# Integrated Principal Components Analysis (iPCA)

Tiffany Tang

University of California, Berkeley

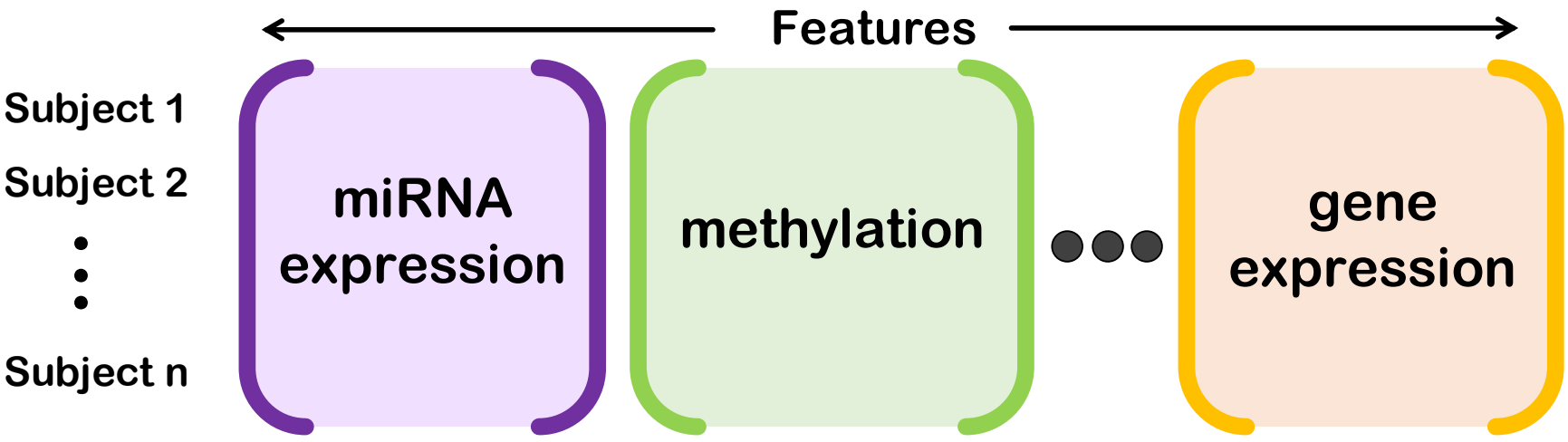
July 30, 2019

Joint Statistical Meetings

Joint work with Genevera Allen (Rice University)

# Motivation

- Data Integration + Unsupervised Learning
- Applications
  - Integrative genomics
  - Multi-modal imaging
  - Multi-sensor data
- Want to find the joint patterns which are common among all of the datasets



## Multiblock PCA

- Concatenated PCA
  - *Westerhuis et al. (1998), Wold et al. (1996)*
- Multiple Factor Analysis
  - *Escofier and Pages (1994), Abdi et al. (2013)*

## Matrix Factorization Methods

- Joint and Individual Variation Explained (JIVE)
  - *Lock et al. (2013)*
- Coupled Matrix Factorizations (CMF)
  - *Singh and Gordon (2008), Acar et al. (2014)*

$$\underset{\mathbf{U},\mathbf{V}_1,\ldots,\mathbf{V}_K}{\operatorname{argmin}} \sum_{k=1}^{K} \|\mathbf{X}_k - \mathbf{U}\mathbf{V}_k^T\|_F^2$$

# Overview of Integrated PCA (iPCA)

**Objective:** Extend a model-based PCA to integrated data

- Exploratory Data Analysis
- Joint Pattern Recognition
- Visualization

**Key tool:** matrix-variate normal distribution

**Why?** PCA can be viewed as maximizing the sample covariance $\hat{\mathbf{\Delta}} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$

# Overview of Integrated PCA (iPCA)

Advantages:

- A unifying framework for the multiblock PCA family
- **U**'s and **V**'s are orthogonal, ordered, and nested
- Convenient visualizations
- Nice theoretical properties
  - Global solution - important for interpretability and reproducibility
  - Provable guarantees

# Outline

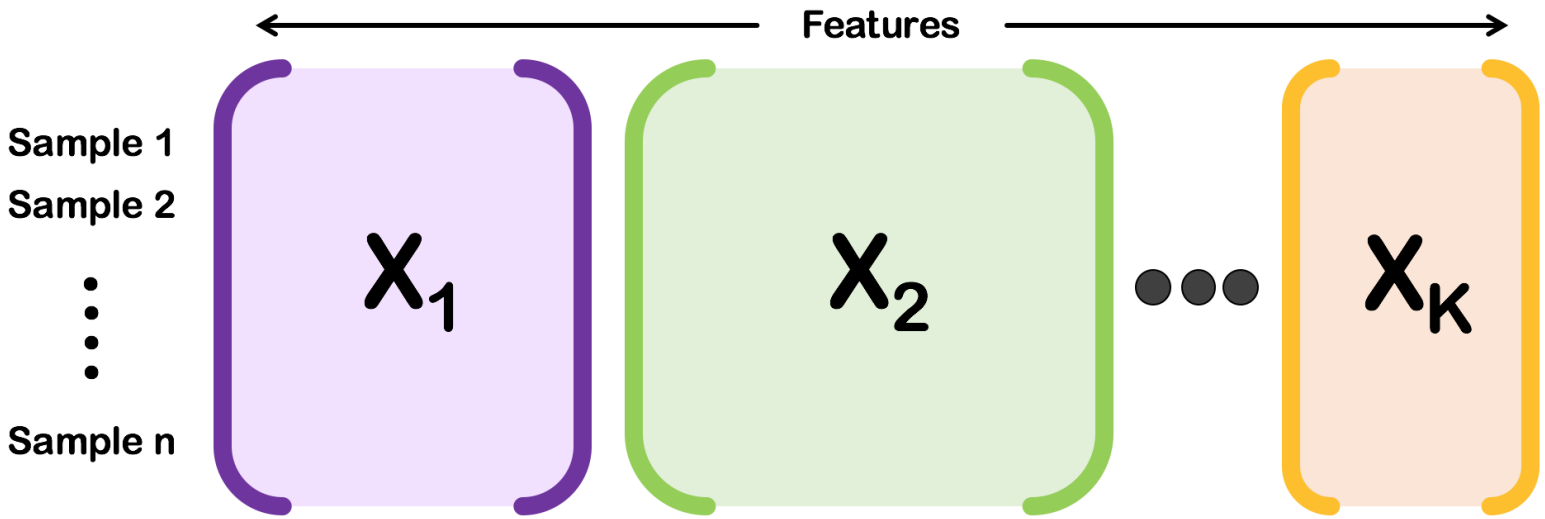
1. iPCA Model
2. Case Study: Alzheimer's Disease

