# 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@b3f1c16 on May 4, 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

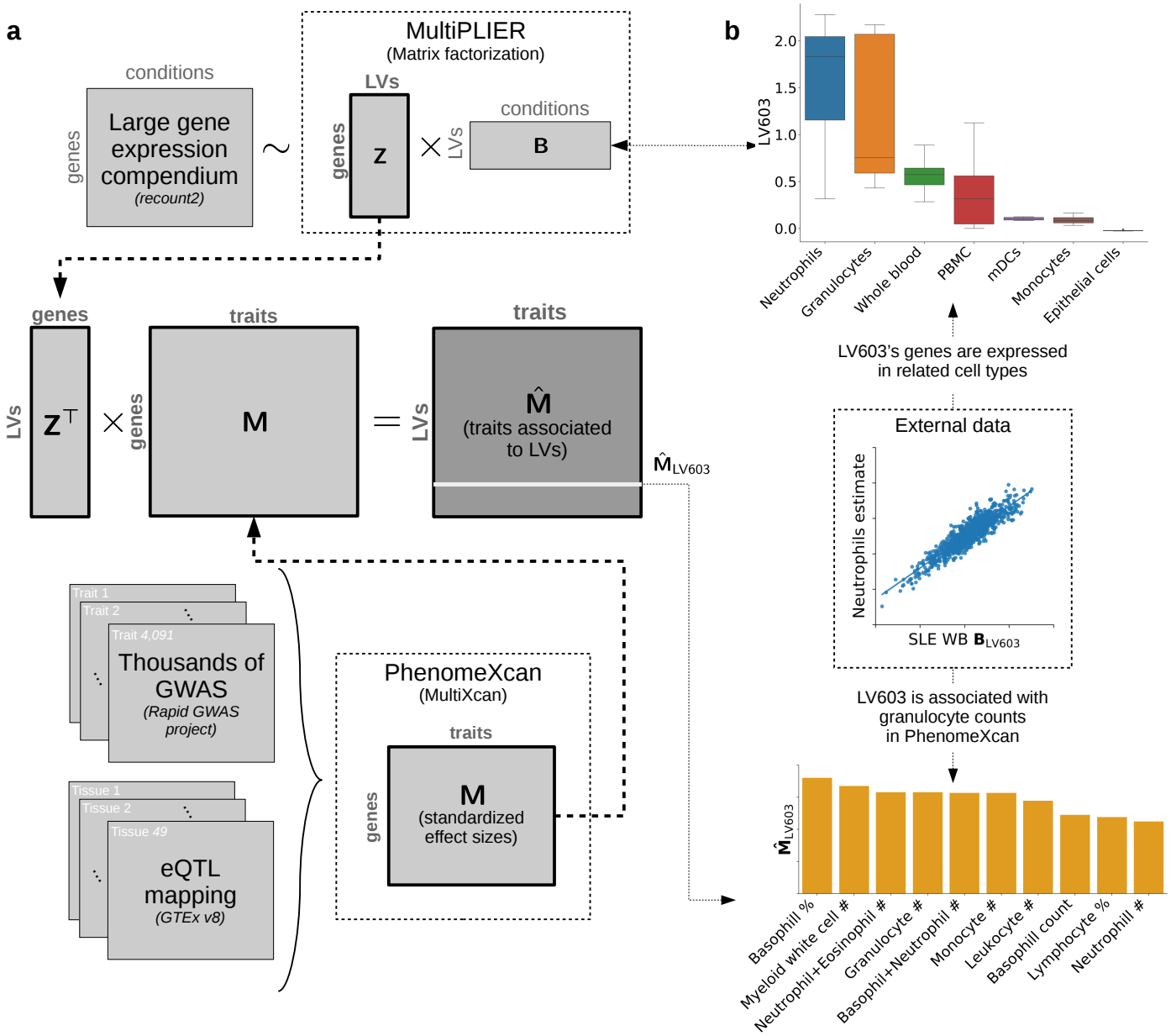


**Figure 1: Schematic of the PhenoPLIER framework. a)** The integration process between gene co-expression patterns from MultiPLIER (top) and TWAS results from PhenomeXcan. PhenoPLIER projects gene-based association results on more than 4,000 traits to a latent space learned from a recount2, a large gene expression compendium. This 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 were found to be 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 (see Methods for more details). The framework combines TWAS results on thousands of phenotypes with gene co-expression patterns by projecting gene-trait associations

on a latent gene expression representation. Each of these latent variables (LVs), obtained with an unsupervised learning method, represents a gene-set or gene module, essentially a group of genes with coordinated expression patterns (i.e. they are expressed together in the same tissues and cell types). Since genes in these modules vary together, we expect that they may also function together [25,26]. Thus, the projection of TWAS results into this latent space might provide context for their interpretation.

For the gene-trait associations we used PhenomeXcan [23], a massive TWAS resource on the UK Biobank [27] and other cohorts that provides results for 4,091 phenotypes, including different diseases and traits. These results were projected to the low-dimensional gene expression representation learned by MultiPLIER using:

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

where $\mathbf{M}^{n \times t}$ is the gene-trait associations matrix from MultiXcan [13] (standardized effect sizes) for $n$ genes and $t$ traits, $\mathbf{Z}^{n \times l}$ are the gene loadings with $l$ latent variables, $\lambda_2$ is the regularization parameter used in the training step, and $\hat{\mathbf{M}}^{l \times t}$ is finally the projection of $\mathbf{M}$ into the latent space: all traits in PhenomeXcan are now described by gene modules. Since the MultiPLIER models also provide the experimental conditions (such as cell types and tissues, represented by matrix $\mathbf{B}$ in Figure 1 a) in which genes in a module are concurringly expressed, our approach also allows inferring the context in which the gene module affects a trait or disease.

In the original MultiPLIER study, the authors found an LV significantly associated with previously known neutrophil gene-sets 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 are expressed in neutrophil cells. 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), and that the approach also allows inferring the context-specific effects of gene modules on complex traits. We will also show how the approach can aid translational efforts by mapping pharmacological perturbations to this latent space, enabling to observe which compounds affect the transcriptional activity of gene modules.

`Notes/questions:`

Ideas/minor:

- In the figure, add a reference to the TWAS plot (genes x traits) for LV603.

## 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.

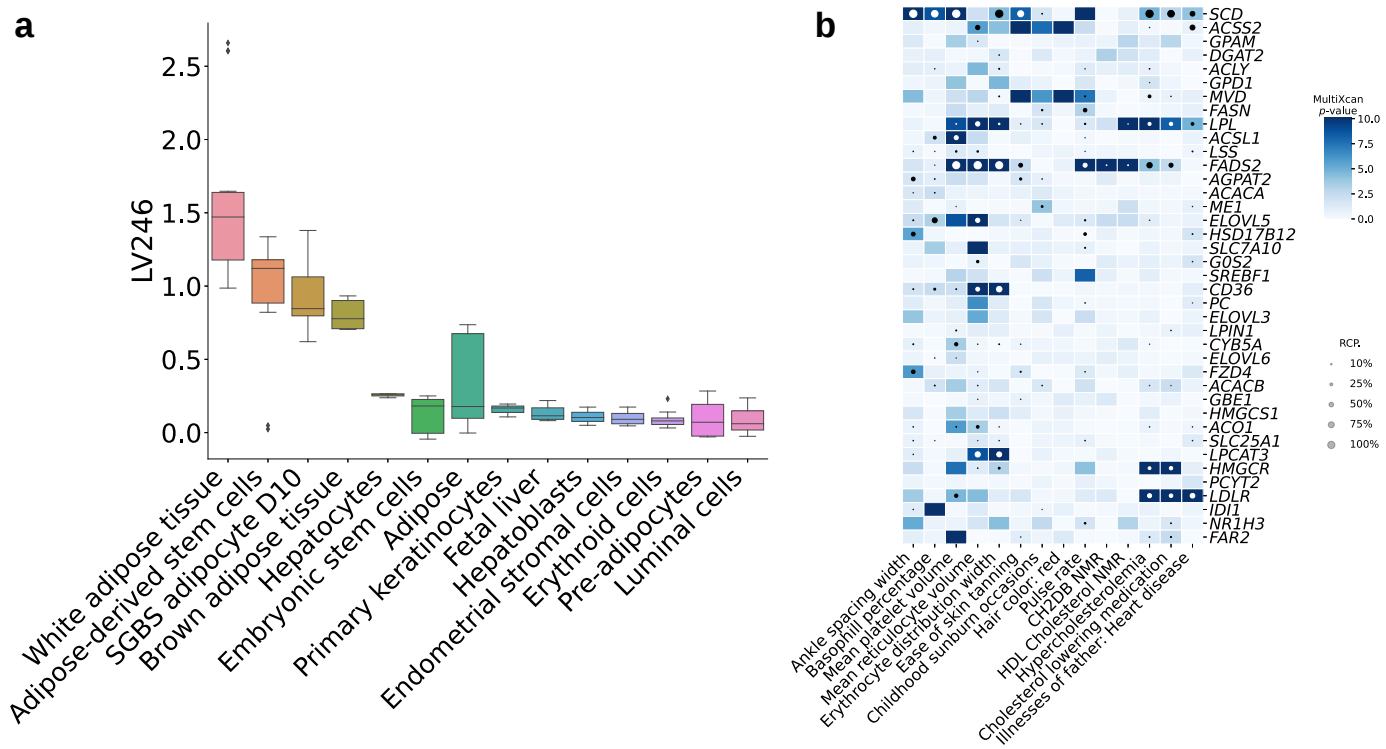


**Figure 2: Tissues and traits associated with a gene module related to lipids metabolism (LV246). a)** Top cell types/tissues where genes in LV246 are expressed on. Values in the $y$-axis come from matrix $\mathbf{B}$ in the MultiPLIER models (Figure 1 a). In the $x$-axis, cell types/tissues are sorted 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.

