Sparse multiple Canonical Correlation Network discovery (SmCCNet)

Date: 6 Dec 2022

Presenter: Zuqi Li

About me

Zuqi Li

Affiliation: PhD student at KU Leuven
Department: Human Genetics

Education: Master in Bioinformatics at the University of Copenhagen
Bachelor in Biology at Beijing Normal University

Background: Machine Learning, computational biology, bioinformatics

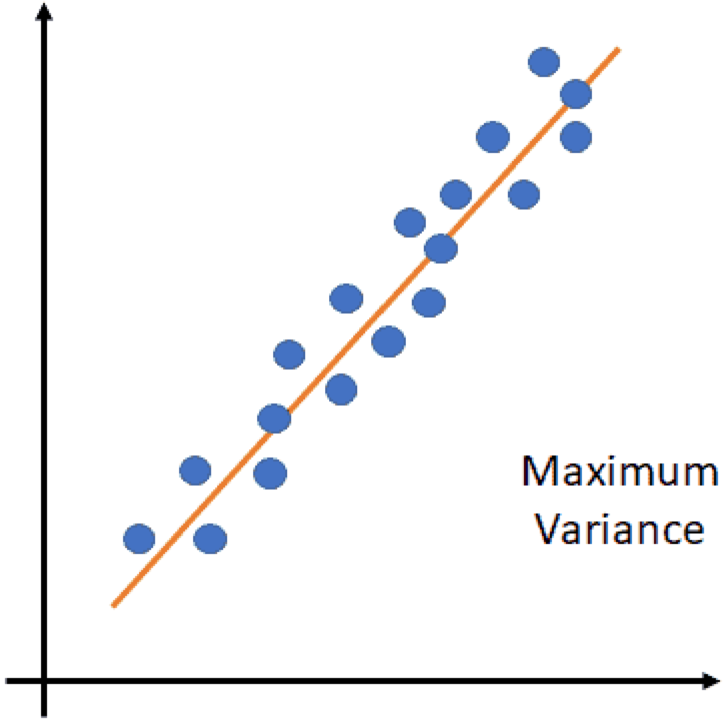
Supervisor: Prof. Dr. Dr. Kristel Van Steen

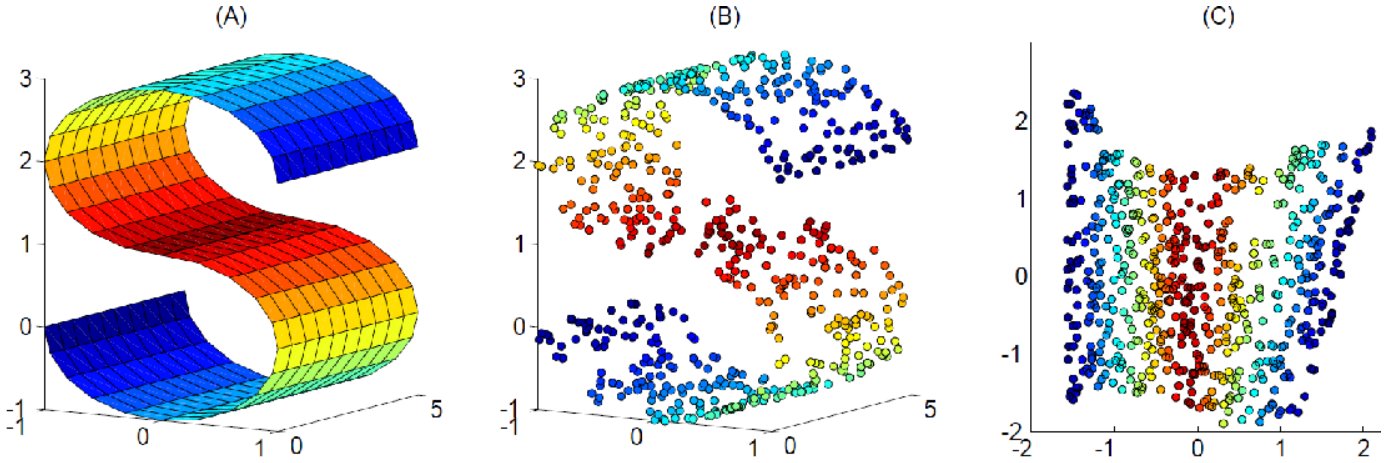
Agenda

- PCA
- CCA
- Sparse CCA
- Sparse multiple CCA
- SmCCNet
- Applications

Dimensionality reduction

- With the development of technologies and reduction in cost, high-dimensional datasets are being collected.
- Features may be correlated or noisy.





1) https://machinelearninggeek.com/dimensionality-reduction-using-pca/ 2) Engel, Daniel, Lars Hüttenberger, and Bernd Hamann. "A survey of dimension reduction methods for high-dimensional data analysis and visualization." Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

