BIO214 Lecture 7

Bioinformatics-II

Finding Patterns in Normalized Omics Data - 1

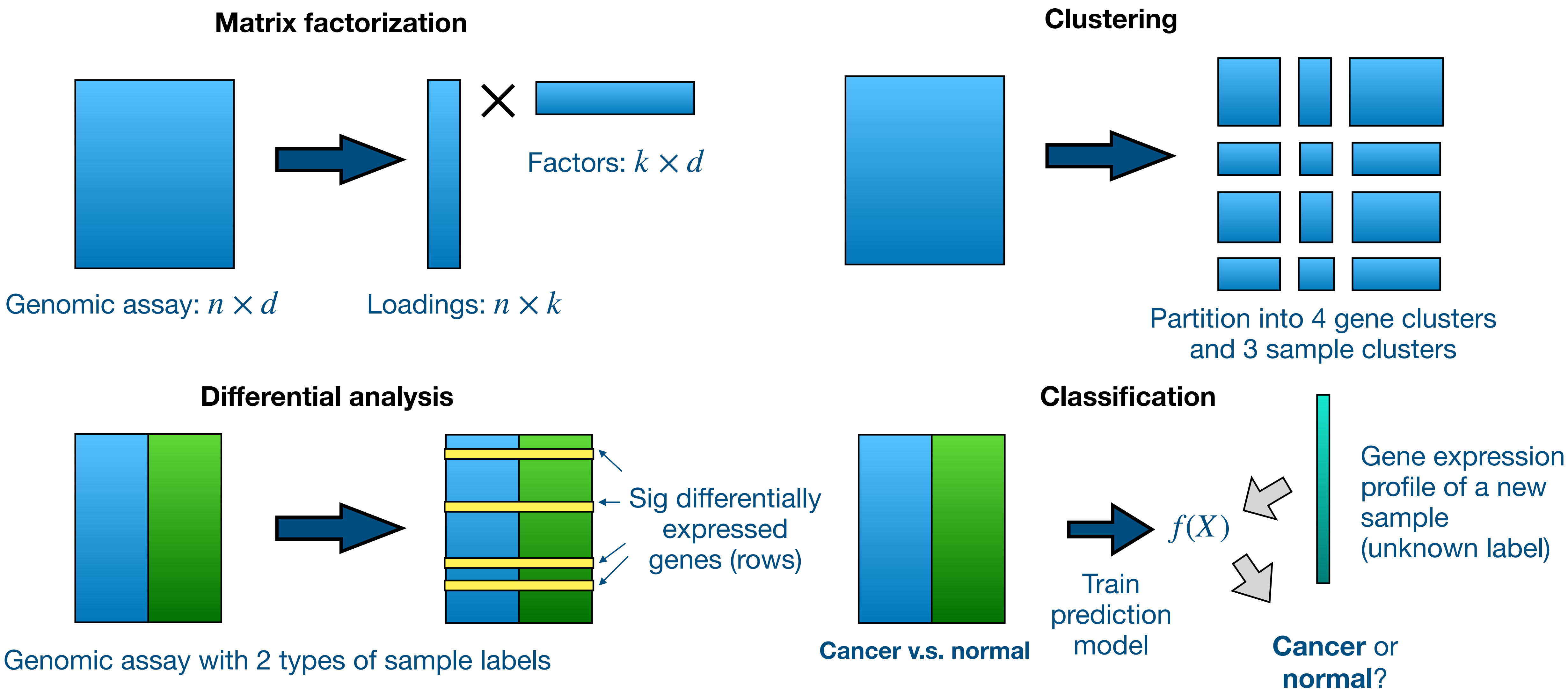
Zhen Wei; 2023-Feb-14

Outline

- The major computational questions asked in genome-wide data
- PCA explained by cases
- Applications of PCA in genomics
- Dimensional reduction techniques

The major computational questions asked in genome-wide data

What are the major biological questions asked given a normalized genomic data matrix?


Matrix factorization
Genomic assay: $n \times d$
Loadings: $n \times k$
Factors: $k \times d$
Clustering
Partition into 4 gene clusters and 3 sample clusters
Differential analysis
Sig differentially expressed genes (rows)
Genomic assay with 2 types of sample labels
Classification
Cancer v.s. normal
Train prediction model
$f(X)$
Gene expression profile of a new sample (unknown label)
Cancer or normal?

The four basic genomic question types:

- **Matrix factorization**: finding latent gene expression factors.
- **Clustering**: dividing samples and genes into different parts.
- **Differential analysis:** identifying statistically significant changes between groups of samples.
- **Classification**: predict sample label given a previously unseen sample.

PCA explained by cases

PCA / Matrix factorization motivation

Recommendation system

Rating <-> movie matrix in amazon

	Movie 1	**Movie 2**	**Movie 3**	**Movie 4**
User 1	5 star	3 star	1 star	5 star
User 2	1 star	2 star	3 star	3 star
User 3	3 star	1 star	2 star	2 star
User 4	2 star	2 star	3 star	4 star

