Network-based Methods for Analysis of Microbiome and Metabolomic Data

Jing Ma

Division of Public Health Sciences
Fred Hutchinson Cancer Research Center

3 Nov 2021

Outline

Outline

Microbiome Science

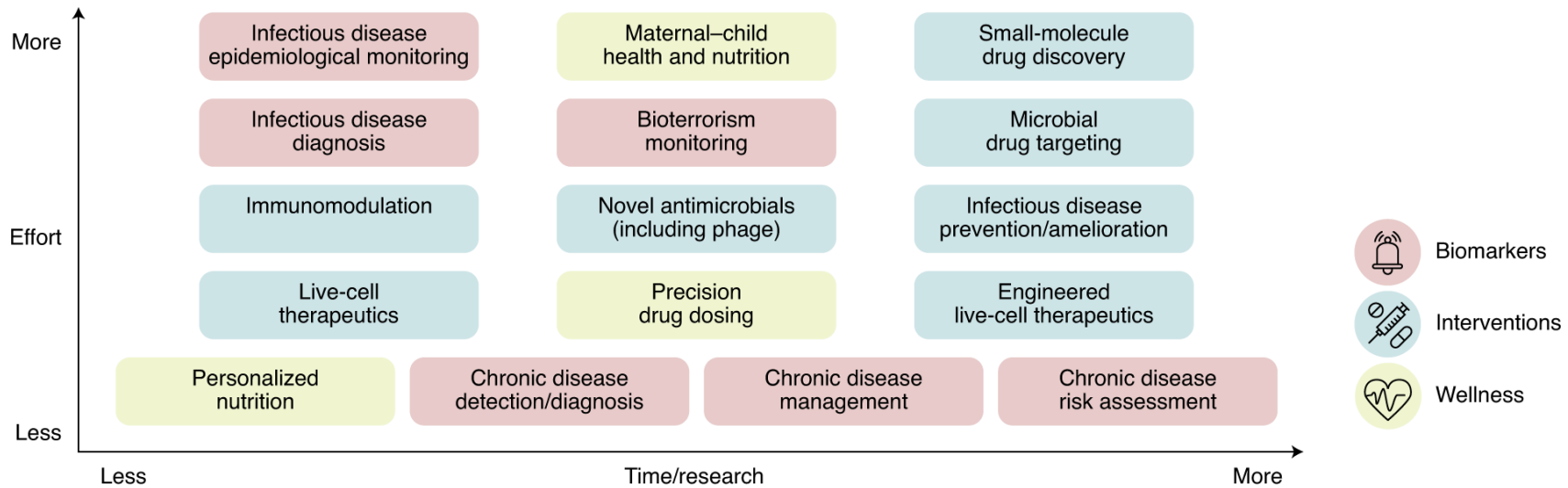
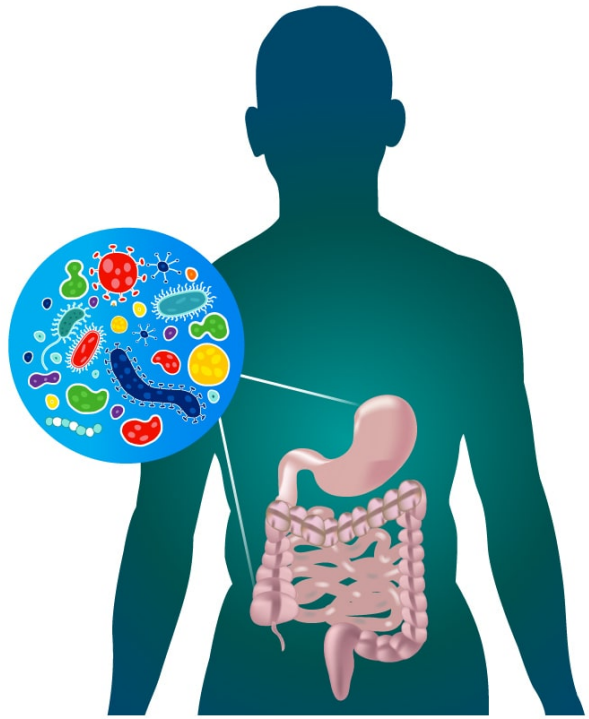


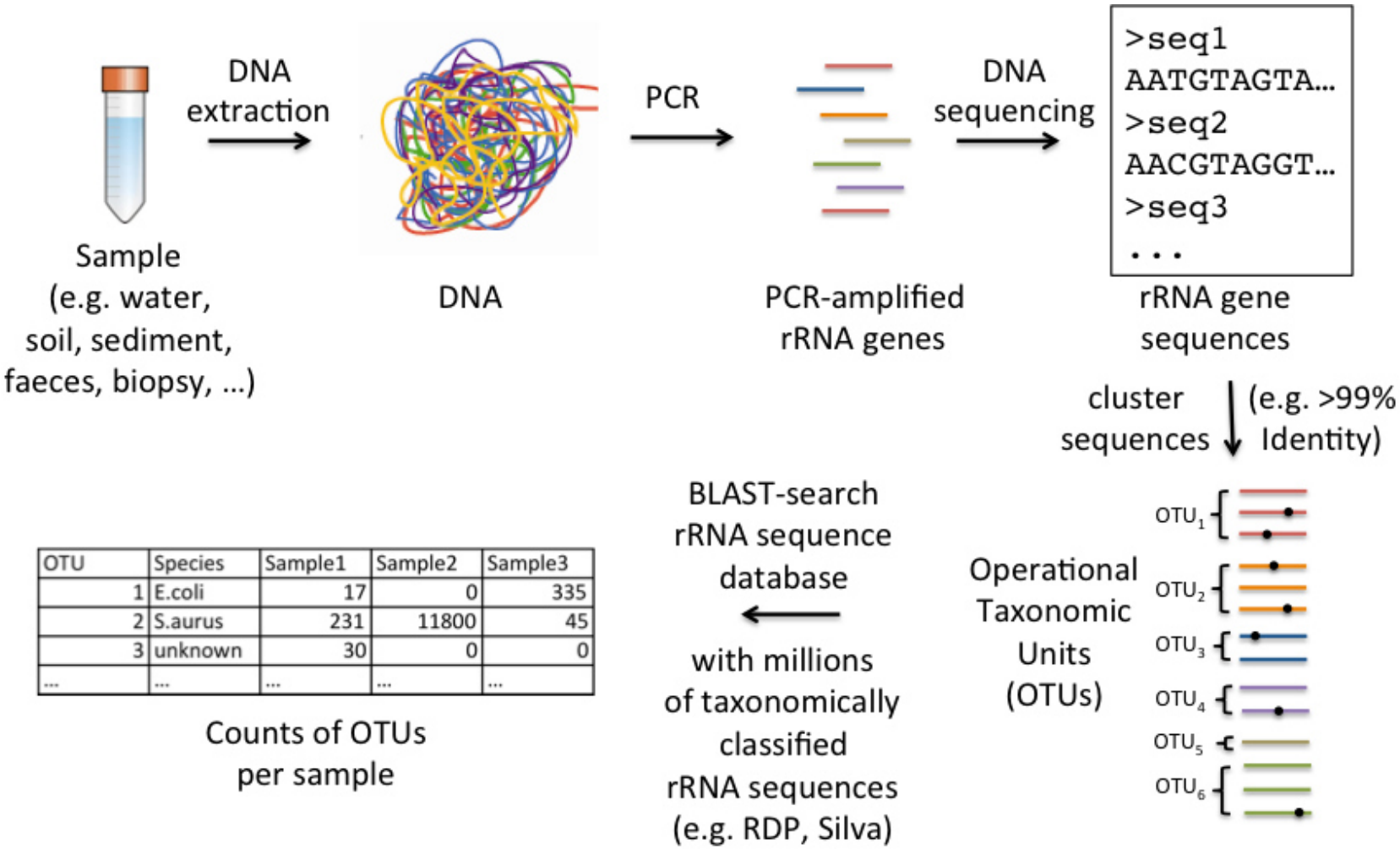
Figure: Plot from Wilkinson et al. (21') Nat Medicine.

The Gut Microbiome



Interest in microorganisms (their genome and products) that live in the digestive tracts of humans and other animals.

Microbiome Data

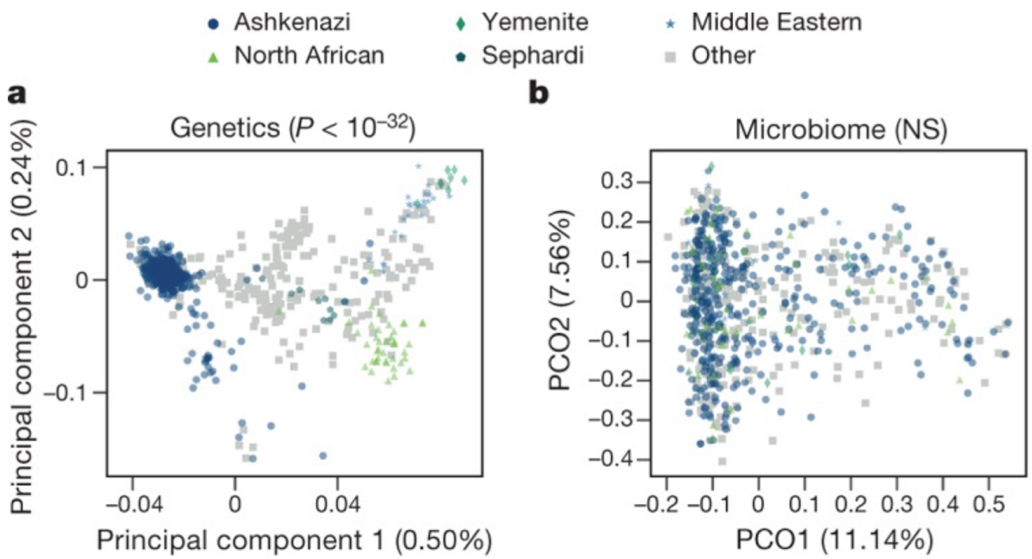


OTU	Species	Sample1	Sample2	Sample3
1	E.coli	17	0	335
2	S.aurus	231	11800	45
3	unknown	30	0	0
...	...	...	...	...

