PhenoPLIER: integrating transcriptome-wide association studies with machine learning

Draft manuscript
Text in red or with red background are internal comments

This manuscript (permalink) was automatically generated from greenelab/phenoplier_manuscript@835bbfc on November 18, 2020.

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)

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · GreeneScientist
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA 19102, 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

Results

Integrating gene expression patterns with transcription-wide association studies

Here we introduce our framework to perform the integration.

1. Brief background on MultiPLIER and PhenomeXcan.
2. Explanation of the framework depicted in Figure 1 and its components.
3. Description of initial set of results matching previous findings (neutrophil counts).

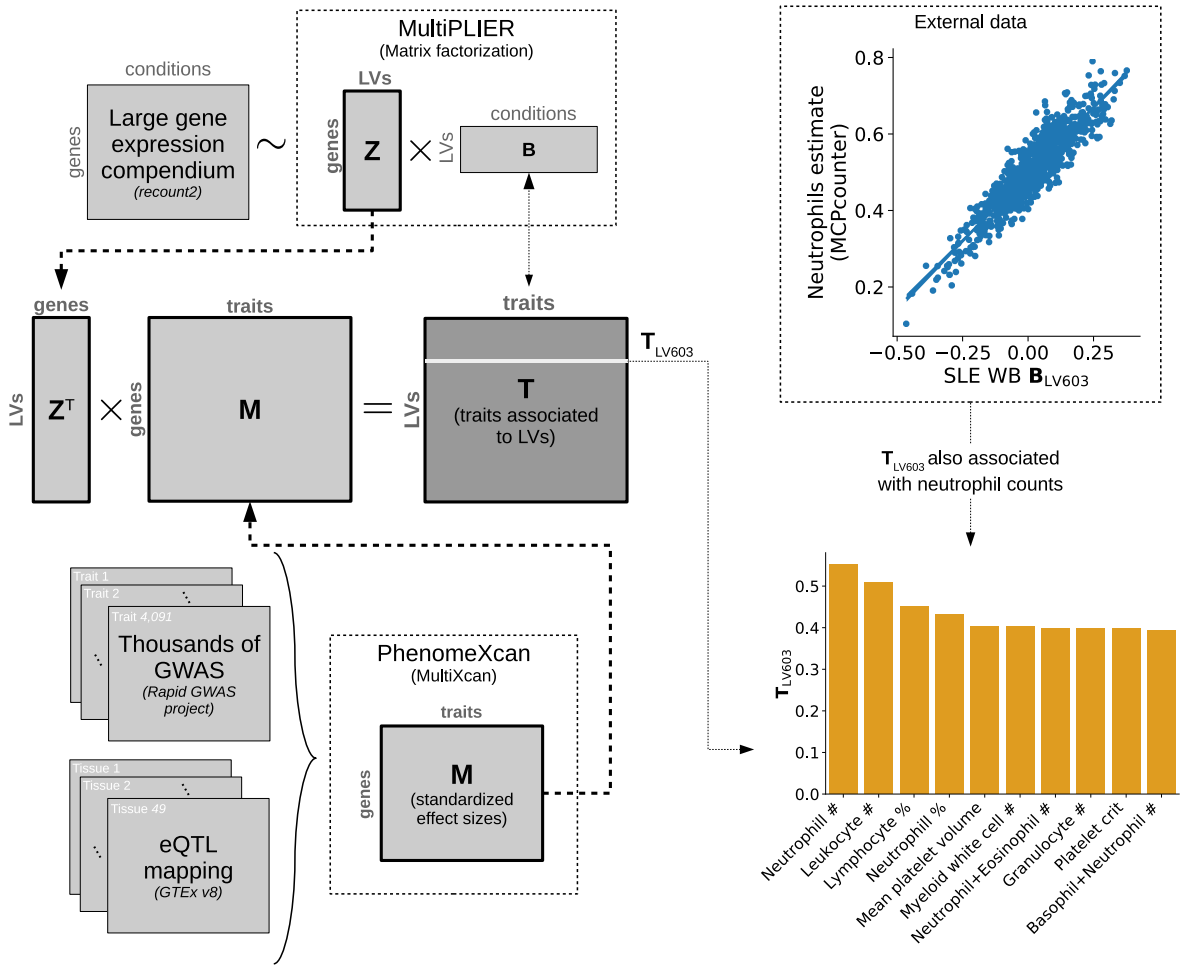

Figure 1: Schematic of the PhenoPLIER framework. The integration process (middle) between MultiPLIER [1] (top) and PhenomeXcan [2] (bottom) outputs matrix $\mathbf{T}$. LV603 was previously found to be strongly associated with neutrophil estimates when projecting a dataset of systemic lupus erythematosus (SLE) whole blood (WB) into MultiPLIER. After integration, this LV is also highly associated with blood count traits, with neutrophil counts at the top.