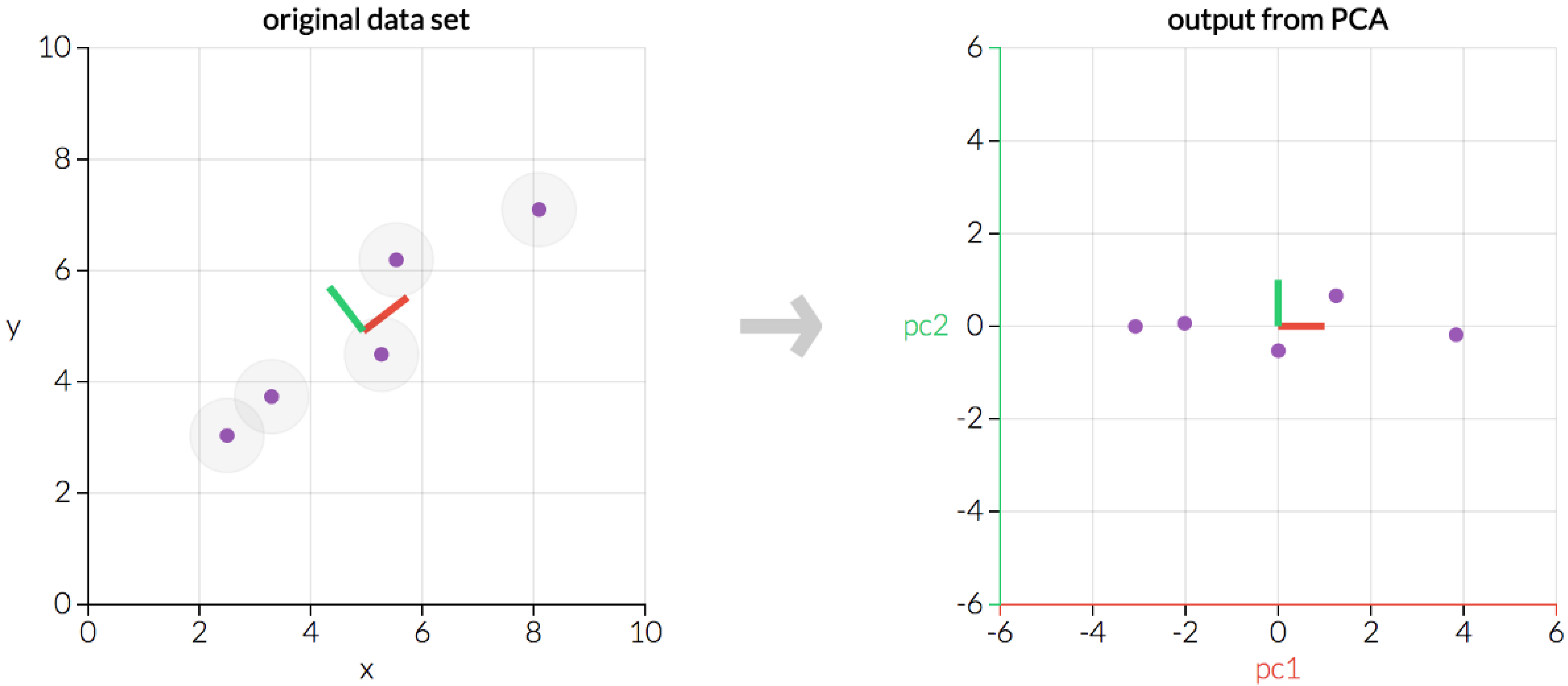
Principal Component Analysis

- Reduce dimensionality of large datasets and increase interpretability with minimal information loss.
- By creating new uncorrelated variables that successively maximize variance, i.e. principal components.

$$X \in \mathbb{R}^{n \times p} \longrightarrow X^T X \longrightarrow W \in \mathbb{R}^{p \times p} \longrightarrow T = XW$$

$$X = U\Sigma W^T \longrightarrow T = XW = U\Sigma W^T W = U\Sigma$$

Principal Component Analysis



An example from setosa.io

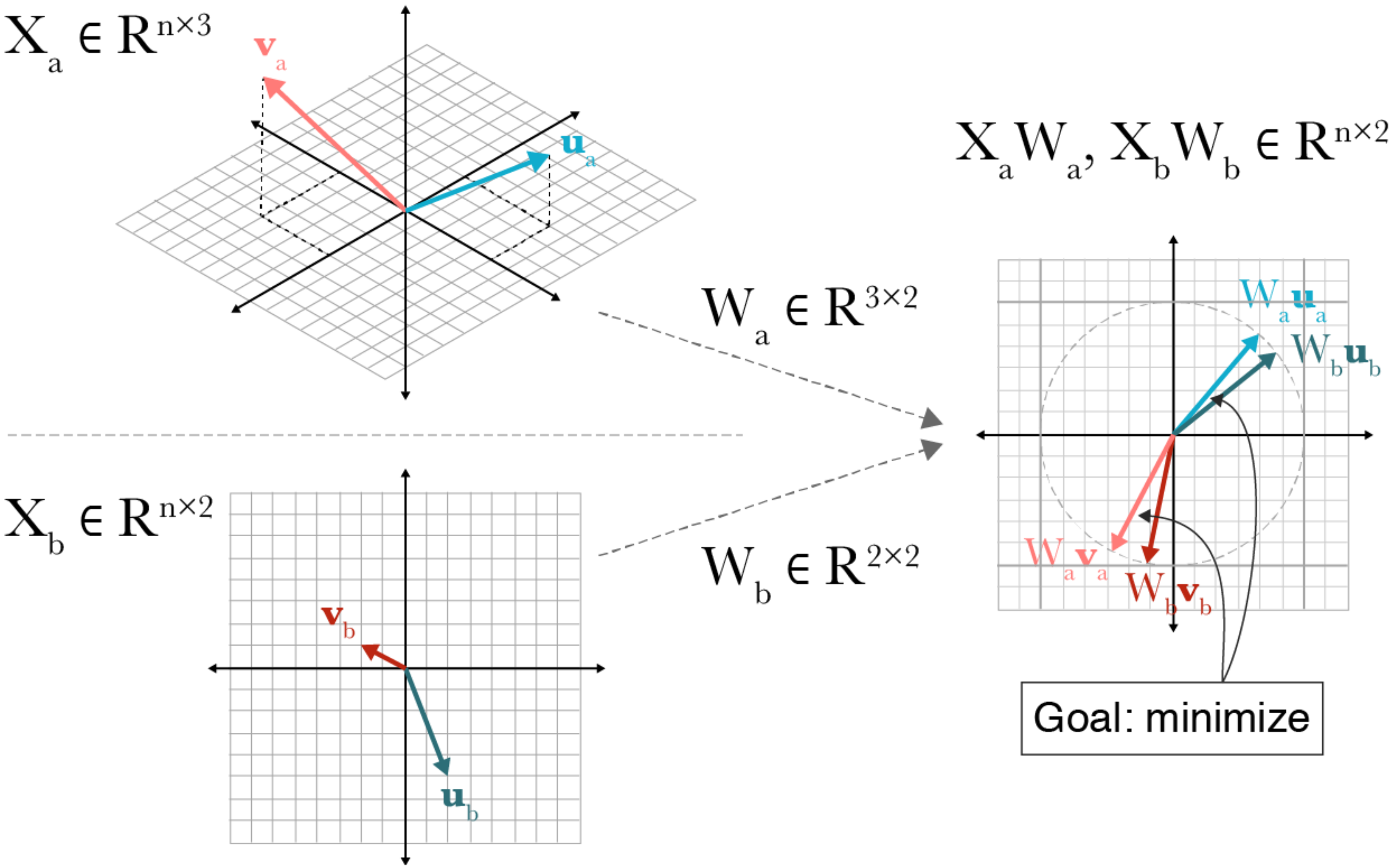
Canonical Correlation Analysis

- CCA measures the relatedness of 2 sets of features by maximizing the correlation between some linear transformations of the 2 sets.
- Given 2 standardized data matrices $X_1 \in R^{n \times p_1}$ and $X_2 \in R^{n \times p_2}$, their canonical correlation is $Corr(X_1 w_1, X_2 w_2) = w_1^T X_1^T X_2 w_2$.

