EDA: Principal Component Analysis (PCA)

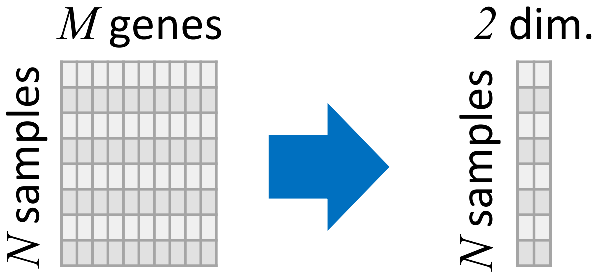
RNA-seq data analysis

Paulo Czarnewski
Julie Lorent

Why PCA?

Simplify complexity, so it becomes easier to work with.

Reduce number of features (genes)



“Remove” redundancies in the data

Identify the most relevant information

Find and filter noise

Detect data quality outliers or batches

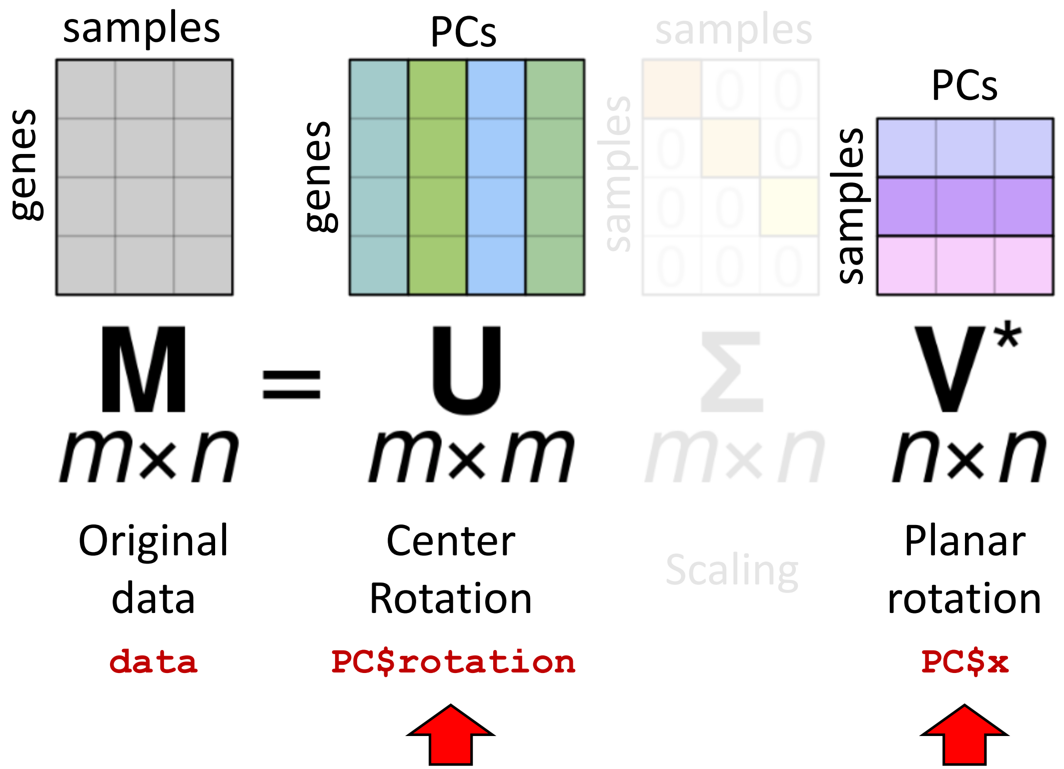
Data visualization

How PCA works

It is an algebraic method of dimensionality reduction (only numerical variables).

Relationships between variables need to be linear to apply PCA. For RNAseq data, it is recommended to have variables on a log-scale with VST or RLOG transformation.