Projecting associations into a gene expression latent space reveals expected and novel trait clusters

In this section, we already introduced some evidence that the proposal could work. The next step (this section) is to cluster traits using the LVs as features.

1. Briefly explain idea of clustering traits using LVs.

2. Explain methods here, and then we see whether that should be moved elsewhere.
    1. Dimentionality reduction using UMAP.
    2. Spectral clustering approach.
    3. Stability measure to detect best number of clusters.
    4. Approach to interpret clusters by training a decision tree classifier.
3. We should have a set of expected results. For instance:
    1. Keratometry measurements are clustered together.
    2. Heel bone mineral density are also clustered together.
    3. See other clusters.
4. We should also include some “novel” clusters, or expected clusters with some novel traits in it.
    1. Lipids cluster? I’m working on this now.
5. Once we focus on a cluster, we provide an interpreation of why those traits are clustered together (which LVs are driving the association). Here we’ll need to explain the method we might be using for this (decision trees on the original data). This could be a separate section.

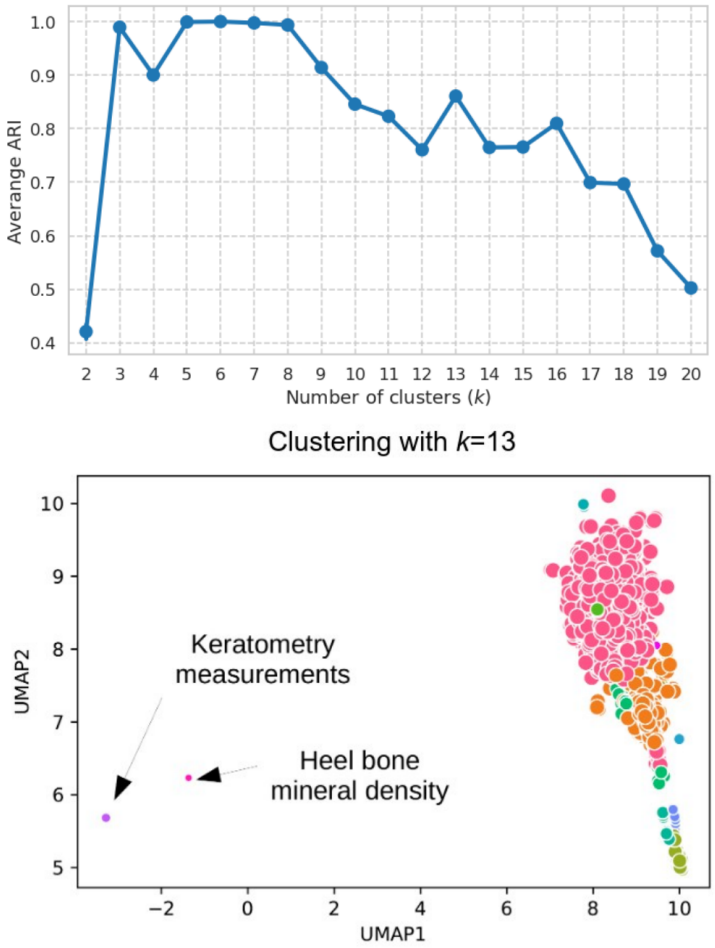

Figure 2: Clustering of traits using matrix $\mathbf{T}^\top$. Estimation of number of clusters using the consensus index method [3]. A spectral clustering approach was used on matrix $\mathbf{T}^\top$ to group traits. The algorithm was run 100 times for each k value from 2 to 20, and the averaged adjusted Rand-index is reported in y-axis (top). A partition with $k = 13$ was obtained from $\mathbf{T}^\top$ (bottom), where ketatometry and heel bone mineral density measurements clearly separate from the rest of traits. `Add svg version`

