

# 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@c62c48b](#) on April 11, 2021.

## Authors

---

- **Milton Pividori**

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

Department of Systems Pharmacology and Translational Therapeutics, 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, USA

- **Casey S. Greene**

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

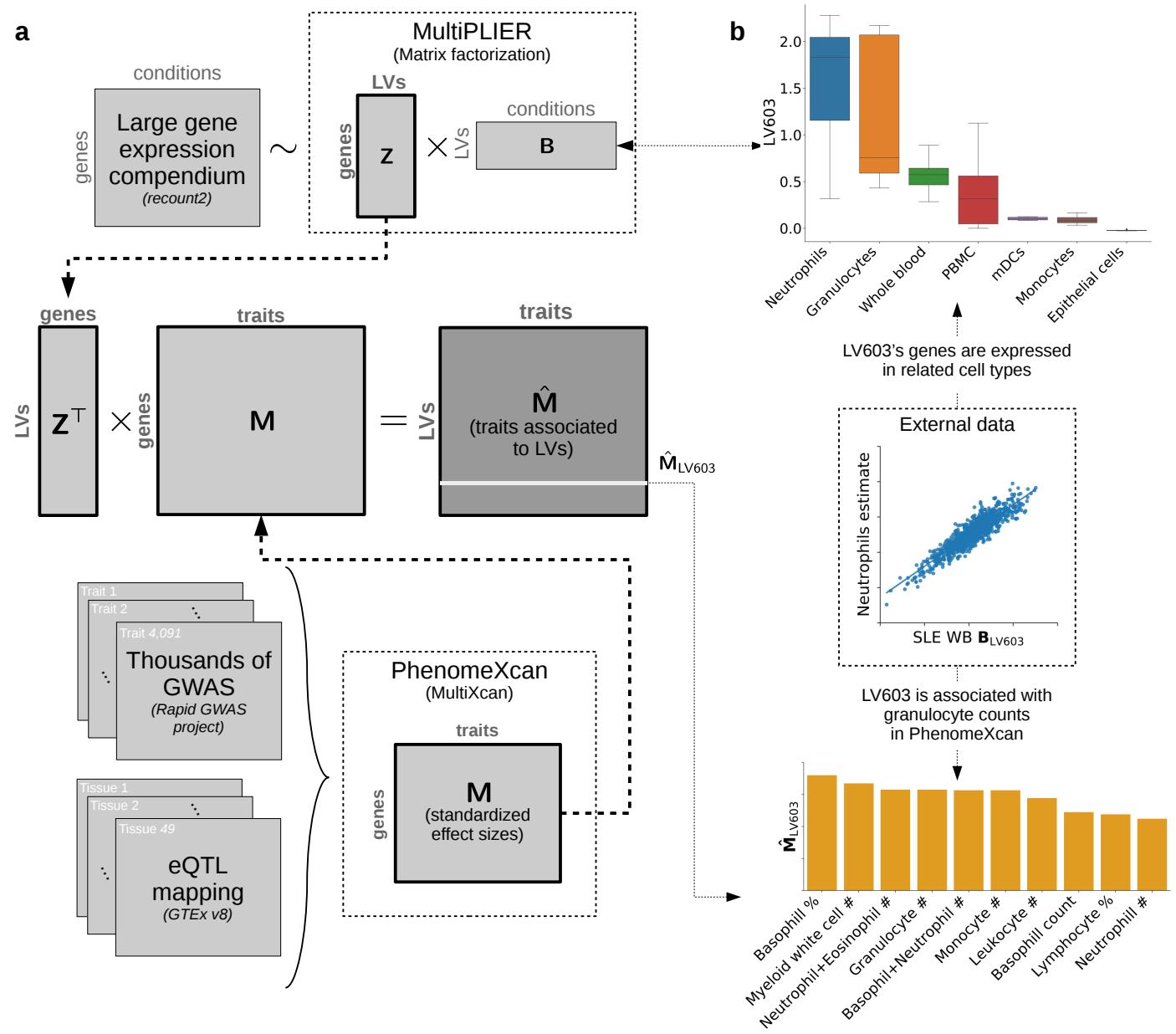
Center for Health AI, University of Colorado School of Medicine, Aurora, CO, 80045, USA; Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO, 80045, USA · Funded by The Gordon and Betty Moore Foundation (GBMF 4552); The National Human Genome Research Institute (R01 HG010067); The National Cancer Institute (R01 CA237170)

# Abstract

## Introduction

## Results

### PhenoPLIER integrates TWAS with gene co-expression patterns



**Figure 1: Schematic of the PhenoPLIER framework.** **a)** The integration process between gene co-expression patterns from MultiPLIER [1] (top) and TWAS results from PhenomeXcan [2]. PhenoPLIER projects gene-based association results on ~4,000 traits to a latent space learned from a large gene expression compendium (recount2 [3]). This generates matrix  $\hat{M}$ , where each trait is now described by latent variables/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). PMBC: Explain; mDCs: Explain.

(This paragraph has probably too much detail about MultiPLIER) MultiPLIER [1] is a recent computational strategy that extracts patterns of co-expressed genes from large gene

expression datasets. The approach uses an unsupervised matrix factorization method that was employed to extract latent variables from recount2 [3]. The latent variables (LVs), essentially gene modules, then revealed biological processes associated with disease severity in rare disease datasets that were too small for effective model training. These gene sets aligned well with known biological pathways and predicted cell type composition, even though the approach was not explicitly designed for this goal.

Although the authors showed that certain patterns learned by MultiPLIER resemble known biology, most of the LVs identified are completely unknown. Since genes in these modules vary together in certain cell types and tissues, it's expected that they may also function together [4,5]. To test whether patterns in the expression space match those in the TWAS space, we used PhenomeXcan [2], a massive transcriptome-wide association studies (TWAS) resource obtained from the UK Biobank [6] and other cohorts (Figure 1 a). These results were projected to the low-dimensional gene 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  $\mathbf{M}^{n \times t}$  has gene-trait associations from MultiXcan [7] (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 LVs, thus we can potentially infer the effects of gene modules on different human traits. Since the MultiPLIER models also provide the experimental conditions (such as cell types and tissues) in which genes in a module are concurredly expressed, our approach would also allow inferring the context in which the gene module affects a trait or disease.