$$(w_1, w_2) = \underset{\widetilde{w}_1, \widetilde{w}_2}{\arg\max}\, (\widetilde{w}_1^T X_1^T X_2 \widetilde{w}_2) \;\; s.t. \|\widetilde{w}_1\|^2 = \|\widetilde{w}_2\|^2 = 1$$

$$u = X_1 w_1$$
$$v = X_2 w_2$$

Canonical Correlation Analysis



$X_a \in R^{n\times3}$

$X_b \in R^{n\times2}$

$W_a \in R^{3\times2}$

$W_b \in R^{2\times2}$

$X_aW_a, X_bW_b \in R^{n\times2}$

Goal: minimize

https://gregorygundersen.com/blog/2018/07/17/cca/

Sparse CCA

- In practice, not all features contribute to the overall canonical correlation. Or we may want to reduce the feature set.
- Sparsity is imposed to the canonical weights by adding convex penalty functions $P_1(\cdot)$, $P_2(\cdot)$.

$$(w_1, w_2) = \underset{\widetilde{w}_1, \widetilde{w}_2}{\arg\max} \left(\widetilde{w}_1^T X_1^T X_2 \widetilde{w}_2 \right)$$

$$s.t. \|\widetilde{w}_1\|^2 = \|\widetilde{w}_2\|^2 = 1,$$

$$P_1(\widetilde{w}_1) \leq c_1, P_2(\widetilde{w}_2) \leq c_2$$

Sparse multiple CCA

- Apart from X_1, X_2, we may want to take an extra dataset X_3 into account of the canonical correlation.
- Use coefficients to prioritize the pairwise correlations.

$$(w_1, w_2, w_3) = \underset{\tilde{w}_1, \tilde{w}_2, \tilde{w}_3}{\arg\max} \left(a\tilde{w}_1^T X_1^T X_2 \tilde{w}_2 + b\tilde{w}_1^T X_1^T X_3 \tilde{w}_3 + c\tilde{w}_2^T X_2^T X_3 \tilde{w}_3\right)$$

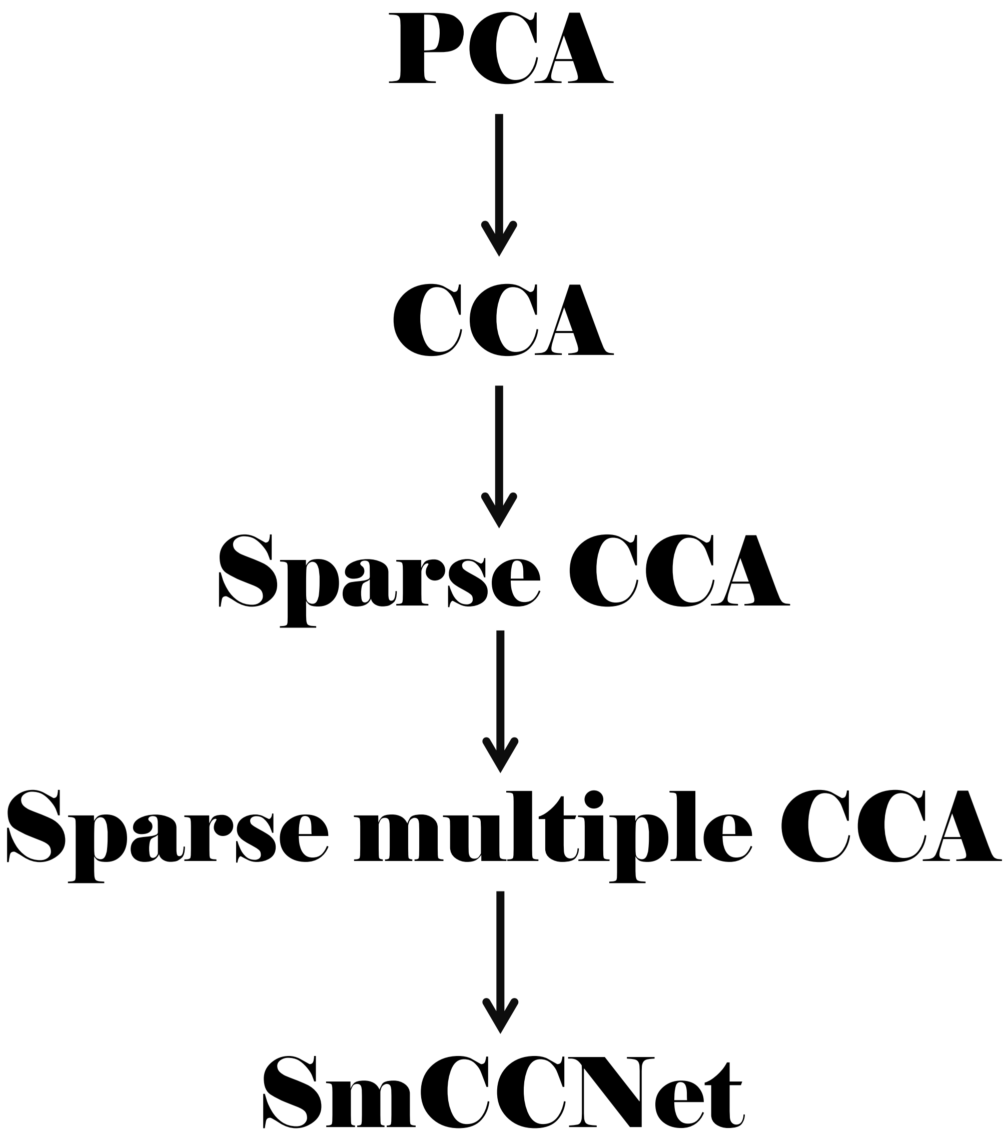
$$s.t. \|\tilde{w}_s\|^2 = 1, P_s(\tilde{w}_s) \leq c_s, s = 1,2,3$$

Sparse multiple CCA (2)

- Apart from X_1, X_2, there's also a phenotype of interest Y that has been measured for the same n subjects.
- The 3rd dataset is a phenotype, i.e. column vector.

$$(w_1, w_2) = \underset{\tilde{w}_1, \tilde{w}_2}{\arg\max} \left(a\tilde{w}_1^T X_1^T X_2 \tilde{w}_2 + b\tilde{w}_1^T X_1^T Y + c\tilde{w}_2^T X_2^T Y\right)$$

$$s.t. \|\tilde{w}_s\|^2 = 1, P_s(\tilde{w}_s) \leq c_s, s = 1,2$$



PCA

CCA

Sparse CCA

Sparse multiple CCA

SmCCNet

Unsupervised discovery of phenotype specific multi-omics networks

W Jenny Shi, Yonghua Zhuang, et al.

Bioinformatics, Volume 35, Issue 21, 1 November 2019, Pages 4336–4343, https://doi.org/10.1093/bioinformatics/btz226
Published: 08 April 2019

Background and Motivation

- Different quantitative omics measurements on the same subjects.
- Combine multiple omics data types to study complex traits.