# iPCA Model

Given datasets $\mathbf{X}_1, \ldots, \mathbf{X}_K$, assume that each dataset $\mathbf{X}_k$ arises from the matrix-variate normal model:

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{M}_k, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k).$$

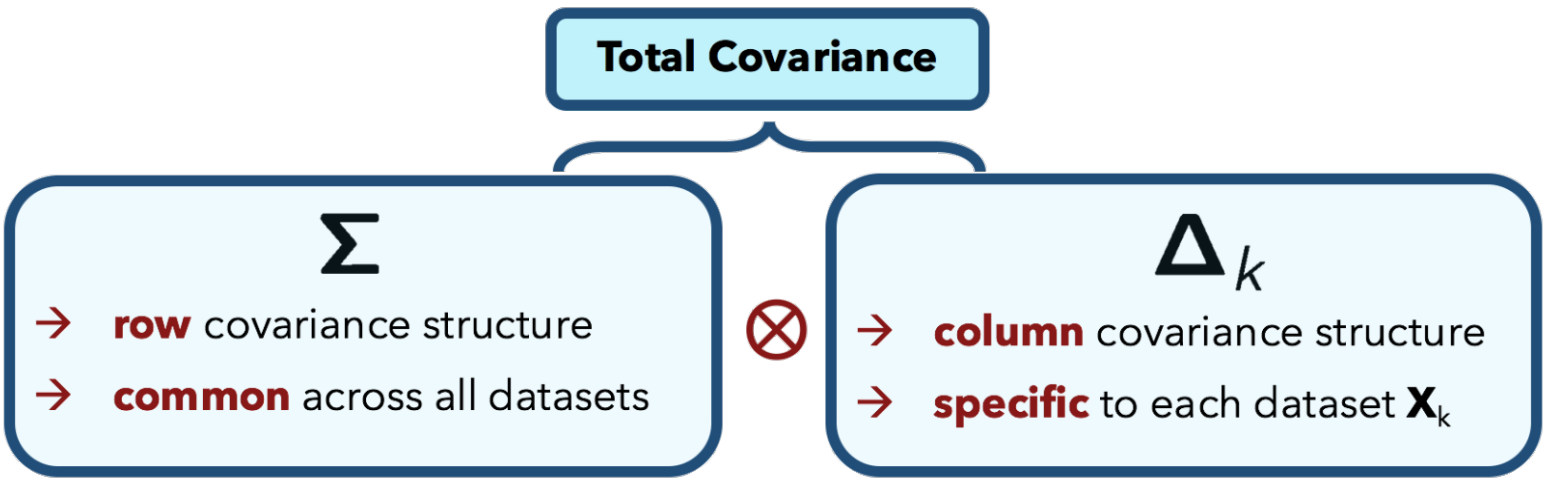
$$\mathbf{\Sigma} \otimes \mathbf{\Delta}_k = \begin{bmatrix} \sigma_{11}\mathbf{\Delta}_k & \cdots & \sigma_{1n}\mathbf{\Delta}_k \\ \vdots & \ddots & \vdots \\ \sigma_{n1}\mathbf{\Delta}_k & \cdots & \sigma_{nn}\mathbf{\Delta}_k \end{bmatrix}$$



# iPCA Model

Given datasets $\mathbf{X}_1, \ldots, \mathbf{X}_K$, assume that each dataset $\mathbf{X}_k$ arises from the matrix-variate normal model:

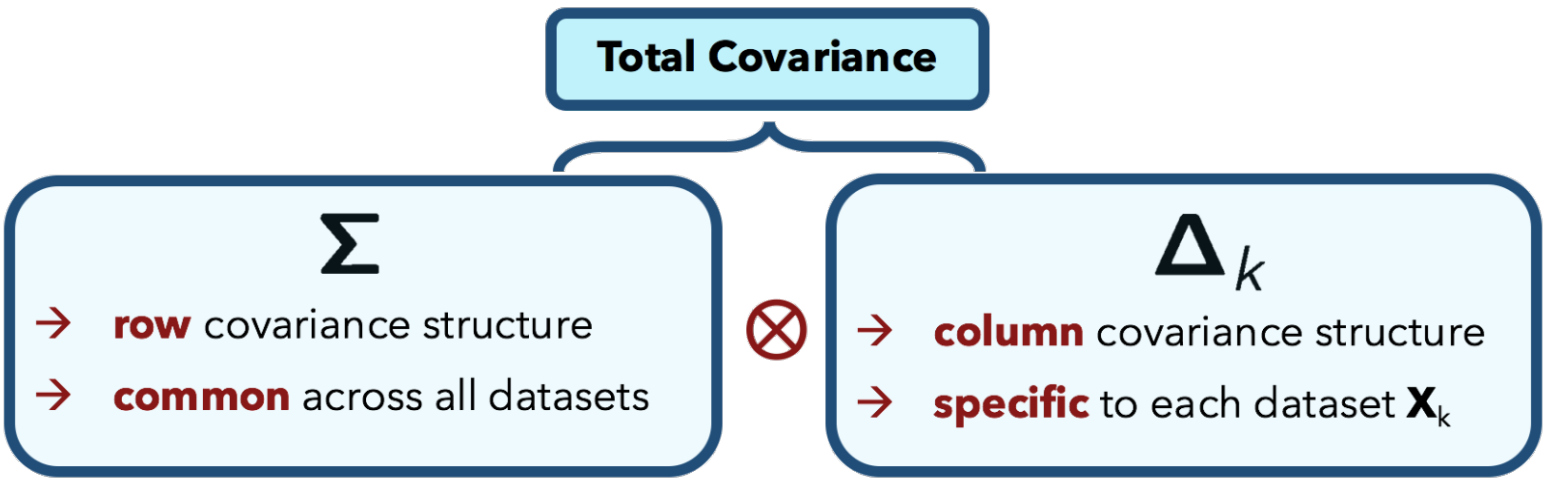
$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{M}_k, \boldsymbol{\Sigma} \otimes \boldsymbol{\Delta}_k).$$