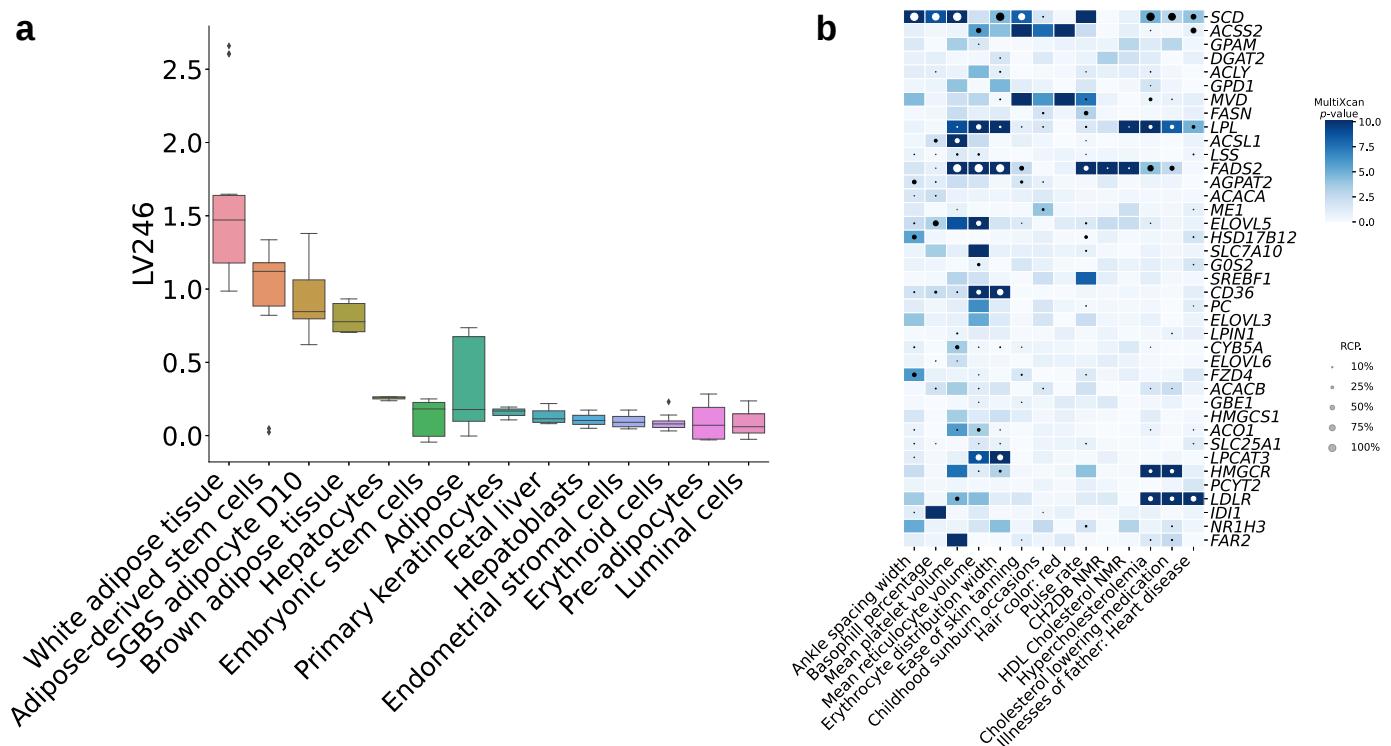
In the original MultiPLIER study [1], the authors found an LV significantly associated with previously known neutrophil gene-sets and highly correlated with neutrophil estimates from gene expression. We analyzed this LV using our approach (Figure 1 b), and found that 1) neutrophil counts and other white blood cell traits from PhenomeXcan [2] were ranked among the top 10 traits for this LV, and 2) that the genes in this LV are expressed in neutrophil cells (see more details in the supplementary material). These initial results strongly suggest 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 to infer 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.

- Minor: LV603 is neutrophil-associated, but it is not significantly associated with other myeloid lineage cell types (see Figure S2A in MultiPLIER study). Maybe we can add a genes-trait figure of MultiXcan and fastENLOC results to see this better.

## 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 single 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 identifying 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 **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 weight, from largest (top gene *SCD*) to smallest (gene *FAR2*). SGBS: Simpson Golabi Behmel Syndrome; CH2DB: CH<sub>2</sub> 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.