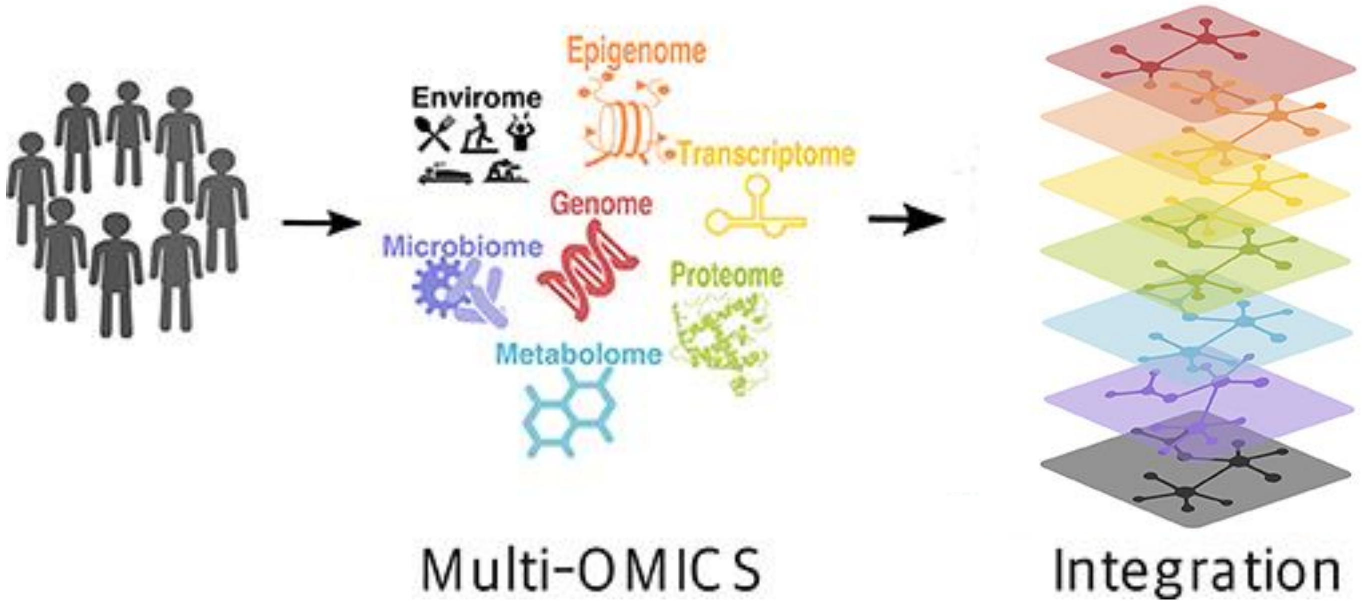
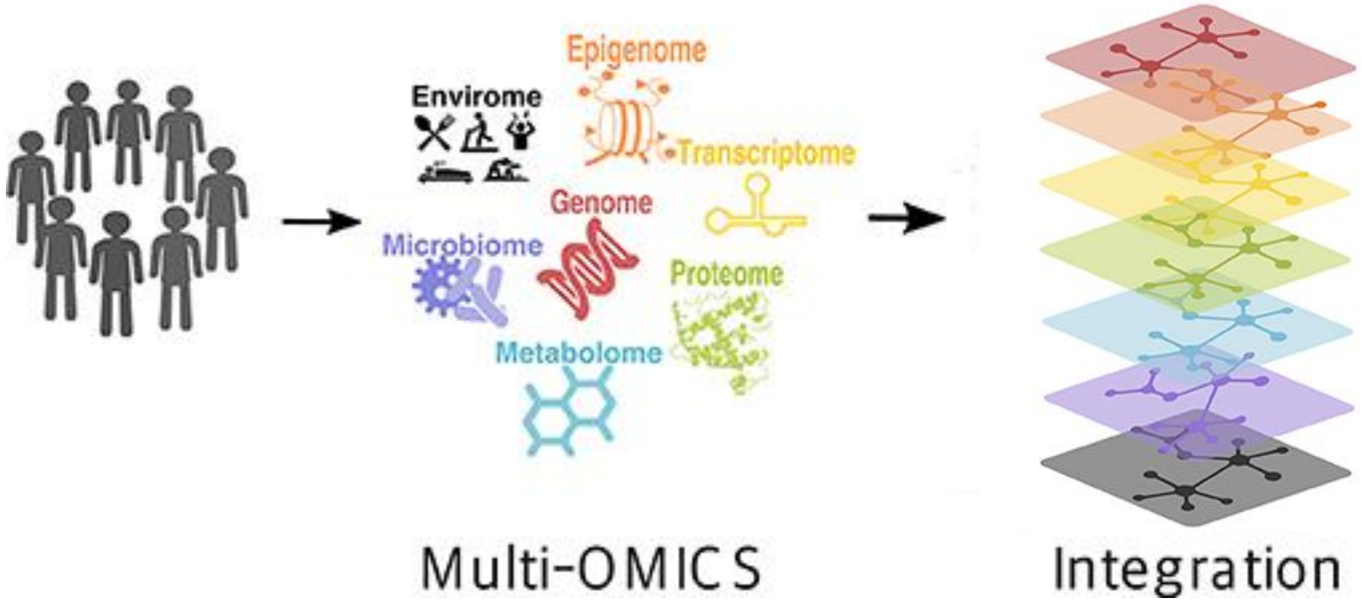


Image: Vilne, B., & Schunkert, H. (2018). Integrating genes affecting coronary artery disease in functional networks by multi-OMICs approach. *Frontiers in cardiovascular medicine*, *5*, 89.