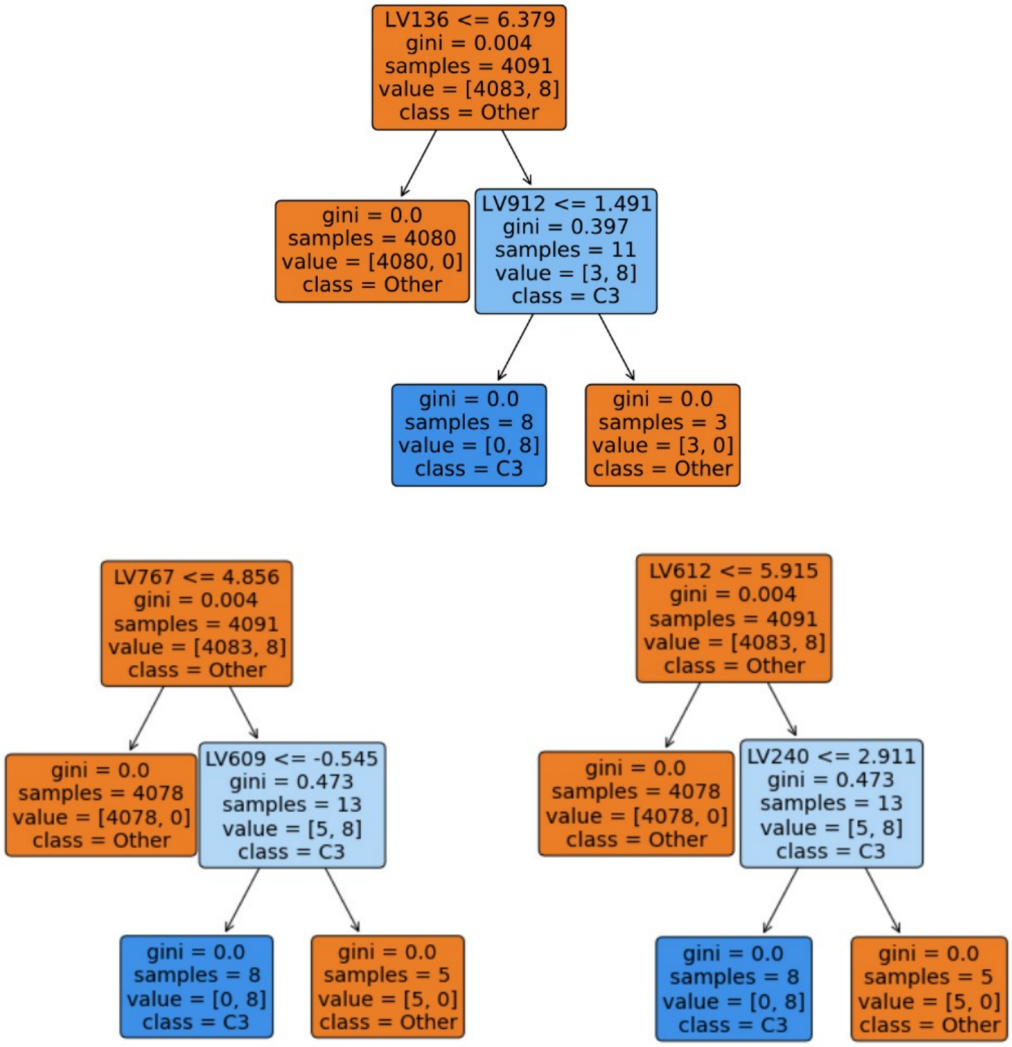

Figure 3: The top discriminating latent variables for the keratometry cluster. For clustering interpretation, a decision tree classifier was trained using the original data ($\mathbf{T}^\top$), the keratometry cluster as positive class, and the rest of traits as negative class. The top associated latent variable was LV136, followed by LV767 and LV612 (which were detected by removing LV136 and LV767 from the training data, respectively). `Add svg version.`

Analysis of keratometry cluster reveals groups of genes associated with cardiovascular-related traits

In this section, we build on our previous results to analyze a single or a couple of LVs. The idea is to show raw MultiXcan/fastENLOC results and how individual genes in the module are associated with the traits. Based on this, we could provide an interpretation (if possible) on how genes could be affecting different and related traits/diseases.

These are some articles I found relating the trait categories in Figure 4 (LV136).

1. Ocular problems and cardiovascular diseases:
    1. (2017) The Relationship Between Cardiovascular Autonomic Dysfunction and Ocular Abnormality in Chinese T2DM
    2. (2018) Looking into the eye of patients with chronic obstructive pulmonary disease: an opportunity for better microvascular profiling of these complex patients
        1. This one also relates spirometry.
    3. (2018) Residual Vision Activation and the Brain-Eye-Vascular Triad: Dysregulation, Plasticity and Restoration in Low Vision and Blindness - A Review
    4. (2017) Evaluating Ocular Blood Flow
2. Spirometry and CVD:

1. (2019) Mysterious link between the restrictive ventilatory impairment in spirometry and cardiovascular disease
2. (2018) Declining Lung Function and Cardiovascular Risk
3. (2018) Restrictive Spirometry and Cardiovascular Risk: Cause or Comorbidity?
4. (2011) Assessment of pulmonary function tests in cardiac patients
5. It would be nice to see if the direction of effect of these genes are positive for CVD and negative for FEV1.