It is a case inside Singular Value Decomposition (SVD) method (data compression)



$$\underset{m\times n}{\mathbf{M}} = \underset{m\times m}{\mathbf{U}}\;\underset{m\times n}{\Sigma}\;\underset{n\times n}{\mathbf{V}^{*}}$$

Original data

data

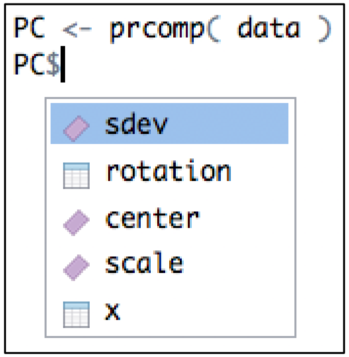
Center Rotation

PC$rotation

Scaling

Planar rotation

PC$x



```
PC <- prcomp( data )
PC$
```

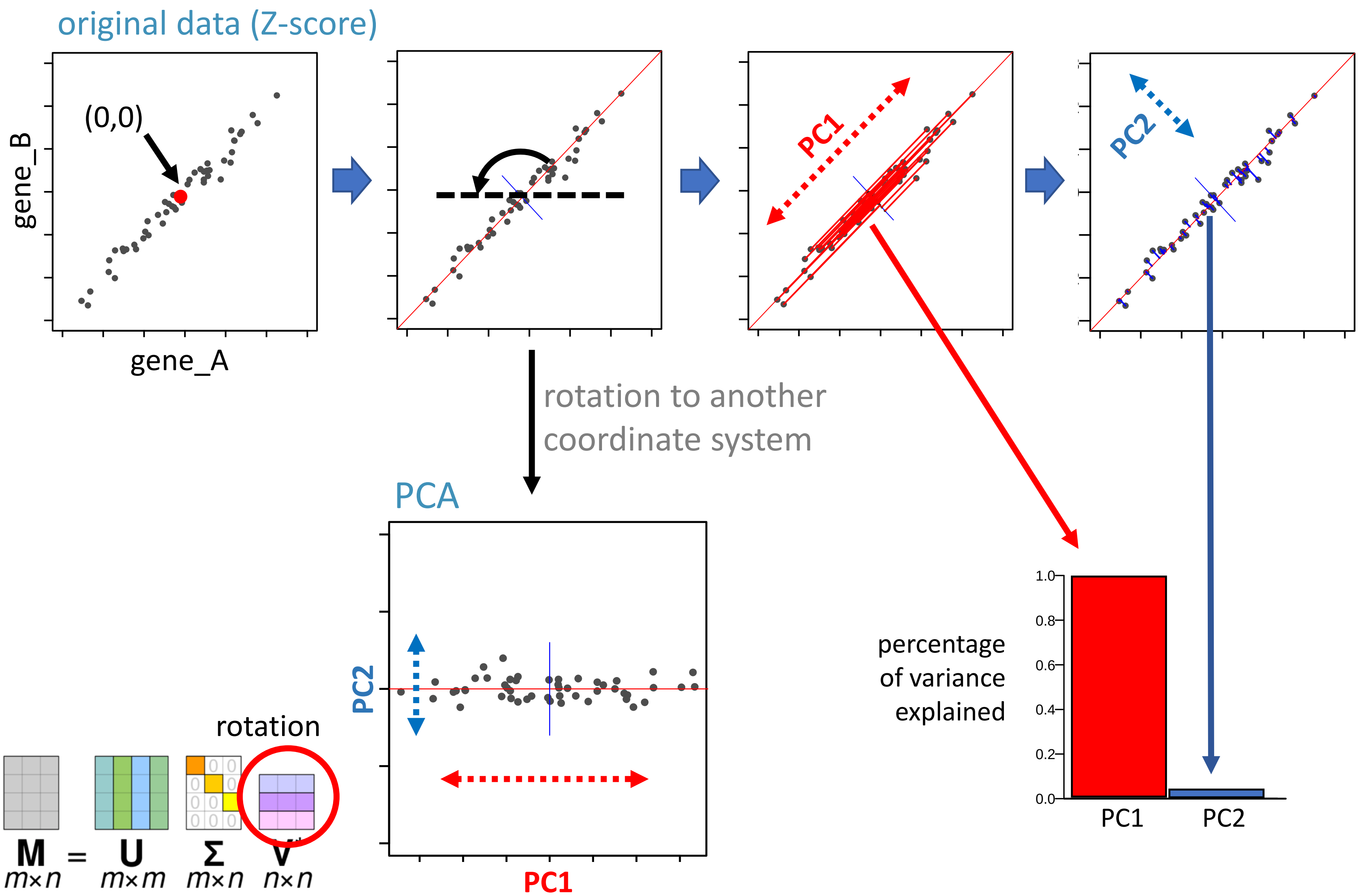
How PCA works

Transform the data into a space of smaller dimension which would *summarize* the data in the most *relevant* way.