Background and Motivation

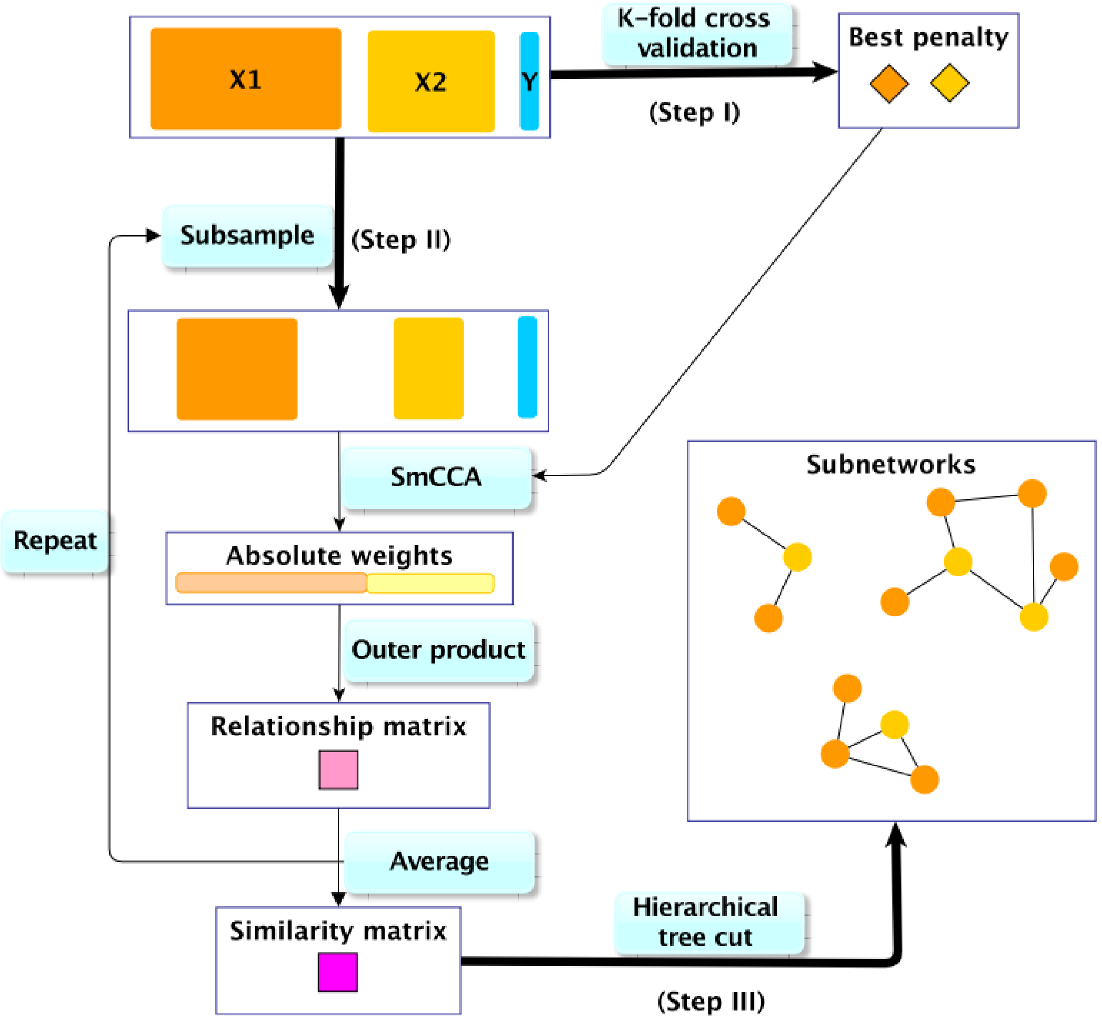


- Supervised methods identify features that are most predictive of the phenotype.
- Inform the integrated feature network with phenotpye.

Sparse multiple Canonical Correlation Network discovery (SmCCNet)

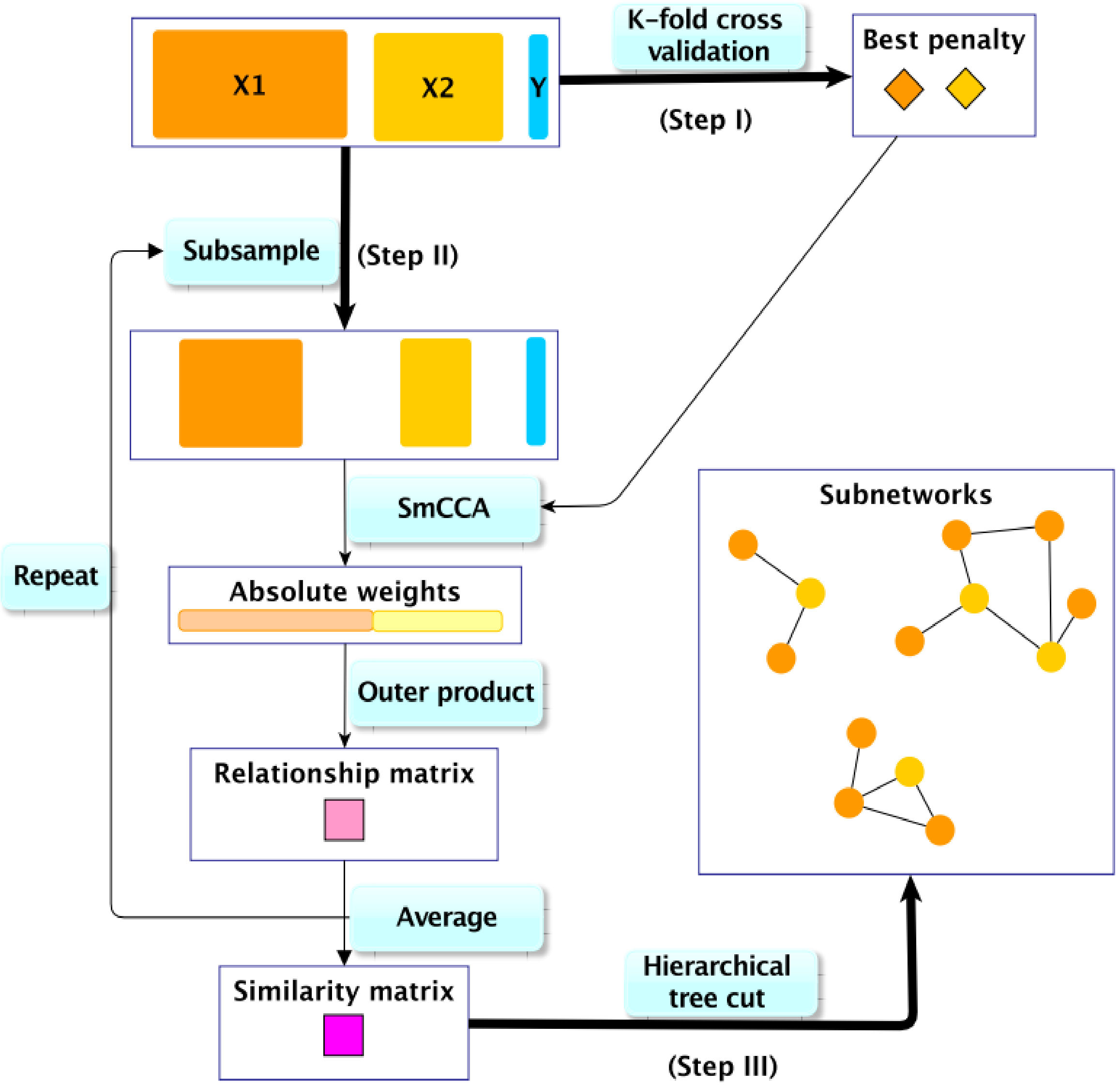
Simultaneously integrates multiple omics profiles and phenotype information to build interpretable network that models the underlying mechanisms.