Figure: Plot courtesy to Anders Andersson

Samples are Correlated

- Microbiome compositions are shaped by environment, diet, and culture
- Sample similarity is often available (e.g. UniFrac distance)



D Rothschild *et al. Nature* **555**, 210–215 (2018) doi:10.1038/nature25973

Variables are Correlated

- Taxon similarity can be computed by the patristic distance between tips of the tree

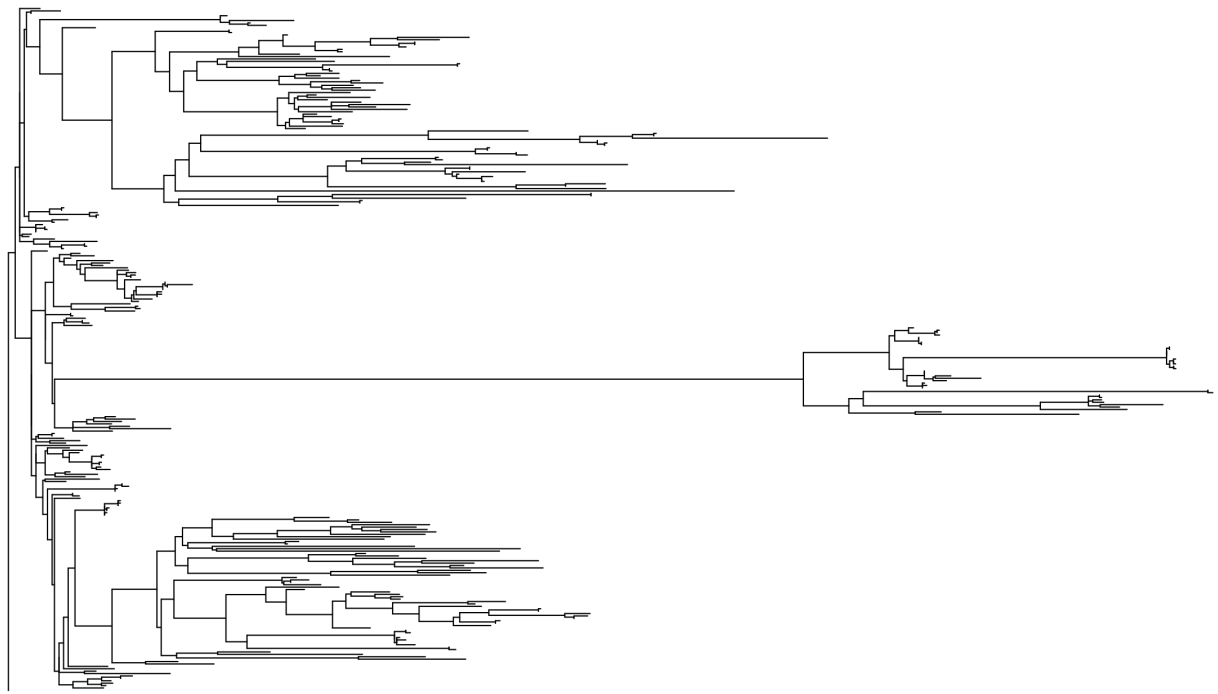


Figure: A phylogenetic tree based on data from Wilmanski et al. (21') Nat Metabolism.

Variables are Correlated

- Taxon similarity can be computed by measuring metabolic functional similarities

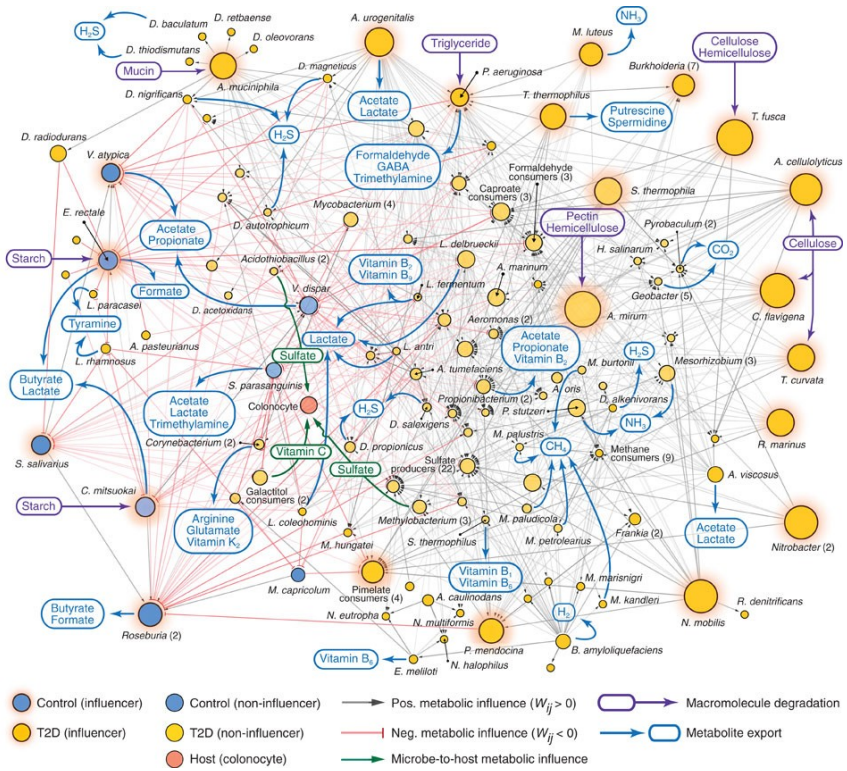
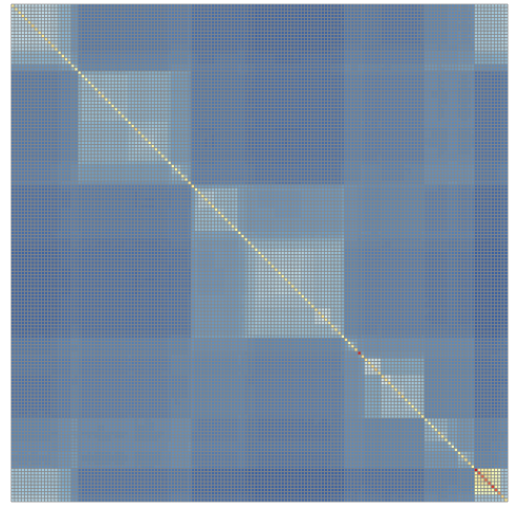
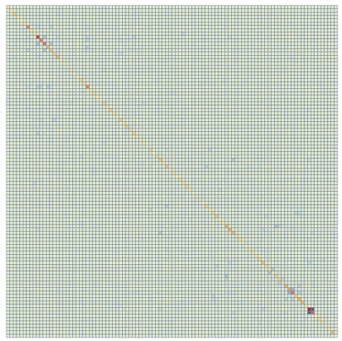


Figure: A metabolic interaction network from Sung et al. (17') Nat Commun.

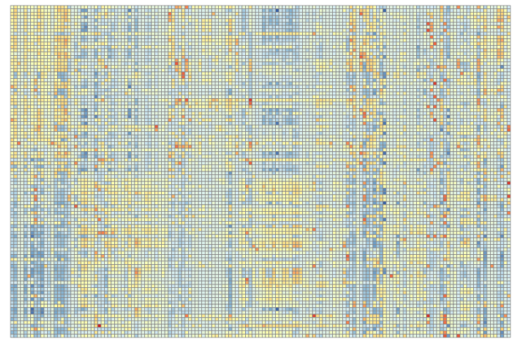
Two-way Structured Data



$Q : p \times p$



$H : n \times n$



$X : n \times p$

Biplot via GMD

- Generalized matrix decomposition (GMD[1]) extends PCA/SVD to two-way structured data

[1] Allen et al. (14') JASA

Biplot via GMD

- Generalized matrix decomposition (GMD[1]) extends PCA/SVD to two-way structured data
- SVD solves

$$\underset{\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}\|_F$$

where $\|A\|_F = \text{trace}(A^{\mathsf{T}}A)$.

[1] Allen et al. (14') JASA

Biplot via GMD

- Generalized matrix decomposition (GMD[1]) extends PCA/SVD to two-way structured data
- SVD solves

$$\underset{\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}\|_F$$

  where $\|A\|_F = \text{trace}(A^{\mathsf{T}}A)$.
- GMD solves

$$\underset{\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}\|_{\boldsymbol{H},\boldsymbol{Q}}$$

  where $\|A\|_{\boldsymbol{H},\boldsymbol{Q}} = \text{trace}(A^{\mathsf{T}}\boldsymbol{H}A\boldsymbol{Q})$.

[1] Allen et al. (14') JASA

GMD Improves Visualization of Samples

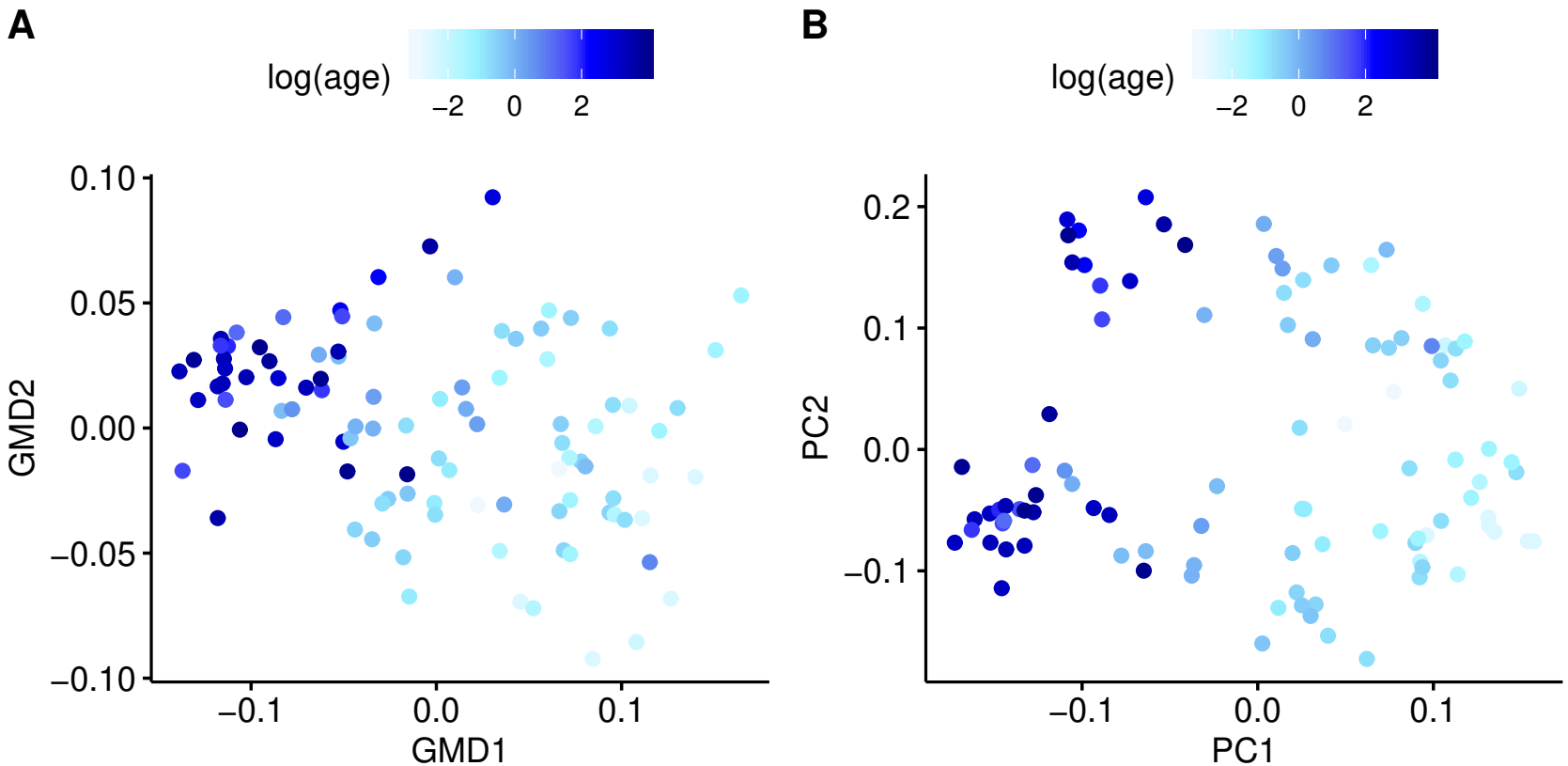
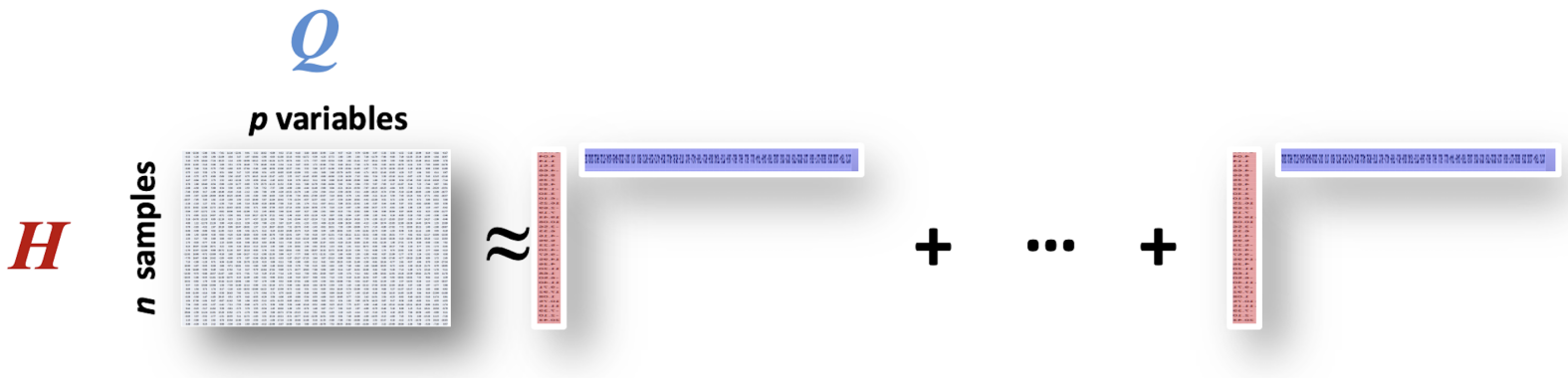