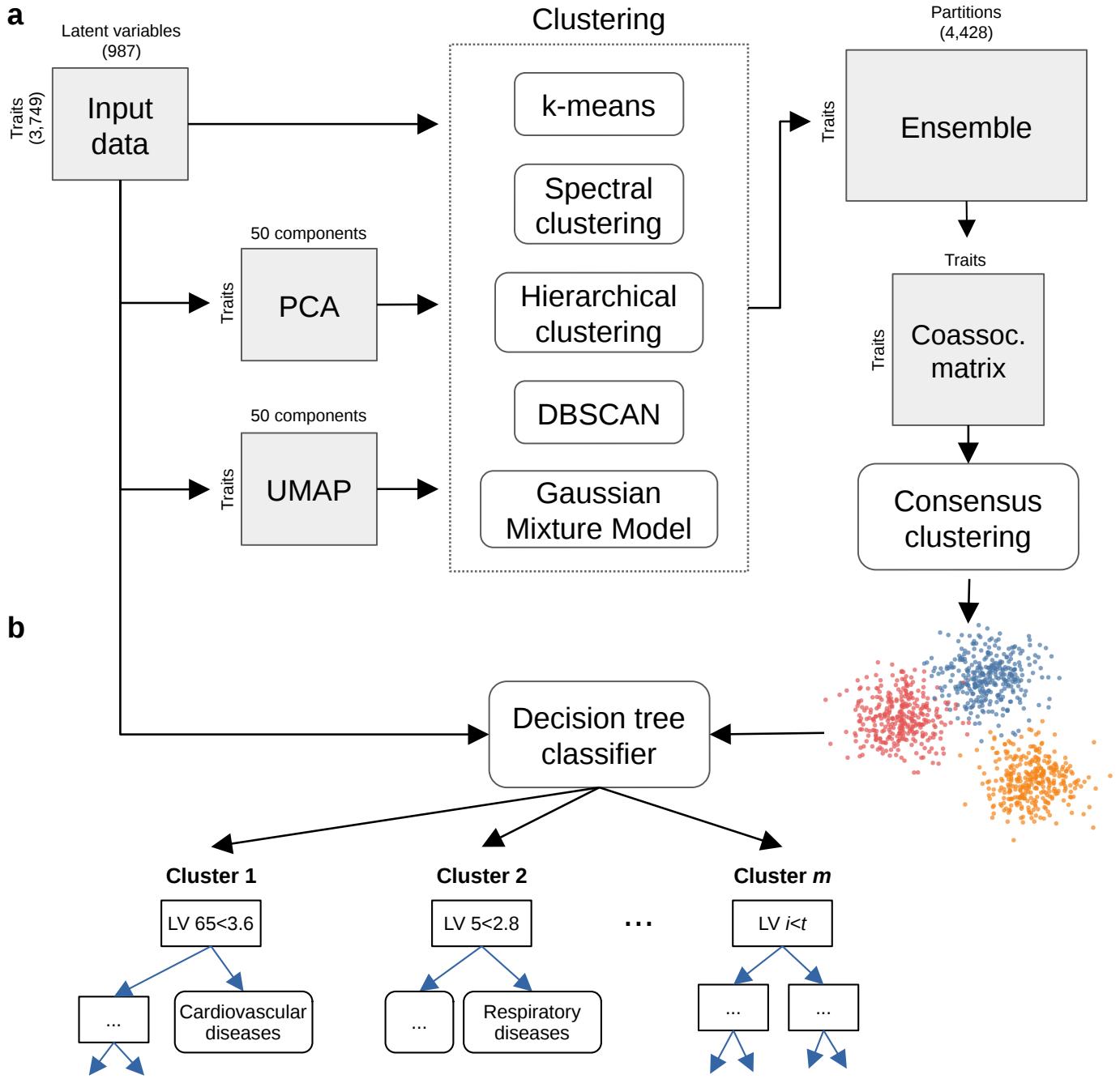
#### Notes:

- 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  $-\log_{10}(p\text{-value})$  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.

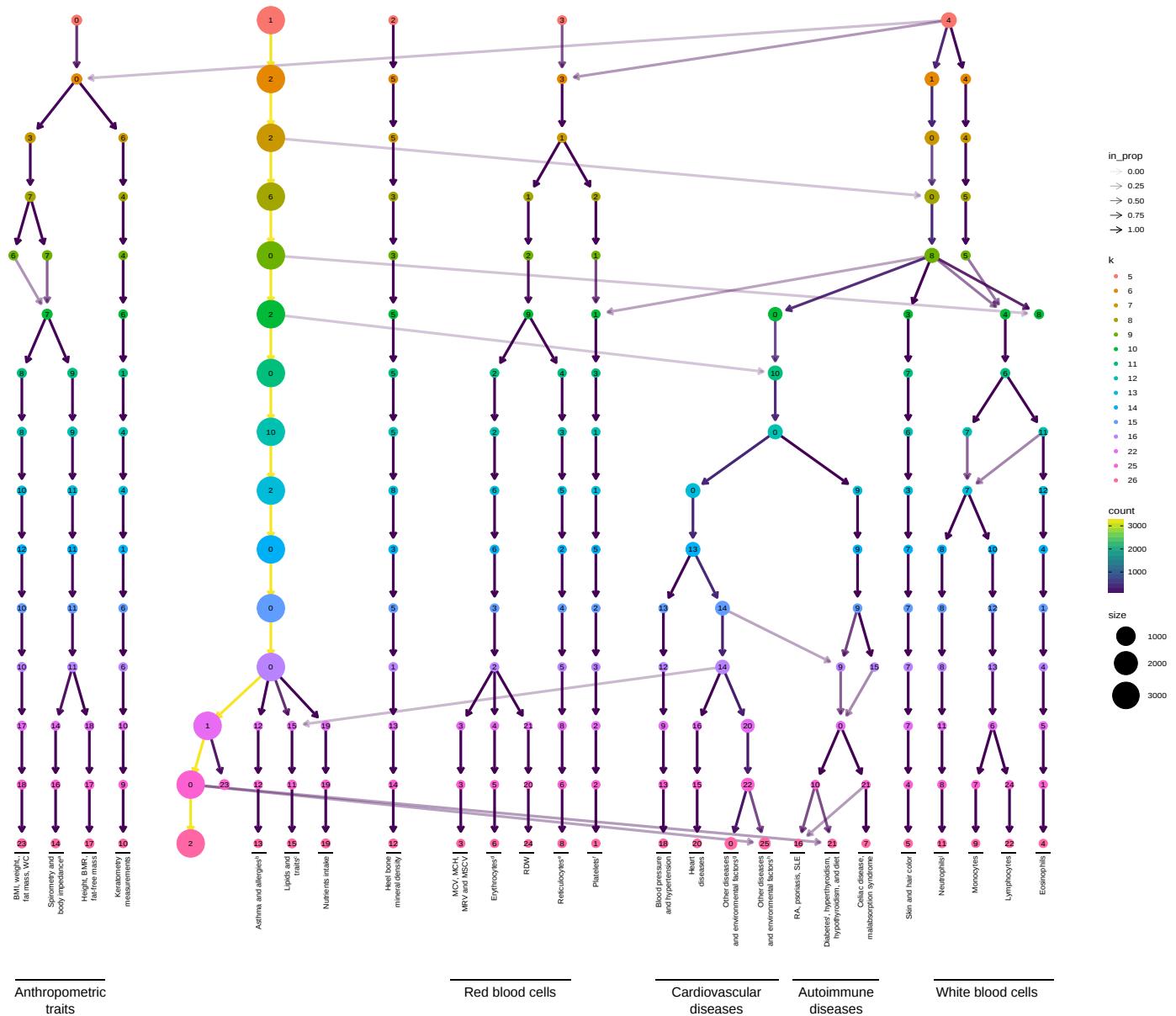
## Clusters of traits in the gene module space are associated with relevant transcriptional processes