# iPCA Model

Given datasets $\mathbf{X}_1, \ldots, \mathbf{X}_K$, assume that each dataset $\mathbf{X}_k$ arises from the matrix-variate normal model:

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{M}_k, \boldsymbol{\Sigma} \otimes \boldsymbol{\Delta}_k).$$




Equivalently, can rewrite Kronecker product covariance model as

$$\left[\mathbf{X}_1 \boldsymbol{\Delta}_1^{-1/2}, \ldots, \mathbf{X}_K \boldsymbol{\Delta}_K^{-1/2}\right]_{\cdot j} \overset{iid}{\sim} N(\mathbf{o}, \boldsymbol{\Sigma})$$

# iPCA

1. **iPCA Model**: For each $k = 1, \ldots, K$, assume

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{1}_n \, \mu_k^T, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k)$$

   - $\mathbf{\Sigma}$ is the common row covariance structure
   - $\mathbf{\Delta}_k$ is the separate column covariance structure

# iPCA

1. **iPCA Model**: For each $k = 1, \ldots, K$, assume

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{1}_n \, \mu_k^T, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k)$$

   - $\mathbf{\Sigma}$ is the common row covariance structure
   - $\mathbf{\Delta}_k$ is the separate column covariance structure

2. **Estimate** covariances $\mathbf{\Sigma}, \mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_K$ to obtain $\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \ldots, \hat{\mathbf{\Delta}}_K$

   ⋮

# iPCA

1. **iPCA Model**: For each $k = 1, \ldots, K$, assume

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{1}_n \, \mu_k^T, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k)$$

   - $\mathbf{\Sigma}$ is the common row covariance structure
   - $\mathbf{\Delta}_k$ is the separate column covariance structure

2. **Estimate** covariances $\mathbf{\Sigma}, \mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_K$ to obtain $\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \ldots, \hat{\mathbf{\Delta}}_K$

   ⋮

3. **Maximize** covariances

# iPCA

1. **iPCA Model**: For each $k = 1, \ldots, K$, assume

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{1}_n \, \mu_k^T, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k)$$

   - $\mathbf{\Sigma}$ is the common row covariance structure
   - $\mathbf{\Delta}_k$ is the separate column covariance structure

2. **Estimate** covariances $\mathbf{\Sigma}, \mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_K$ to obtain $\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \ldots, \hat{\mathbf{\Delta}}_K$

   ⋮

3. **Maximize** covariances

$\mathbf{U} \leftarrow$ eigenvectors of $\hat{\mathbf{\Sigma}}$ = joint patterns

$\mathbf{V}_k \leftarrow$ eigenvectors of $\hat{\mathbf{\Delta}}_k$ = individual patterns

# iPCA

1. **iPCA Model**: For each $k = 1, \ldots, K$, assume

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{1}_n \, \mu_k^T, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k)$$

   - $\mathbf{\Sigma}$ is the common row covariance structure
   - $\mathbf{\Delta}_k$ is the separate column covariance structure

2. **Estimate** covariances $\mathbf{\Sigma}, \mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_K$ to obtain $\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \ldots, \hat{\mathbf{\Delta}}_K$

   ⋮

3. **Maximize** covariances

   iPC Scores: $\mathbf{U} \leftarrow$ eigenvectors of $\hat{\mathbf{\Sigma}}$ = joint patterns

   iPC Loadings: $\mathbf{V}_k \leftarrow$ eigenvectors of $\hat{\mathbf{\Delta}}_k$ = individual patterns

# iPCA

1. **iPCA Model**: For each $k = 1, \ldots, K$, assume

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{1}_n \, \mu_k^T, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k)$$

   - $\mathbf{\Sigma}$ is the common row covariance structure
   - $\mathbf{\Delta}_k$ is the separate column covariance structure

2. **Estimate** covariances $\mathbf{\Sigma}, \mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_K$ to obtain $\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \ldots, \hat{\mathbf{\Delta}}_K$

   ⋮

3. **Maximize** covariances

   iPC Scores: $\mathbf{U} \leftarrow$ eigenvectors of $\hat{\mathbf{\Sigma}}$ = joint patterns

   iPC Loadings: $\mathbf{V}_k \leftarrow$ eigenvectors of $\hat{\mathbf{\Delta}}_k$ = individual patterns

4. **Visualize** dominant joint patterns by plotting iPC scores $\mathbf{U}$

# iPCA

1. **iPCA Model**: For each $k = 1, \dots, K$, assume

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{1}_n \, \mu_k^T, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k)$$