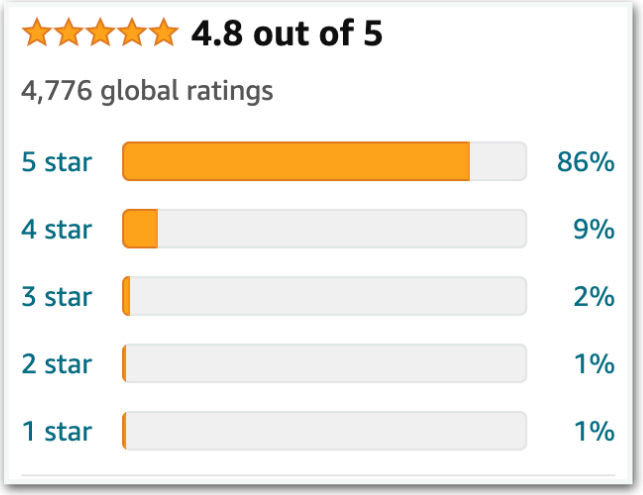
$$= \begin{array}{c|c|c|c|} & \text{Sci} & \text{Horr} & \text{Rom} \\ \hline \text{User 1} & 4.07 & 0.37 & 4.08 \\ \hline \text{User 2} & 1.98 & 3.55 & 0.01 \\ \hline \text{User 3} & 0.78 & 2.49 & 2.57 \\ \hline \text{User 4} & 2.43 & 3.49 & 0.96 \\ \hline \end{array} \overset{\text{Matrix multiplication}}{\times} \begin{array}{|c|c|c|c|c} \text{Movie 1} & \text{Movie 2} & \text{Movie 3} & \text{Movie 4} & \\ \hline 0.26 & 0.68 & 0.17 & 1.04 & \text{Sci} \\ \hline 0.14 & 0.15 & 0.74 & 0.31 & \text{Horr} \\ \hline 0.96 & 0.04 & 0.00 & 0.17 & \text{Rom} \\ \hline \end{array}$$

User loadings of the 3 factors:

Describing how each user like each movie attribute

3 latent factors:

Characterizing movies by 3 attributes: **sci-fi, horror, and romance**



★★★★★ **4.8 out of 5**

4,776 global ratings

5 star 86%

4 star 9%

3 star 2%

2 star 1%

1 star 1%

- **PCA** and other **matrix factorization** techniques can help us estimate latent movie attributes directly from the rating data.
- Using the attributes information (and its user loadings), we can recommend new movies to the users in the future.

Matrix multiplication recall

$$
\begin{array}{c|c|c|c|}
 & \text{Sci} & \text{Horr} & \text{Rom} \\
\hline
\text{User 1} & 4.07 & 0.37 & 4.08 \\
\hline
\text{User 2} & 1.98 & 3.55 & 0.01 \\
\hline
\text{User 3} & 0.78 & 2.49 & 2.57 \\
\hline
\text{User 4} & 2.43 & 3.49 & 0.96 \\
\hline
\end{array}
\times
\begin{array}{|c|c|c|c|c}
\text{Movie 1} & \text{Movie 2} & \text{Movie 3} & \text{Movie 4} & \\
\hline
0.26 & 0.68 & 0.17 & 1.04 & \text{Sci} \\
\hline
0.14 & 0.15 & 0.74 & 0.31 & \text{Horr} \\
\hline
0.96 & 0.04 & 0.00 & 0.17 & \text{Rom} \\
\hline
\end{array}
=
$$

	Movie 1	Movie 2	Movie 3	Movie 4
User 1	4.07 * 0.26 + 0.37 * 0.14 + 4.08 * 0.96	4.07 * 0.68 + 0.37 * 0.15 + 4.08 * 0.04	4.07 * 0.17 + 0.37 * 0.74 + 4.08 * 0.00	4.07 * 1.04 + 0.37 * 0.31 + 4.08 * 0.17
User 2	1.98 * 0.26 + 3.55 * 0.14 + 0.01 * 0.96	1.98 * 0.68 + 3.55 * 0.15 + 0.01 * 0.04	1.98 * 0.17 + 3.55 * 0.74 + 0.01 * 0.00	1.98 * 1.04 + 3.55 * 0.31 + 0.01 * 0.17
User 3	0.78 * 0.26 + 2.49 * 0.14 + 2.57 * 0.96	0.78 * 0.68 + 2.49 * 0.15 + 2.57 * 0.04	…	…
User 4	2.43 * 0.26 + 3.49 * 0.14 + 0.96 * 0.96	…	…	…

PCA / Matrix factorization motivation - 2

Personality psychology

	Behavior 1	Behavior 2	Behavior 3	Behavior 4
Individual 1	5	3	1	5
Individual 2	1	2	3	3
Individual 3	3	1	2	2
Individual 4	2	2	3	4

=

Cons	Open	Extr
4.07	0.37	4.08
1.98	3.55	0.01
0.78	2.49	2.57
2.43	3.49	0.96

×

0.26	0.68	0.17	1.04
0.14	0.15	0.74	0.31
0.96	0.04	0.00	0.17

loadings

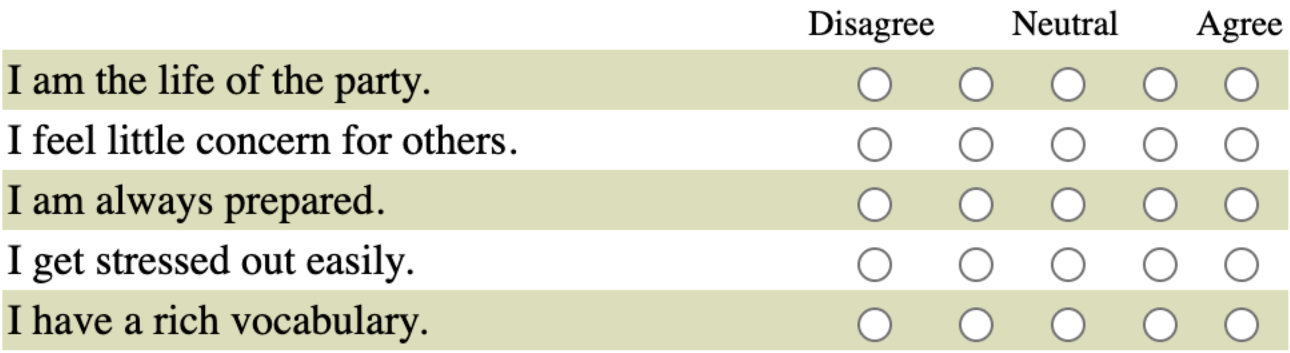
Loadings are each individual's score on each personality trait