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

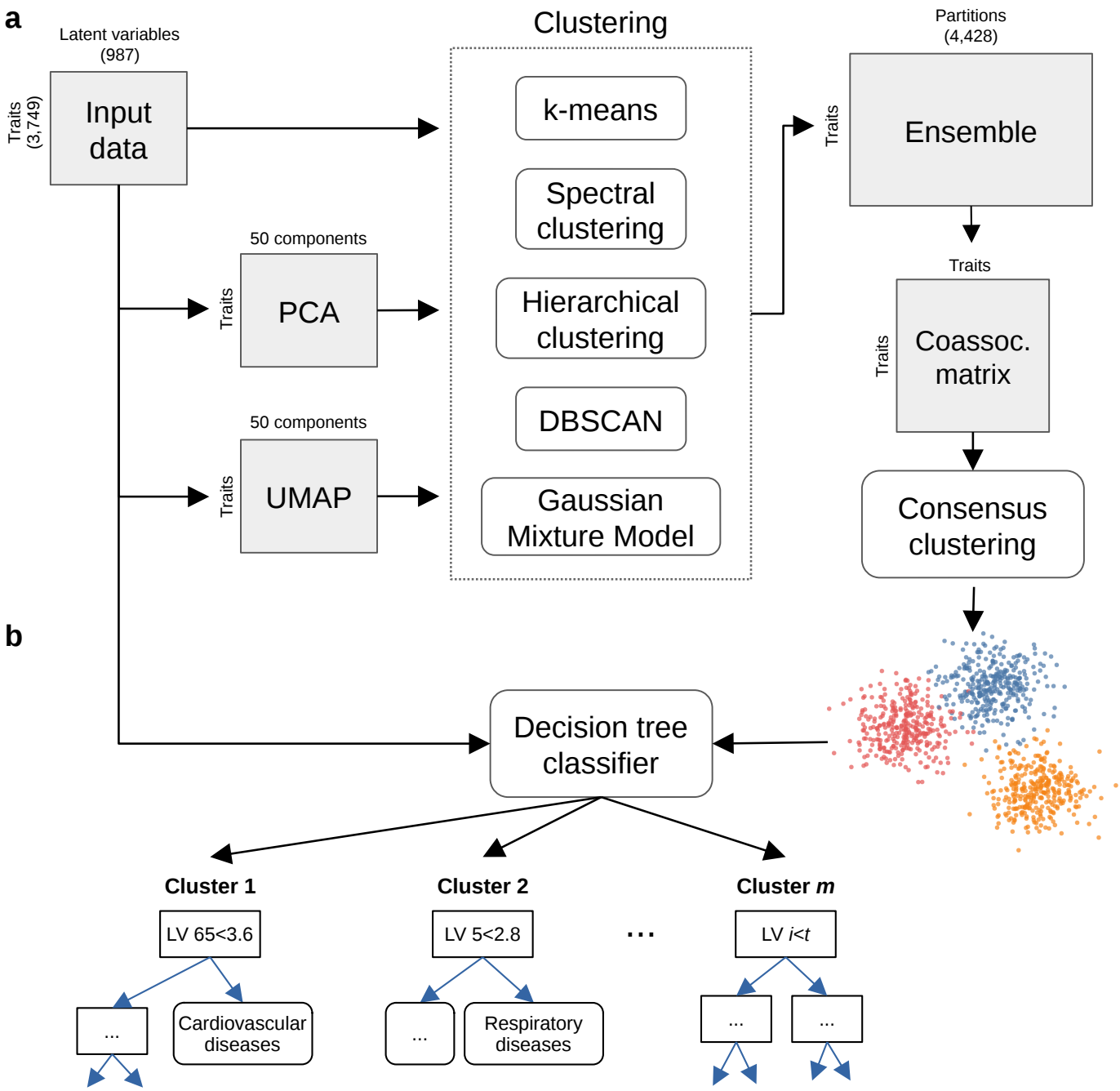


**Figure 3: Cluster analysis on traits from PhenomeXcan. a)** The projection of TWAS results for 3,752 traits to the latent representation learned from recount2 are the input data to the clustering process. A linear (PCA) and non-linear (UMAP) dimensionality reduction techniques are applied to the input 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. A coassociation matrix is derived by counting how many times a pair of traits were grouped together in the ensemble. Finally, a consensus function is applied to the coassociation matrix to generate consolidated partitions with different number of clusters. These final solutions are represented in the clustering tree (Figure 4). **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 most important LVs that differentiate groups of traits.

All traits in PhenomeXcan were projected into the latent space learned from recount2 using Equation (1). We conducted cluster analysis using this new representation 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 are applied on the same data, and later combined into a consolidated solution [28,29,30] (Figure 3). An important property for a successful application of a consensus clustering approach is the diversity of the ensemble, understood as the level of disagreement between the base clustering solutions [28,31,32]. A diverse

set of solutions can be generated by using different data representations (such as dimensionality reduction methods or subsets of features), clustering algorithms with distinct assumptions ($k$-means, for instance, assumes hyperspherical clusters), and a varying set of algorithm's parameters (such as the number of clusters or the initial random seeds). In our approach, we performed cluster analysis using five different clustering algorithms on three representations of the input data (the original data with 987 latent variables, its projection into the top 50 principal components, and the embedding learned by UMAP [33] using 50 components) (see Figure 3 a). The clustering methods used cover a wide range of different assumptions on cluster shapes and a varying set of parameters such as the number of clusters (from 2 to 60), the width of the Gaussian kernel in spectral clustering, and other method-specific parameters (see the supplementary material for more details). The process generated an ensemble with 4,428 clustering solutions for all traits. This ensemble was used to derive a coassociation matrix between traits by counting the number of times a pair of traits was clustered together. A consensus function was applied on the coassociation matrix to derive a consolidated solution using the information in the ensemble. For these final partitions, we did not select a specific number of clusters, but instead used a clustering tree [34] (Figure 4) to examine stable groups of traits at multiple resolutions. Finally, for the interpretation of the clusters, we trained a decision tree classifier (a highly interpretable machine learning model) on the original input data using the clusters found as labels. This approach allowed us to quickly identify the most important gene modules for the groups of traits found. More details of the clustering process are available in the supplementary material.

