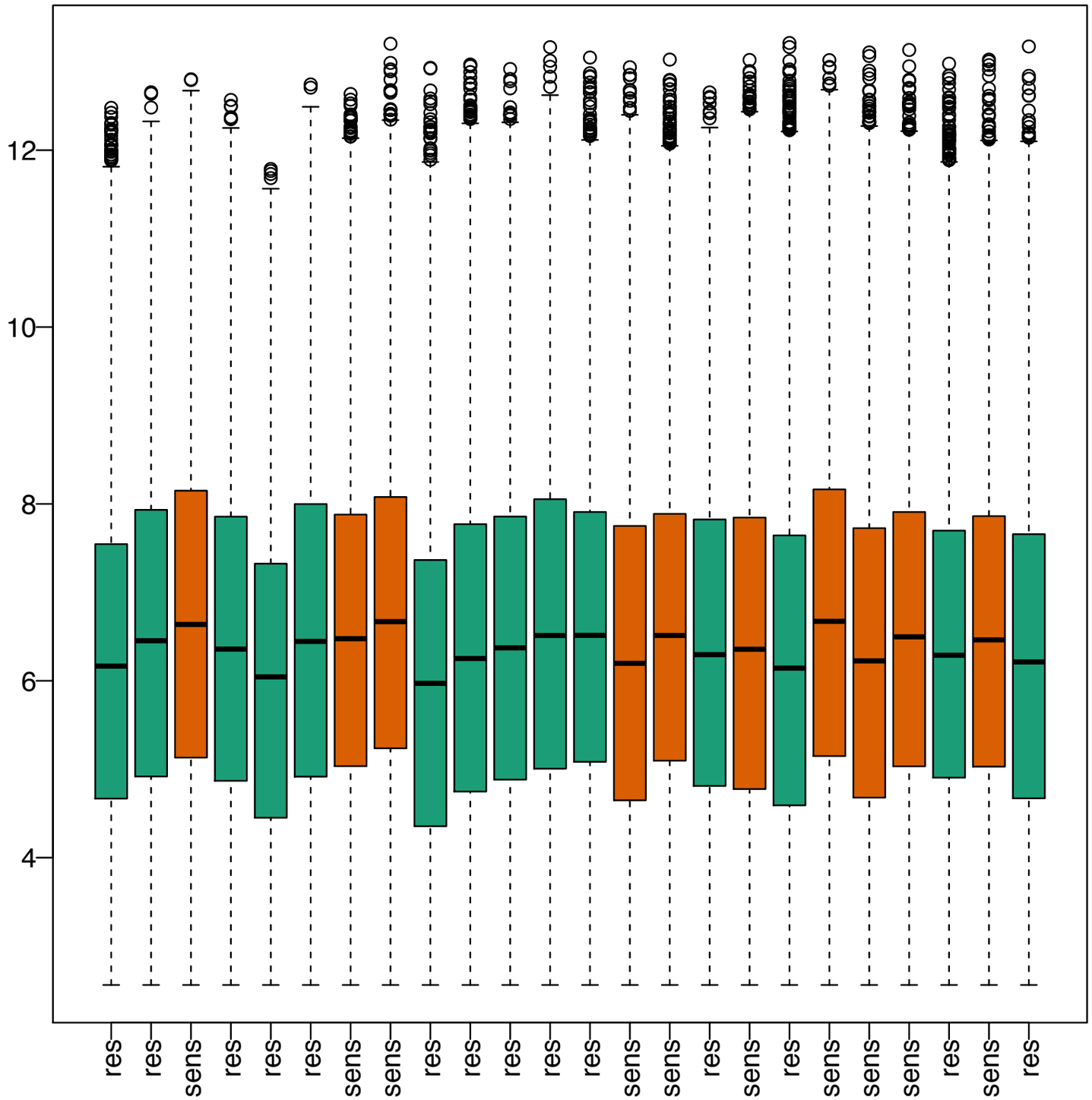
PCA and friends

CMSC702 Spring 2014
Héctor Corrada Bravo

A breast cancer experiment

- Tumors respond differently to drugs used in chemotherapy
- Experiment: measure mRNA from docetaxel resistant and docetaxel sensitive tumors and find if there are differences in expression