factors

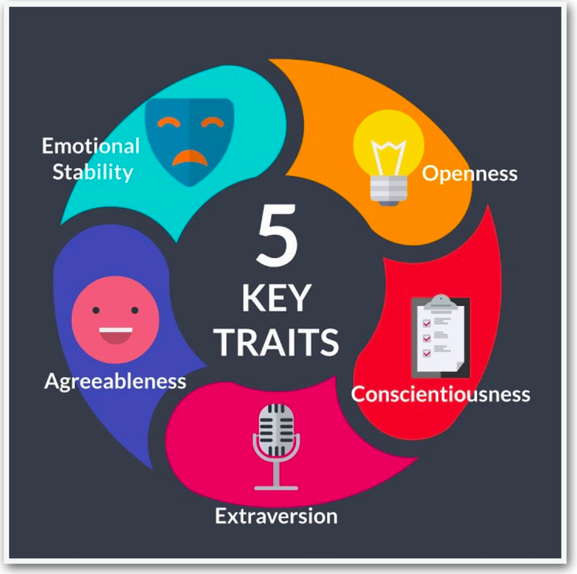
Factors are personality traits (e.x. big five) associated with each behavior

Behaviors

Test items (Agree = 5; Neutral = 3; disagree = 1)



	Disagree		Neutral		Agree
I am the life of the party.	○	○	○	○	○
I feel little concern for others.	○	○	○	○	○
I am always prepared.	○	○	○	○	○
I get stressed out easily.	○	○	○	○	○
I have a rich vocabulary.	○	○	○	○	○



You can try the test at here: https://openpsychometrics.org/tests/IPIP-BFFM/

- In personality psychology, PCA is often used to discover latent factors from individual answers to behavioral questionnaires.
- The identified factors in this situation are personality traits, which are very useful knowledge of individual differences, and it can be used to predict human behaviors in the future.

PCA / Matrix factorization in genomics

Dimensional reduction on genomic assay

	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**
Gene 1	5	3	1	5
Gene 2	1	2	3	3
Gene 3	3	1	2	2
Gene 4	2	2	3	4

Quantified gene expression values across samples

=

EG1	EG2	EG3
4.07	0.37	4.08
1.98	3.55	0.01
0.78	2.49	2.57
2.43	3.49	0.96

loadings

×

0.26	0.68	0.17	1.04	EG1
0.14	0.15	0.74	0.31	EG2
0.96	0.04	0.00	0.17	EG3

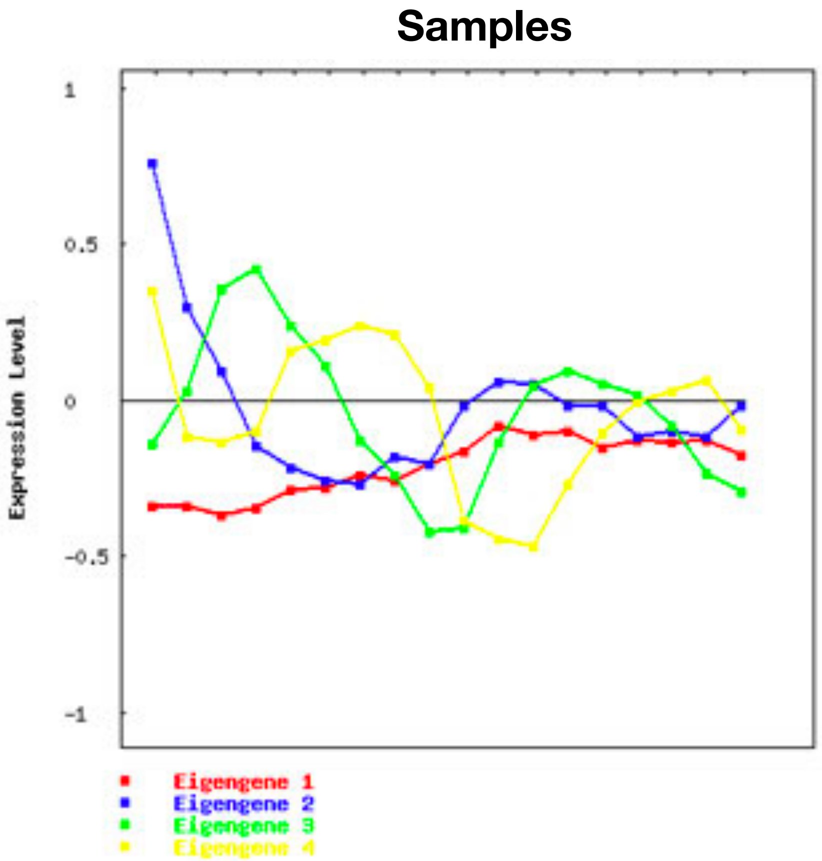
factors

Factors are **eigengenes**, which are the modular gene expression changing patterns across the samples.

- In genomics, PCA and matrix factorizations will return the factors of "**eigengenes**".

- A eigengene can be understood as the characteristic gene expression pattern of a **gene module**.