   - $\mathbf{\Sigma}$ is the common row covariance structure
   - $\mathbf{\Delta}_k$ is the separate column covariance structure

2. **Estimate** covariances $\mathbf{\Sigma}, \mathbf{\Delta}_1, \dots, \mathbf{\Delta}_K$ to obtain $\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \dots, \hat{\mathbf{\Delta}}_K$

$$\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \dots, \hat{\mathbf{\Delta}}_K = \text{argmax } \ell(\mathbf{\Sigma}, \mathbf{\Delta}_1, \dots, \mathbf{\Delta}_K)$$

3. **Maximize** covariances

   iPC Scores: $\mathbf{U} \leftarrow$ eigenvectors of $\hat{\mathbf{\Sigma}}$ = joint patterns

   iPC Loadings: $\mathbf{V}_k \leftarrow$ eigenvectors of $\hat{\mathbf{\Delta}}_k$= individual patterns

4. **Visualize** dominant joint patterns by plotting iPC scores $\mathbf{U}$

# iPCA

1. **iPCA Model**: For each $k = 1, \dots, K$, assume

$$\mathbf{X}_k \sim N_{n,p_k}(\mathbf{1}_n \, \mu_k^T, \mathbf{\Sigma} \otimes \mathbf{\Delta}_k)$$

   - $\mathbf{\Sigma}$ is the common row covariance structure
   - $\mathbf{\Delta}_k$ is the separate column covariance structure

2. **Estimate** covariances $\mathbf{\Sigma}, \mathbf{\Delta}_1, \dots, \mathbf{\Delta}_K$ to obtain $\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \dots, \hat{\mathbf{\Delta}}_K$

$$\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Delta}}_1, \dots, \hat{\mathbf{\Delta}}_K = \text{argmax } \ell(\mathbf{\Sigma}, \mathbf{\Delta}_1, \dots, \mathbf{\Delta}_K) - \sum_{k=1}^{K} \lambda_k || \, \mathbf{\Sigma}^{-1} \otimes \mathbf{\Delta}_k^{-1} \, ||_F^2$$

3. **Maximize** covariances

   iPC Scores: $\mathbf{U} \leftarrow$ eigenvectors of $\hat{\mathbf{\Sigma}}$ = joint patterns

   iPC Loadings: $\mathbf{V}_k \leftarrow$ eigenvectors of $\hat{\mathbf{\Delta}}_k$ = individual patterns

4. **Visualize** dominant joint patterns by plotting iPC scores $\mathbf{U}$

# Case Study: Alzheimer's Disease (AD)

**ROSMAP Study** (*Bennett et al. (2012)*)

- Longitudinal clinical-pathological cohort study of aging and AD
- Genomics on post-mortem brains

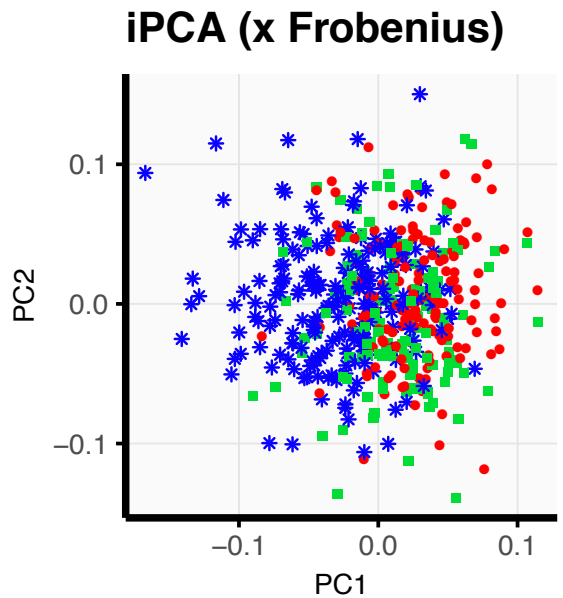
**ROSMAP Genomics Data** ($n = 507$)

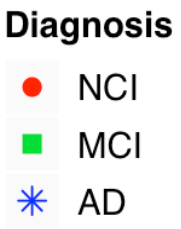
- miRNA Expression: $p_1 = 309$
- Gene Expression via RNASeq (log-transformed): $p_2 = 900$
- DNA Methylation (m-values): $p_3 = 1250$

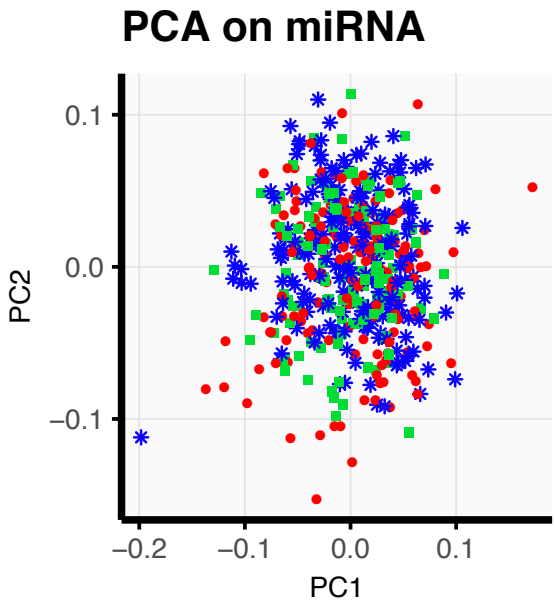
**Clinical Outcomes of Interest:**

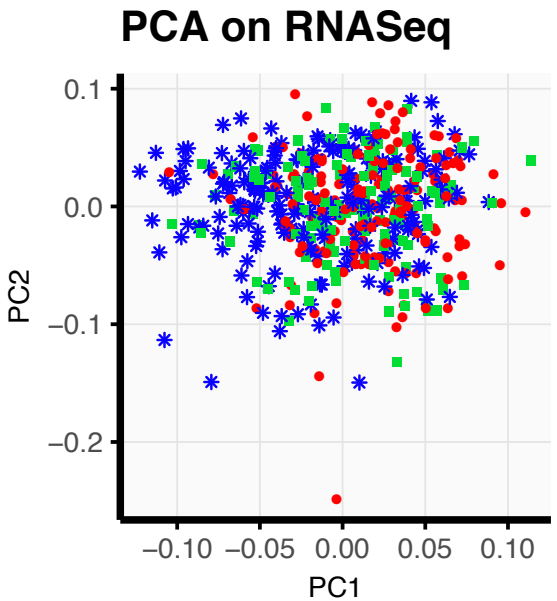
- Clinician's Diagnosis
  - AD, Mild Cognitive Impairment, No Cognitive Impairment
- Global Cognition Score