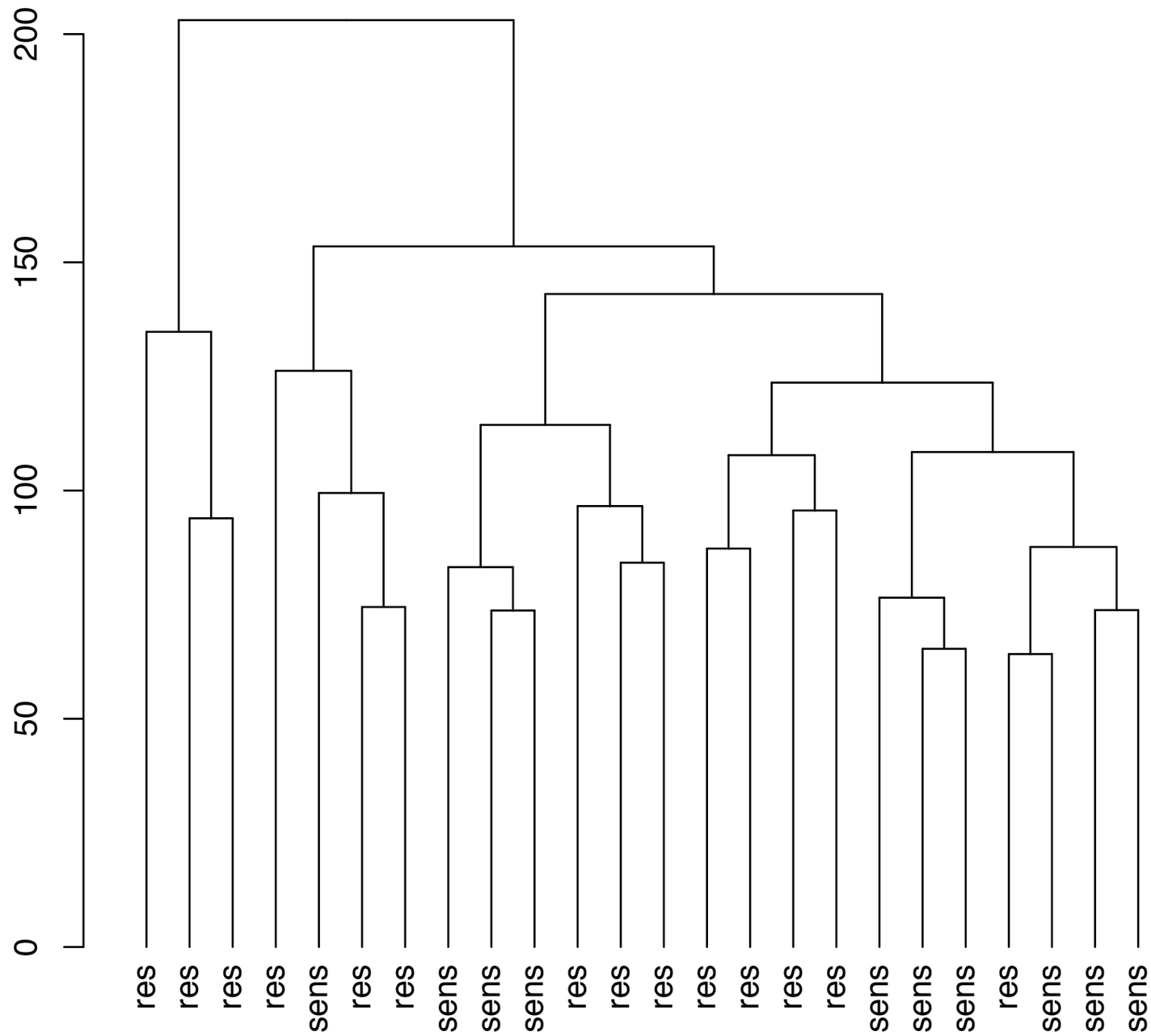
Preprocessed and normalized



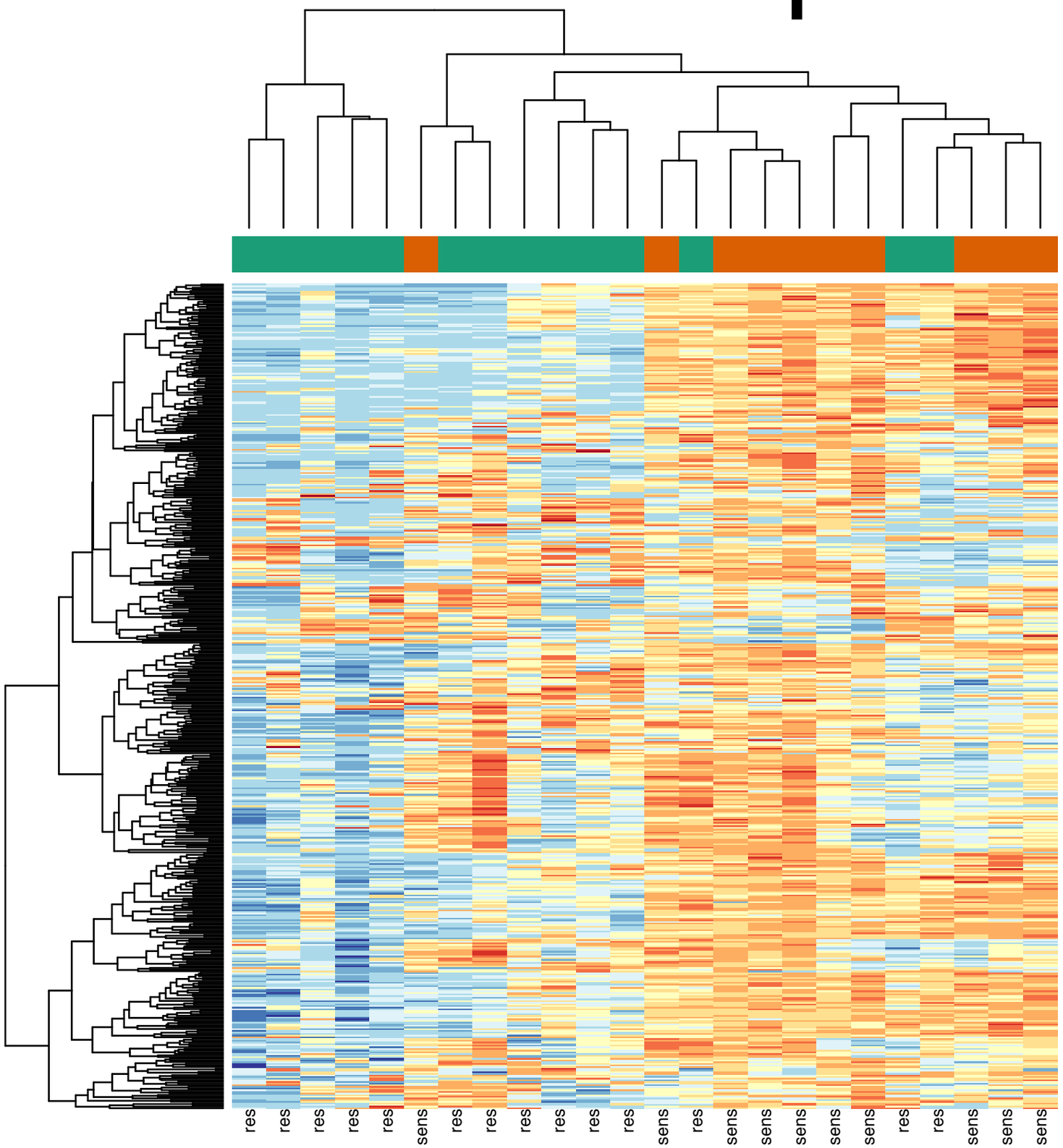
Filtering

- The Affymetrix HGU95aV2 array has ~12K probesets
- For convenience, we'll filter probesets
  - what kind of filters make sense?
  - I chose to use median absolute deviation