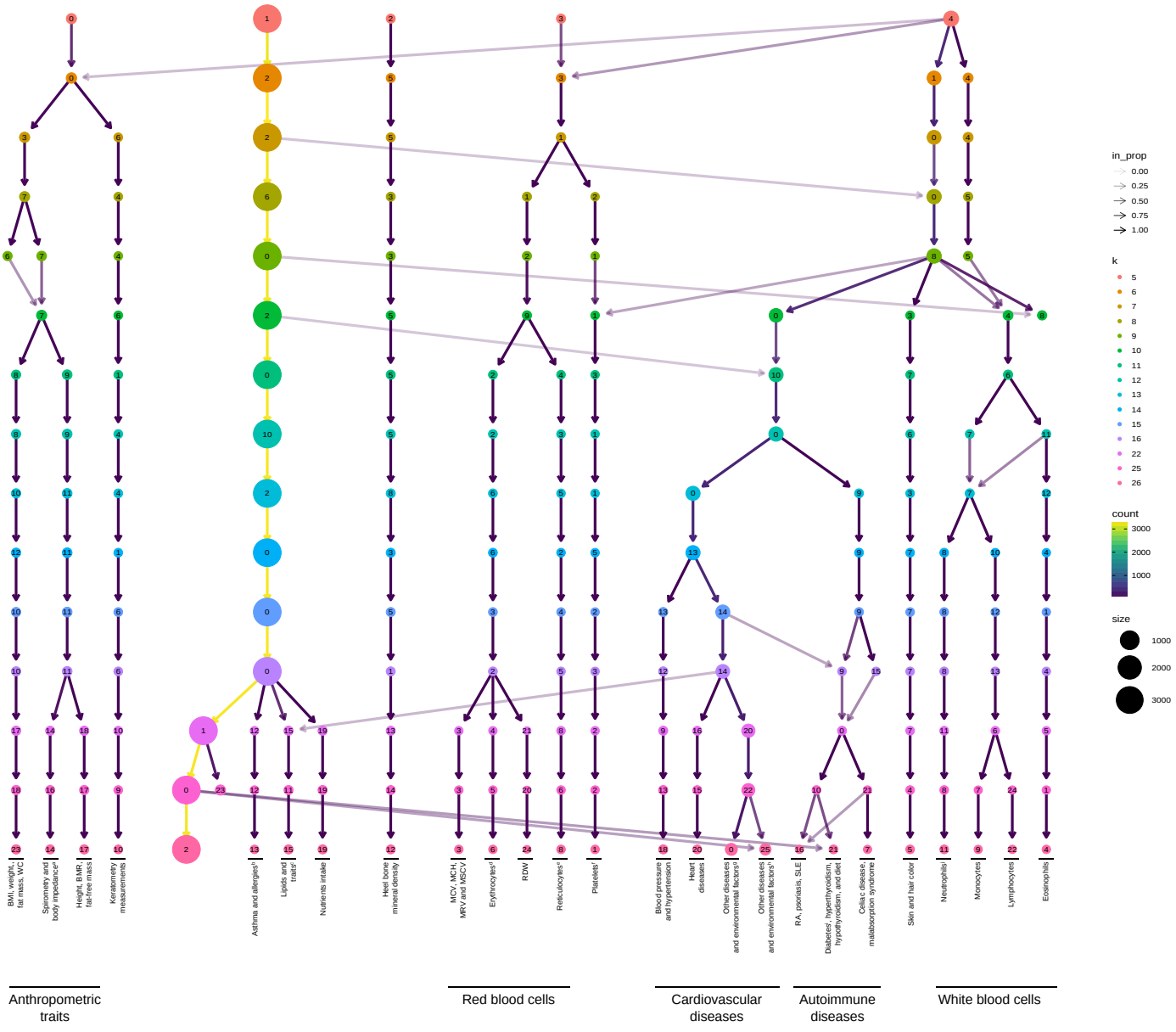

**Figure 4: Clustering tree using multiple resolutions for clusters of traits.** Clustering tree of traits partitions at different resolutions (from 5 to 26 clusters). Each row represents a partition of the traits, and each circle is a cluster from that partition. Arrows indicate how traits in one cluster move across clusters from different partitions. Most of the clusters are preserved across different resolutions, showing a high stability even with independent runs of the clustering algorithm. BMR: basal metabolic rate; WC: waist/hip circumference; MCV: mean corpuscular volume; MCH: mean corpuscular hemoglobin; MRV: mean reticulocyte volume; MRV: mean reticulocyte volume; MSCV: mean sphered cell volume; RDW: red cell (erythrocyte) distribution width; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus; [a] includes whole-body, arms and legs impedances; [b] allergies refer to allergic rhinitis or atopic dermatitis; [c] includes Alzheimer's disease, coronary artery disease, breast cancer, fasting blood glucose and insulin measurements, inflammatory bowel disease, and atopic dermatitis; [d] includes erythrocyte count, hemoglobin concentration, and hematocrit percentage; [e] includes reticulocyte count and percentage, immature reticulocyte fraction, and high light scatter reticulocytes count and percentage; [f] includes platelet count, platelet crit, mean platelet volume, and platelet distribution width. [g] includes diabetes, gout, arthrosis, and respiratory diseases (and related medications such as ramipril, allopurinol, lisinopril, and albuterol), and several environmental/behavioral factors such as intake of a range of common food/drink items including alcohol, time spent outdoors and watching TV, smoking and sleeping habits, early-life factors, education attainment, psychological and mental health, and health satisfaction. [h] includes vascular problems such as angina, deep vein thrombosis (DVT), intraocular pressure, eye and mouth problems, hand-grip strength, several measurements of physical activity, jobs involving heavy physical work, transport type for commuting, intake of common vitamin/mineral supplements, and various types of body pain and medications for pain relief. [i] age when diabetes was first diagnosed; [j] includes neutrophil count, neutrophil+basophil count, neutrophil+eosinophil count, granulocyte count, leukocyte count, and myeloid cell count.

A clustering tree of the consensus solutions at different resolutions is shown in Figure 4. For each $k$ (the number of clusters), the consensus partition that maximized the agreement with the ensemble was selected (see supplementary material). Since it is expected that a subset of resolutions better represents the patterns among traits, we further filtered the consensus partitions by taking those with an agreement value higher than the 75th percentile, which included partitions from 5 to 26 clusters.

The clustering tree shows five clear branches that start at the top with different numerical labels (from left to right): 0) physical measures including anthropometric traits (with fat-free and fat mass measures in separate sub-branches) and eye measures (keratometry), 1) a "large" branch that includes most of the traits that start to be subdivided only at $k = 16$ (with asthma, lipids, and nutrient intake clusters), 2) bone-densitometry measurements, 3) hematological assays on red blood cells and platelets, and 4) a "complex" branch including assays on white blood cells, skin and hair color, autoimmune disorders, and cardiovascular diseases (hypertension, heart problems, and other cardiovascular-related traits such hand-grip strength [35], and environmental/behavioral factors such as physical activity and diet). 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 clusters of different resolutions show how traits move from one group to another, and this only happens between the "complex" branch and the rest, particularly with traits expected to be linked to several others, such as white blood cells, cardiovascular and autoimmune diseases, and other related factors. (`(Be very careful with the following examples, they need to be extremely obvious/expected)`) For example, the arrow from cluster 14 at $k = 16$ (heart problems and related traits) to cluster 15 at $k = 22$ (lipids), indicate the move of coronary artery disease, fasting glucose and blood insulin measurement to the lipids clusters, which are highly related (`CITATION`). Another example is age when diabetes was first diagnosed that is moved from cluster 0 at $k = 25$ to cluster 21 at $k = 26$ (with autoimmune diseases such as hypo and hyperthyroidism), and intraocular pressure and eye problems from the same cluster to cluster 25 at $k = 26$ (with traits related to hypertension and cardiovascular problems). This movement of traits across highly-related clusters indicates the complexity of the relationships, where in these cases the algorithm finds meaningful but changing traits at different resolutions.

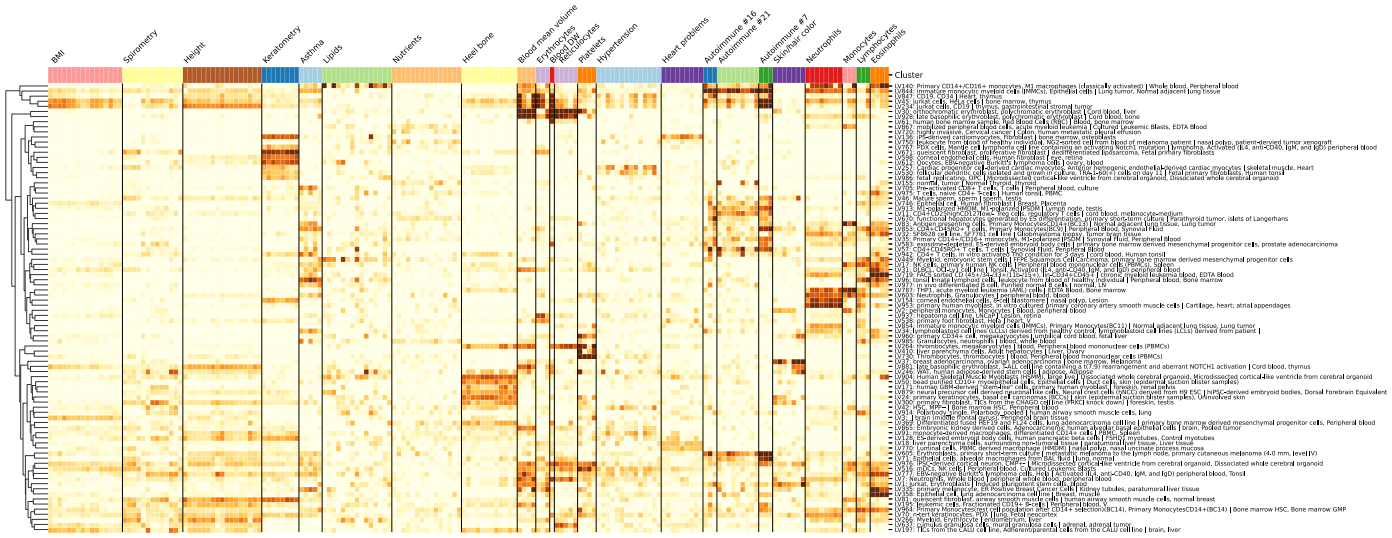


**Figure 5: Gene modules show specific and general transcriptional processes associated with different clusters of traits.** `(Early draft version; we need to improve the y-axis labels on the right)` The plot shows a submatrix of $\hat{\mathbf{M}}$ for the main trait clusters (from the bottom of the clustering tree in Figure 4), considering only gene modules (rows) that align well with at least one known pathway. Labels on the right show the top two cell types and top two tissues where gene modules are expressed in. Values range from -5 (lighter color) to 16 (darker color).