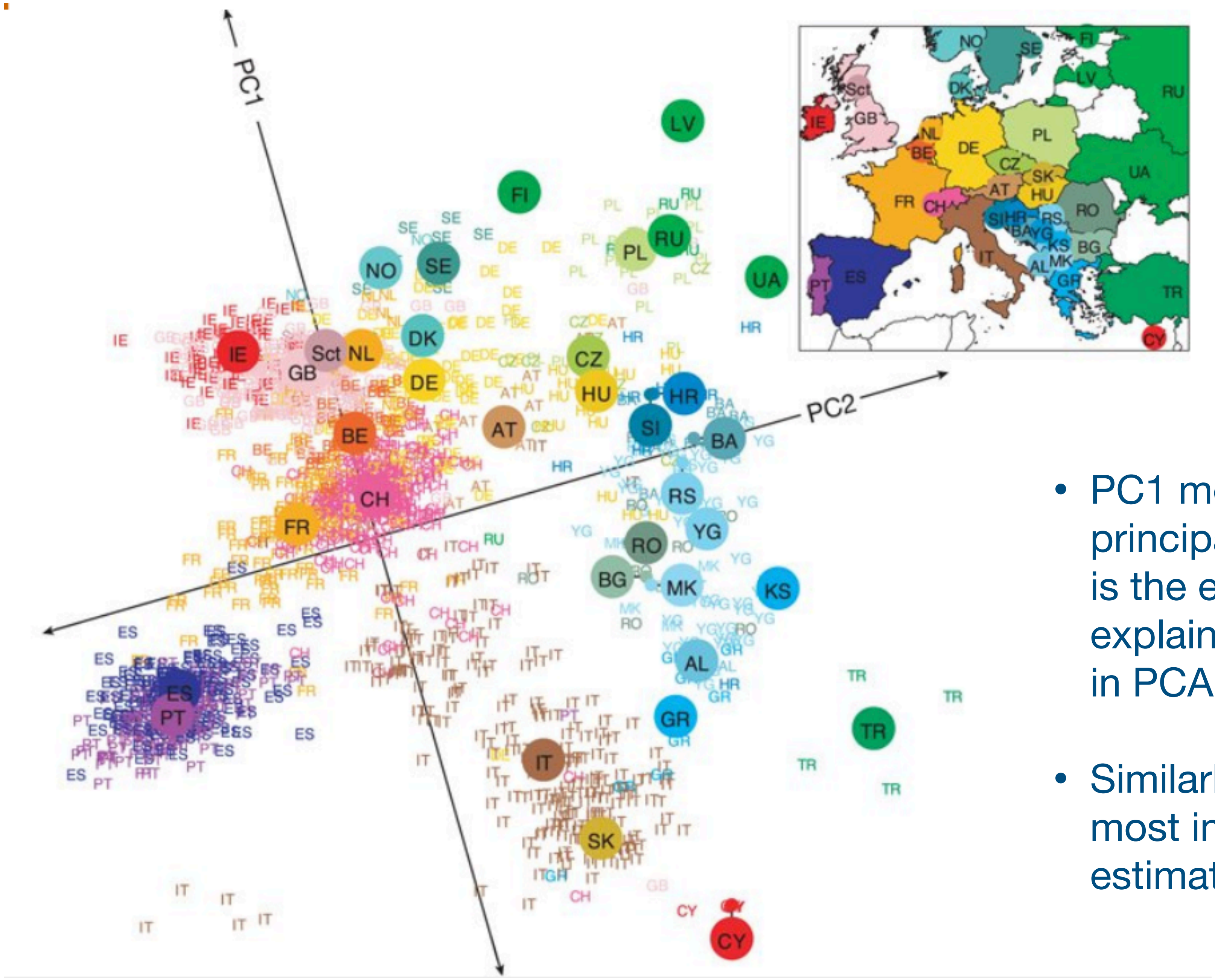
- Eigengenes are low dimensional representations of the gene expression matrix.



Applications of PCA in genomics

How can we use the eigengenes?