**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 of parameters are applied on the same data, and later combined into a consolidated solution [8,9,10] (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 [8,11,12]. A diverse

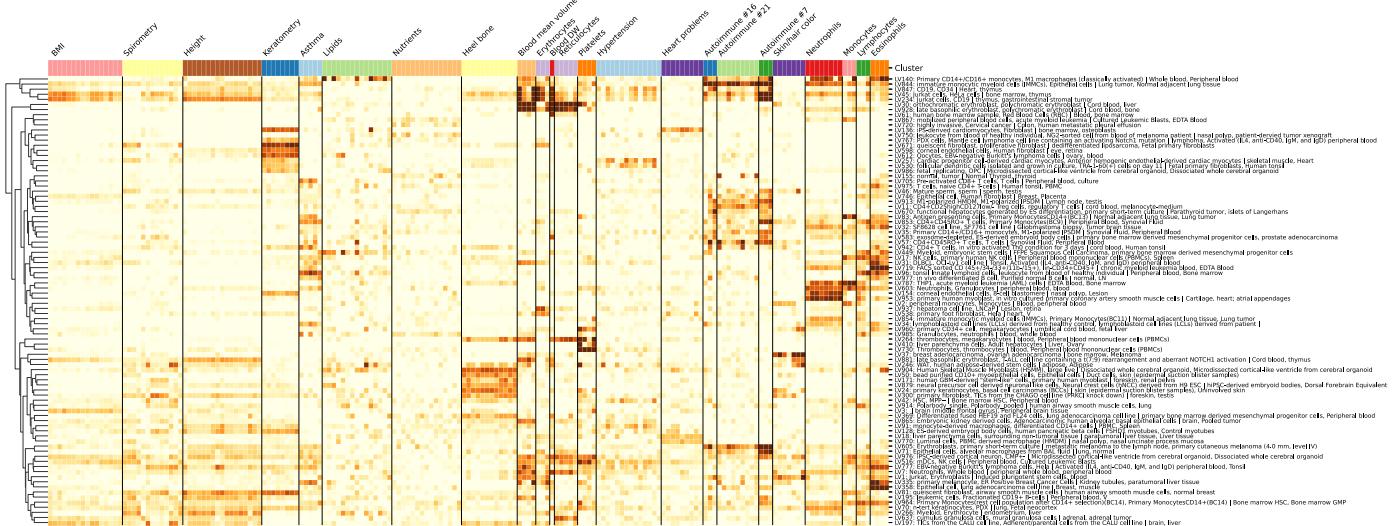
set of solutions can be generated by using different data representations (such as dimensionality reduction methods or subsets of features), clustering algorithms with distinct assumptions ( $k$ -means, for instance, assumes hyperspherical clusters), and a varying set of algorithm's parameters (such 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 [13] 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 [14] (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 spheroid cell volume; RDW: red cell (erythrocyte) distribution width; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus; <sup>a</sup> includes whole-body, arms and legs impedances; <sup>b</sup> allergies refer to allergic rhinitis or atopic dermatitis; <sup>c</sup> includes Alzheimer's disease, coronary artery disease, breast cancer, fasting blood glucose and insulin measurements, inflammatory bowel disease, and atopic dermatitis; <sup>d</sup> includes erythrocyte count, hemoglobin concentration, and hematocrit percentage; <sup>e</sup> includes reticulocyte count and percentage, immature reticulocyte fraction, and high light scatter reticulocytes count and percentage; <sup>f</sup> includes platelet count, platelet crit, mean platelet volume, and platelet distribution width. <sup>g</sup> 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. <sup>h</sup> 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. <sup>i</sup> age when diabetes was first diagnosed; <sup>j</sup> 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 [15], 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 [16] (DMAP\_ERY pathways in Supplementary Figure 11), 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 hemathological assays on platelets; or LV598, whose genes are expressed in corneal endothelial cells, and are almost only related to keratometry measurements.

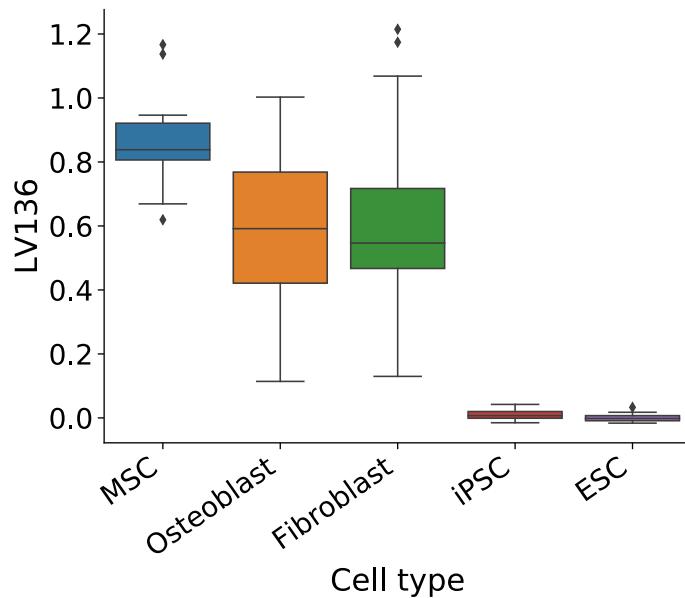
- Maybe include Ben's question about why we expect these clustering results. My answer was more or less that far apart clusters are explained by more specific LVs to those traits, and the complex branch is more related to traits that are highly connected to all biological processes, where more subtle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