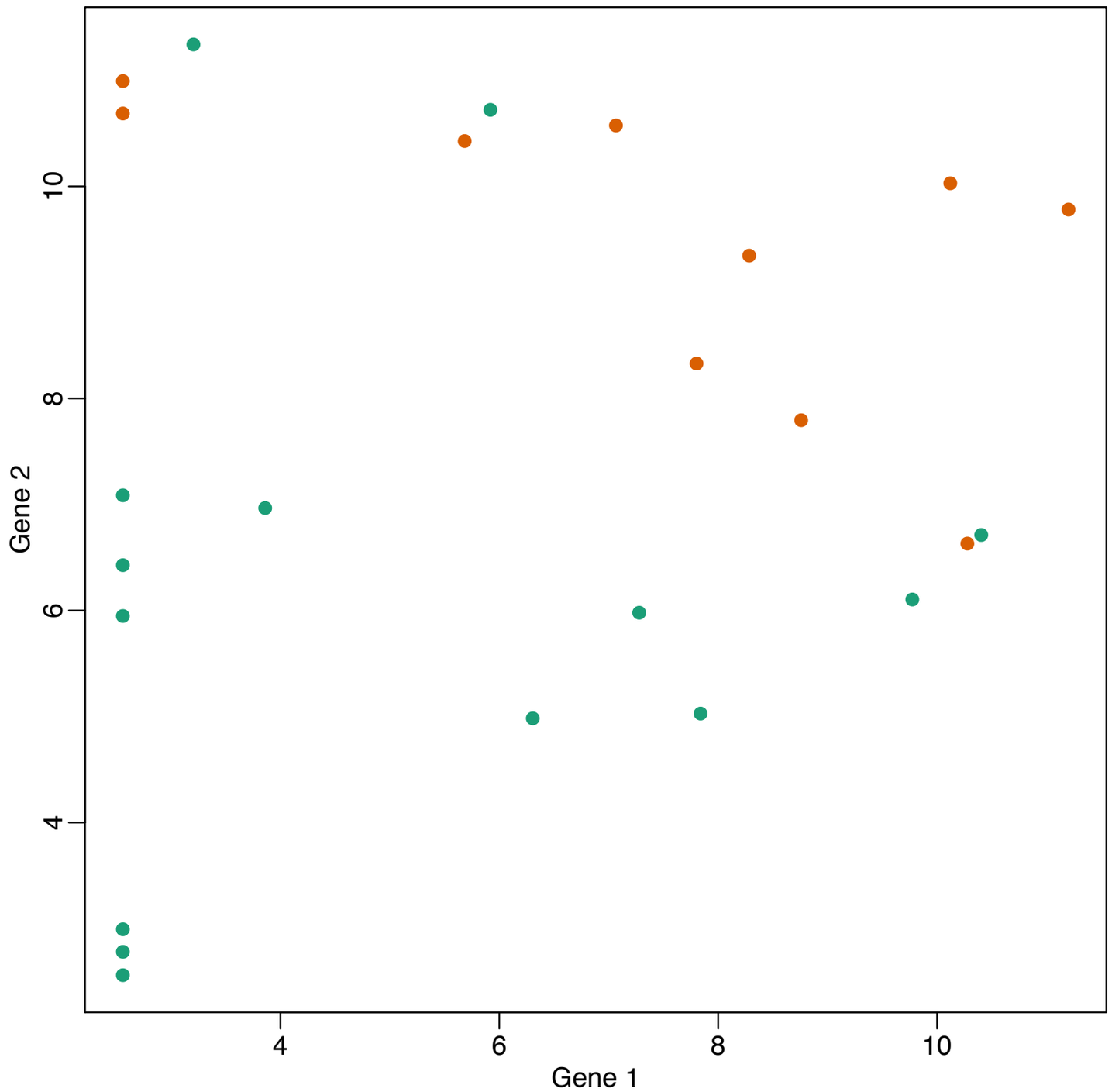
Hierarchical Clustering



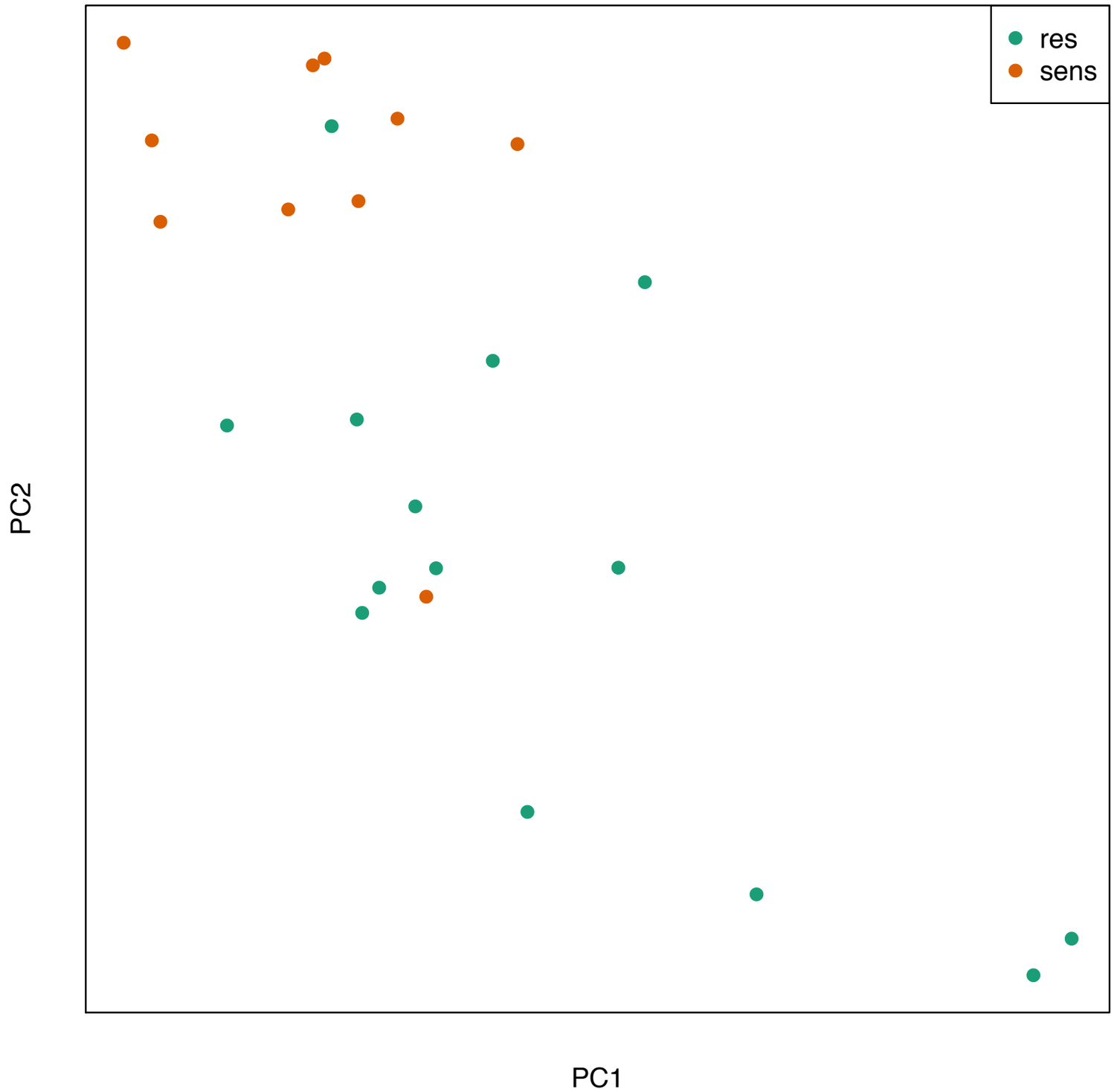
Heatmap


res
res
res
res
res
sens
res
res
res
res
res
res
sens
res
sens
sens
sens
sens
sens
res
res
sens
sens
sens