Next, we analyzed which gene modules are driving these trait clusters. For that, we trained decision tree classifiers on the input data (Figure 3) using each cluster at $k = 26$ (bottom of Figure 4) as labels. This yielded for each cluster the top associated gene modules, where several of them were well-aligned to existing pathways, and other modules were "novel" and expressed in relevant tissues (`REVISE`). Results are shown in Figure 5, 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, that are known to be related to early progenitors of the erythrocytes lineage [36] (`DMAP_ERY` pathways in Supplementary Figure 10), are predominantly expressed in early differentiation stages of erythropoiesis, and strongly associated with different assays on red blood cells (erythrocytes and reticulocytes, including cell counts, mean volumes and distribution width). On the other side, others are highly specific, such as LV730, expressed in thrombocytes, and exclusively associated with hematological assays on platelets; or LV598, whose genes are expressed in corneal endothelial cells, and are almost only related to keratometry measurements.

- Maybe include Ben's question about why we expect these clustering results. My answer was more or less that far apart clusters are explained by more specific LVs to those traits, and the complex branch is more related to traits that are highly connected to all biological processes, where more subtle differences are captured only at higher resolutions.

## Cardiovascular traits

See if we are going to analyze this cluster.

The figures below include examples of the type of figure we can include here:

- Show the LVs that are distinct for this cluster (maybe a small decision tree for some of the clusters).
- For those LV, show which cell types/tissues are important (such as Figure 6 below).
- For some of these LVs, we can include the list of other traits also related and gene association results (such as Figure 7).

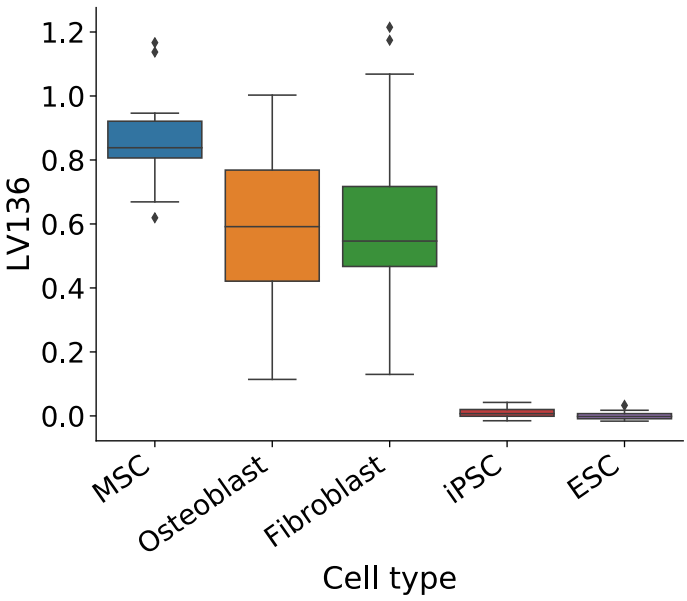


**Figure 6: Cell types/tissues associated with genes in LV136.**

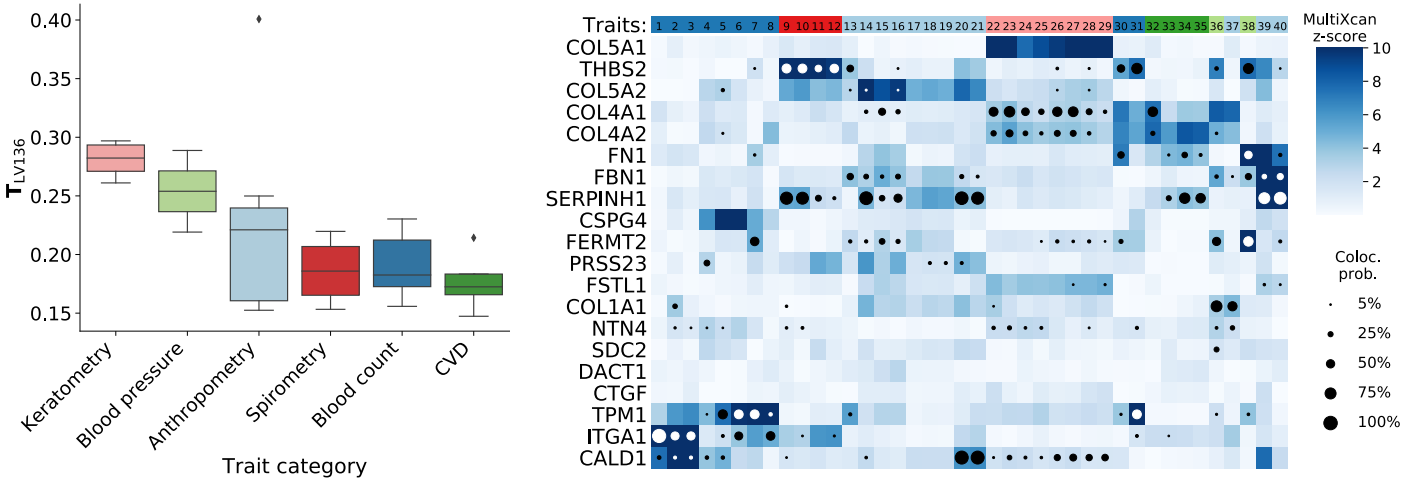


**Figure 7: Traits associated with genes in LV136.**

## Schizophrenia, educational attainment and intelligence

See if we are going to analyze this cluster.

## Asthma and allergies

See if we are going to analyze this cluster.

## Replication using Penn Medicine BioBank

Maybe we can incorporate **Binglan's TWAS results on PMBB** and see if expected traits-clusters are correctly predicted.

## 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 [37], which were further processed and mapped to DrubBank IDs [38,39,40]. Based on the established drug repurposing strategy that looks for reversed transcriptome patterns between genes and drug-induced perturbations [41,42], we used a framework for prioritizing drug candidates that uses imputed transcriptomes from GWAS [43]. 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 1). We used a manually-curated gold standard of drug-disease medical indications [39,44] across 53 diseases and 322 compounds to evaluate and compare the prediction performance.