## **Cardiovascular traits**

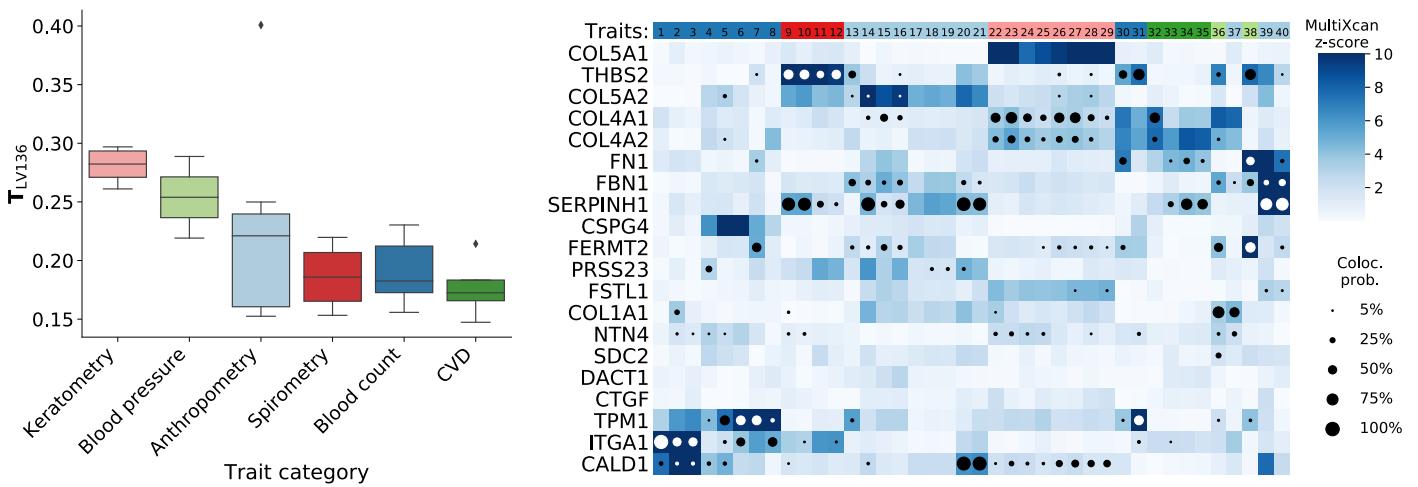
See if we are going to analyze this cluster.

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.

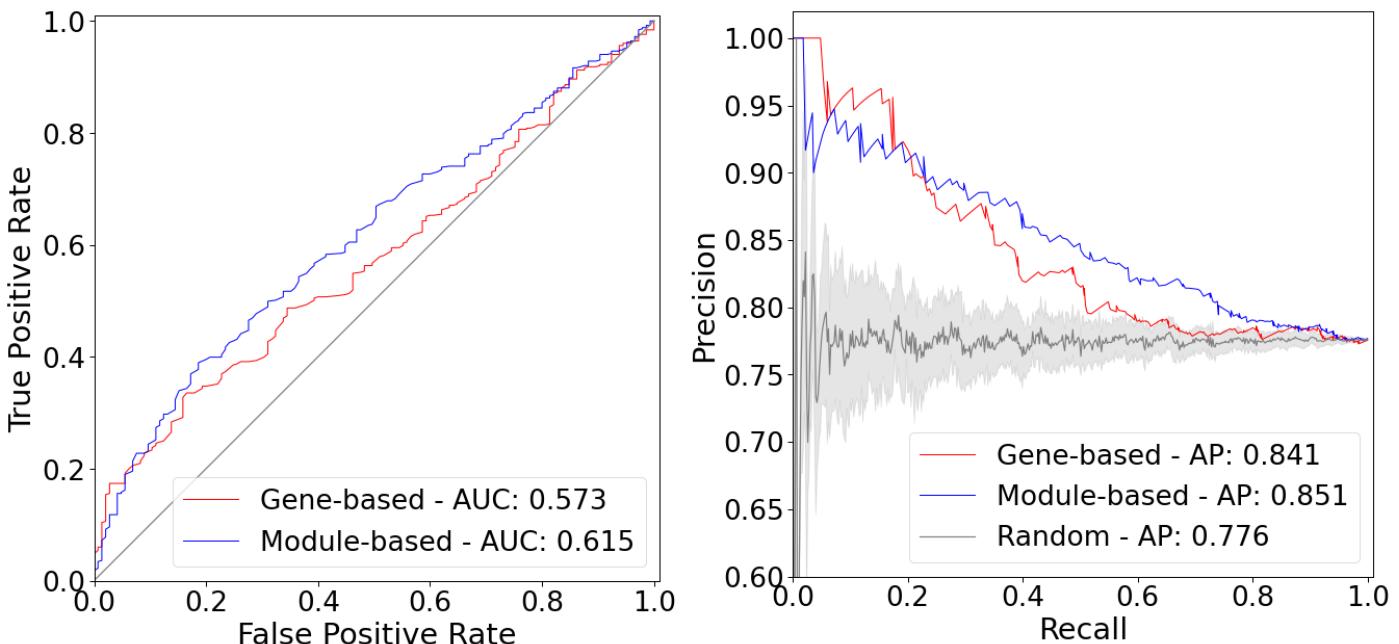
## Asthma and allergies

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.

## PhenoPLIER improves drug-disease prediction



**Figure 8: Drug-disease prediction.** Explain. AUC: area under the curve; AP: average precision.

We showed that the latent representation learned in a large gene expression data set is helpful to link novel gene sets to complex diseases and the cell types where they are expressed. We reasoned that this representation might be also useful to more accurately predict the potential therapeutic effects of drugs. If the gene patterns captured by MultiPLIER represent real and possibly unknown biological pathways, then our approach might actually produce a more accurate estimation of the effects of a perturbed molecular mechanisms on disease, and also how pharmacological perturbations affect the process activity. This would, in addition to connecting gene modules with their context-specific effects on complex traits, identify which compounds might provide an avenue to alter a dysregulated process activity for therapeutical utility.

To test this hypothesis, we used a gold standard of drug-disease medical indications [17,18] to evaluate and compare the prediction performance of both the original gene-disease associations from PhenomeXcan, and its projection representing gene module-disease associations. To test this, we used the transcriptional responses to small molecule perturbations profiled in LINCS L1000 [19], which were further processed to obtain consensus signatures and map to DrubBank IDs [17,20,21]. To compute a drug-disease score, we followed a similar procedure used previously [22] to anti-correlates gene-trait associations from TWAS and expression profiles of drugs using their signed  $z$ -scores (see the supplementary material). Here we used the dot product between gene-trait  $z$ -scores and the consensus  $z$ -scores in LINCS L1000, which led to a score for each drug-disease pair. (Add more details?). 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-trait scores and module expression profiles of drugs.