Figure: Plot of samples based on data from Yatsunenko et al. (12') Nature.

GMD Regression



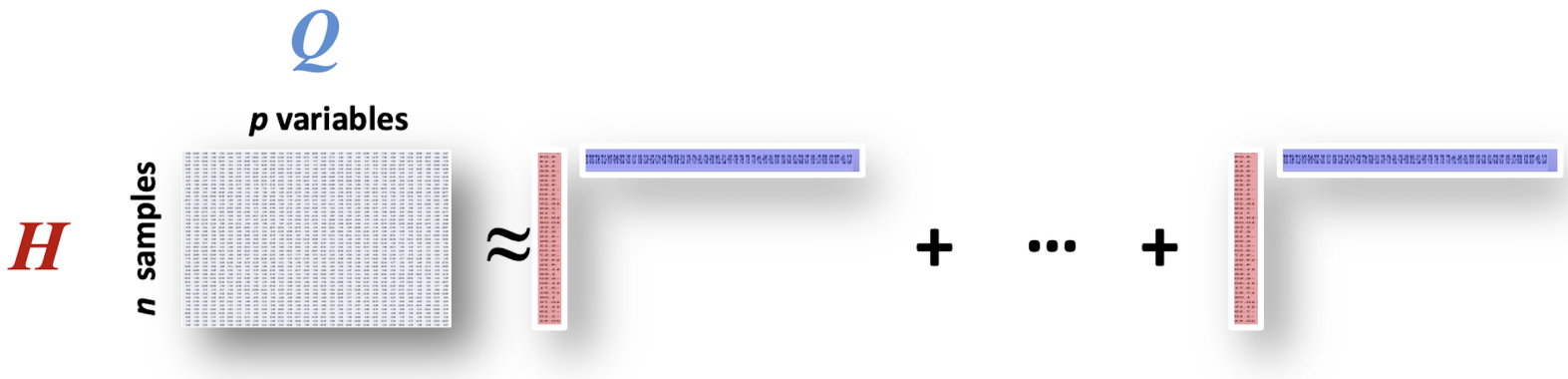
GMD:

$$\boldsymbol{X} \approx \sigma_1 U_1 V_1^{\mathsf{T}} + \cdots + \sigma_K U_K V_K^{\mathsf{T}}$$

Weighted LS:

$$\hat{\gamma} = \text{argmin}_{\gamma} \|y - \sigma_1 U_1 \gamma_1 - \cdots - \sigma_K U_K \gamma_K\|_{\boldsymbol{H}}^2$$

GMD Regression



GMD:

$$\boldsymbol{X} \approx \sigma_1 U_1 V_1^{\mathsf{T}} + \cdots + \sigma_K U_K V_K^{\mathsf{T}}$$

Weighted LS:

$$\hat{\gamma} = \text{argmin}_{\gamma} \|y - \sigma_1 U_1 \gamma_1 - \cdots - \sigma_K U_K \gamma_K\|_{\boldsymbol{H}}^2$$

Remark: rank is selected based on variable importance and generalized cross-validation.

GMD Regression

GMDR estimator:

$$\hat{\boldsymbol{\beta}}_{GMDR} = \boldsymbol{Q}\mathcal{V}\hat{\boldsymbol{\gamma}}$$

[2] Randolph et al. (18') AOAS

GMD Regression

GMDR estimator:

$$\hat{\boldsymbol{\beta}}_{GMDR} = \boldsymbol{Q}\mathcal{V}\hat{\boldsymbol{\gamma}}$$

KPR estimator[2]:

$$\hat{\boldsymbol{\beta}}_{KPR} = \arg\min_{\boldsymbol{\beta}} \{\|y - \boldsymbol{X}\boldsymbol{\beta}\|^2_{\boldsymbol{H}} + \lambda\|\boldsymbol{\beta}\|^2_{\boldsymbol{Q}^{-1}}\}$$

[2] Randolph et al. (18') AOAS

GMD Regression

GMDR estimator:

$$\hat{\boldsymbol{\beta}}_{GMDR} = \boldsymbol{Q}\mathcal{V}\hat{\boldsymbol{\gamma}}$$

KPR estimator[2]:

$$\hat{\boldsymbol{\beta}}_{KPR} = \arg\min_{\boldsymbol{\beta}}\{\|y - \boldsymbol{X}\boldsymbol{\beta}\|^2_{\boldsymbol{H}} + \lambda\|\boldsymbol{\beta}\|^2_{\boldsymbol{Q}^{-1}}\}$$

The GMD class of estimators

$$\hat{\boldsymbol{\beta}}^{\mathcal{W}} = \boldsymbol{Q}\mathcal{V}\mathcal{W}\mathcal{S}^{-1}\mathcal{U}^{\mathsf{T}}\boldsymbol{H}y,$$

for a diagonal weight matrix $\mathcal{W} = \text{diag}(w_1, \ldots, w_K)$, $w_j \geq 0$.

- $w_j = 1_{j\in\mathcal{J}} \rightarrow \hat{\boldsymbol{\beta}}_{GMDR}$
- $w_j = \sigma_j^2(\sigma_j^2 + \lambda)^{-1} \rightarrow \hat{\boldsymbol{\beta}}_{KPR}$

[2] Randolph et al. (18') AOAS

GMD Inference

GMD Inference

- Find the estimation bias

$$B_j = \mathbb{E}[\hat{\beta}_j^{\mathcal{W}} - \beta_j^*] = (\boldsymbol{Q}\mathcal{V}\mathcal{W}\mathcal{V}^\intercal\boldsymbol{\beta}^*)_j - \beta_j^*$$

GMD Inference

- Find the estimation bias

$$B_j = \mathbb{E}[\hat{\beta}_j^{\mathcal{W}} - \beta_j^*] = (\boldsymbol{Q}\mathcal{V}\mathcal{W}\mathcal{V}^\intercal\boldsymbol{\beta}^*)_j - \beta_j^*$$

- Correct the estimation bias via an initial estimator $\hat{\boldsymbol{\beta}}^{init} = D\tilde{\boldsymbol{\beta}}(\lambda)$ where

$$\boldsymbol{Q} = D\Delta D^\intercal, \quad \tilde{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}}\{\|y - \boldsymbol{X}D\boldsymbol{\beta}\|_{\boldsymbol{H}}^2 + \lambda\|\Delta^{-1/2}\boldsymbol{\beta}\|_1\}.$$

GMD Inference

- Find the estimation bias

$$B_j = \mathbb{E}[\hat{\beta}_j^{\mathcal{W}} - \beta_j^*] = (\boldsymbol{Q}\mathcal{V}\mathcal{W}\mathcal{V}^{\mathsf{T}}\boldsymbol{\beta}^*)_j - \beta_j^*$$

- Correct the estimation bias via an initial estimator $\hat{\boldsymbol{\beta}}^{init} = D\tilde{\boldsymbol{\beta}}(\lambda)$ where

$$\boldsymbol{Q} = D\Delta D^{\mathsf{T}}, \quad \tilde{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}}\{\|y - \boldsymbol{X}D\boldsymbol{\beta}\|_{\boldsymbol{H}}^2 + \lambda\|\Delta^{-1/2}\boldsymbol{\beta}\|_1\}.$$

- Obtain inference for $H_{0,j} : \beta_j^* = 0$ using de-biased $\tilde{\beta}_j^{\mathcal{W}} = \hat{\beta}_j^{\mathcal{W}} - \hat{B}_j$.

Model Assumptions

Linear model: $y = \boldsymbol{X}\beta^* + \varepsilon$ with $\mathrm{Cov}(\varepsilon \mid \boldsymbol{X}) = \Psi = L_\Psi^\top L_\Psi$. There exists sub-Gaussian $\tilde{\varepsilon}$ such that $\varepsilon = L_\Psi^\top \tilde{\varepsilon}$.

- $\|L_\Psi \boldsymbol{H} L_\Psi^\top - I_n\|_2 = o(1)$ as $n \to \infty$.
- β^* is $\boldsymbol{Q}$-smooth:

$$\|\boldsymbol{Q}^{-1/2}\beta^*\|_0 = o\{(n/\log p)^r\} \text{ for } r \in (0, 1/2).$$

- A compatibility assumption w.r.t $\boldsymbol{Q}$ and $\boldsymbol{H}$

$$0 < \underline{c} \leq \frac{\|\widetilde{\boldsymbol{X}}_A v\|^2}{n\|v\|^2} \leq \bar{c} < \infty,$$

where $\widetilde{\boldsymbol{X}} = \boldsymbol{H}^{1/2}\boldsymbol{X}\boldsymbol{Q}^{1/2}$ and $|A| \geq M_1^* s_0 + 1$.

Microbiome Data from Yatsunenko et al. (12)

Data

- $\boldsymbol{X}$: 100 healthy individuals from Venezuela, Malawi & US $\times$ 149 OTUs
- $\boldsymbol{Q}$: variable similarity from the phylogenetic tree
- $\boldsymbol{H}$: sample similarity from functional profiling
- y: age (proxy to aging)

Questions

Predict age from microbiome data and identify age-associated bacteria.