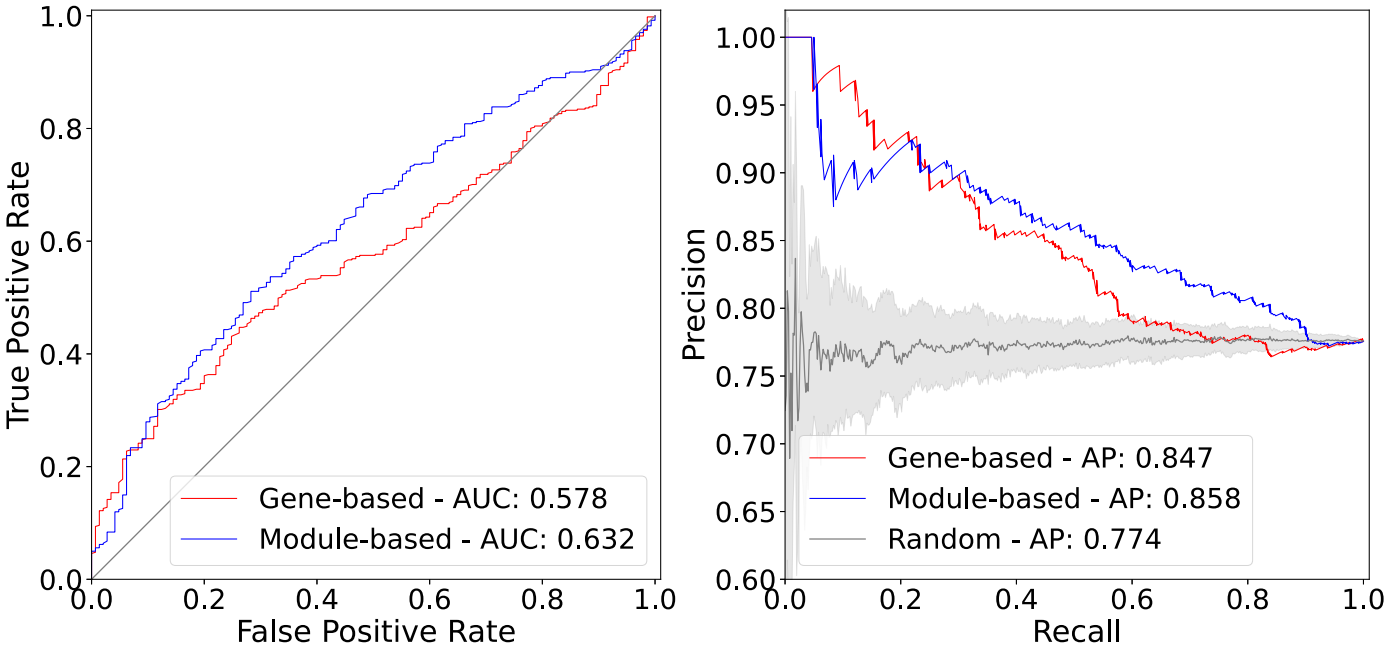


**Figure 8: Drug-disease prediction performance for gene-based and module-based approaches.** The receiver operating characteristic (ROC) (left) and the precision-recall curves (right) for a gene-based and our module-based approach. AUC: area under the curve; AP: average precision.

The ROC and precision-recall (PR) curves comparing both approaches are shown in Figure 8. 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 [45,46]. 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 [45]. 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).

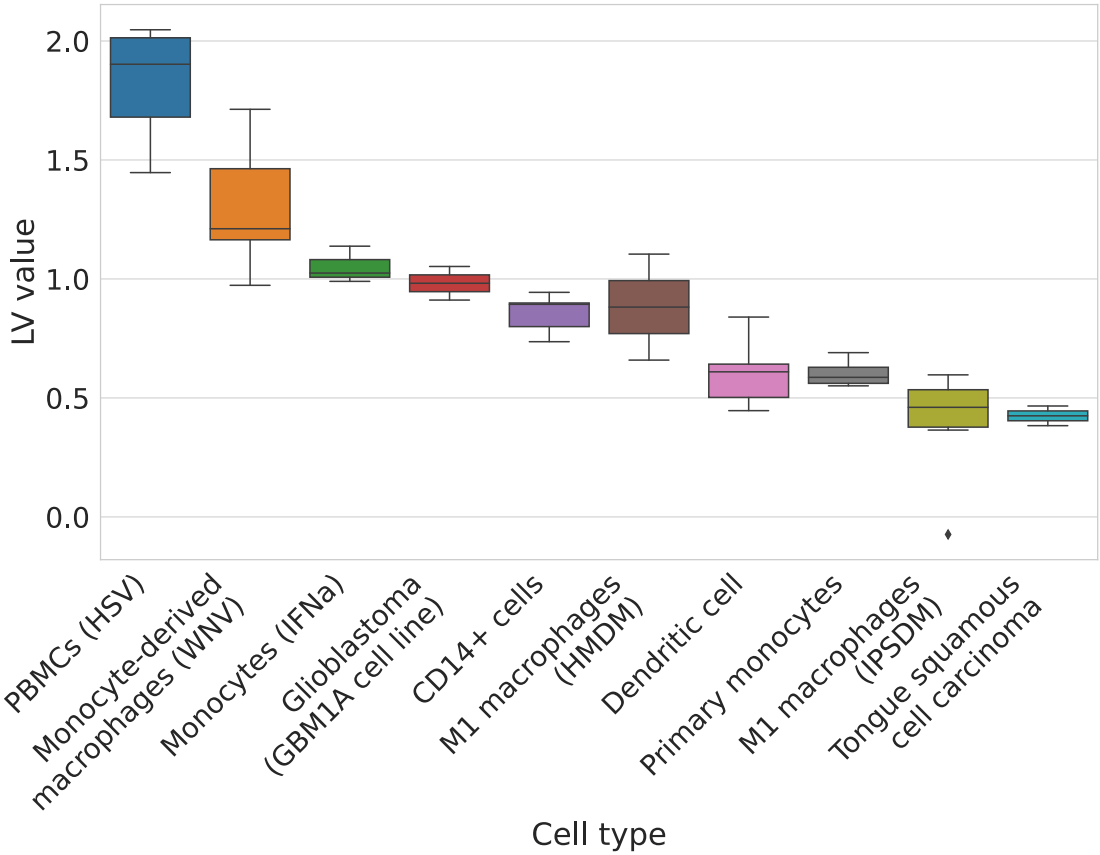


**Figure 9: Cell types where the top 10 modules contributing for niacin-atherosclerosis prediction are expressed.** Average module expression ($y$-axis) of different cell types ($x$-axis) across the top 10 latent variables/modules with a positive contribution for the niacin-AT prediction. The figure shows a clear immune cells signature, driven mainly by the top 2 modules: LV116 and LV931 (see Supplementary Figures 13 and 14). 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 [47,48]. 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 [49], thus boosting anti-inflammatory processes and reversing cholesterol transport [50]. As shown in Figure 9, 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 15) and LV885/LV840 expressed in kidneys (Supplementary Figures 16 and 17)

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 18) is mainly expressed in ovarian granulosa cells. This compound has been very recently considered as a potential therapeutic for ovarian diseases [51,52], 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 9 ? 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.

## 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. **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

26. **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

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 · PMID: 30305743 · PMCID: PMC6786975

28. **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

29. **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

30. **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

31. **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 · PMID: 21576752

32. **Hybrid clustering solution selection strategy**
Zhiwen Yu, Le Li, Yunjun Gao, Jane You, Jiming Liu, Hau-San Wong, Guoqiang Han
*Pattern Recognition* (2014-10) https://doi.org/ghtzwt
DOI: 10.1016/j.patcog.2014.04.005

33. **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

34. **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

35. **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

36. **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

37. **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

38. **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

39. **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

40. **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

41. **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

42. **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

43. **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

44. **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

45. **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

46. **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

47. **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

48. **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

49. **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

50. **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

51. **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

52. **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

53. **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

# 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 [53]), PLIER finds $\mathbf{U}$, $\mathbf{Z}$, and $\mathbf{B}$ minimizing

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

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 $\lambda_i$ 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

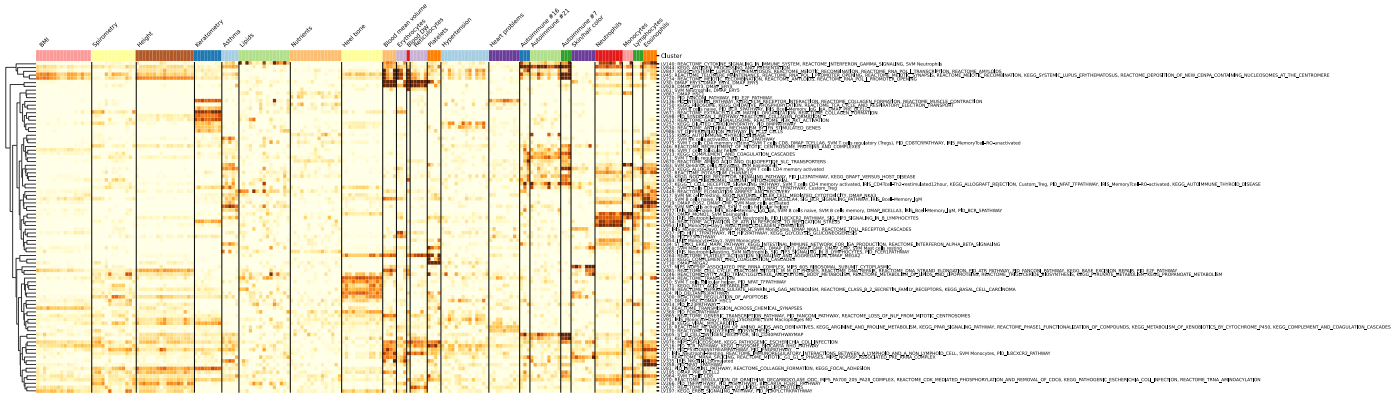

**Figure 10: Pathways associated with gene modules in Figure 5.** `(Early draft version; instead of this figure, we might want to just add a table with information about LVs)` This figure is equivalent to Figure 5 but, instead of cell types, labels on the right show all the associated pathways for each 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 (1), 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.