Visualizing genomic assay in 2D

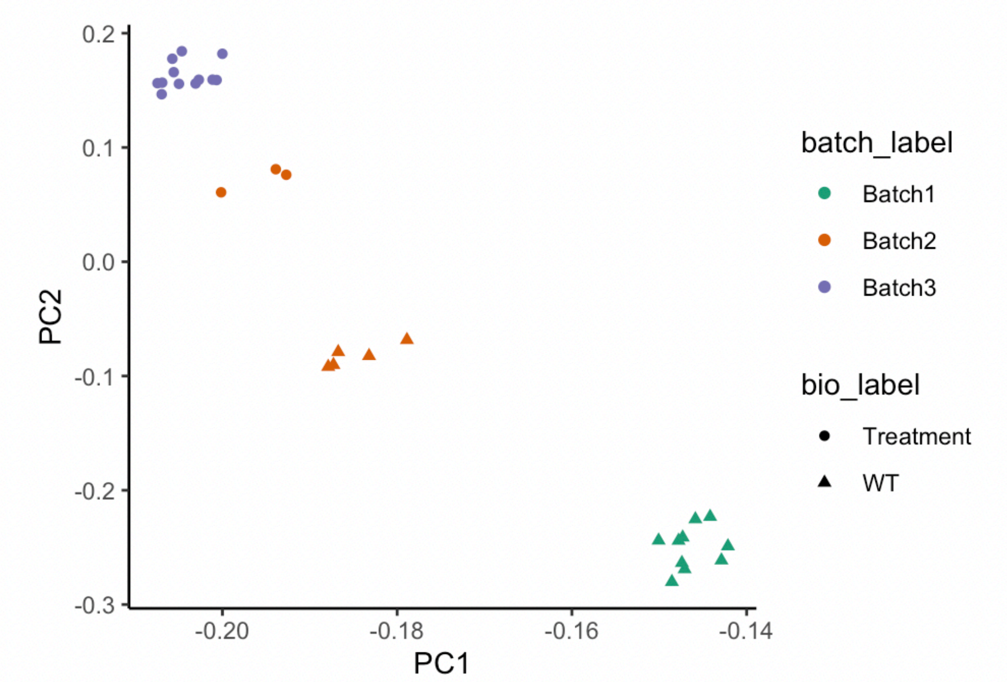


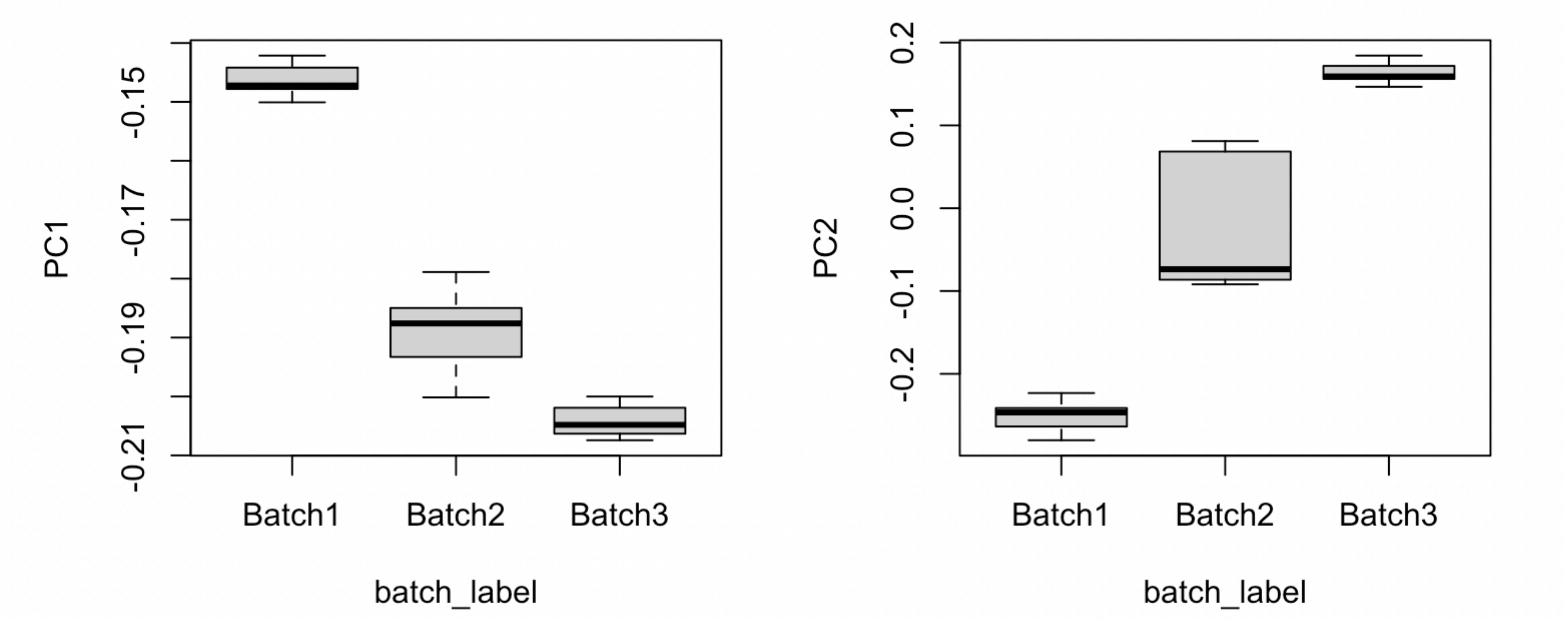
- PC1 means the first principal component, which is the eigengene that can explain the most variances in PCA.
- Similarly, PC2 is the 2nd most important eigengene estimated by PCA.

- Visualization of gene expression profiles among 1400 europeans; colors represent individuals of different countries & ethnicity groups.

How can we use the eigengenes?

Estimation and correction for the batch effect



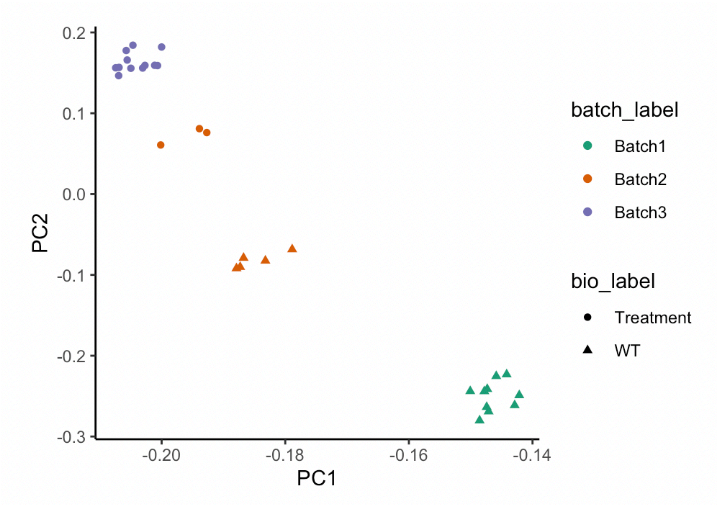


The idea is that, in heterogenous data set, the top eigengenes are often batch factors.

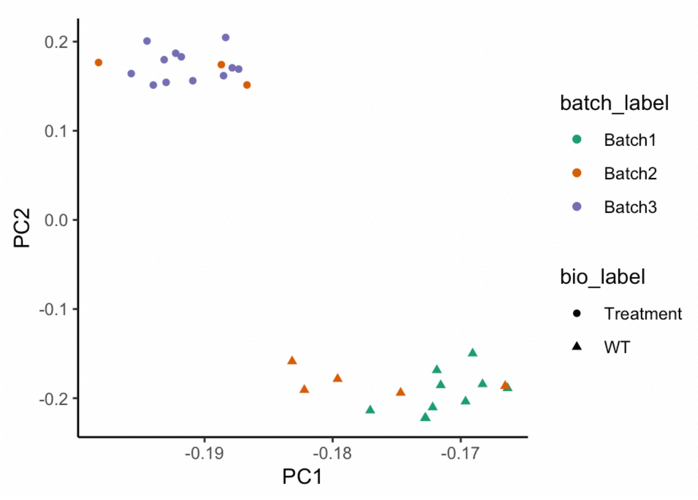
How can we use the eigengenes?

Estimation and correction for the batch effect

Before PCC



After PCC



The principal component correction (PCC) is a method used to correct for batch effects in gene expression data.