(Add number of drugs, diseases, and final number of mappings)

The ROC and Precision-Recall (PR) curves comparing both approaches are shown in Figure 8. Notably, the gene module-based approach proposed here clearly outperformed the gene-based one, with an area under the curve (AUC) of 0.615 vs 0.573, and an average precision (AP) of 0.851 vs 0.841. This is particularly striking given that the projected TWAS results represent a reduced or compressed version of the complete gene-based associations, suggesting that a gene module perspective can be more informative, for instance, for drug-repurposing scenarios using genetic studies.

Another analysis here could be:

- Try the prediction again by keeping well-aligned LVs only. Do we get the same prediction performance?

## Discussion

---

## Conclusions

---

# References

---

- 1. MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease**  
Jaclyn N. Taroni, Peter C. Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A. Merkel, Casey S. Greene  
*Cell Systems* (2019-05) <https://doi.org/gf75g5>  
DOI: [10.1016/j.cels.2019.04.003](https://doi.org/10.1016/j.cels.2019.04.003) · PMID: [31121115](https://pubmed.ncbi.nlm.nih.gov/31121115/) · PMCID: [PMC6538307](https://pubmed.ncbi.nlm.nih.gov/PMC6538307/)
- 2. PhenomeXcan: Mapping the genome to the phenotype through the transcriptome**  
Milton Pividori, Padma S. Rajagopal, Alvaro Barbeira, Yanyu Liang, Owen Melia, Lisa Bastarache, YoSon Park, GTEx Consortium, Xiaoquan Wen, Hae K. Im  
*Science Advances* (2020-09) <https://doi.org/ghbvbf>  
DOI: [10.1126/sciadv.aba2083](https://doi.org/10.1126/sciadv.aba2083) · PMID: [32917697](https://pubmed.ncbi.nlm.nih.gov/32917697/)
- 3. Reproducible RNA-seq analysis using recount2**  
Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead, Jeffrey T Leek  
*Nature Biotechnology* (2017-04-11) <https://doi.org/gf75hp>  
DOI: [10.1038/nbt.3838](https://doi.org/10.1038/nbt.3838) · PMID: [28398307](https://pubmed.ncbi.nlm.nih.gov/28398307/) · PMCID: [PMC6742427](https://pubmed.ncbi.nlm.nih.gov/PMC6742427/)
- 4. Finding function: evaluation methods for functional genomic data**  
Chad L Myers, Daniel R Barrett, Matthew A Hibbs, Curtis Huttenhower, Olga G Troyanskaya  
*BMC Genomics* (2006-07-25) <https://doi.org/fg6wnk>  
DOI: [10.1186/1471-2164-7-187](https://doi.org/10.1186/1471-2164-7-187) · PMID: [16869964](https://pubmed.ncbi.nlm.nih.gov/16869964/) · PMCID: [PMC1560386](https://pubmed.ncbi.nlm.nih.gov/PMC1560386/)
- 5. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens**  
Naihui Zhou, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, George Georghiou, Huy N. Nguyen, Md Nafiz Hamid, ... Iddo Friedberg  
*Genome Biology* (2019-11-19) <https://doi.org/ggnxpz>  
DOI: [10.1186/s13059-019-1835-8](https://doi.org/10.1186/s13059-019-1835-8) · PMID: [31744546](https://pubmed.ncbi.nlm.nih.gov/31744546/) · PMCID: [PMC6864930](https://pubmed.ncbi.nlm.nih.gov/PMC6864930/)
- 6. The UK Biobank resource with deep phenotyping and genomic data**  
Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, ... Jonathan Marchini  
*Nature* (2018-10-10) <https://doi.org/gfb7h2>  
DOI: [10.1038/s41586-018-0579-z](https://doi.org/10.1038/s41586-018-0579-z) · PMID: [30305743](https://pubmed.ncbi.nlm.nih.gov/30305743/) · PMCID: [PMC6786975](https://pubmed.ncbi.nlm.nih.gov/PMC6786975/)
- 7. Integrating predicted transcriptome from multiple tissues improves association detection**  
Alvaro N. Barbeira, Milton Pividori, Jiamao Zheng, Heather E. Wheeler, Dan L. Nicolae, Hae Kyung Im  
*PLOS Genetics* (2019-01-22) <https://doi.org/ghs8vx>  
DOI: [10.1371/journal.pgen.1007889](https://doi.org/10.1371/journal.pgen.1007889) · PMID: [30668570](https://pubmed.ncbi.nlm.nih.gov/30668570/) · PMCID: [PMC6358100](https://pubmed.ncbi.nlm.nih.gov/PMC6358100/)
- 8. Diversity control for improving the analysis of consensus clustering**  
Milton Pividori, Georgina Stegmayer, Diego H. Milone  
*Information Sciences* (2016-09) <https://doi.org/ghtqbk>  
DOI: [10.1016/j.ins.2016.04.027](https://doi.org/10.1016/j.ins.2016.04.027)

## 9. Clustering ensembles: models of consensus and weak partitions

A. Topchy, A. K. Jain, W. Punch

*IEEE Transactions on Pattern Analysis and Machine Intelligence* (2005-12) <https://doi.org/c8z32x>

DOI: [10.1109/tpami.2005.237](https://doi.org/10.1109/tpami.2005.237) · PMID: [16355656](#)

## 10. 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>

## 11. A Link-Based Approach to the Cluster Ensemble Problem

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

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

DOI: [10.1109/tpami.2011.84](https://doi.org/10.1109/tpami.2011.84) · PMID: [21576752](#)

## 12. 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/gtzwt>

DOI: [10.1016/j.patcog.2014.04.005](https://doi.org/10.1016/j.patcog.2014.04.005)

## 13. 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>

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

Luke Zappia, Alicia Oshlack

*GigaScience* (2018-07) <https://doi.org/gfzqf5>

DOI: [10.1093/gigascience/giy083](https://doi.org/10.1093/gigascience/giy083) · PMID: [30010766](#) · PMCID: [PMC6057528](#)

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

Darryl P Leong, Koon K Teo, Sumathy Rangarajan, Patricio Lopez-Jaramillo, Alvaro Avezum, Andres Orlandini, Pamela Seron, Suad H Ahmed, Annika Rosengren, Roya Kelishadi, ...