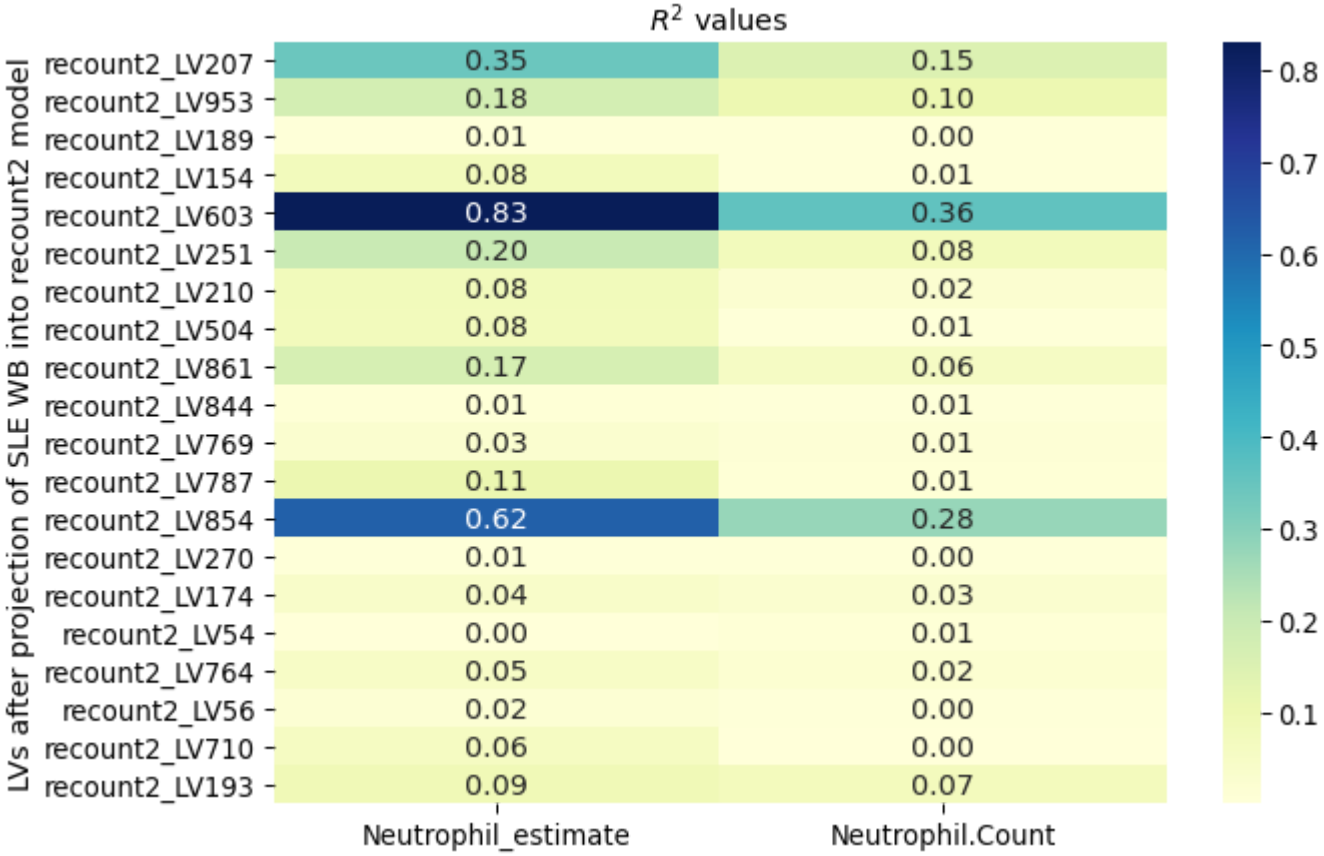


**Figure 11: Correlation of neutrophil counts with top LVs associated with neutrophils traits.**

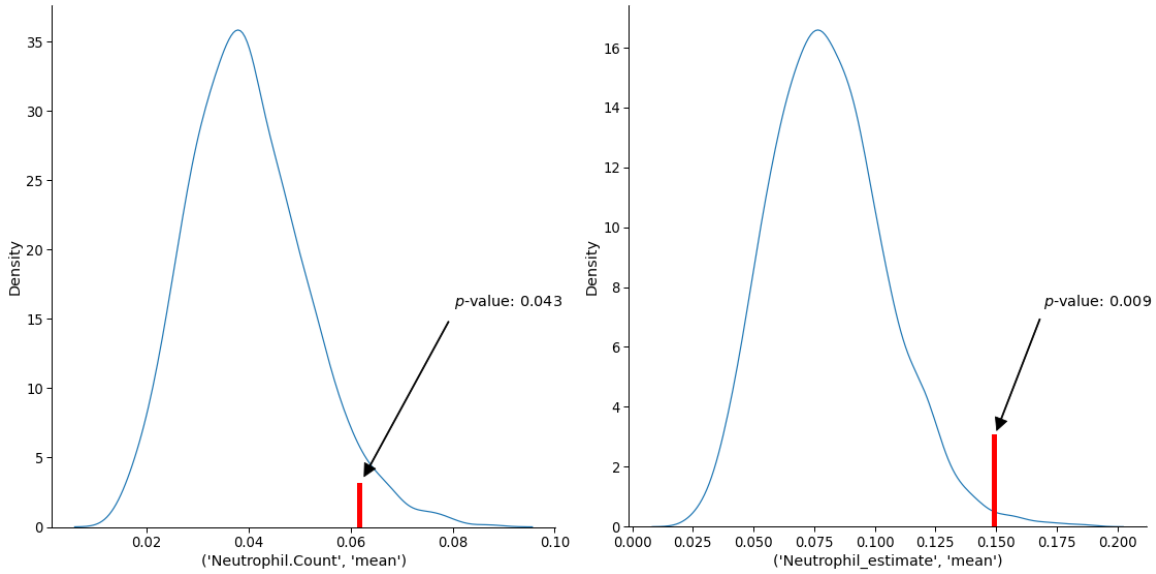


**Figure 12: Significance of neutrophil counts correlation.**

# LV116 cell types

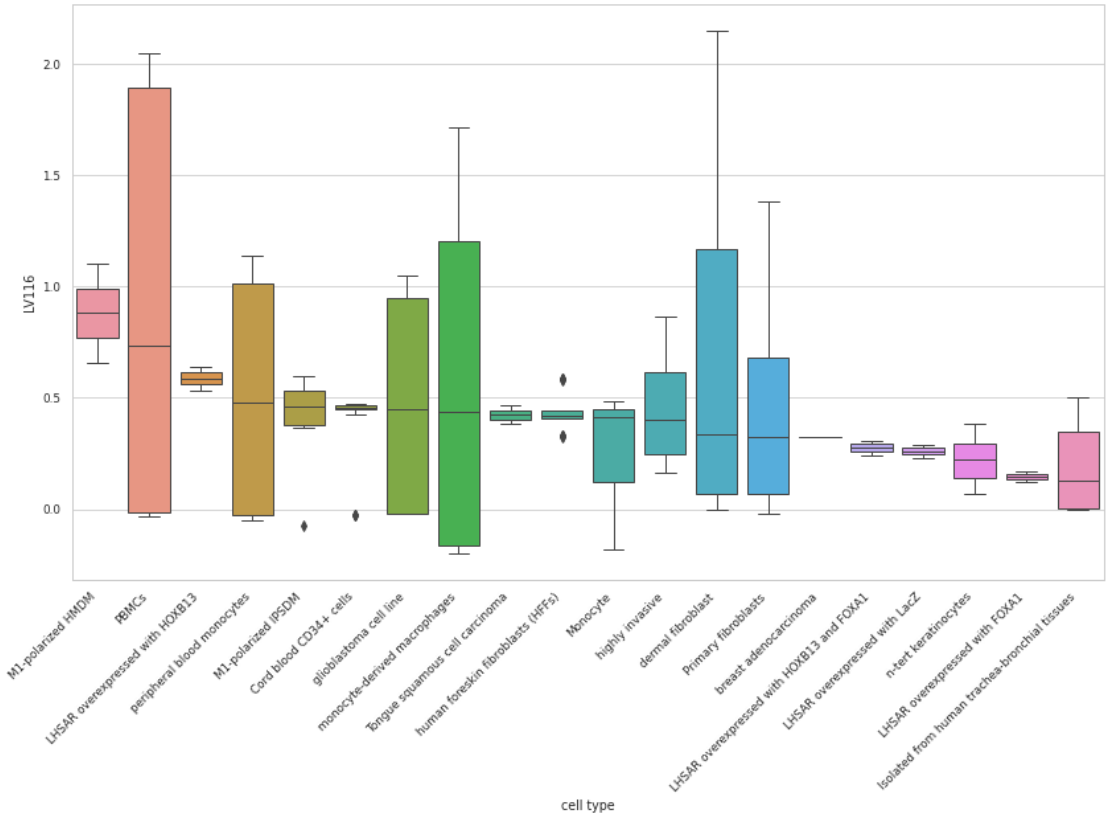