- The PCC involves two main steps:

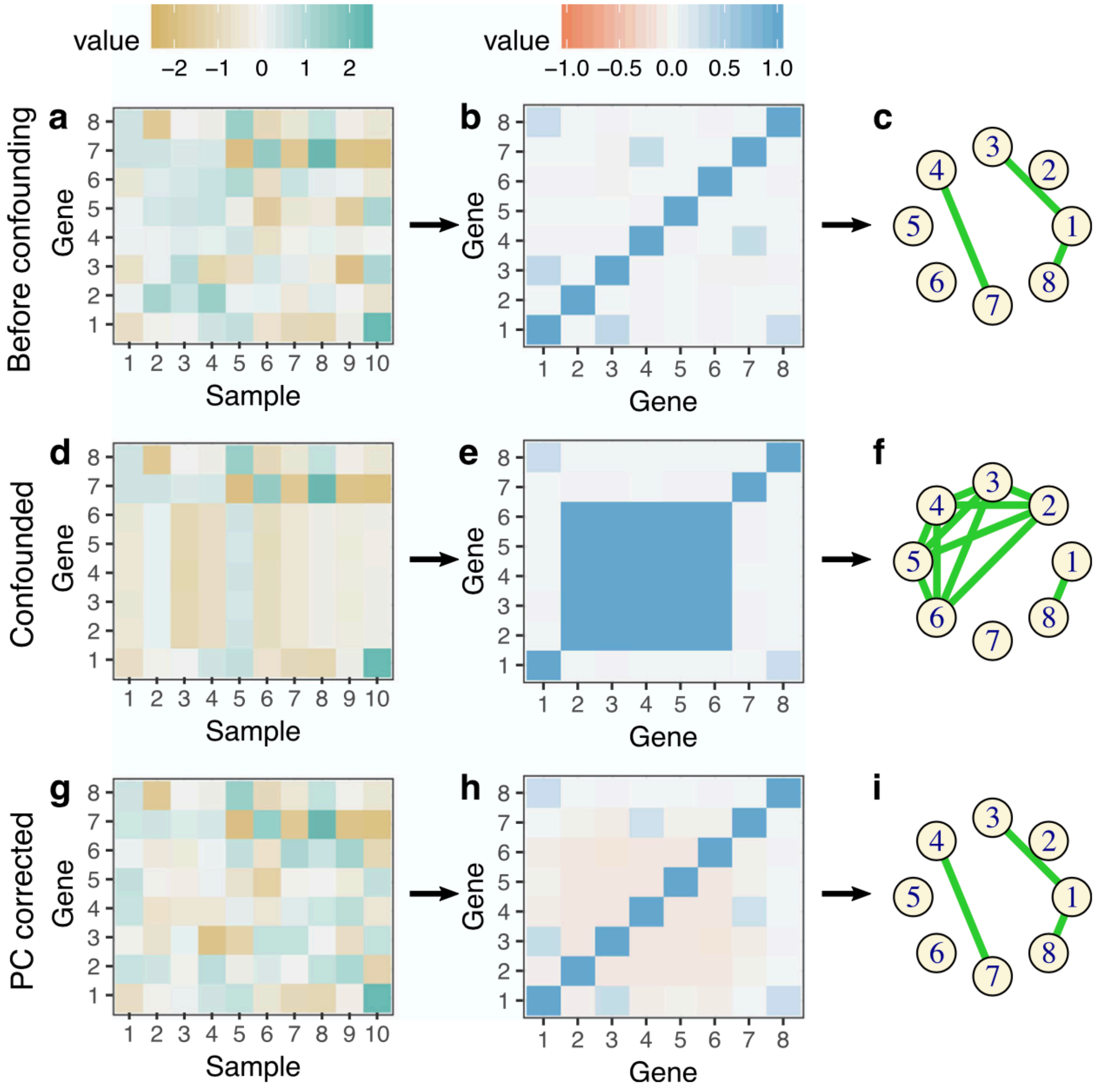
  Step 1: Perform a principal component analysis (PCA) on the normalized expression matrix to obtain the principal components (PCs). The number of top PCs (p) to use is usually determined by a method in the SVA (surrogate variable analysis) package.

  Step 2: For each gene, regress the top p PCs using multiple linear regression. The corrected expression values are the residuals of the fitted models.

- The PCC is an effective way to correct for batch effects and other unwanted technical variation in gene expression data, and is widely used in genomic research.

How can we use the eigengenes?

Estimation and correction for the batch effect

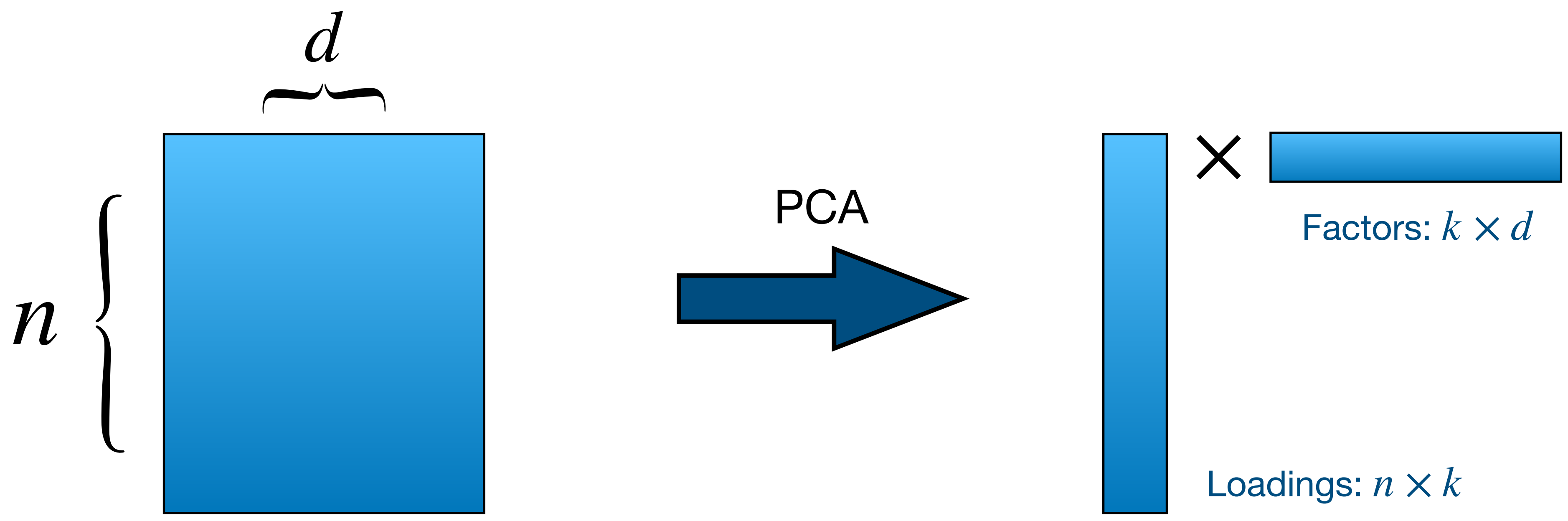


Parsana, Princy, et al. "Addressing confounding artifacts in reconstruction of gene co-expression networks." Genome biology 20.1 (2019): 1-6.