Age Prediction

- Lasso
- Ridge regression
- KPR1/GMDR1: only $\boldsymbol{Q}$
- KPR2/GMDR2: only $\boldsymbol{H}$
- KPR3/GMDR3: both $\boldsymbol{Q}$ and $\boldsymbol{H}$

Leave-one-out cross-validation:

Method	Lasso	Ridge	KPR1	GMDR1	KPR3	GMDR3
MSE	0.71	1.13	0.61	0.57	0.56	0.52

Significant Marginal Associations

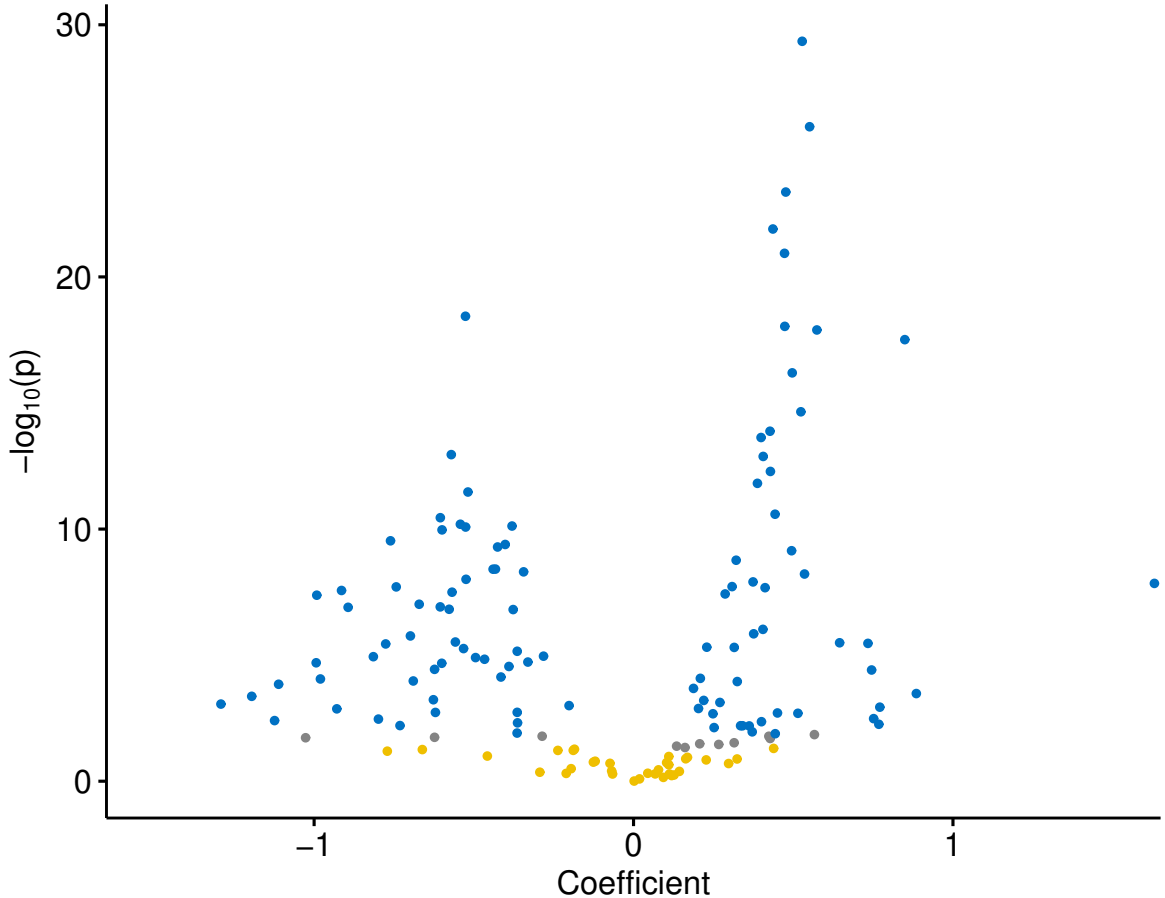


Figure: Univariate association analysis using data in Yatsunenko et al. Nature. (12') (FDR=0.1). Results suggest that either the true signals are not sparse or taxa are highly correlated.

Significant Conditional Associations

Table: Genera found to be associated with age when controlling for FDR=0.1

	Genus	Total
Ridge	(none)	0
LDPE	*Atopobium*, *Campylobacter*, *Peptoniphilus*	3
Grace	Atopobium, Campylobacter	2
GMDI	*Acidaminococcus*, *Aggregatibacter*, *Akkermansia*, *Alkalimonas*, *Anaerofustis*, *Anaerostipes*, *Arthrobacter*, *Atopobium*, *Bilophila*, *Brochothrix*, *Bulleidia*, *Butyrivibrio*, *Caloramator*, *Campylobacter*, *Citrobacter*, *Coprobacillus*, *Desulfitobacterium*, *Enterobacter*, *Epulopiscium*, *Facklamia*, *Lachnobacterium*, *Lactobacillus*, *Leclercia*, *Megasphaera*, *Methylophaga*, *Mitsuokella*, *Peptococcus*, *Peptoniphilus*, *Roseburia*, *Ruminococcus*, *Rummeliibacillus*, *Slackia*, *Sphingobacterium*, *Sphingomonas*, *Streptococcus*, *Thiomonas*, *Xenorhabdus*	37

Outline

Chronic Kidney Disease

- An estimated 37 million adults (15% of the population) in the US have CKD[3].
- CKD leads to marked metabolic changes that can be studied via lipidomics.

[3] Johansen et al. (21') American Journal of Kidney Diseases

Chronic Kidney Disease

- An estimated 37 million adults (15% of the population) in the US have CKD[3].
- CKD leads to marked metabolic changes that can be studied via lipidomics.

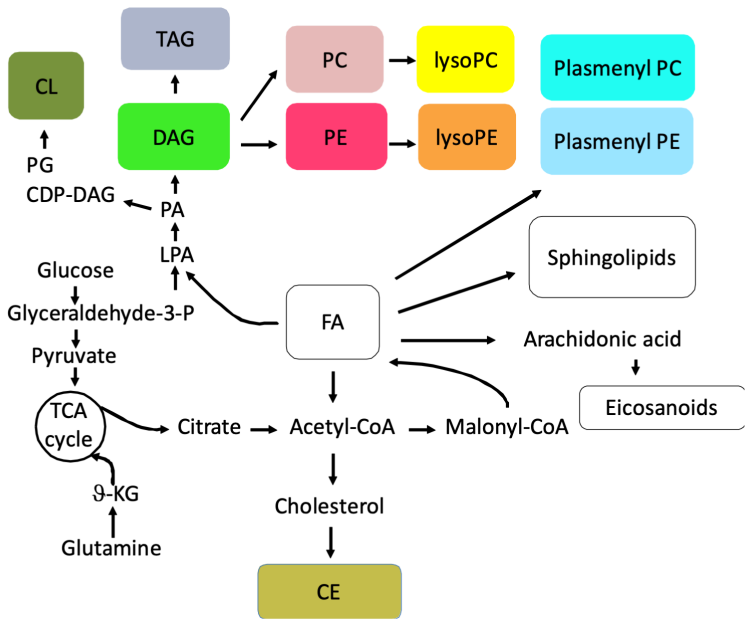


Figure: Overview of lipid biosynthesis from Ma et al. (19') Bioinformatics.

[3] Johansen et al. (21') American Journal of Kidney Diseases

Data and Study Population

Study cohorts

- Clinical Phenotyping Resource and Biobank Core (CPROBE[4]): 135 early vs. 79 advanced.
- Chronic Renal Insufficiency Cohort (CRIC[5]): 121 non-progressors vs. 79 progressors. Baseline only.

Objective

To identify lipidomic signatures of CKD progression.

[5] Afshinnia et al. (18') J. Am. Soc. Nephrol.
[5] Afshinnia et al. (16') Kidney Int. Rep.

Lipids are Correlated!

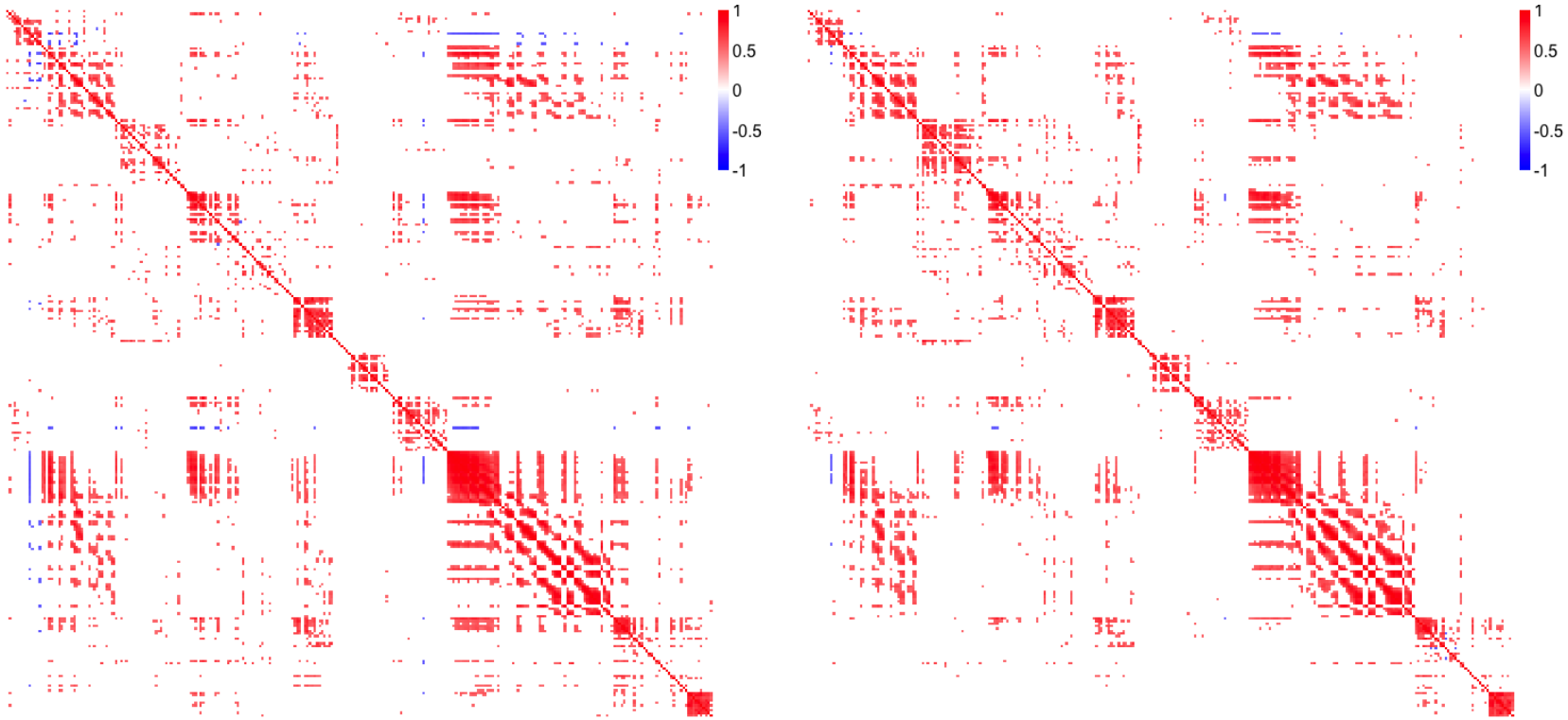


Figure: Heatmap of empirical correlations in non-progressors (left) and progressors (right). Correlation magnitudes are thresholded at 0.5.