**Figure 13: Cell types for LV116.**

# LV931 cell types

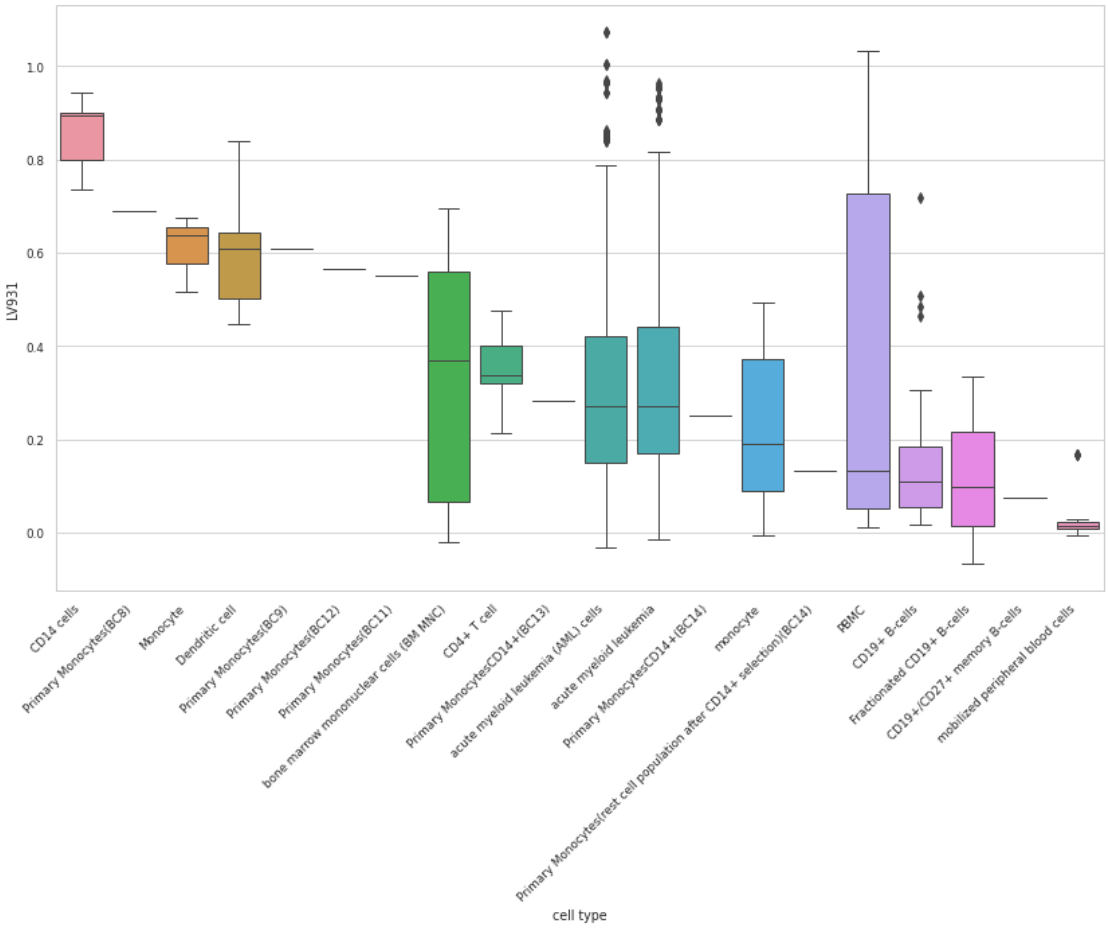


**Figure 14: Cell types for LV931.**

# LV536 cell types

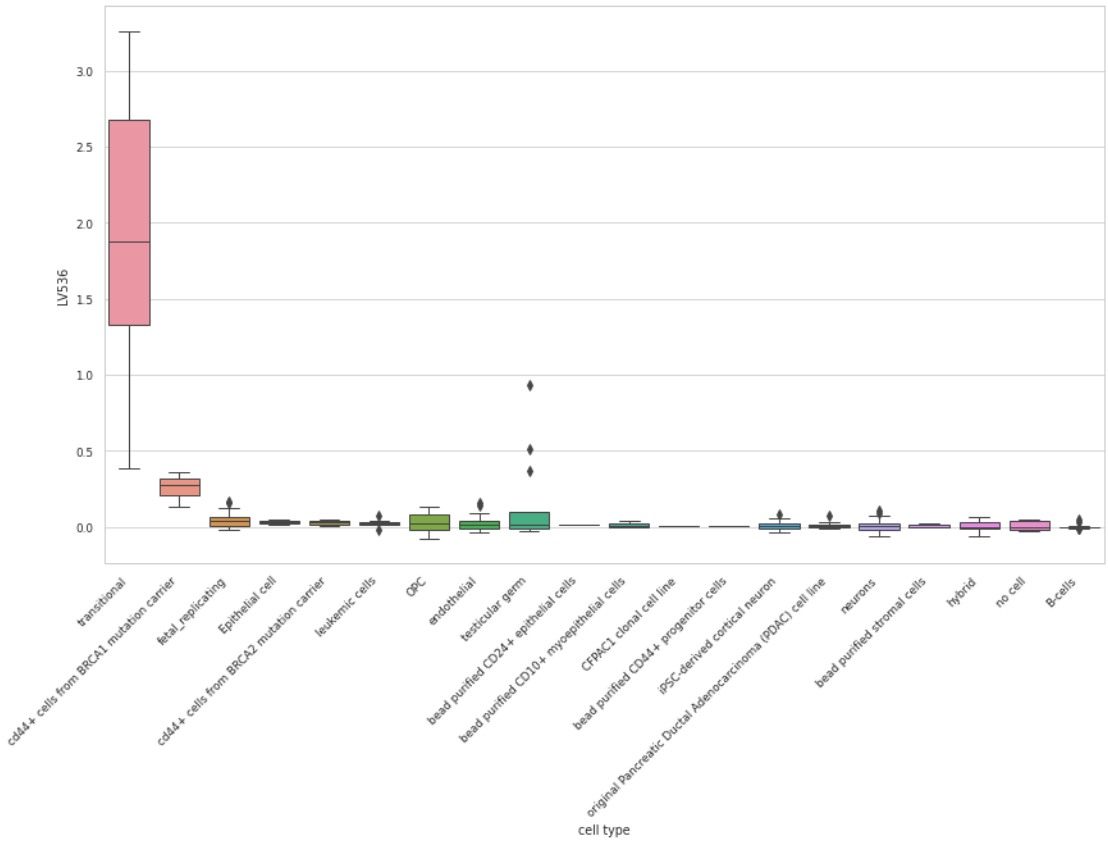


**Figure 15: Cell types for LV536.** FIXME: *transitional* here refers to bladder: https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP007947