In PCA, *relevance* is measured by the variance (spread)
Among the n-dimensional space, PCA will find the direction with highest variance
Why does PCA consider that higher variance = more relevant?

Typical question	PCA perspective
Is there any significant differences in expression of some genes between my study groups?	Would an "axis" with large spread in my data distinguish samples from one study group to another?
Is there any major outlier among my samples?	Would an "axis" with large spread in my data distinguish "failed experiment" samples from successful ones?
Is there some batch effect in my data (are technical differences larger than biological differences)?	Would an "axis" with large spread in my data distinguish samples from one batch to another?

How PCA works



original data (Z-score)

rotation to another coordinate system

PCA

percentage of variance explained

rotation

How PCA works

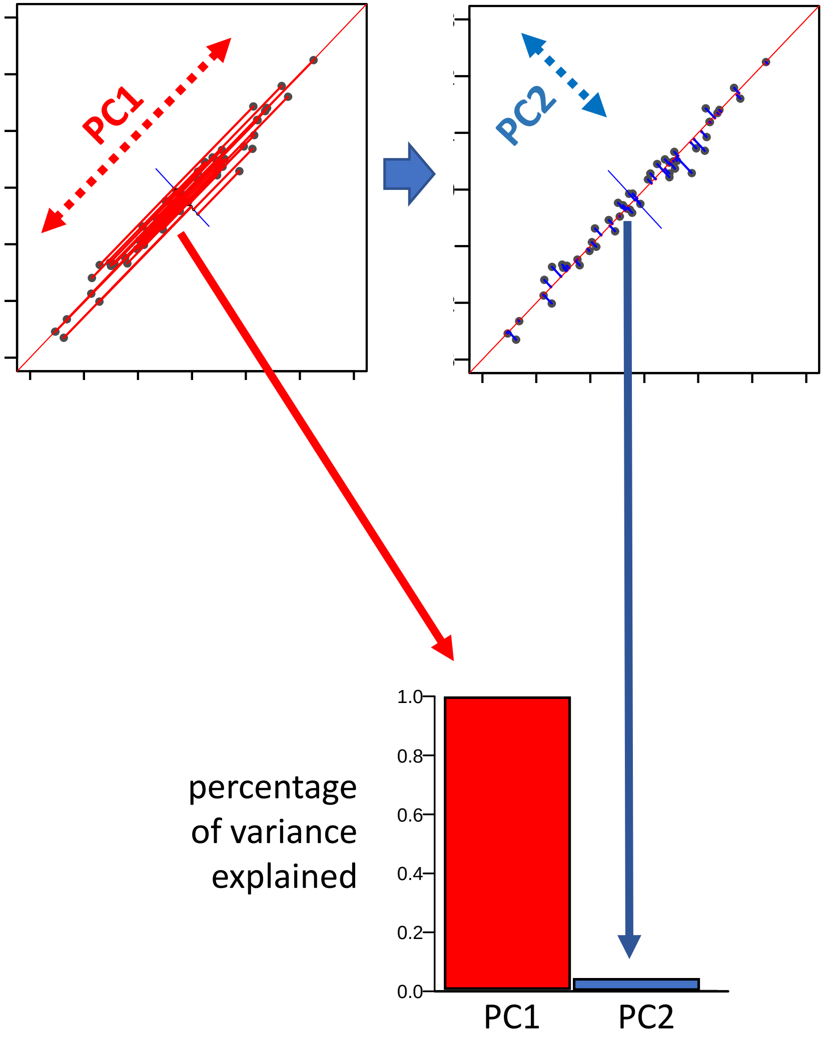
PC1 explains >98% of the variance

1 PC thus represents 2 genes very well

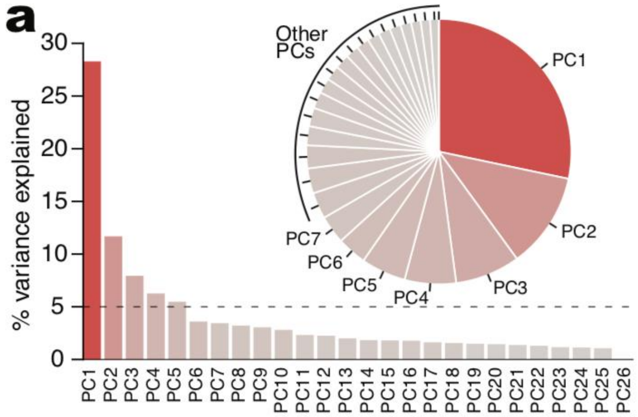
“Removing” redundancy

PC2 is nearly insignificant in this example

Could be disregarded



In real life …

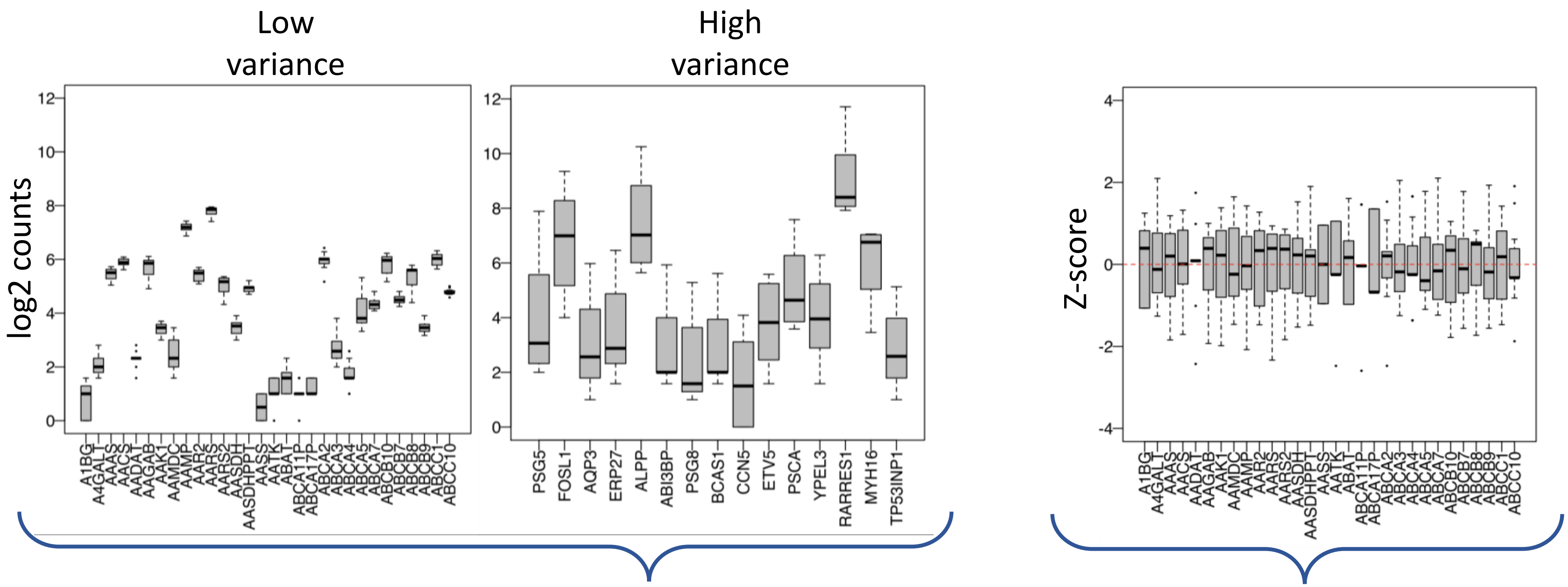


Czarnewski *et al* 2019

Data transformation and scaling

Before applying PCA, the data should be first transformed VST or RLOG

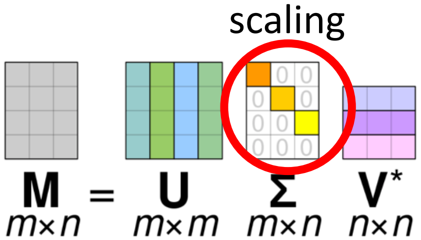
Each feature can be centered and scaled to have a similar center (zero) and similar deviation.