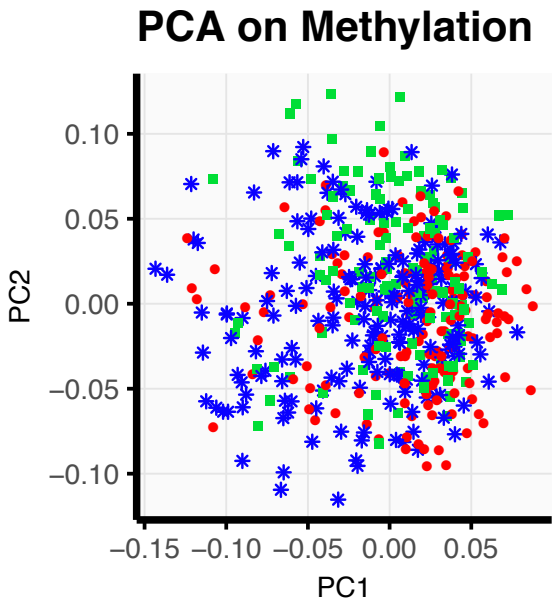
# ROSMAP: Clinician's Diagnosis



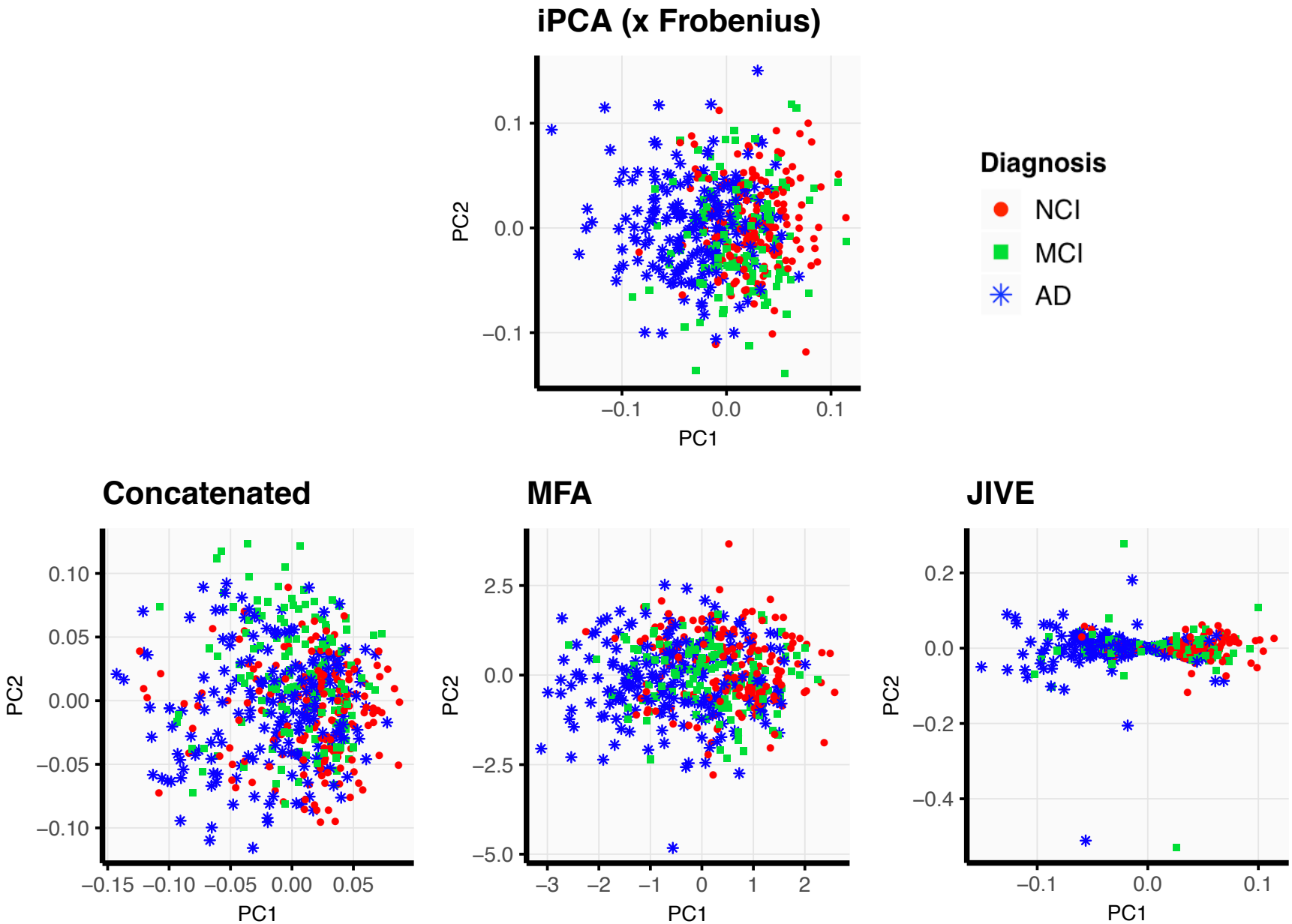



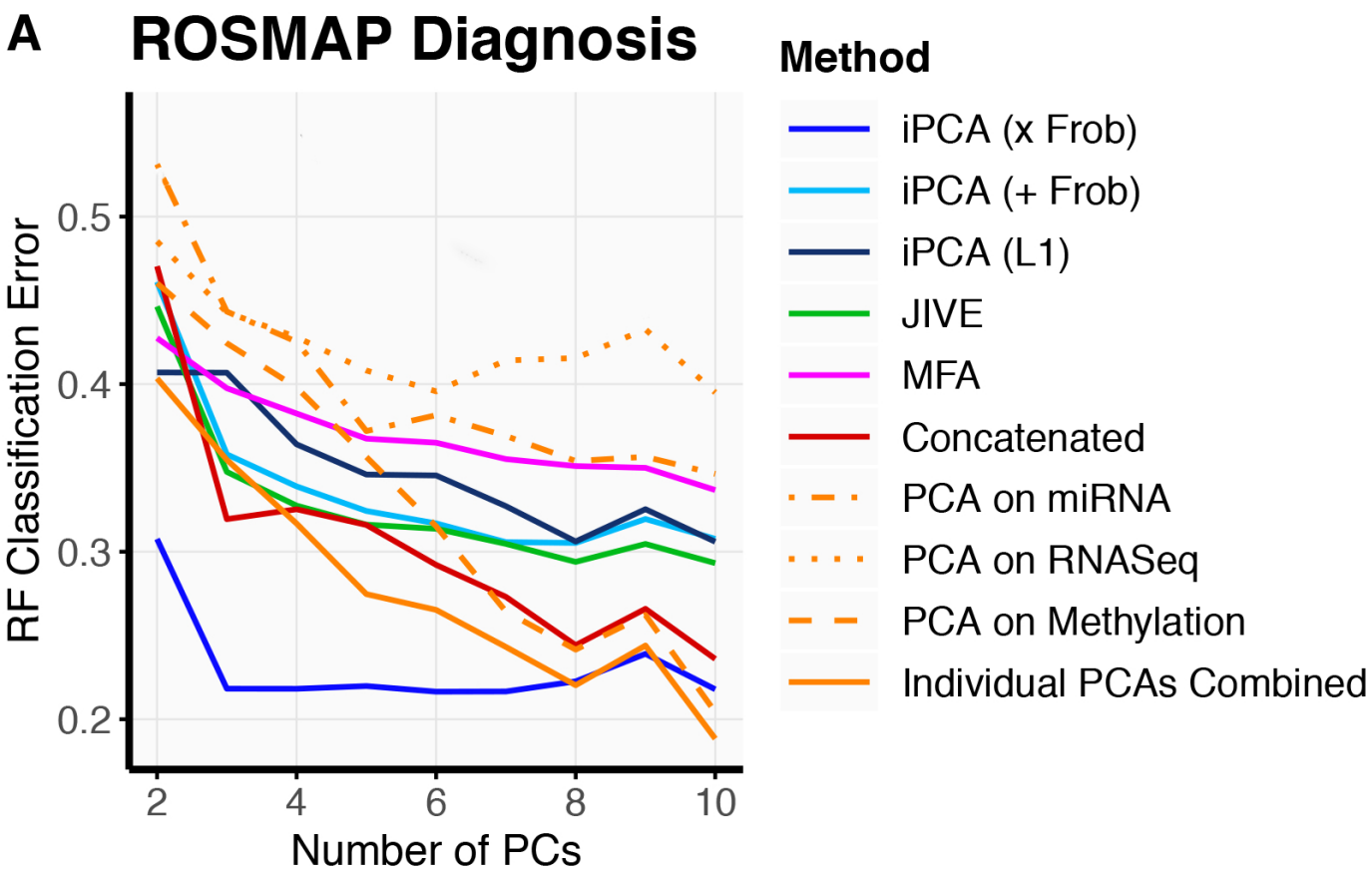



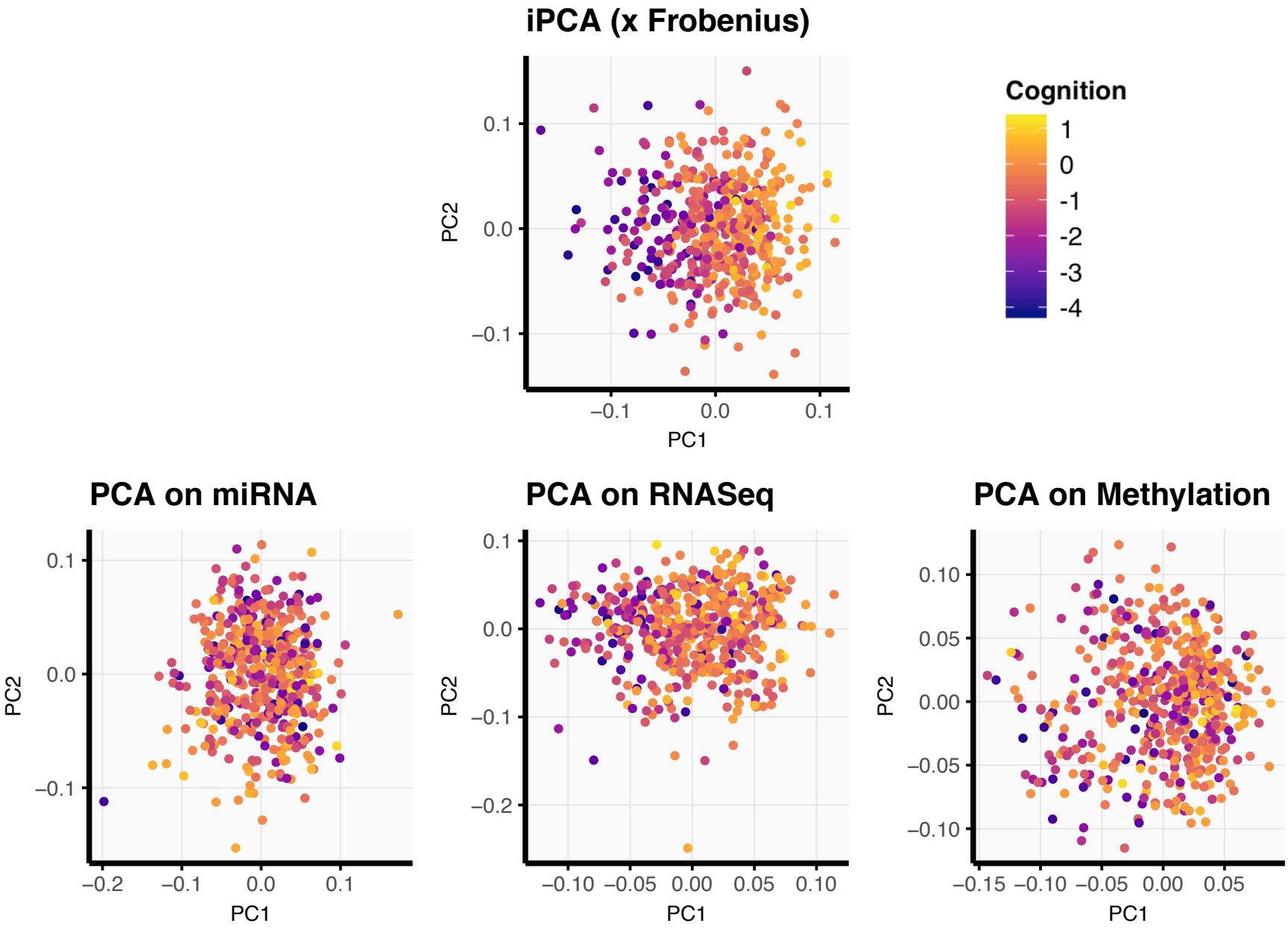
# ROSMAP: Clinician's Diagnosis



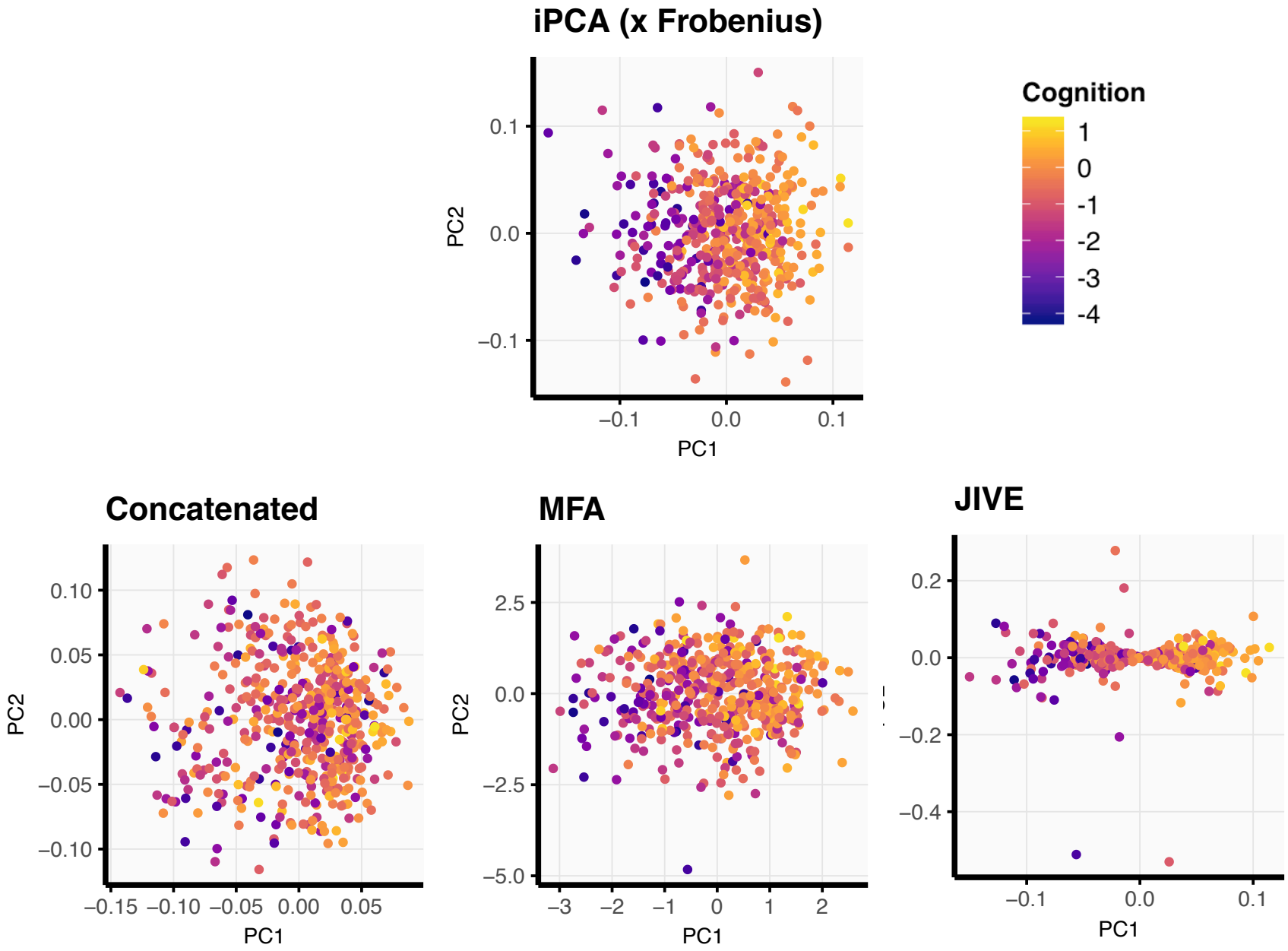
# ROSMAP: Clinician's Diagnosis



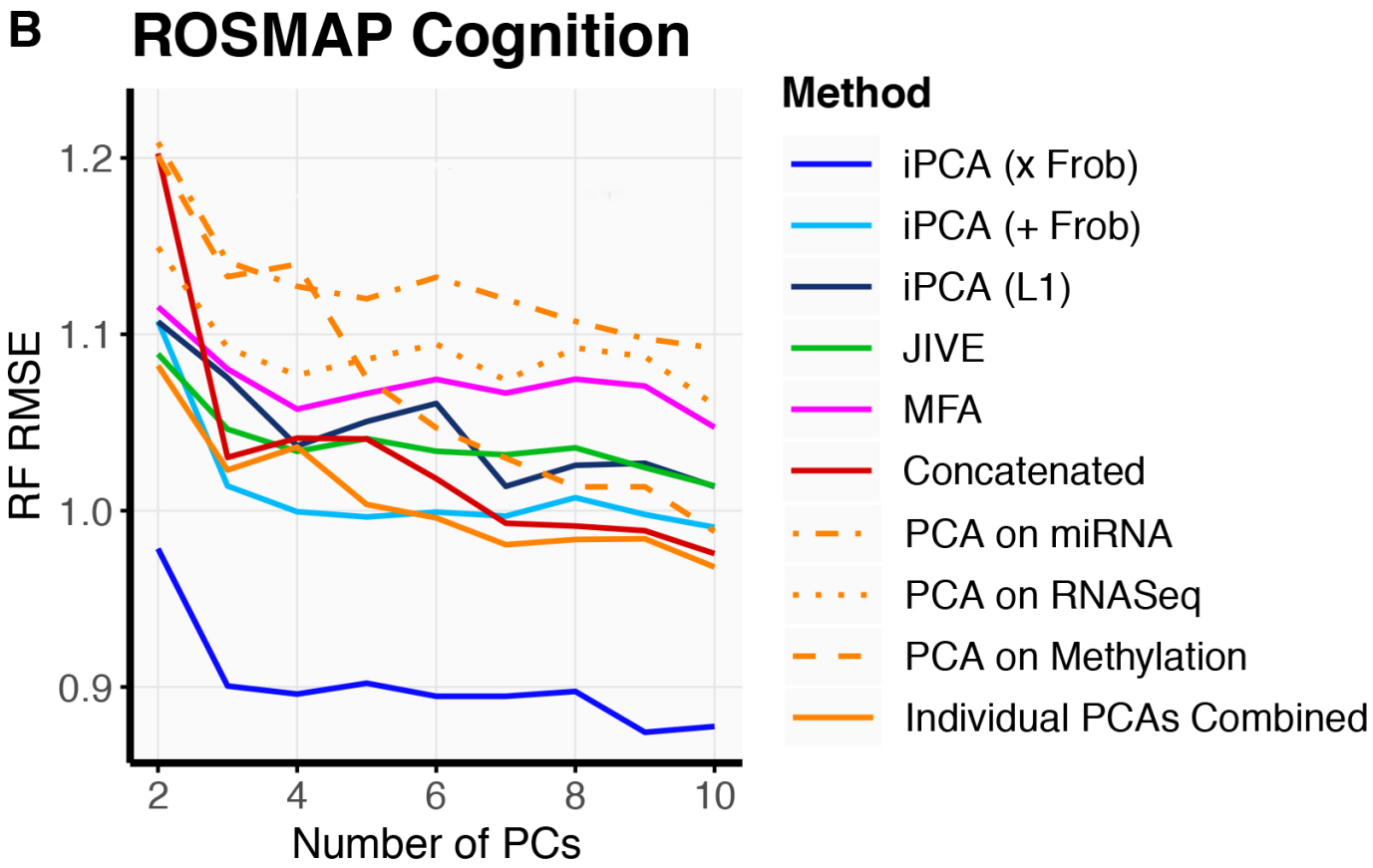