But I want to look at data



PCA

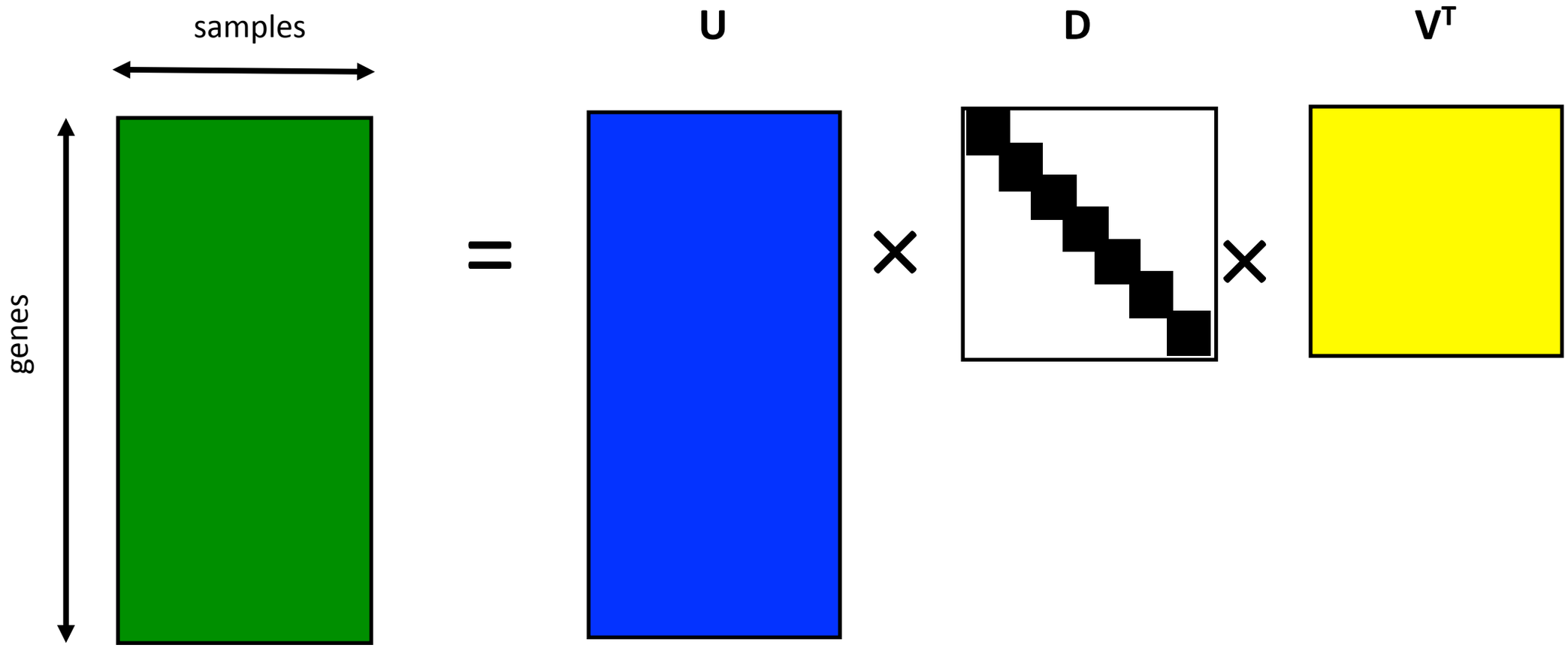


Principal Components Analysis / Singular Value Decomposition

- A method to identify patterns in the data that explain a large percentage of the variation