PCA on raw counts will separate genes with <u>higher counts</u> in the first PCs

(higher distance to 0)

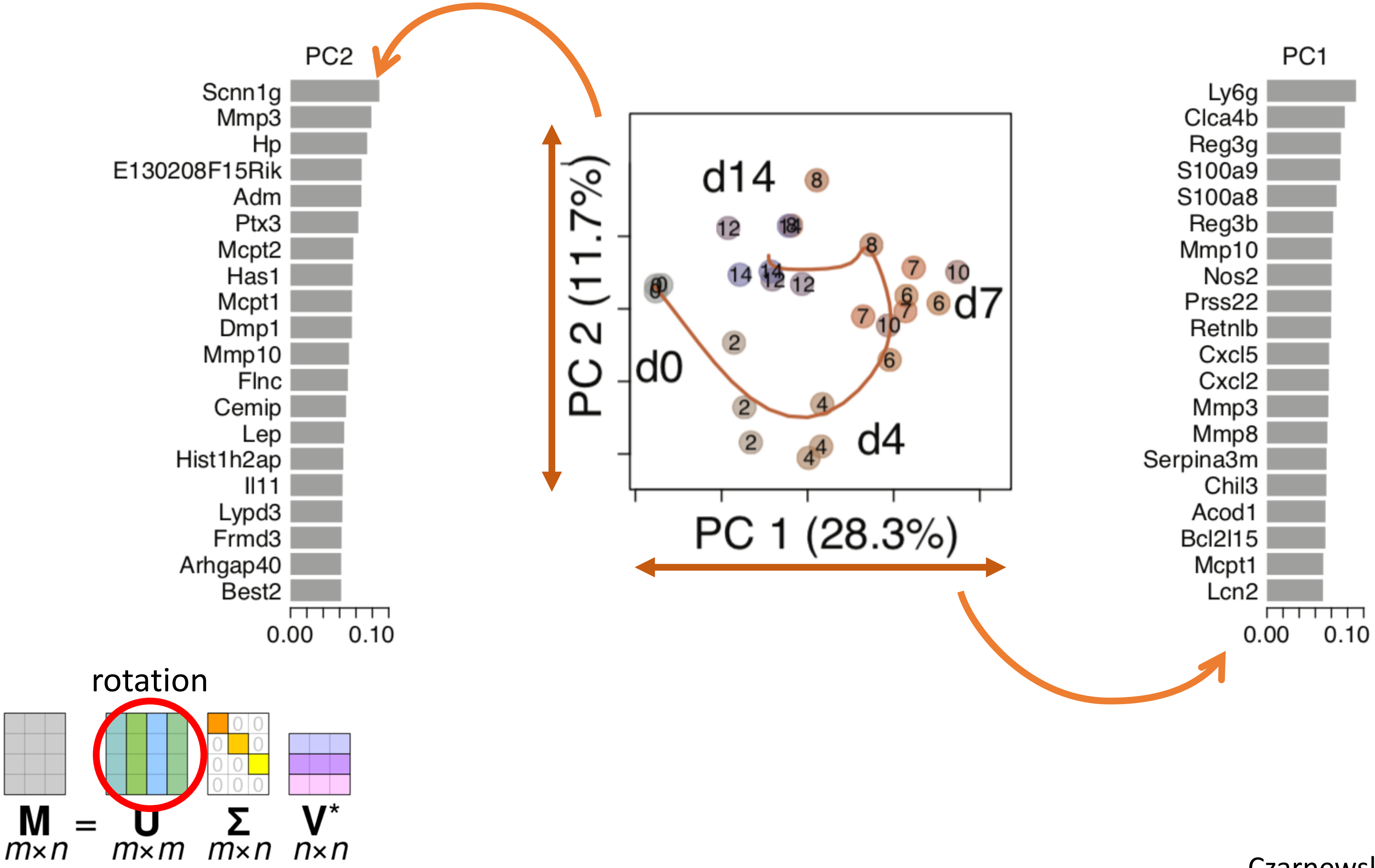
PCA on Z-score will separate genes with most <u>common expression trends</u> in the first PCs



$$\underset{m\times n}{\mathbf{M}} = \underset{m\times m}{\mathbf{U}}\ \underset{m\times n}{\mathbf{\Sigma}}\ \underset{n\times n}{\mathbf{V}^*}$$