# ROSMAP: Global Cognition



# ROSMAP: Global Cognition



# ROSMAP: Global Cognition



# ROSMAP: Interpretation

**Table 1:** Top genetic features obtained by applying Sparse PCA to each $\hat{\Delta}_k$ in ROSMAP analysis (using the multiplicative Frobenius iPCA estimator)

| | miRNA | RNASeq | Methylation |
|---|---|---|---|
| 1 | miR 216a | VGF | TMCO6 |
| 2 | miR 127 3p | SVOP | PHF3 |
| 3 | miR 124 | PCDHGC5 | BRUNOL4 |
| 4 | miR 30c | ADCYAP1 | OSCP1 |
| 5 | miR 143 | LINC01007 | GRIN2B |

# Conclusion

- iPCA is a new practical tool for discovering and visualizing interesting joint patterns which occur in multiple datasets
- In order to fit iPCA, we propose using the multiplicative Frobenius estimator.
- Though we impose a model, the assumptions are analogous to those in classical PCA
- Tang, T. M., & Allen, G. I. (2018). Integrated Principal Components Analysis. *arXiv preprint arXiv:1810.00832*.

# Acknowledgments


Collaborator

- **Genevera Allen**

Thank you also to...

- ROSMAP Team
  - Joshua Shulman
  - David Bennett
  - Zhandong Liu
  - Ying-Wooi Wan
- Bin Yu

Funding

- National Science Foundation Graduate Research Fellowship
- Office of Naval Research

# Thank you!