- First proposed for genomics by Alter et al. (2000) PNAS

Singular Value Decomposition



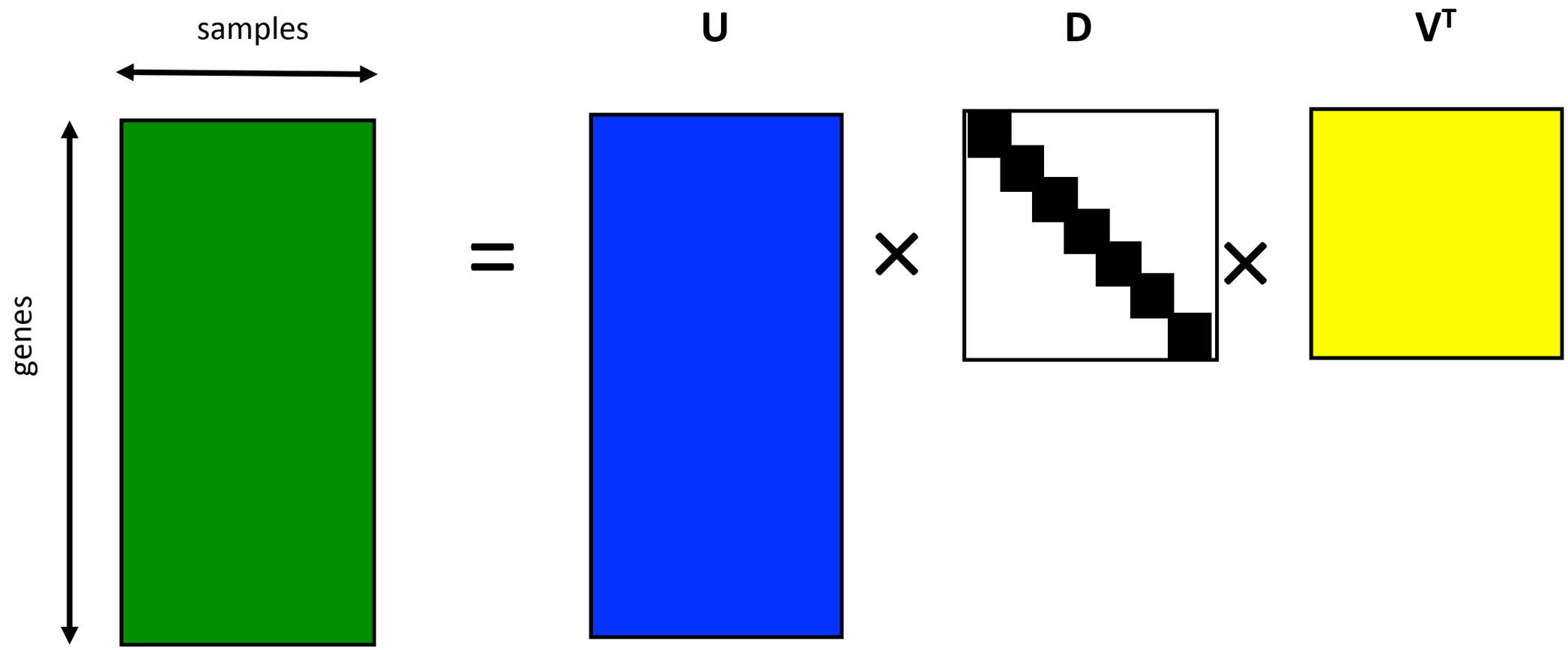
Data

eigenarrays/
left singular vectors/
loadings

singular
values

eigengenes/
right singular vectors/
principal components

Properties of SVD



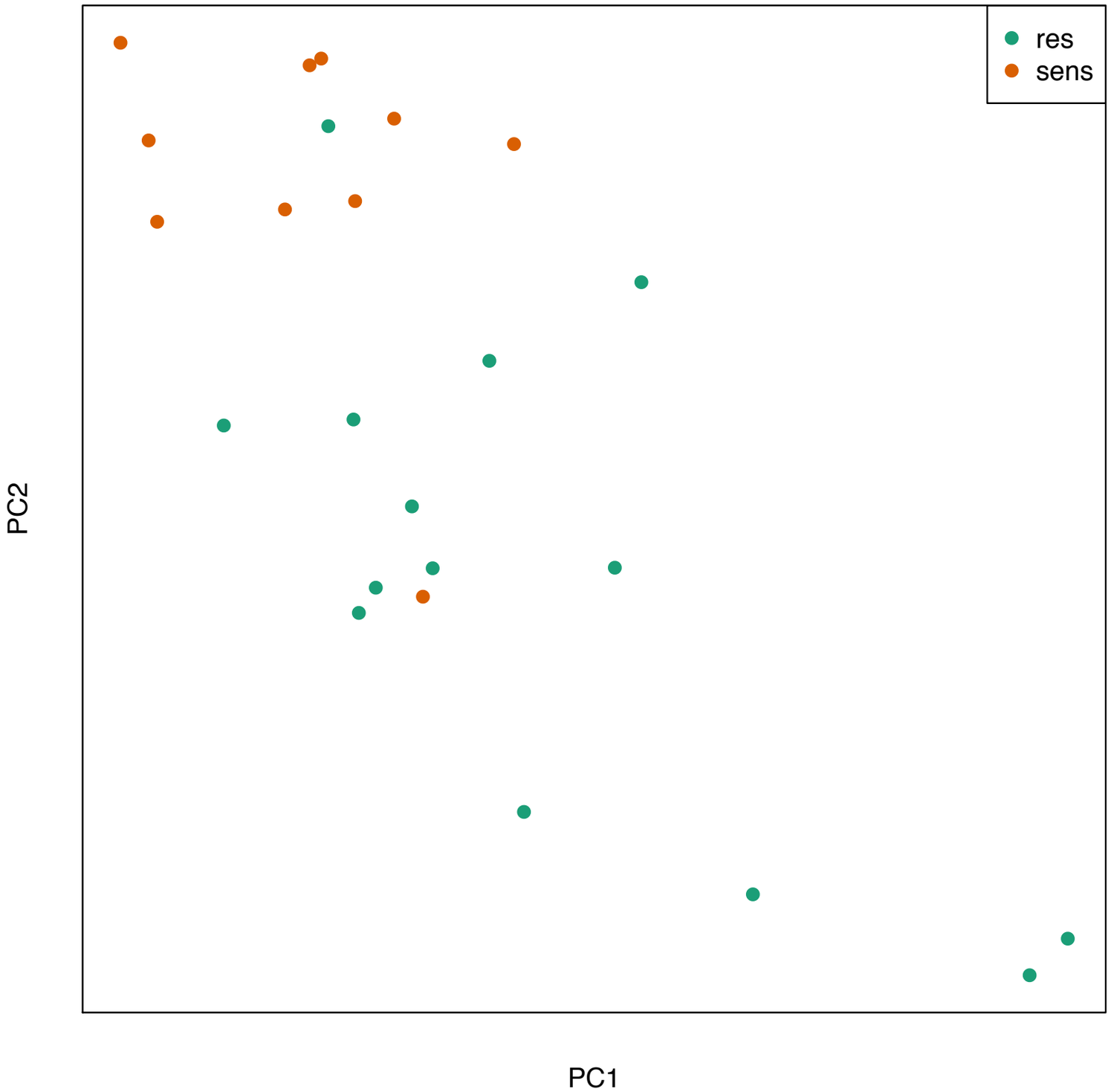
Columns of V^T/rows of U are orthogonal and calculated one at a time