- Figures a-e show that the reconstruction of gene co-expression networks is affected by confounders (batch effects); the edges in the network are formed by Pearson correlation between genes > a threshold.
- Figures g-h demonstrate the true underlying network structure can be reconstructed after principal component correction of the gene expression data.

Dimensional reduction techniques

Understanding PCA as a linear “compaction” algorithm



A high dimensional data of $n \times d$, where for example $d > 1000$

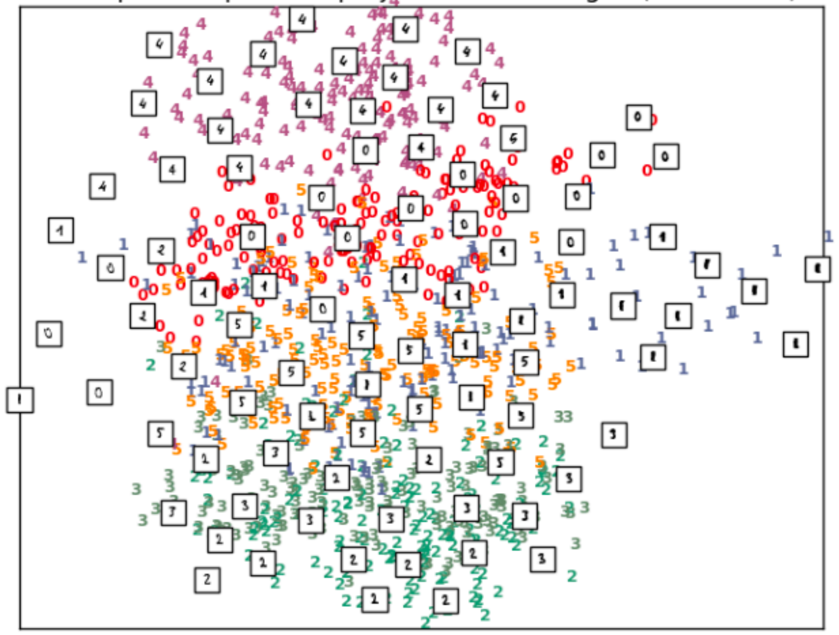
Compacting data into the linear combination of loadings and factors, which are two low dimensional matrixes ($k << d$).

- We don’t need that many movies, all we need is a few key movie attributes.
- We don’t need that many behavioral indicators, all we need is a few key personality traits.
- We don’t need that many records of genes, all we need is a few key enigengenes / gene modules.

What about nonlinear dimensional reduction?

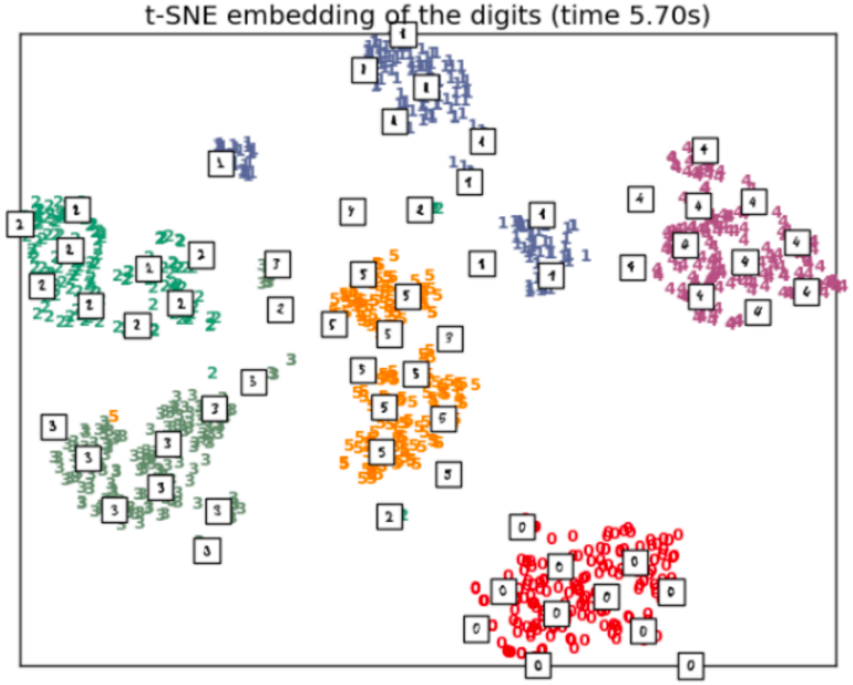
Data of hand-written digits

PCA



Principal Components projection of the digits (time 0.01s)