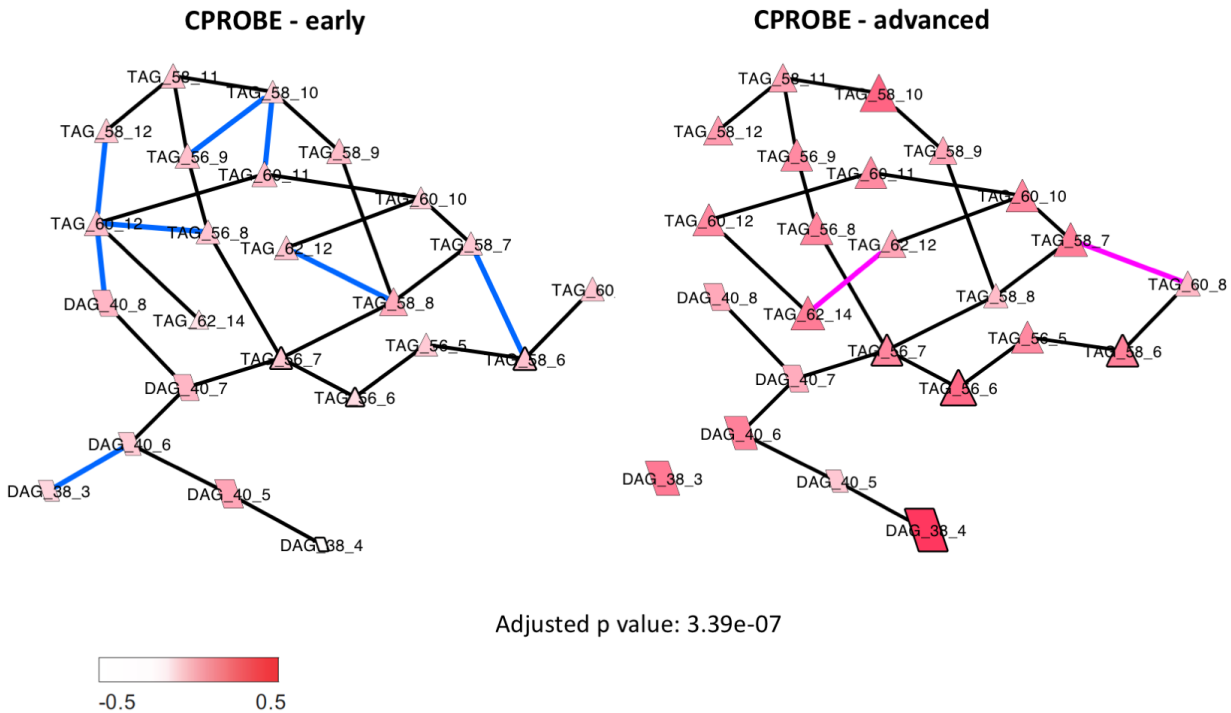
Differential Network Enrichment Analysis

Unlike genetic and genomic data, no *a priori* defined lipid pathways.

We propose, DNEA[6], a data-driven approach for subnetwork discovery and enrichment analysis.

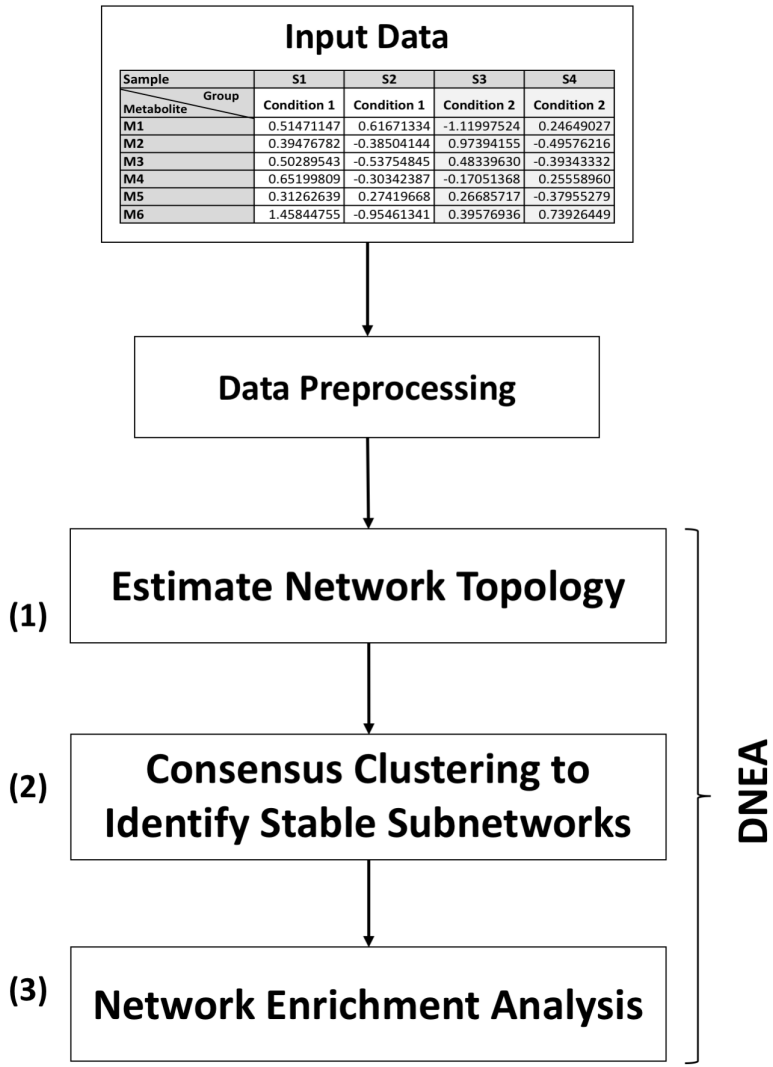
[6] Ma et al. (19') Bioinformatics.

Novel Lipid Signatures in CKD



- Higher abundance of triacylglycerol (TAG) in advanced CKD
- Fewer edges were associated with advanced CKD

Differential Network Enrichment Analysis



Step I: Estimate Network Topology

We want two lipid partial correlation networks for early and advanced stage, respectively.

[7] Guo et al. (11') Biometrika

Step I: Estimate Network Topology

We want two lipid partial correlation networks for early and advanced stage, respectively.

☹ Sample size is smaller than the number of features (285 lipids).

[7] Guo et al. (11') Biometrika

Step I: Estimate Network Topology

We want two lipid partial correlation networks for early and advanced stage, respectively.

☹ Sample size is smaller than the number of features (285 lipids).

☺ Penalized joint estimation strategy to estimate sparse networks[7]:

$$\min_{\Theta_1,\Theta_2} \sum_{k=1}^{K=2} \{\text{trace}(\hat{\Sigma}_k \Theta_k) - \log\det(\Theta_k)\} + P_\lambda(\Theta_1, \Theta_2).$$

[7] Guo et al. (11') Biometrika

Step I: Estimate Network Topology

We want two lipid partial correlation networks for early and advanced stage, respectively.

☹ Sample size is smaller than the number of features (285 lipids).

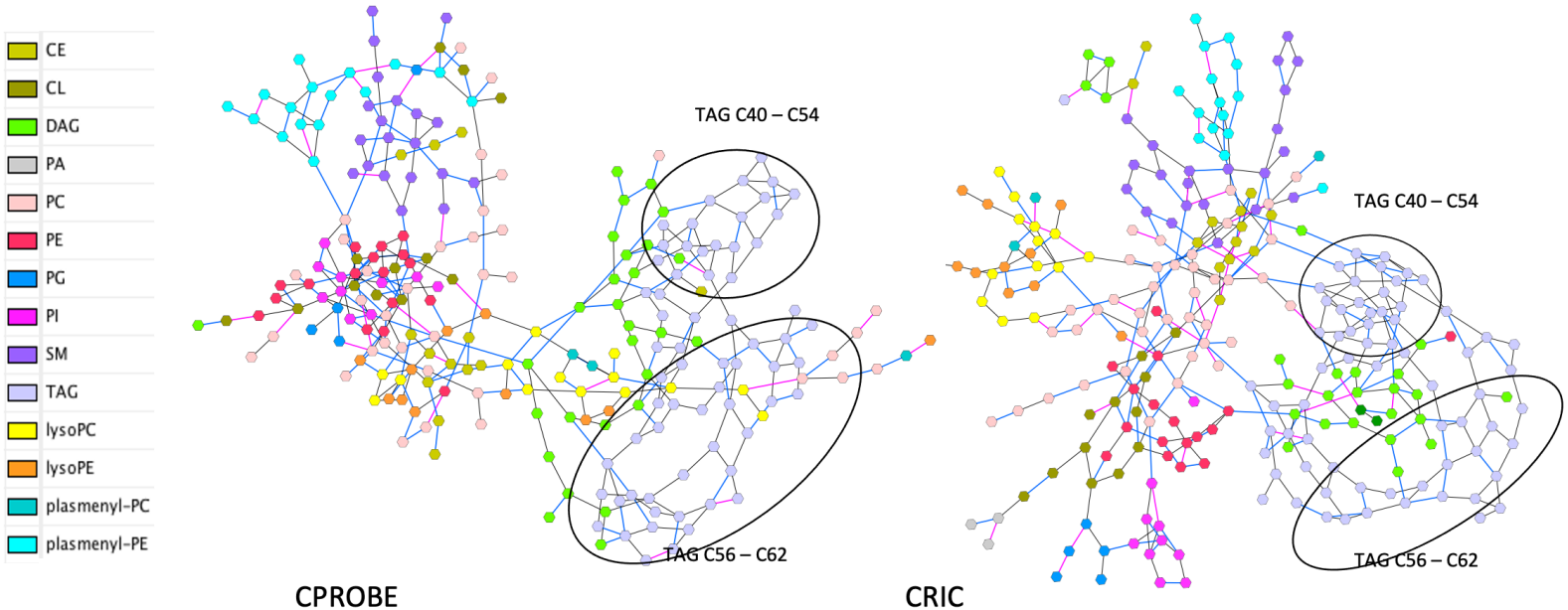
☺ Penalized joint estimation strategy to estimate sparse networks[7]:

$$\min_{\Theta_1,\Theta_2} \sum_{k=1}^{K=2} \{\text{trace}(\hat{\Sigma}_k \Theta_k) - \log\det(\Theta_k)\} + P_\lambda(\Theta_1, \Theta_2).$$