Columns of V^T describe patterns across genes

Columns of U describe patterns across arrays

$d_i^2 / \sum_{i=1}^{n} d_i^2$ is the percent of variation explained by the ith column of V

PCA



PCA

- A method to identify patterns in the data that explain a large percentage of the variation
  - What patterns? *Rotation, reflection, and scaling of original data*
  - i.e. linear transformations, projection

$$X_1 = Xv$$

Centered Data

PCA

- A method to identify patterns in the data that explain a large percentage of the variation
  - What patterns? *Rotation, reflection, and scaling of original data*
  - i.e. linear transformations, projection

$$X_1 = Xv$$

  - objective: maximize variance

$$\max_u \; X_1' X_1 = v' X' X v$$

PCA

- A method to identify patterns in the data that explain a large percentage of the variation
  - What patterns? *Rotation, reflection, and scaling of original data*
  - i.e. linear transformations, projection

$$X_1 = Xv$$

  - objective: maximize variance

$$\max_{u} \quad X_1'X_1 = v'X'Xv$$

$$\text{s.t.} \qquad v'v = 1$$

Another Interpretation

- Can we find a low dimensional approximation to the data?
  - Find a basis (orthonormal vectors)
  - Find parameters to represent data in that basis
  - Approximate? Use least squares

Another Interpretation

- Can we find a low dimensional approximation to the data?
  - –Find a basis (orthonormal vectors)
  - –Find parameters to represent data in that basis

$$f(\lambda) = \mu + V_q \lambda$$

- –Approximate? Use least squares

$$\min_{\mu, \{\lambda_i\}, V_q} \sum_{i=1}^{N} \|x_i - \mu - V_q \lambda_i\|^2$$

Another Interpretation

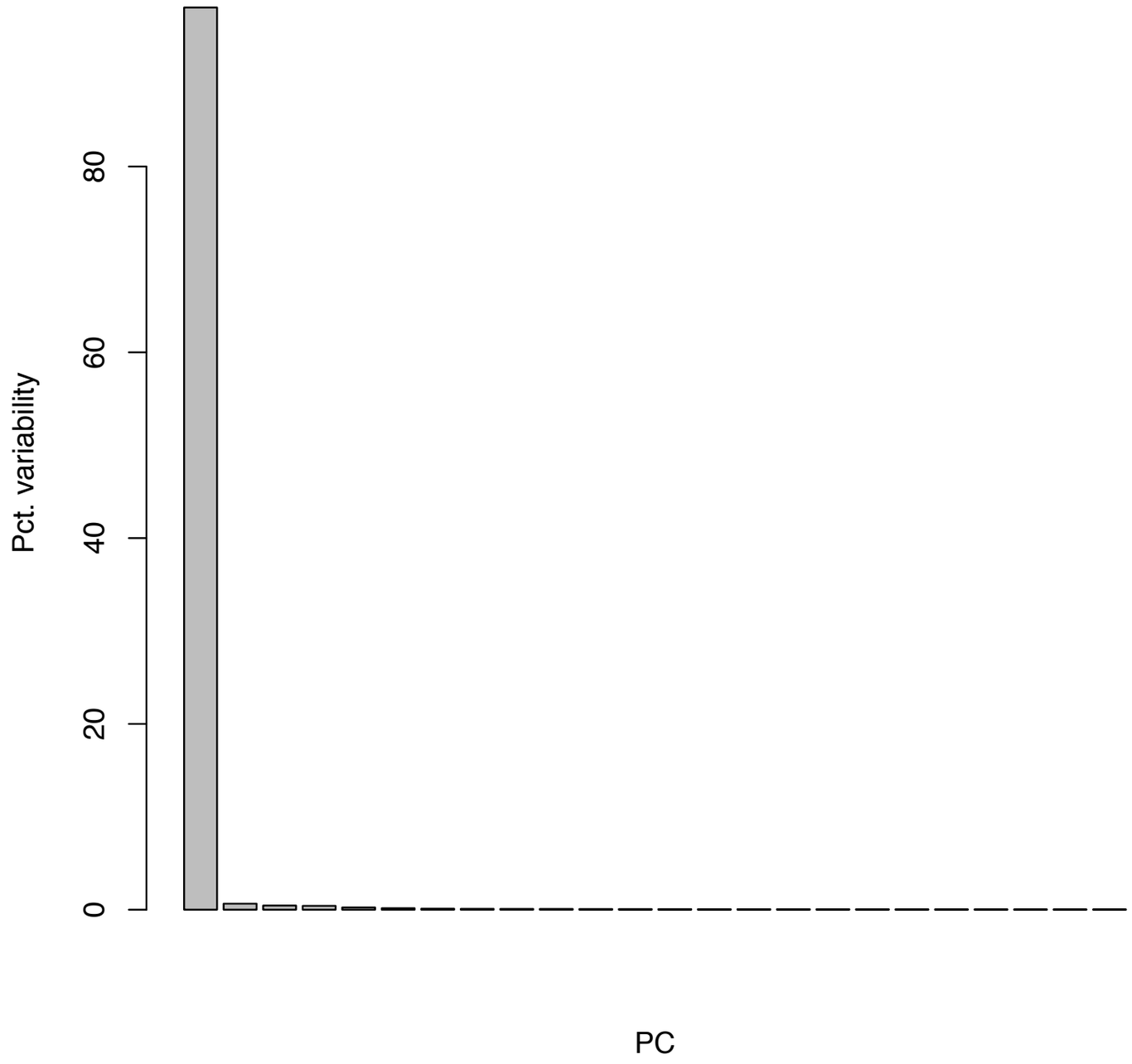
- Can we find a low dimensional approximation to the data?
- Partially solve and get