How PCA works

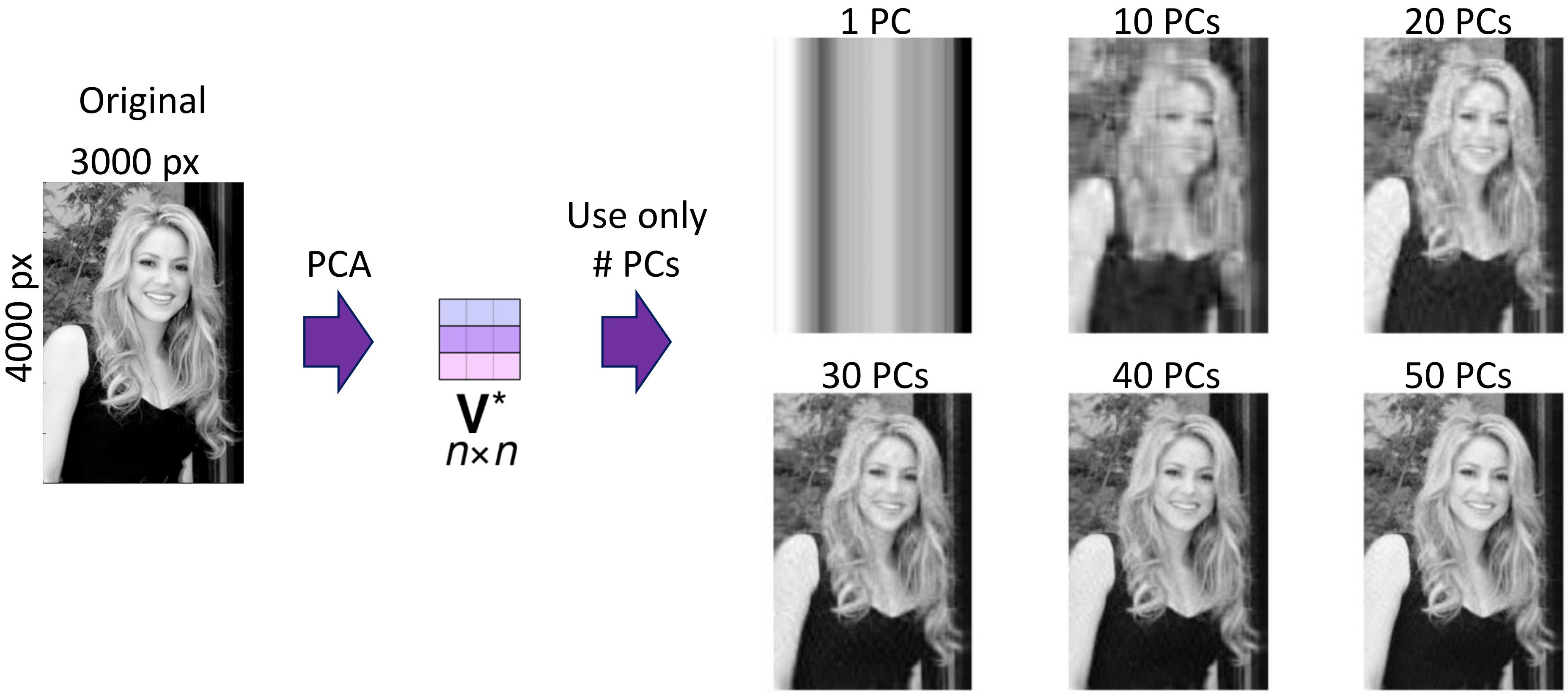
Interpretability of principal components



Czarnewski *et al* 2019

A visual intuition of PCA

The top principal components store more important ~~Shakira~~ information



Original

3000 px

4000 px

PCA

$\mathbf{V}^{*}_{n \times n}$

Use only # PCs

1 PC

10 PCs

20 PCs

30 PCs

40 PCs

50 PCs

https://glowingpython.blogspot.com/2011/07/pca-and-image-compression-with-numpy.html

PCA summary

PCA is a dimensionality reduction method for numerical variables

Linear combinations -> when relationships between variables are non-linear, PCA is not recommended.

The data is usually SCALED and TRANSFORMED (i.e. VST/RLOG) prior to PCA

It is an interpretable dimensionality reduction

The top principal components contain higher variance from the data and PCA preserves the whole variance

Can be used as filtering, by selecting only the top significant PCs

- PCs that explain at least 1% of variance
- The first 5-10 PCs

Thank you. Questions?