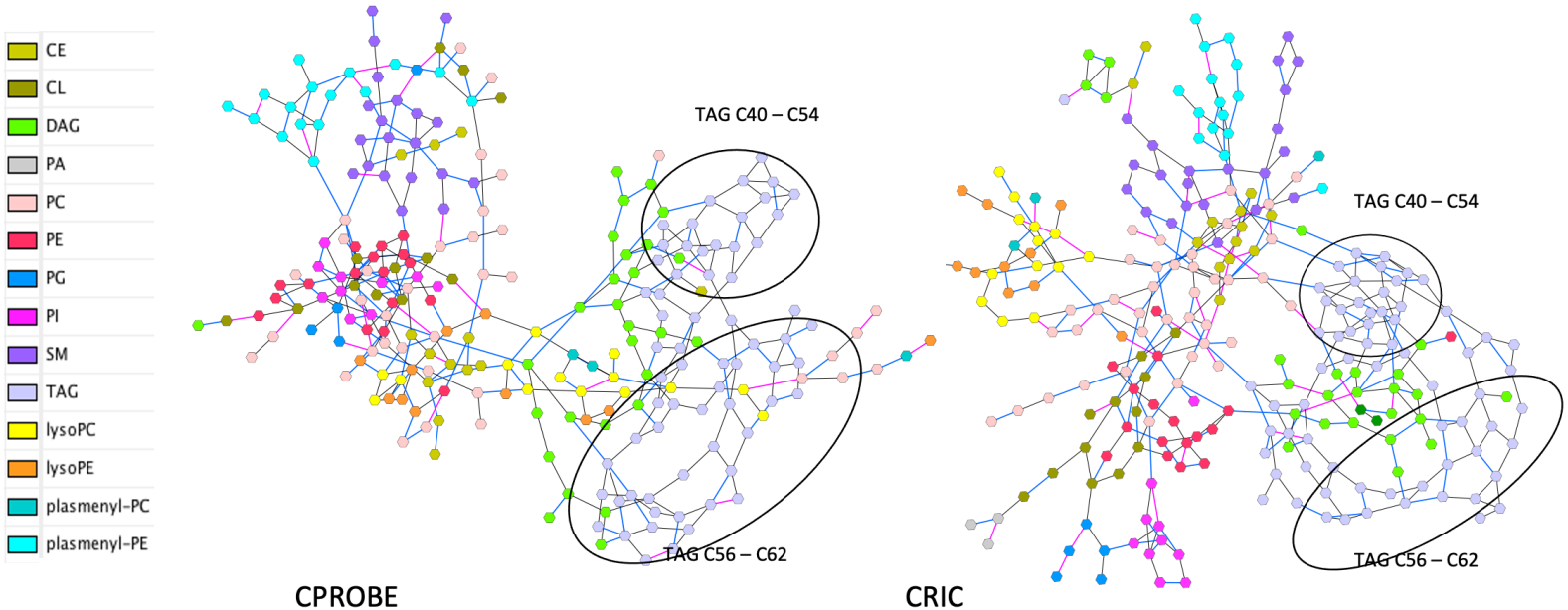
Model selection is done via stability selection.

[7] Guo et al. (11') Biometrika

Step II: Consensus Clustering



Step II: Consensus Clustering



Apply consensus clustering to extract stable subnetworks (proxy to pathways)

Step III: Network Enrichment Analysis

Apply NetGSA[8] to detect enriched subnetworks.

[9] Ma et al. (16') Bioinformatics

[9] Hellstern et al. (21') PLoS Comp Bio; Yue et al. (21') Genet Epidemiol

Step III: Network Enrichment Analysis

Apply NetGSA[8] to detect enriched subnetworks.

$$
\begin{aligned}
Y_j^{(1)} &= \Lambda^{(1)}\mu^{(1)} + \Lambda^{(1)}\gamma_j + \epsilon_j, \quad (j = 1, \ldots, n_1) \\
Y_j^{(2)} &= \Lambda^{(2)}\mu^{(1)} + \Lambda^{(2)}\gamma_j + \epsilon_j, \quad (j = n_1 + 1, \ldots, n)
\end{aligned}
$$

Obtain significance for each subnetwork $S \subset \{1, \ldots, p\}$ by testing the null

$$H_{0,S} : \mathbf{1}^\top \Lambda_{S,S}^{(1)} \mu_S^{(1)} = \mathbf{1}^\top \Lambda_{S,S}^{(2)} \mu_S^{(2)}$$

[9] Ma et al. (16') Bioinformatics

[9] Hellstern et al. (21') PLoS Comp Bio; Yue et al. (21') Genet Epidemiol

Step III: Network Enrichment Analysis

Apply NetGSA[8] to detect enriched subnetworks.

$$Y_j^{(1)} = \Lambda^{(1)}\mu^{(1)} + \Lambda^{(1)}\gamma_j + \epsilon_j, \quad (j = 1, \ldots, n_1)$$
$$Y_j^{(2)} = \Lambda^{(2)}\mu^{(1)} + \Lambda^{(2)}\gamma_j + \epsilon_j, \quad (j = n_1 + 1, \ldots, n)$$

Obtain significance for each subnetwork $S \subset \{1, \ldots, p\}$ by testing the null

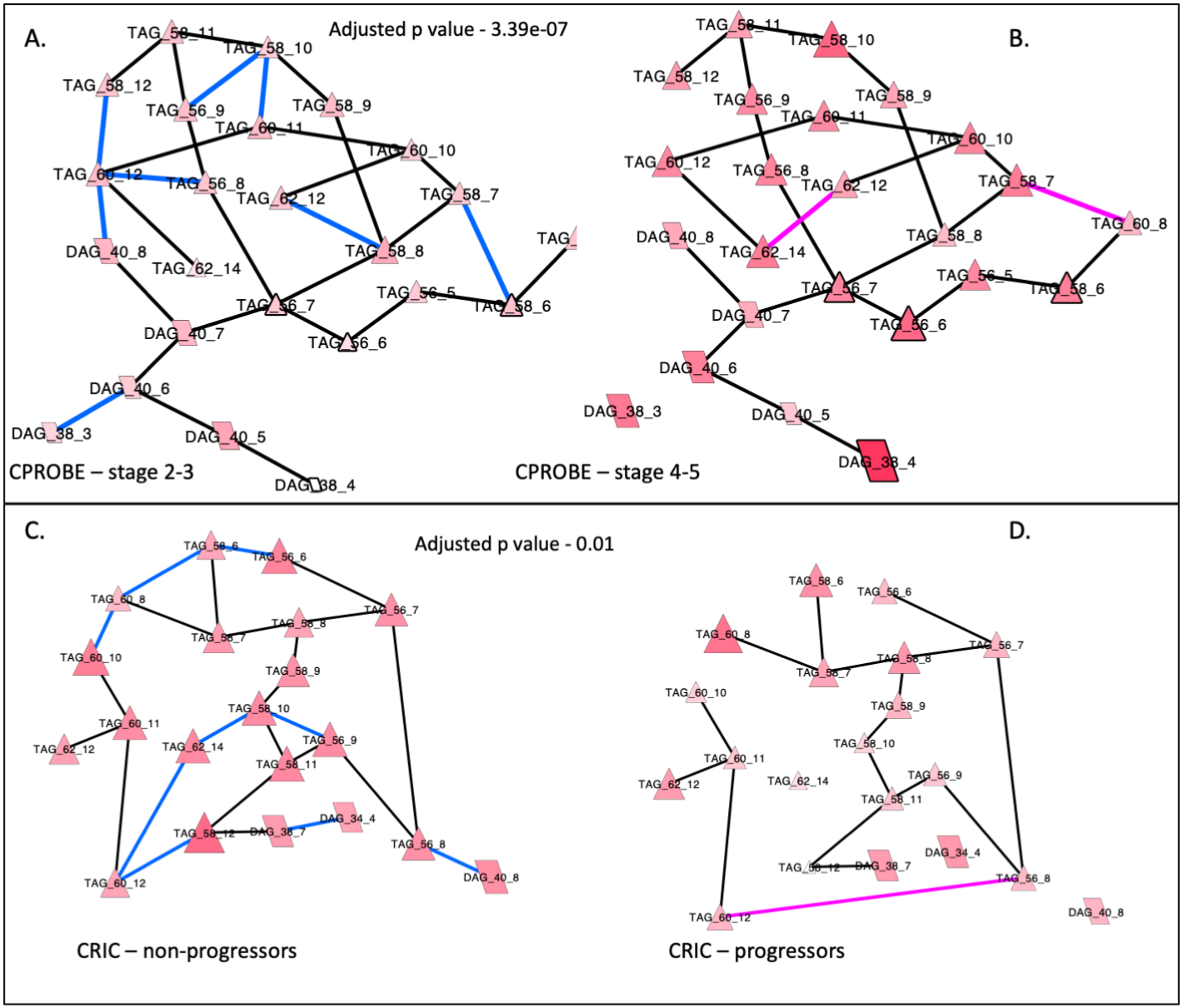
$$H_{0,S} : \mathbf{1}^\top \Lambda_{S,S}^{(1)} \mu_S^{(1)} = \mathbf{1}^\top \Lambda_{S,S}^{(2)} \mu_S^{(2)}$$

NetGSA is now scalable to large datasets[9].

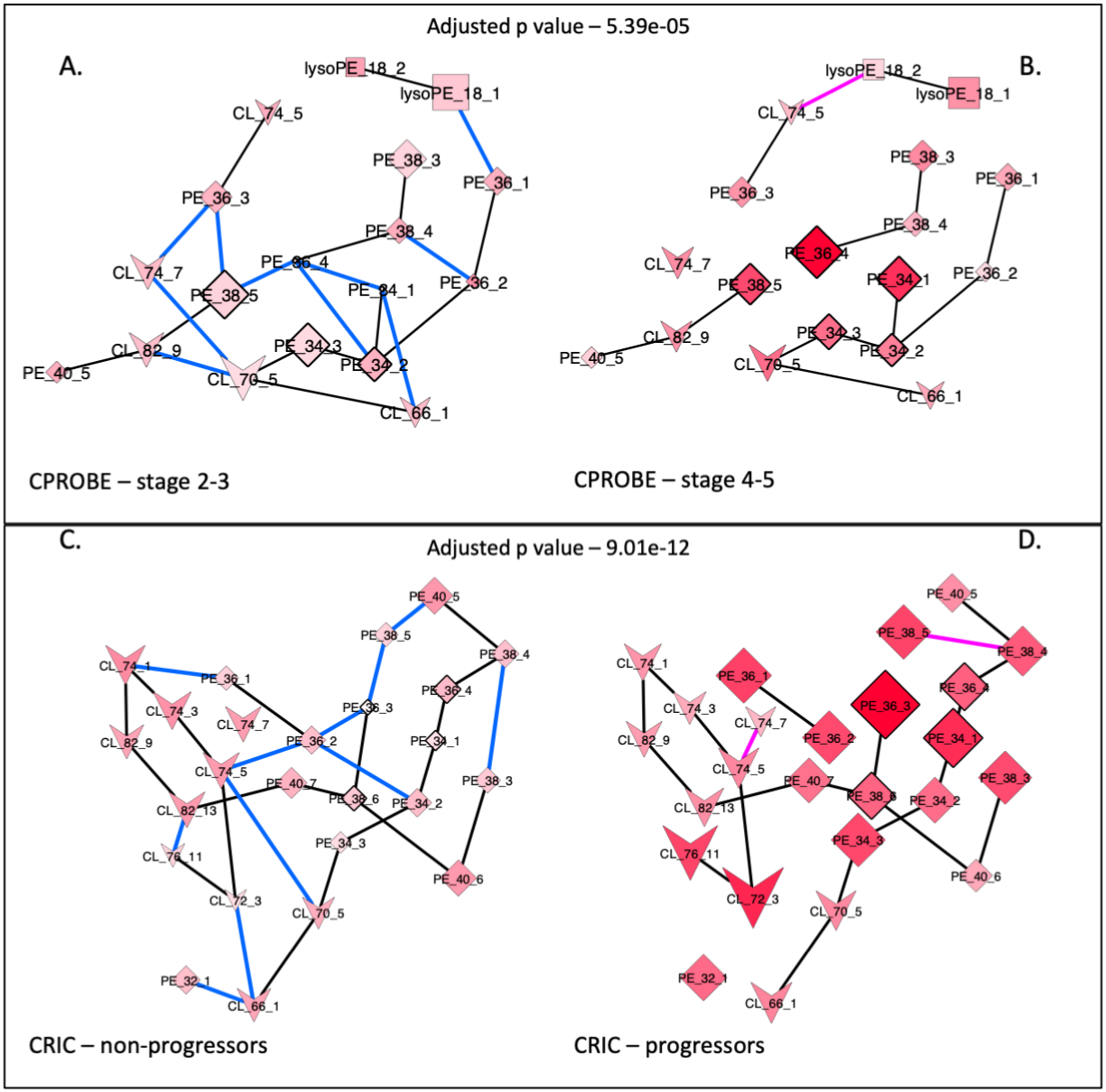
[9] Ma et al. (16') Bioinformatics

[9] Hellstern et al. (21') PLoS Comp Bio; Yue et al. (21') Genet Epidemiol

Lipid Pathways Associated with CKD



Lipid Pathways Associated with CKD



Outline

Summary

The GMD framework

- encompasses classical methods, both unsupervised (PCA, PCoA/MDS, biplots) and supervised (ridge, GLS) methods
- extends them to non-standard settings (multi-view data)

DNEA allows data-driven discovery of pathway signatures of CKD progression.