*Lancet (London, England)* (2015-05-13) <https://www.ncbi.nlm.nih.gov/pubmed/25982160>

DOI: [10.1016/s0140-6736\(14\)62000-6](https://doi.org/10.1016/s0140-6736(14)62000-6) · PMID: [25982160](#)

## 16. Densely Interconnected Transcriptional Circuits Control Cell States in Human Hematopoiesis

Noa Novershtern, Aravind Subramanian, Lee N. Lawton, Raymond H. Mak, W. Nicholas Haining, Marie E. McConkey, Naomi Habib, Nir Yosef, Cindy Y. Chang, Tal Shay, ... Benjamin L. Ebert

*Cell* (2011-01) <https://doi.org/cf5k92>

DOI: [10.1016/j.cell.2011.01.004](https://doi.org/10.1016/j.cell.2011.01.004) · PMID: [21241896](#) · PMCID: [PMC3049864](#)

## 17. Systematic integration of biomedical knowledge prioritizes drugs for repurposing

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

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

DOI: [10.7554/elife.26726](https://doi.org/10.7554/elife.26726) · PMID: [28936969](#) · PMCID: [PMC5640425](#)

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

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

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

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

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

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, ... Todd R. Golub  
*Cell* (2017-11) <https://doi.org/cgwt>  
DOI: [10.1016/j.cell.2017.10.049](https://doi.org/10.1016/j.cell.2017.10.049) · PMID: [29195078](https://pubmed.ncbi.nlm.nih.gov/29195078/) · PMCID: [PMC5990023](https://pubmed.ncbi.nlm.nih.gov/PMC5990023/)

## 20. DrugBank 4.0: shedding new light on drug metabolism

Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, ... David S. Wishart  
*Nucleic Acids Research* (2014-01) <https://doi.org/f3mn6d>  
DOI: [10.1093/nar/gkt1068](https://doi.org/10.1093/nar/gkt1068) · PMID: [24203711](https://pubmed.ncbi.nlm.nih.gov/24203711/) · PMCID: [PMC3965102](https://pubmed.ncbi.nlm.nih.gov/PMC3965102/)

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

Daniel Himmelstein, Leo Brueggeman, Sergio Baranzini  
*Zenodo* (2016-03-08) <https://doi.org/f3mqvr>  
DOI: [10.5281/zenodo.47223](https://doi.org/10.5281/zenodo.47223)

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

Hon-Cheong So, Carlos Kwan-Long Chau, Wan-To Chiu, Kin-Sang Ho, Cho-Pong Lo, Stephanie Ho-Yue Yim, Pak-Chung Sham  
*Nature Neuroscience* (2017-08-14) <https://doi.org/gbrssh>  
DOI: [10.1038/nn.4618](https://doi.org/10.1038/nn.4618) · PMID: [28805813](https://pubmed.ncbi.nlm.nih.gov/28805813/)

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

Weiguang Mao, Elena Zaslavsky, Boris M. Hartmann, Stuart C. Sealfon, Maria Chikina  
*Nature Methods* (2019-06-27) <https://doi.org/gf75g6>  
DOI: [10.1038/s41592-019-0456-1](https://doi.org/10.1038/s41592-019-0456-1) · PMID: [31249421](https://pubmed.ncbi.nlm.nih.gov/31249421/) · PMCID: [PMC7262669](https://pubmed.ncbi.nlm.nih.gov/PMC7262669/)

# Methods

---

## Pathway-level information extractor (PLIER)

MultiPLIER [1], the computational strategy used in this work, extracts patterns of co-expressed genes on large gene expression datasets. MultiPLIER applies the pathway-level information extractor method (PLIER) [23] to the recount2 data [3], which performs unsupervised learning using prior knowledge to reduce technical noise. Via a matrix factorization approach, PLIER deconvolutes the gene expression data into a set of latent variables (LV) that each represent a gene module (i.e. a set of genes with coordinated expression patterns).

