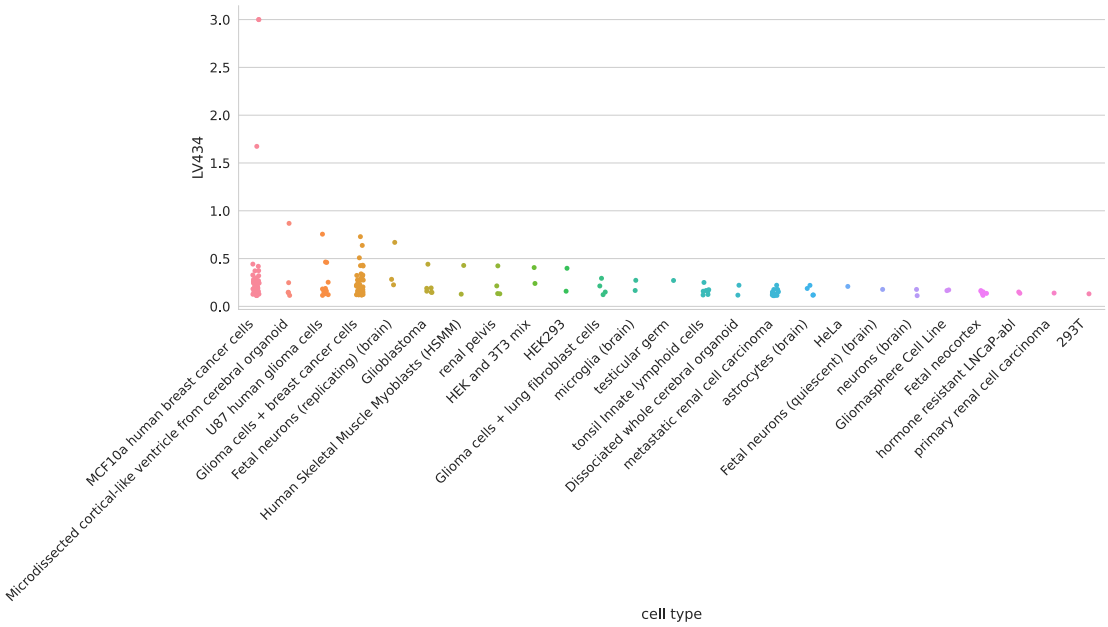


Figure 32: Cell types for LV434. `FIXME:` create a section for all LVs and clarify how figures were generated, like: The top samples in human breast cancer cells have larger LV values that were thresholded at 3.0 for visualization purposes. HEK293 is a cell line derived from human embryonic kidney cells; 3T3 is a cell line derived from mouse embryonic fibroblasts.

Table 24: 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 25: 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

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

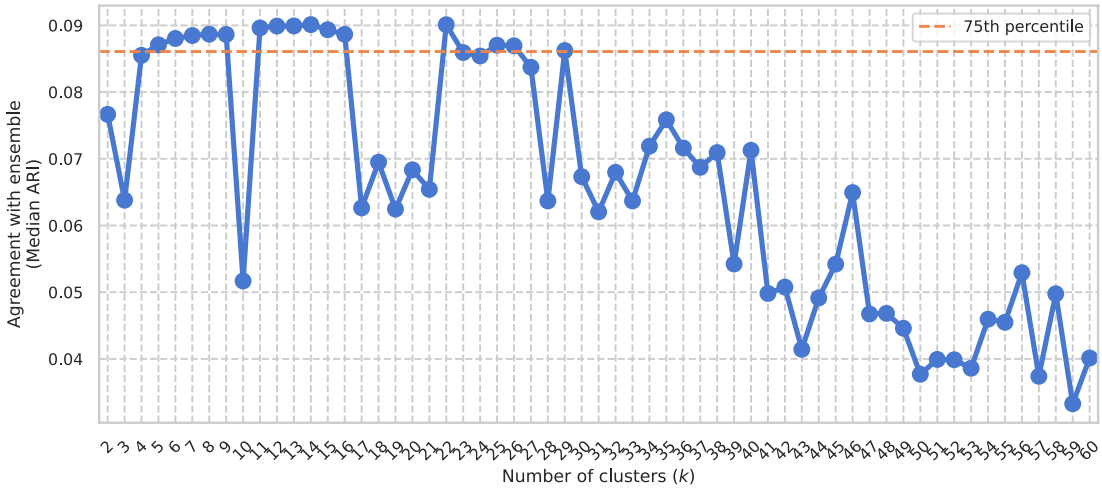


Figure 33: Final selected partitions for follow-up analysis. From all consensus clustering partitions generated with k from 2 to 60, we selected those with a median adjusted Rand index (ARI) with the ensemble members greater the 75th percentile.