Aknowledgement

Yue Wang
Kun Yue
Michael Hellstern

Tim Randolph
Ali Shojaie
George Michailidis
Alla Karnovsky
Farsad Afshinnia

- The GMD-biplot and its application to microbiome data. *mSystems.* 2019
- Generalized matrix decomposition: estimation and inference for two-way structured data. *arXiv:2104.08408*
- Differential network enrichment analysis reveals novel lipid pathways in Chronic Kidney Disease. *Bioinformatics.* 2019
- netgsa: Fast computation and interactive visualization for topology-based pathway enrichment analysis. *PLoS Comp Bio.* 2021
- REHE fast variance components estimation for linear mixed models. *Genet Epidemiol.* 2021

Thank You!

`http://drjingma.com/`

Simulation Examples

Data generated from linear model with $p = 300$ & $n = 200$

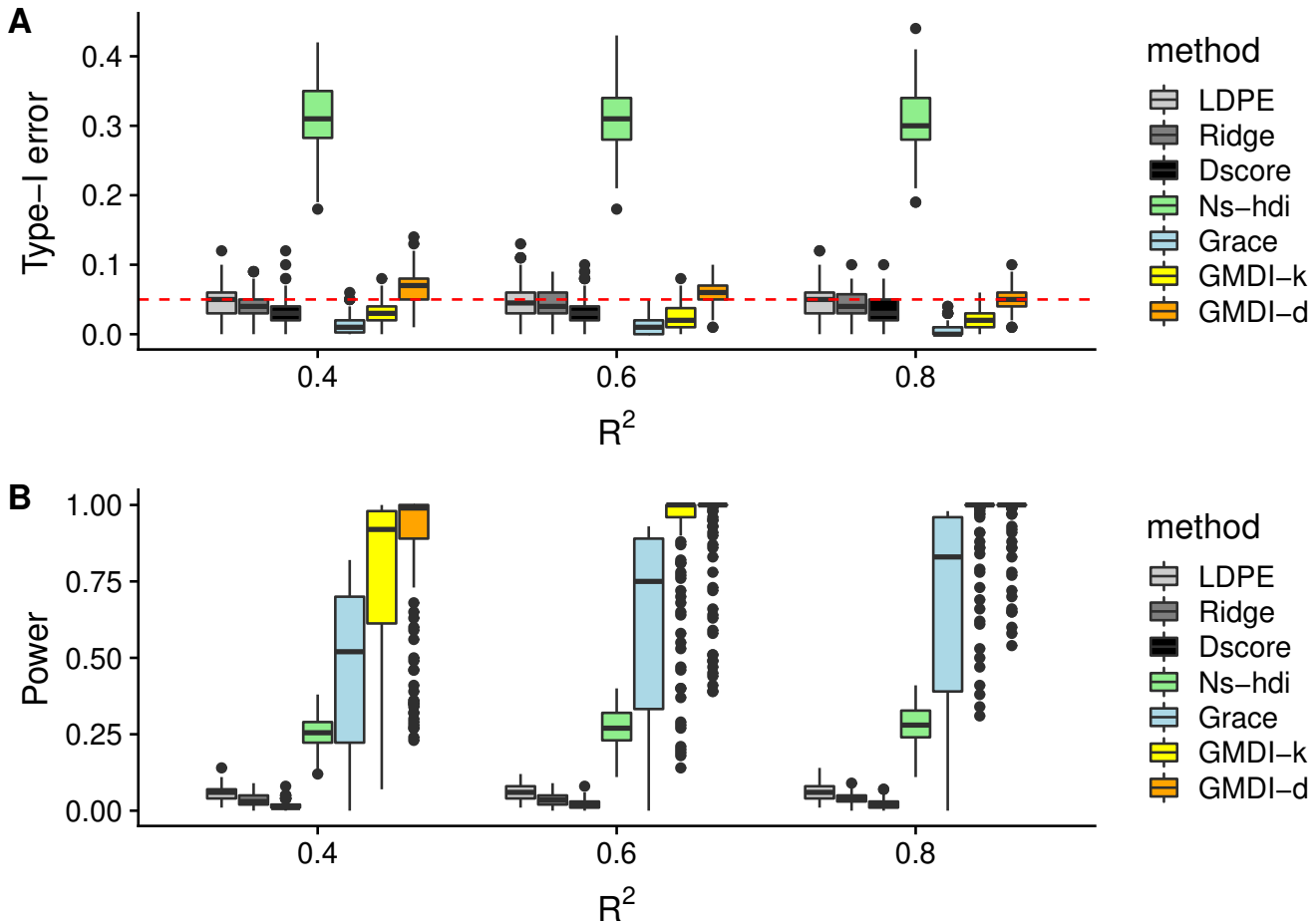
Compare GMDI-k and GMDI-d with

- Low-dimension Projection Estimator (LDPE)[10]
- Ridge-based inference[11]
- Decorrelated score test (Dscore)[12]
- Non-sparse high-dimensional inference (Ns-hdi)[13]
- Grace test[14]

[9] van de Geer et al. AoS, 14'; Zhang & Zhang. JRSSB, 14'
[10] Bühlmann. Bernoulli, 13'
[11] Ning & Liu. AoS, 17'
[12] Zhu & Bradic. JASA, 18'
[13] Zhao & Shojaie. Biometrics,15'

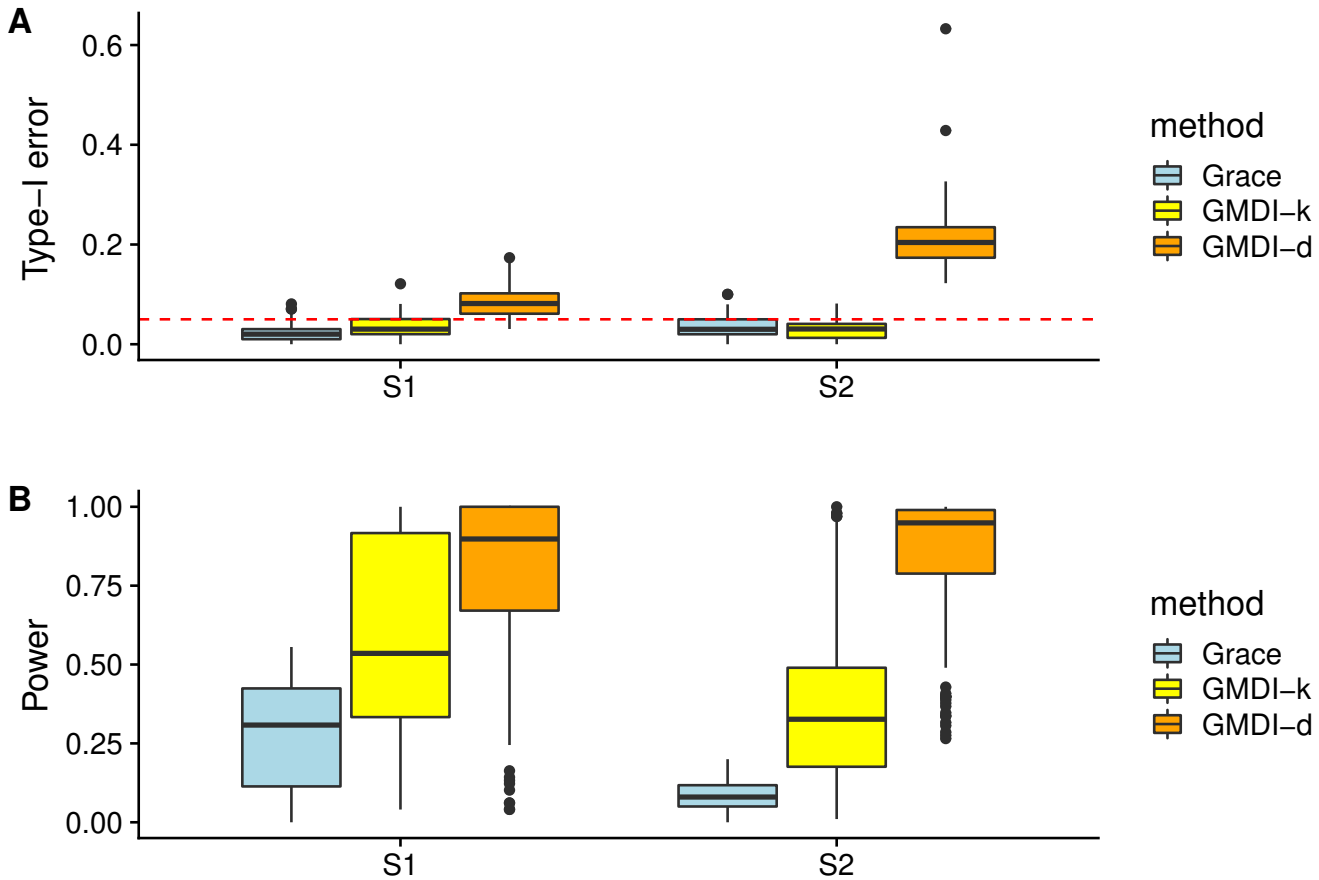
Simulation Example I

- $\boldsymbol{H} = I_n, \|\boldsymbol{Q}^{-1/2}\beta^*\|_0 = 10$ but β^* is not sparse.