- Sparsity
- Additional trait
- Module detection



SmCCNet workflow

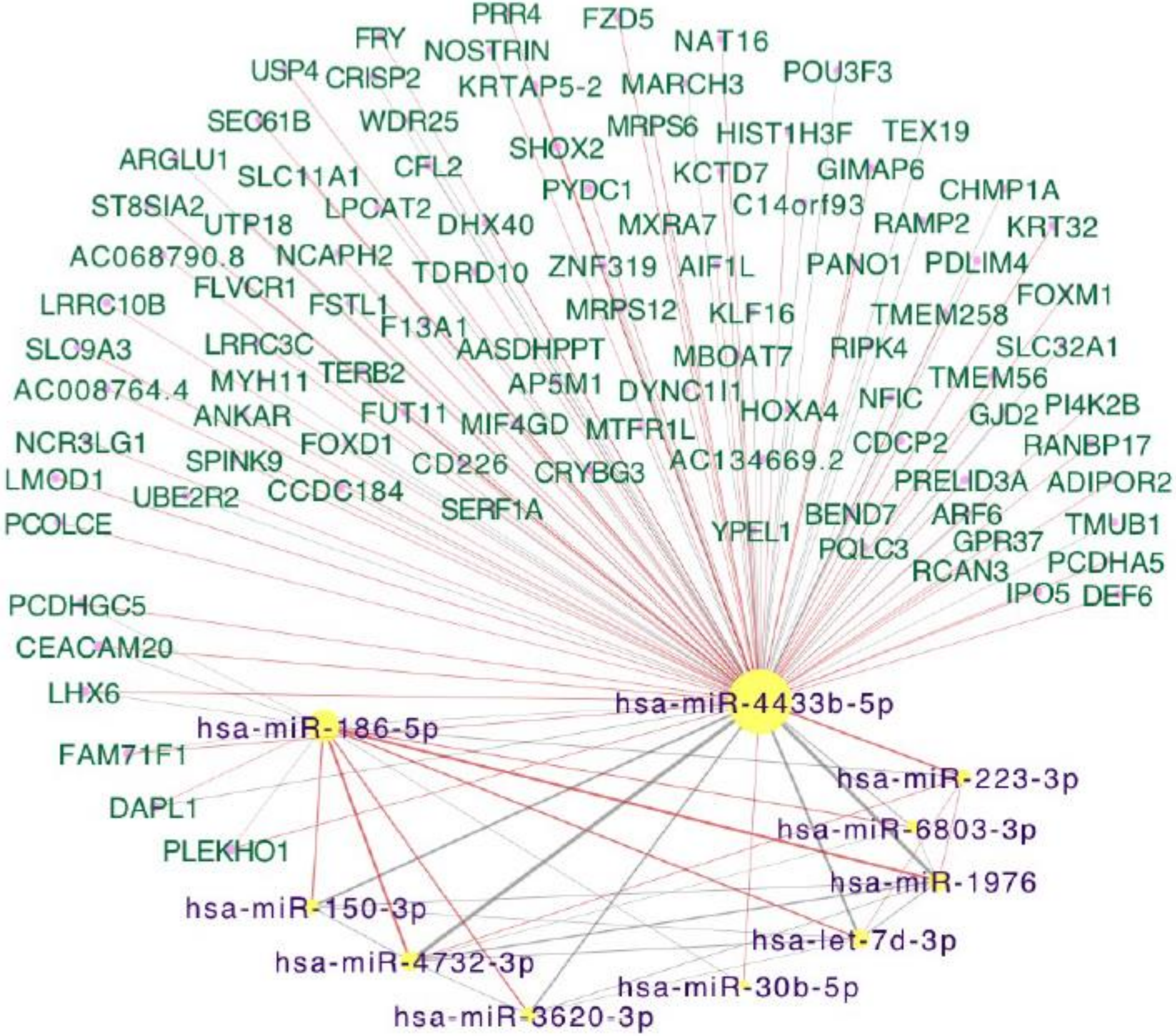
1. Identify the best sparsity penalty parameters.
2. Generate robust canonical weights
   A. Feature subsampling
   B. Relationship matrices
   C. Similarity matrix
3. Hierarchical tree cut
   A. Modules
   B. Edge threshold



Evaluation: COPD data

- 27 subjects (13 controls and 14 cases)
  - 414 miRNAs and 5001 mRNAs
  - Forced expiratory volume during the first second
- 12 connected miRNA-mRNA modules
  - 14,694 negative connections
  - 147 miRNA-mRNA targets have been validated
  - 988 additional targets have been predicted using MultiMir
- SmCCNet identified a higher percentage of predicted and validated miRNA-mRNA pairs than SsCCA