COMPLETE

## Top latent variables associated with neutrophils

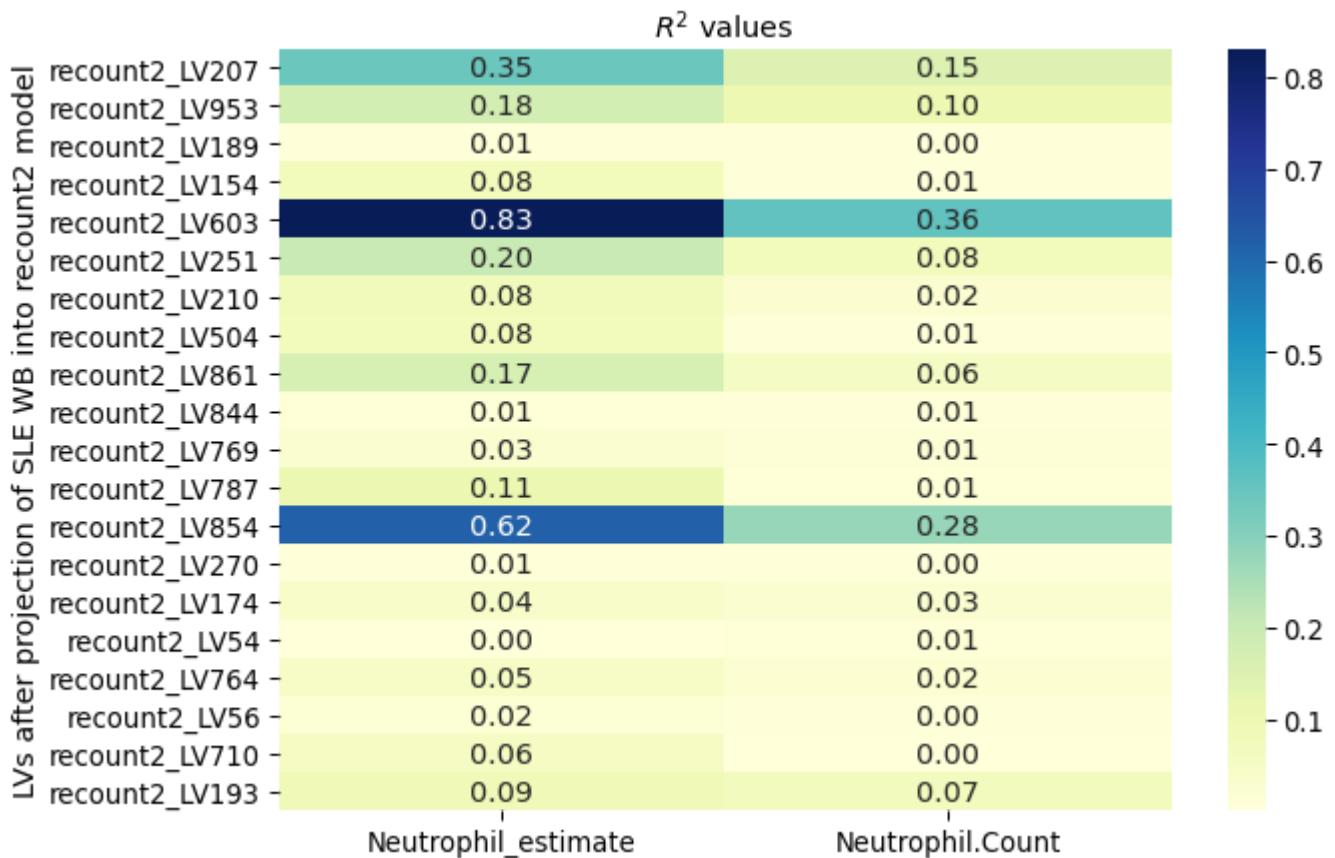


Figure 9: Correlation of neutrophil counts with top LVs associated with neutrophils traits.

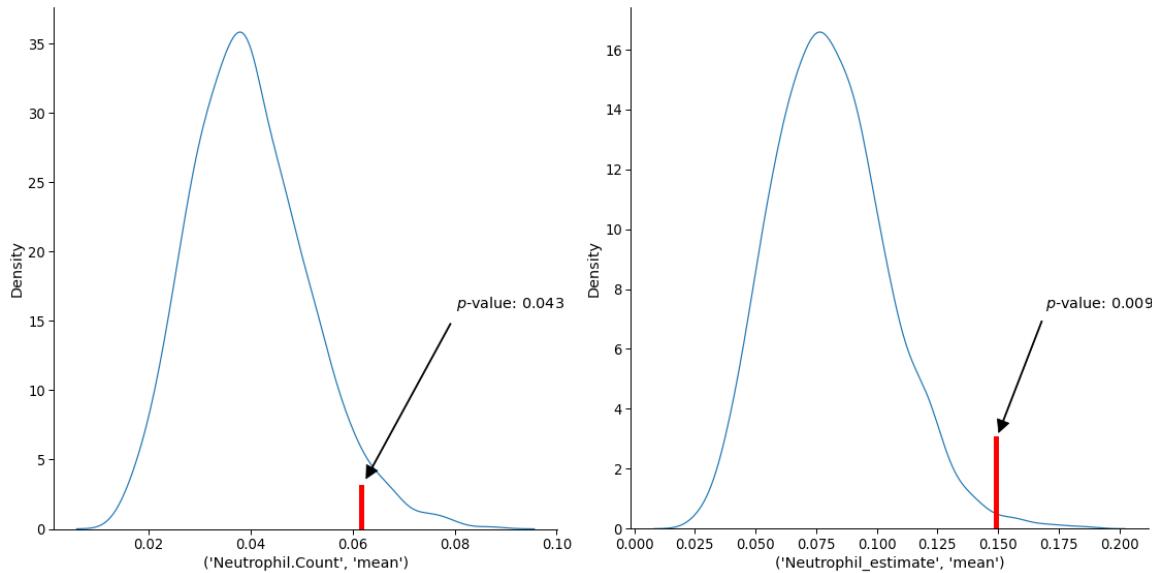


Figure 10: Significance of neutrophil counts correlation.

## CRISPR-Cas9 screening

[Add details](#)

## Consensus clustering of traits in PhenomeXcan

### Dimensionality reduction

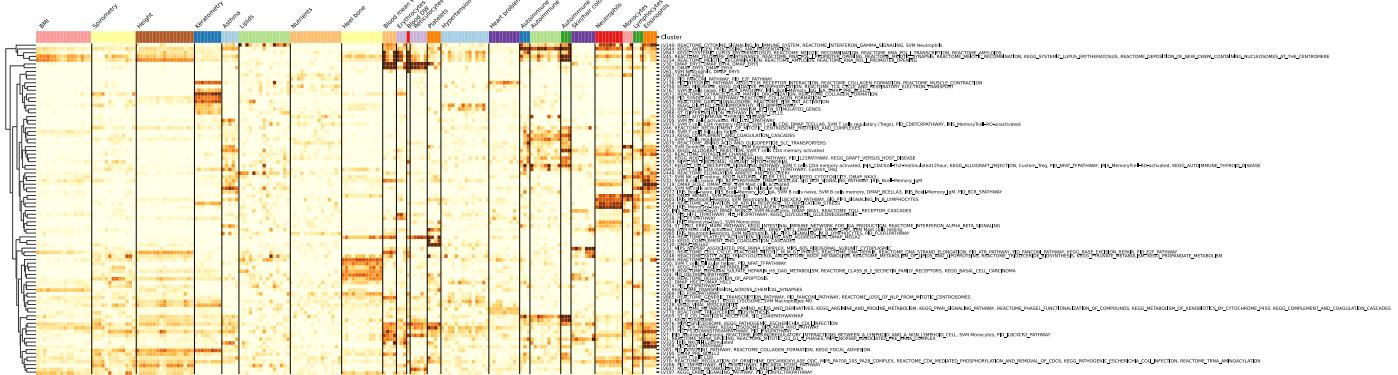
### Ensemble creation: clustering algorithms and parameters

- 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 11: 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

- anti-correlation using dot product of s-predixcan on all tissues and lincs

## Supplementary material

---