# LV885 cell types

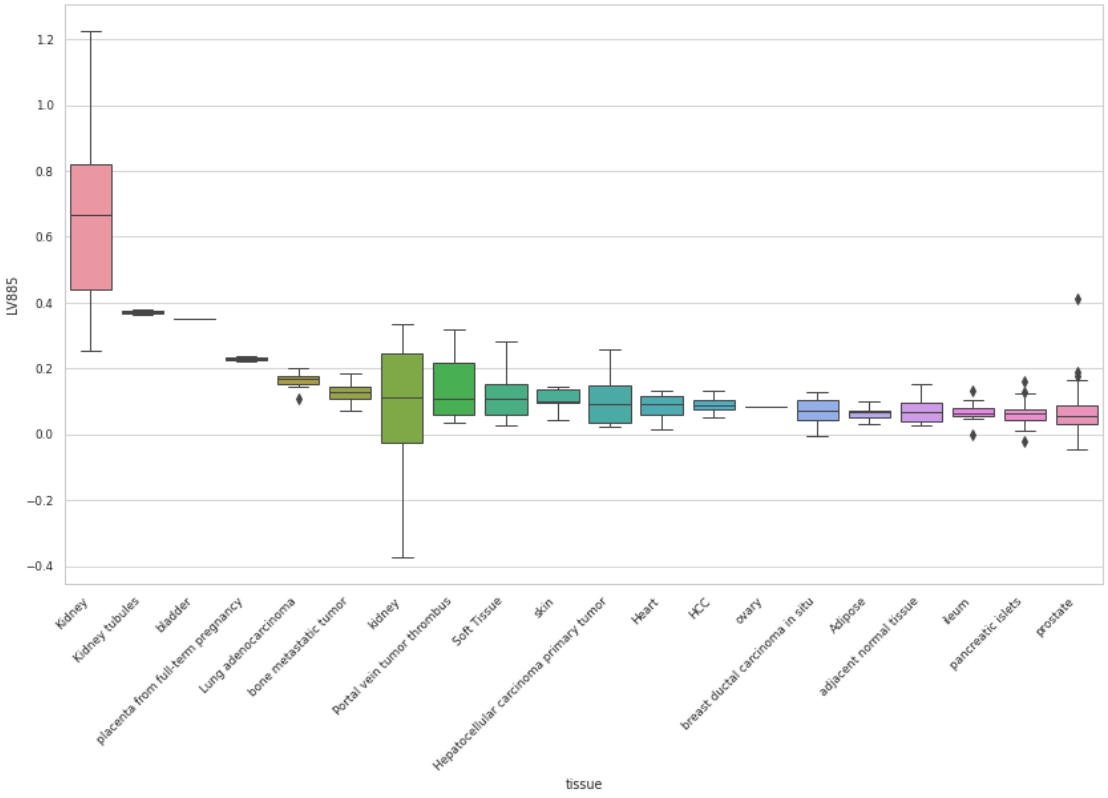


**Figure 16: Cell types for LV885.**

# LV840 cell types

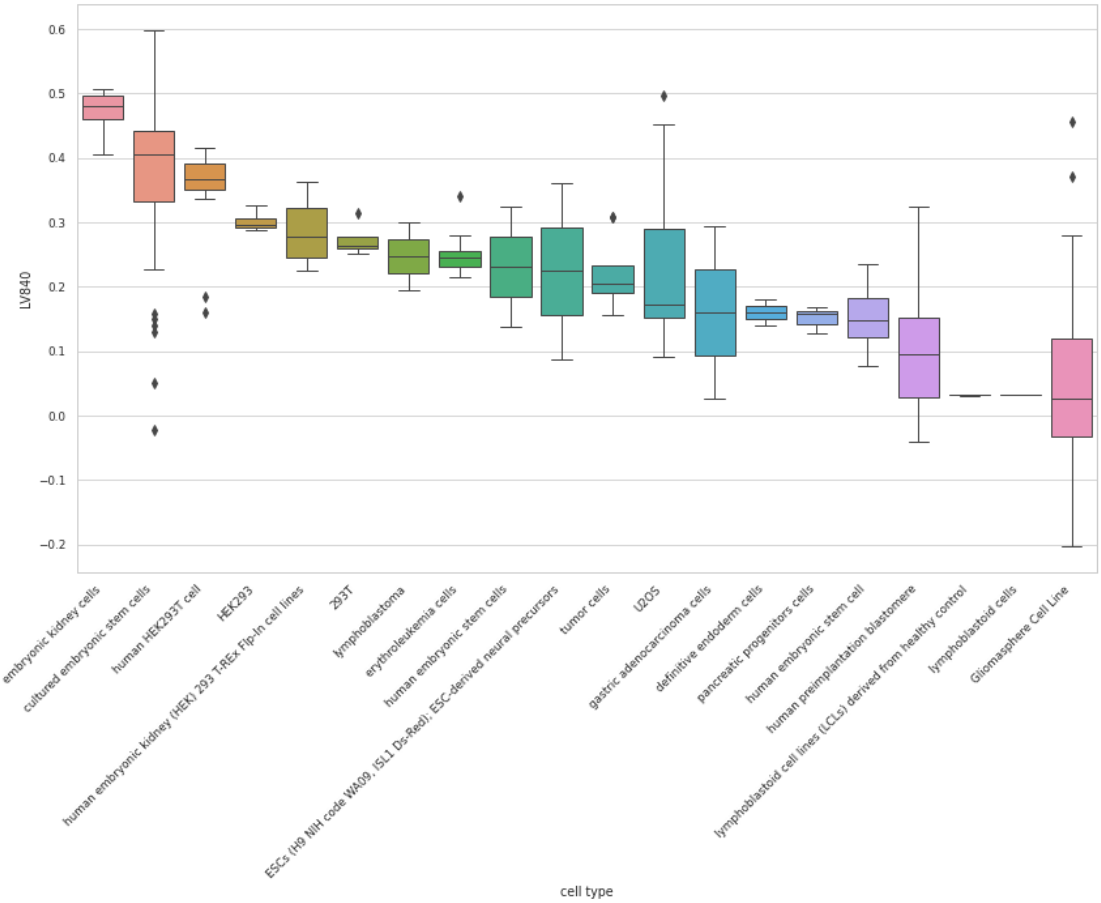


**Figure 17: Cell types for LV840.**

# LV66 cell types

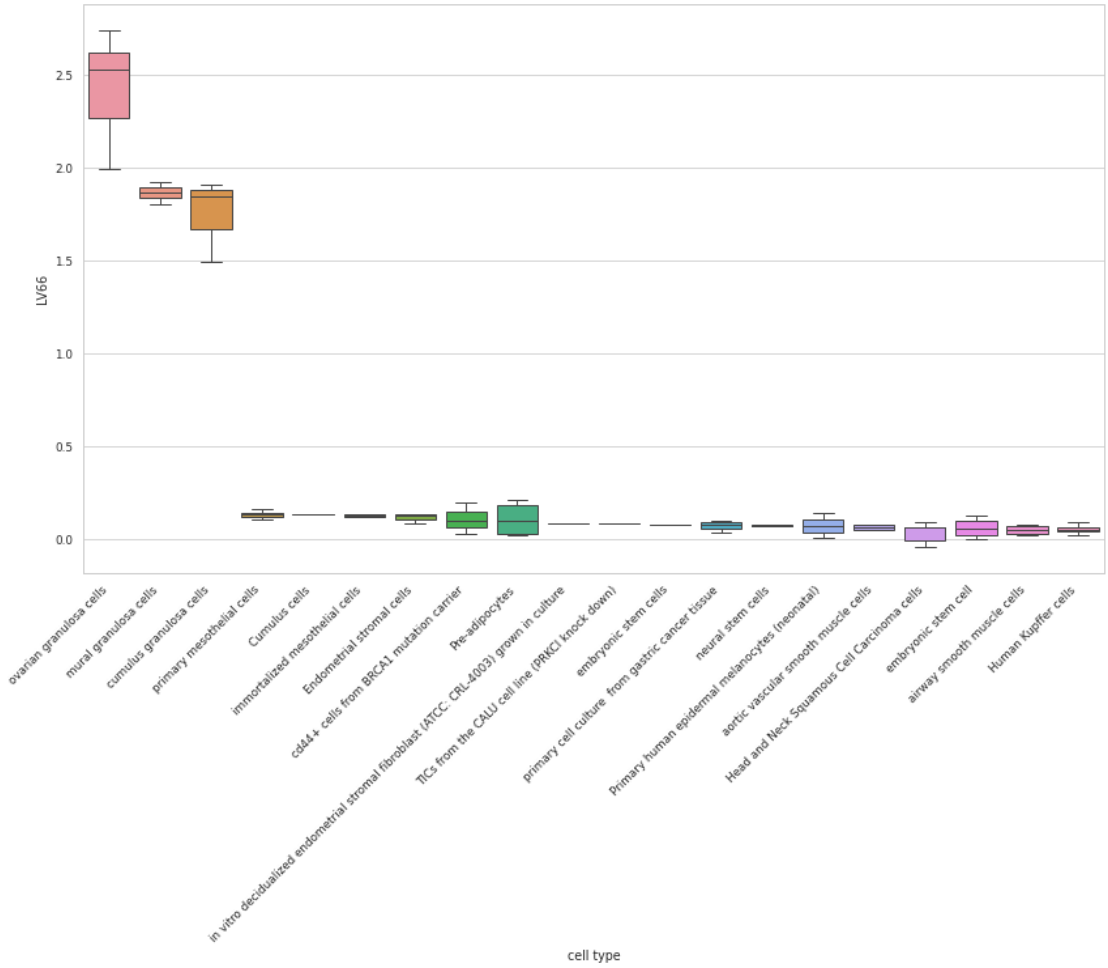


**Figure 18: Cell types for LV66.**