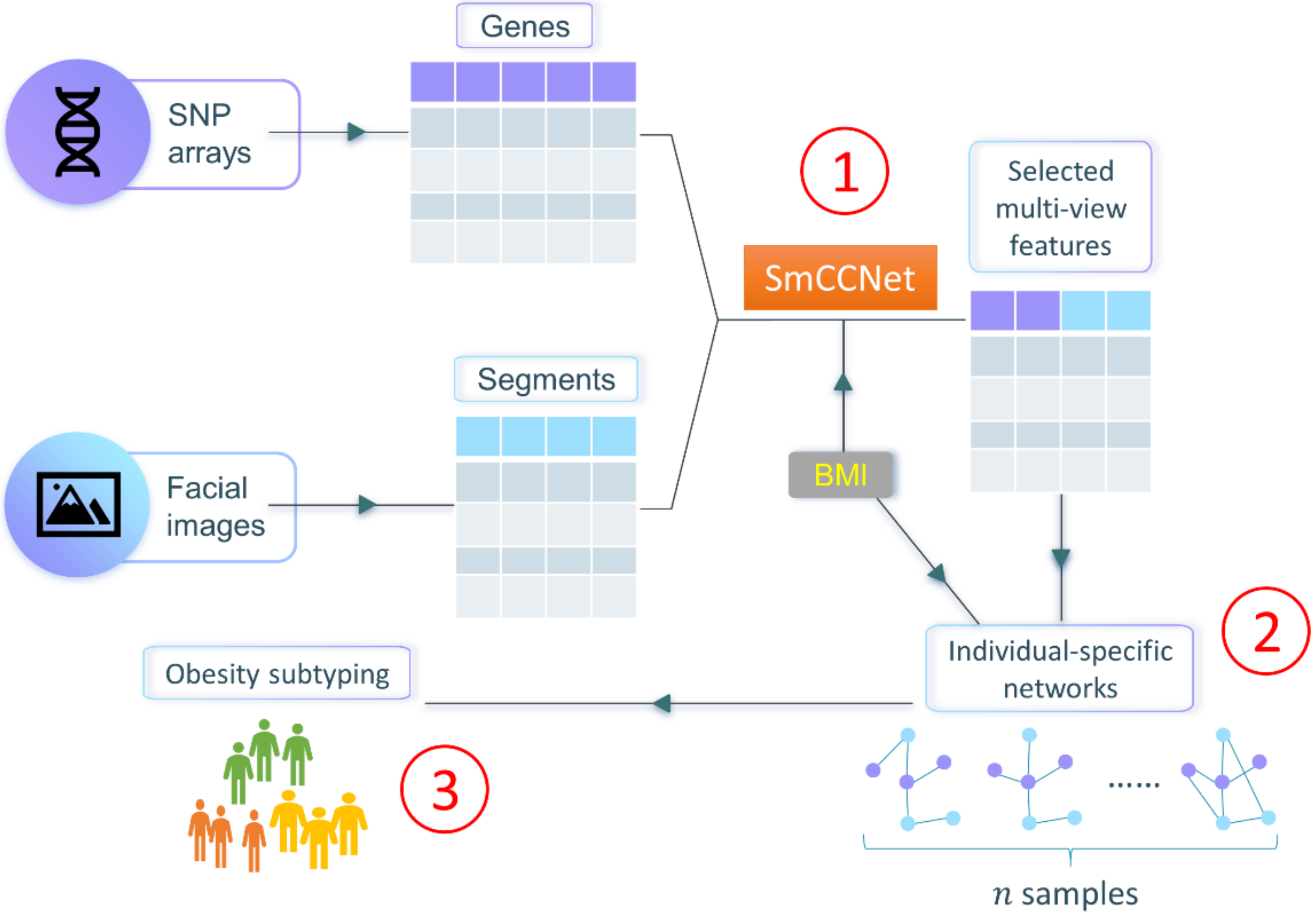
- After edge thresholding, only module 2 remains (10 miRNA and 97 genes)
- miR-4433b-5p is a hub, connected to all 97 genes and 10 miRNAs
    - A biomarker for multidrug-resistance tuberculosis.
- miR-186-5p has been found to be up-regulated in COPD patients.



Applications of SmCCNet

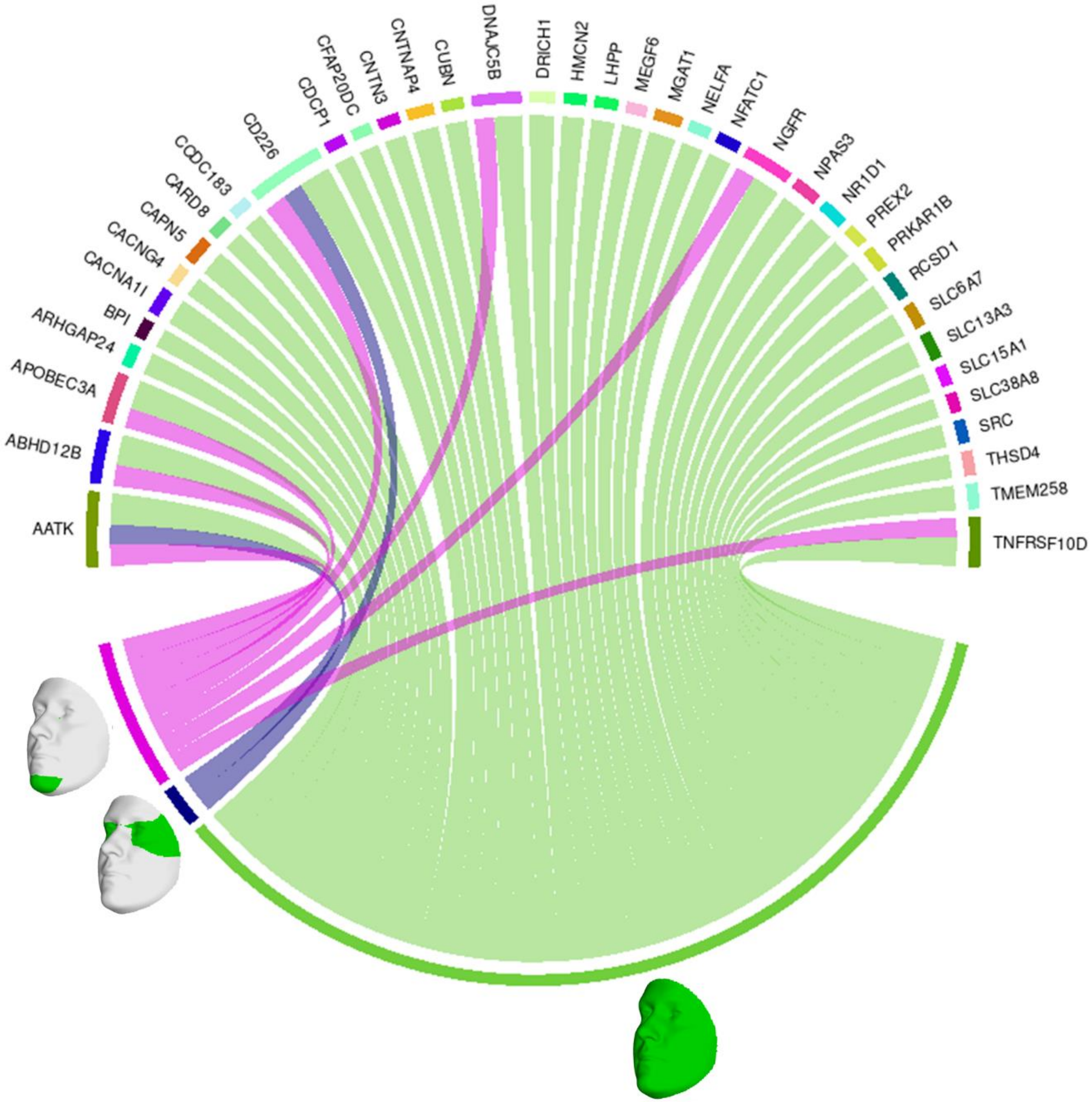
- SmCCNet can effectively construct phenotype-specific multi-omics network and detect informative modules.
- Features contained in such informed modules can be used for subsequent analyses.
- “A novel network-guided multi-view clustering workflow: dissecting genetic and facial heterogeneity” (netMUG)
    - Data: genomics and facial images
    - Trait: BMI
    - Goal: obesity sutyping