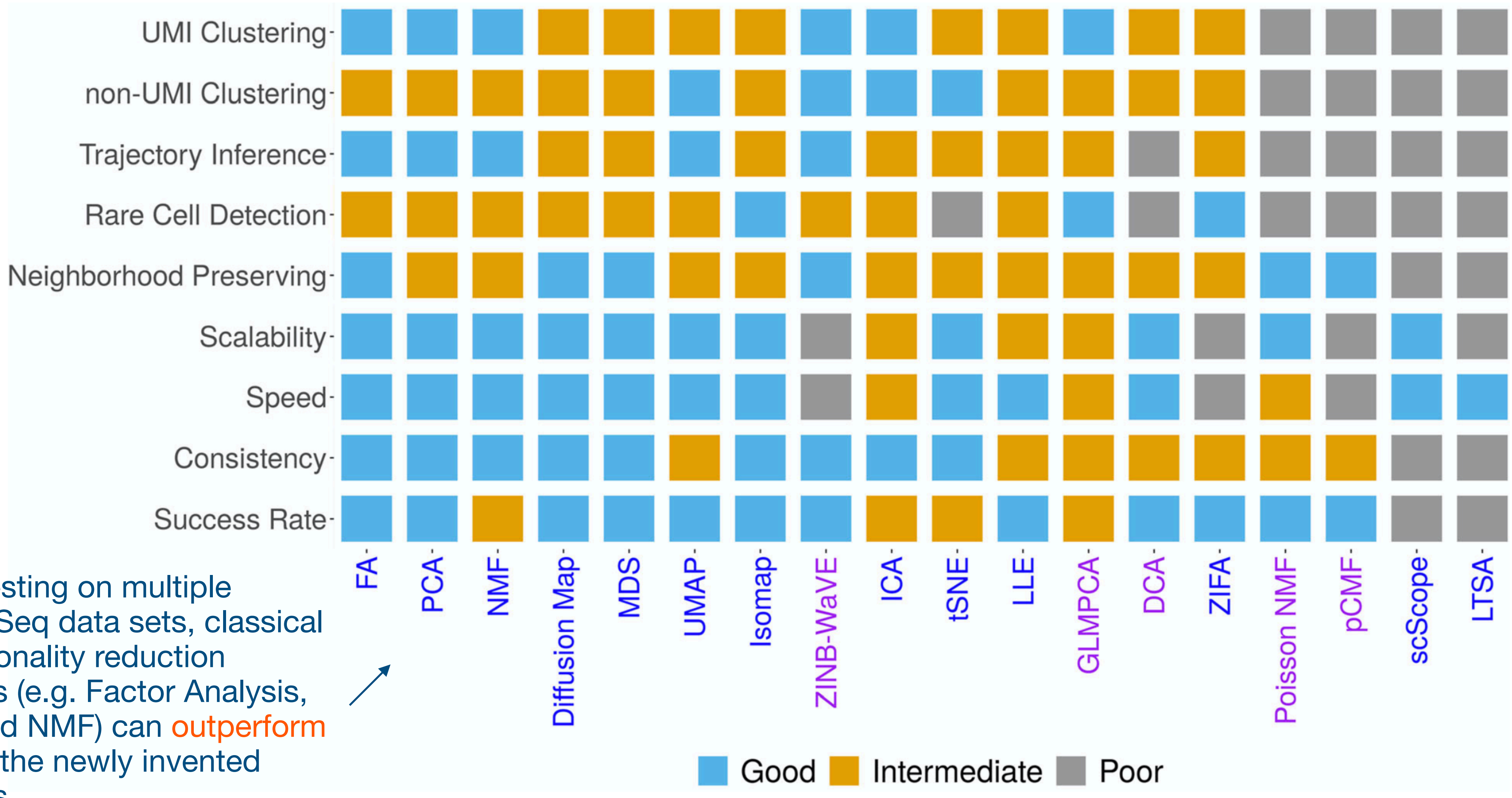
t-SNE



t-SNE embedding of the digits (time 5.70s)

Method	Principle	Advantage	Disadvantage
PCA	Finding low dimensional projections that spread data as much as possible.	High interpretability as factor analysis	Work less well for non-linear patterns
tSNE / UMAP	Non-linear embedding that keep close-by points close using a probabilistic objective.	Can learn complex non-linear relationships	Axes have no meanings

Performance comparison of dimensional reduction methods in scRNA-Seq



	FA	PCA	NMF	Diffusion Map	MDS	UMAP	Isomap	ZINB-WaVE	ICA	tSNE	LLE	GLMPCA	DCA	ZIFA	Poisson NMF	pCMF	scScope	LTSA
UMI Clustering	🟦	🟦	🟦	🟧	🟧	🟧	🟧	🟦	🟦	🟧	🟧	🟦	🟧	🟧	⬛	⬛	⬛	⬛
non-UMI Clustering	🟧	🟧	🟧	🟧	🟧	🟦	🟧	🟦	🟦	🟦	🟧	🟧	🟧	🟧	⬛	⬛	⬛	⬛
Trajectory Inference	🟦	🟦	🟦	🟧	🟧	🟦	🟧	🟦	🟧	🟧	🟧	🟧	⬛	🟧	⬛	⬛	⬛	⬛
Rare Cell Detection	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟧	🟧	⬛	🟧	🟦	⬛	🟦	⬛	⬛	⬛	⬛
Neighborhood Preserving	🟦	🟧	🟧	🟦	🟦	🟧	🟧	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟦	⬛	⬛
Scalability	🟦	🟦	🟦	🟦	🟦	🟦	🟦	⬛	🟧	🟦	🟧	🟧	🟦	⬛	🟦	⬛	🟦	⬛
Speed	🟦	🟦	🟦	🟦	🟦	🟦	🟦	⬛	🟧	🟦	🟦	🟧	🟦	⬛	🟧	⬛	🟦	🟦
Consistency	🟦	🟦	🟦	🟦	🟦	🟧	🟦	🟦	🟦	🟦	🟧	🟧	🟧	🟧	🟧	🟧	⬛	⬛
Success Rate	🟦	🟦	🟧	🟦	🟦	🟦	🟦	🟦	🟧	🟧	🟦	🟧	🟦	🟦	🟦	🟦	⬛	⬛

🟦 Good 🟧 Intermediate ⬛ Poor

When testing on multiple scRNA-Seq data sets, classical dimensionality reduction methods (e.g. Factor Analysis, PCA, and NMF) can outperform most of the newly invented methods.

- Most mathematically complex methods fail to perform well in new data sets.

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome biology, 20(1), 1-16.