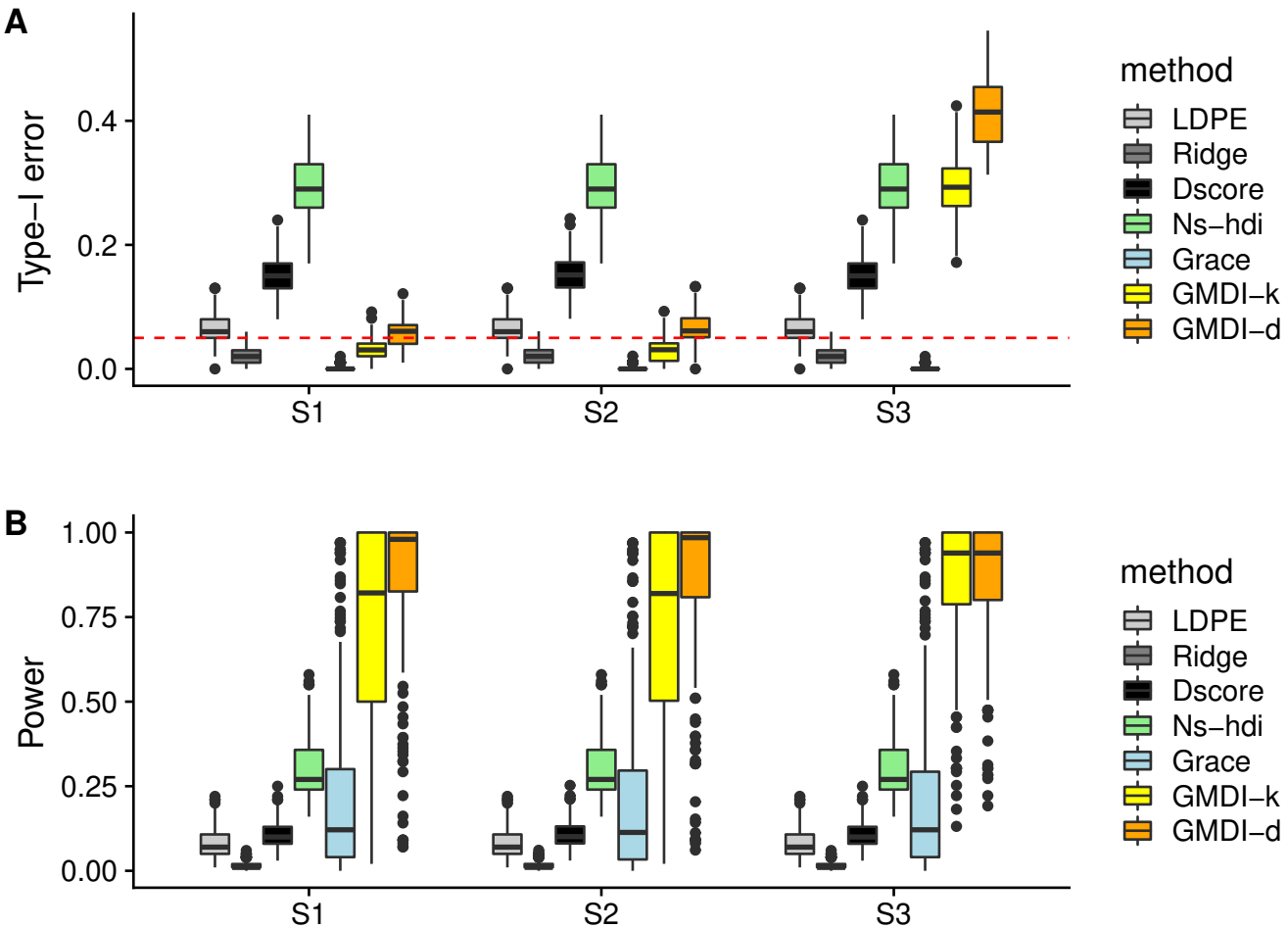
Simulation Example II

- Similar to Setting 1, but with Q perturbed. S1: small perturbation; S2: large perturbation.



Simulation Example III

- $\boldsymbol{H}$ is block diagonal, $\boldsymbol{Q}$ is the same as in setting 1. S1: correct $\boldsymbol{H}$; S2: perturbed $\boldsymbol{H}$; S3: I_n.



Lipids are Correlated!

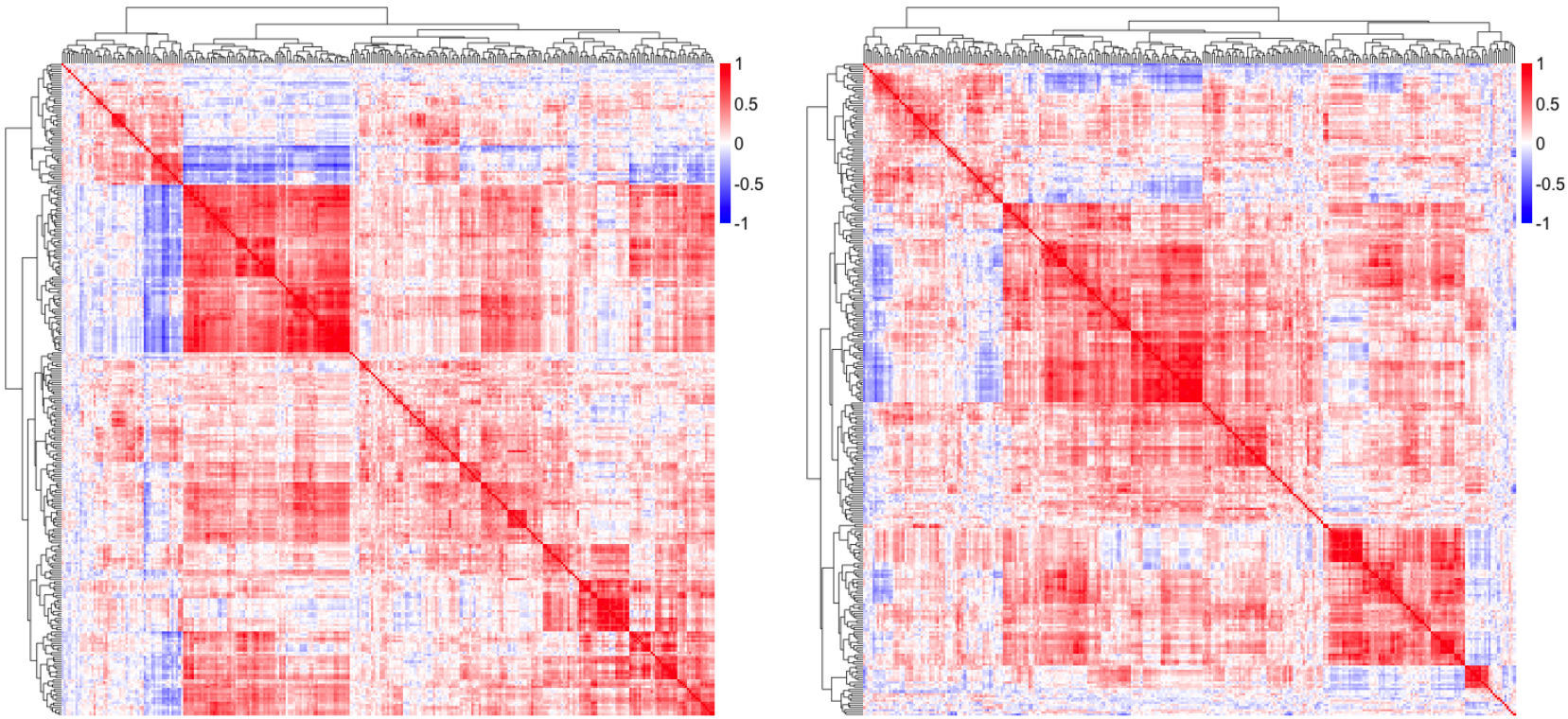


Figure: Heatmap of empirical correlations in non-progressors (left) and progressors (right).

Lipids are Correlated!

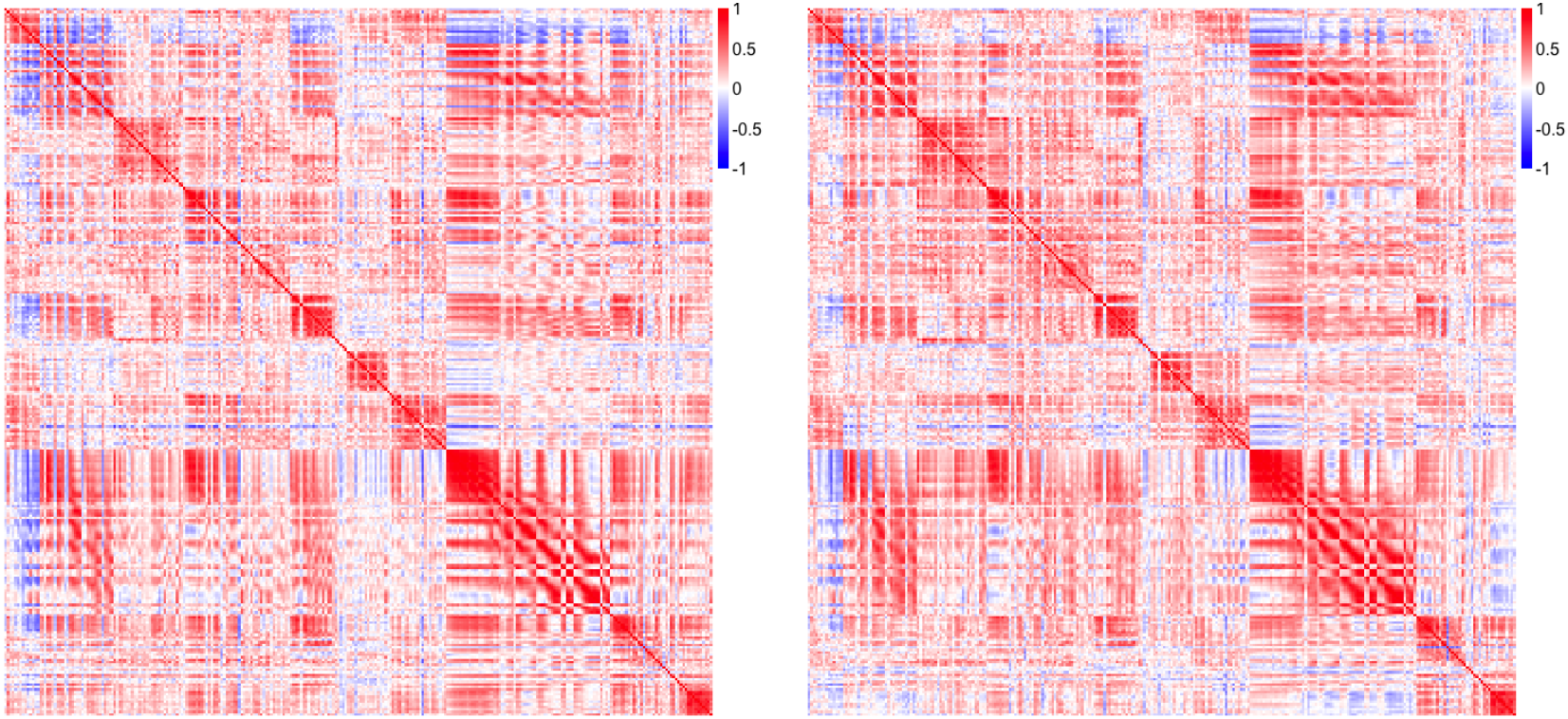


Figure: Heatmap of empirical correlations in non-progressors (left) and progressors (right).