netMUG workflow



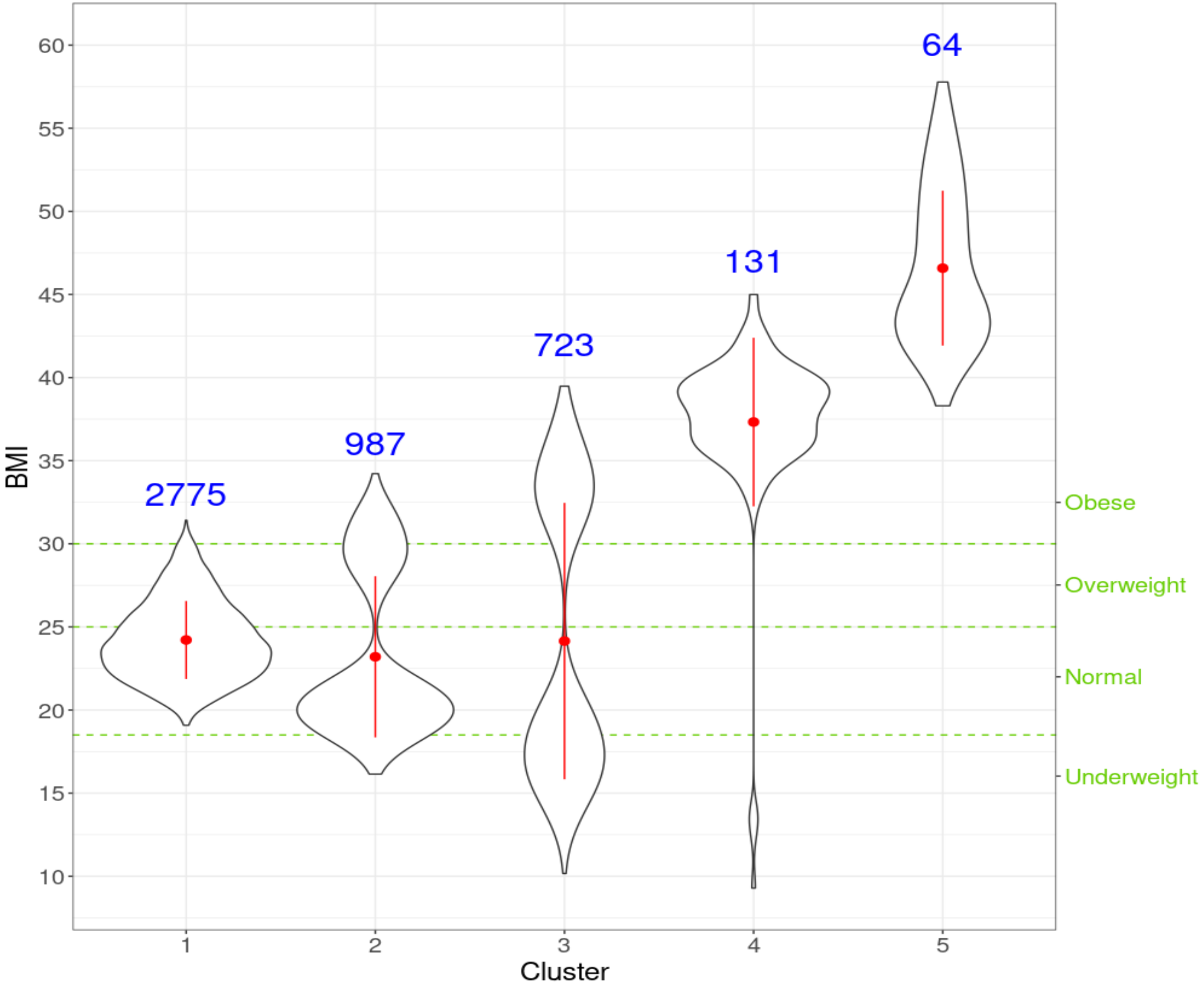
Features selected by SmCCNet

- SmCCNet selected 278 genes and 26 facial segments.
  - 150 genes are also found by GWAS
  - 39 genes are also highlighted in DisGeNET
- Top 1% connections in the SmCCNet network:
  - AATK is known highly associated with BMI
  - APOBEC3A, DNAJC5B and NGFR all affect body height

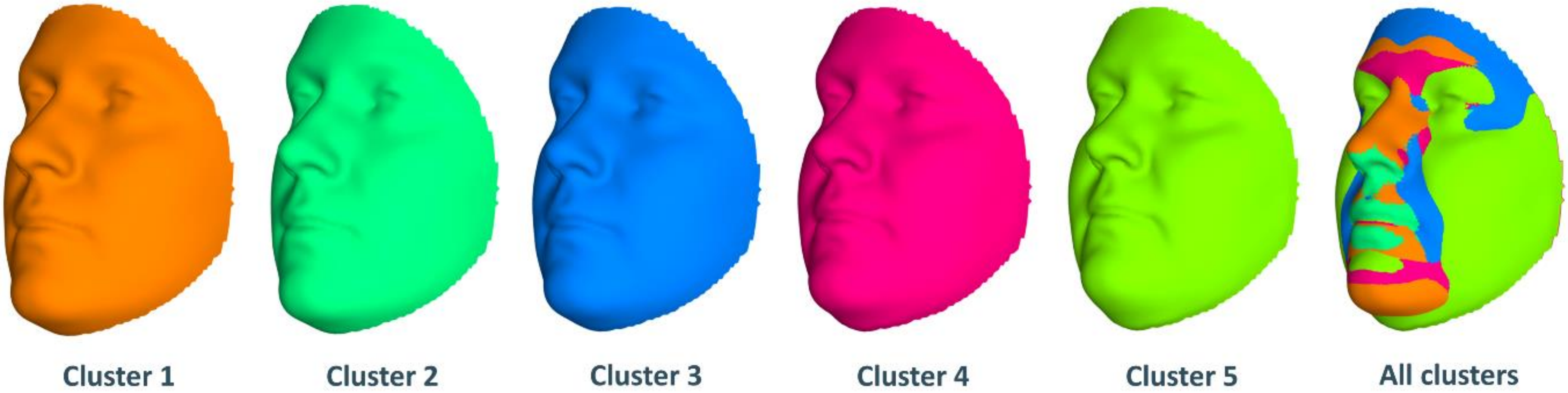


Obesity subtyping

- netMUG found 5 clusters
- The clusters are significantly associated with BMI
- Our subtyping is complementary to the classic BMI categories.



Obesity subtyping



Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5 All clusters

MARIE CURIE ACTIONS

TRANSYS PERSONALIZED MEDICINE

Take-home messages

- PCA is a dimensionality reduction method for a single dataset.
- CCA works with two datasets and maximizes their canonical correlation.
- Sparse CCA prioritizes features most contributing to the correlation.
- Sparse multiple CCA can incorporate more than two datasets.
- SmCCNet constructs phenotype-specific multi-omic network and can be used as feature selector.

Thank you!

Q&A