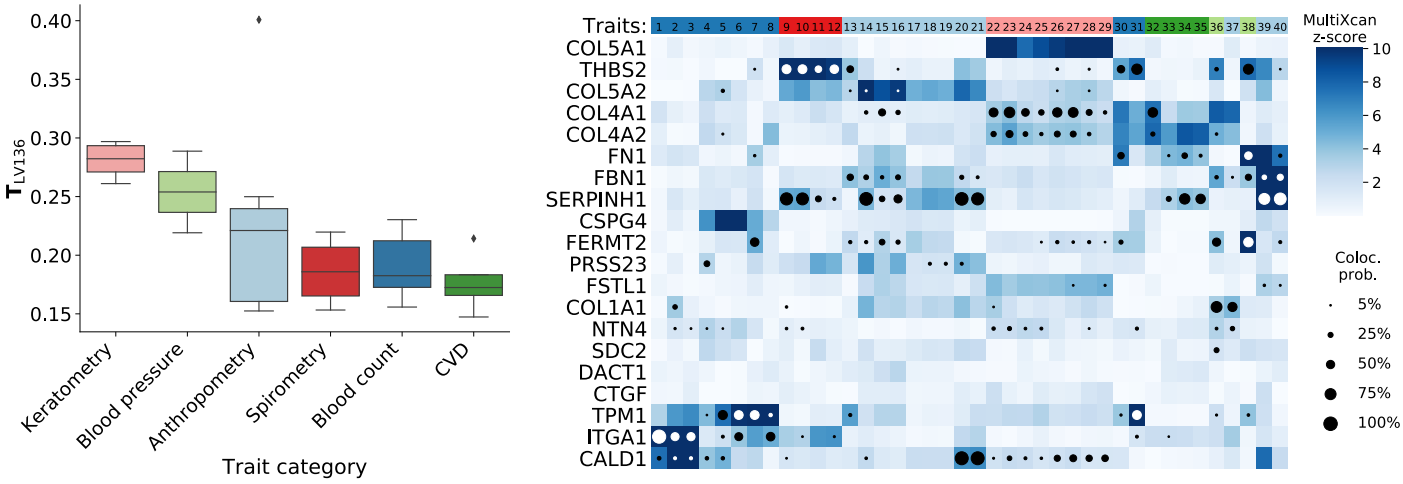

Figure 4: Traits and diseases associated with top genes in LV136. Categories of the top 40 traits associated with genes in LV136 (left), and associations of traits with the top 20 genes in LV136 (right): S-MultiXcan associations ($-\log_{10}(p-\text{value})$, thresholded at 10) are shown with gradients, whereas fastENLOC colocalization probabilities are depicted with different circle sizes (only for $> 5\%$). Colors used for trait categories are the same in both subfigures.

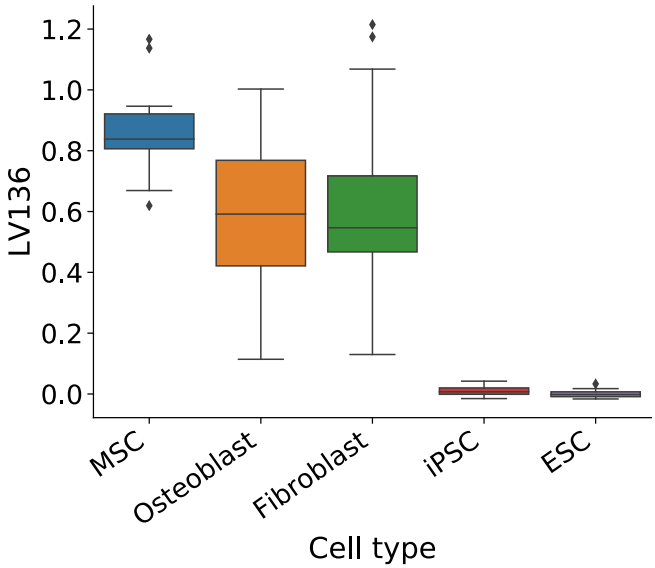

Figure 5: Cell types found in top five studies (recount2) associated with LV136. Genes associated with LV136 are highly expressed in MSC, osteoblast and fibroblast when considering all conditions for the given cell types in the top five studies in recount2. MSC: mesenchymal stem cells; iPSC: induced pluripotent stem cells; ESC: embryonic stem cells. `We should consider more studies maybe, not just the top 5.` `Remove colors.`

Clustering of LVs (gene modules) with similar trait associations

I'm not sure if it's a good idea to include this, but I leave it here just in case.

The idea here is to cluster LVs by seeing how are they associated with different traits. So, taking the keratometry cluster shown before as an example, here LV136, LV767, LV612 (and possibly others)

would be clustered together. The use case would be to start from a trait of interest and see the cluster of LVs associated with it.

Drugs associated with gene modules

This section includes the projection of Connectivity Map into the MultiPLIER space.

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 · PMID: 31121115 · PMCID: PMC6538307

2. **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, Xiaoquan Wen, Hae K. Im, The GTEx Consortium
   bioRxiv (2019-11-06) https://doi.org/gg334k
   DOI: 10.1101/833210

3. **Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance**
   Nguyen Xuan Vinh, Julien Epps, James Bailey
   Journal of Machine Learning Research http://www.jmlr.org/papers/v11/vinh10a.html