$$\min_{V_q} \quad \sum_{i=1}^{N} \|(x_i - \bar{x}) - V_q V_q'(x_i - \bar{x})\|^2$$

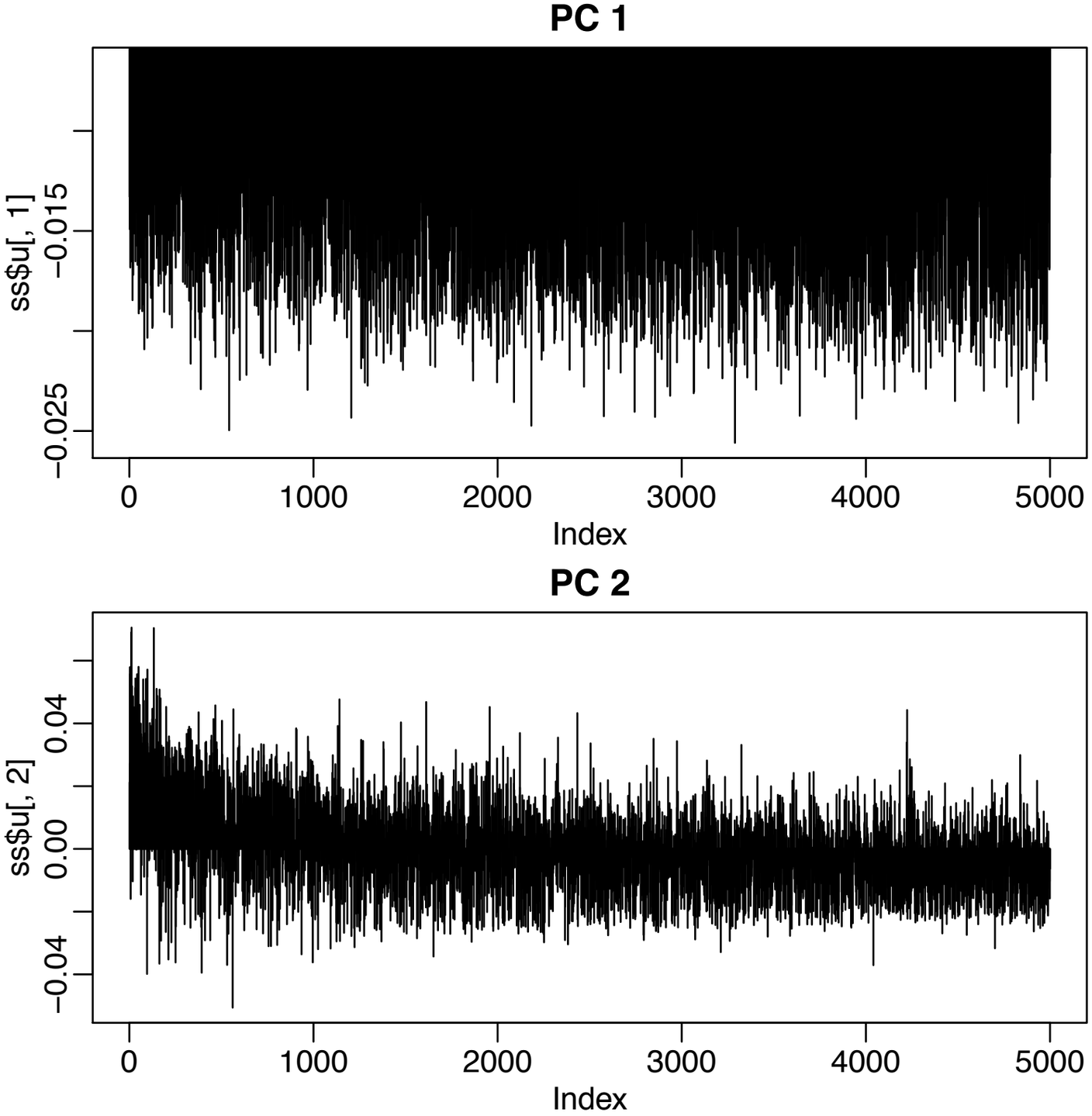
$$\text{s.t.} \qquad V_q' V_q = I$$

- What is the solution to this? The SVD…

Interpreting PCA



Interpreting PCA



Sparse PCA

- We found a low dimensional approximation to data
- We know one direction explains most of the variation
- What can we say about genes though?
- Idea: can we get away with using a small number of genes to get approximation

Sparse PCA via L1 regularization

- Idea, penalize basis vectors, so only a few genes are included

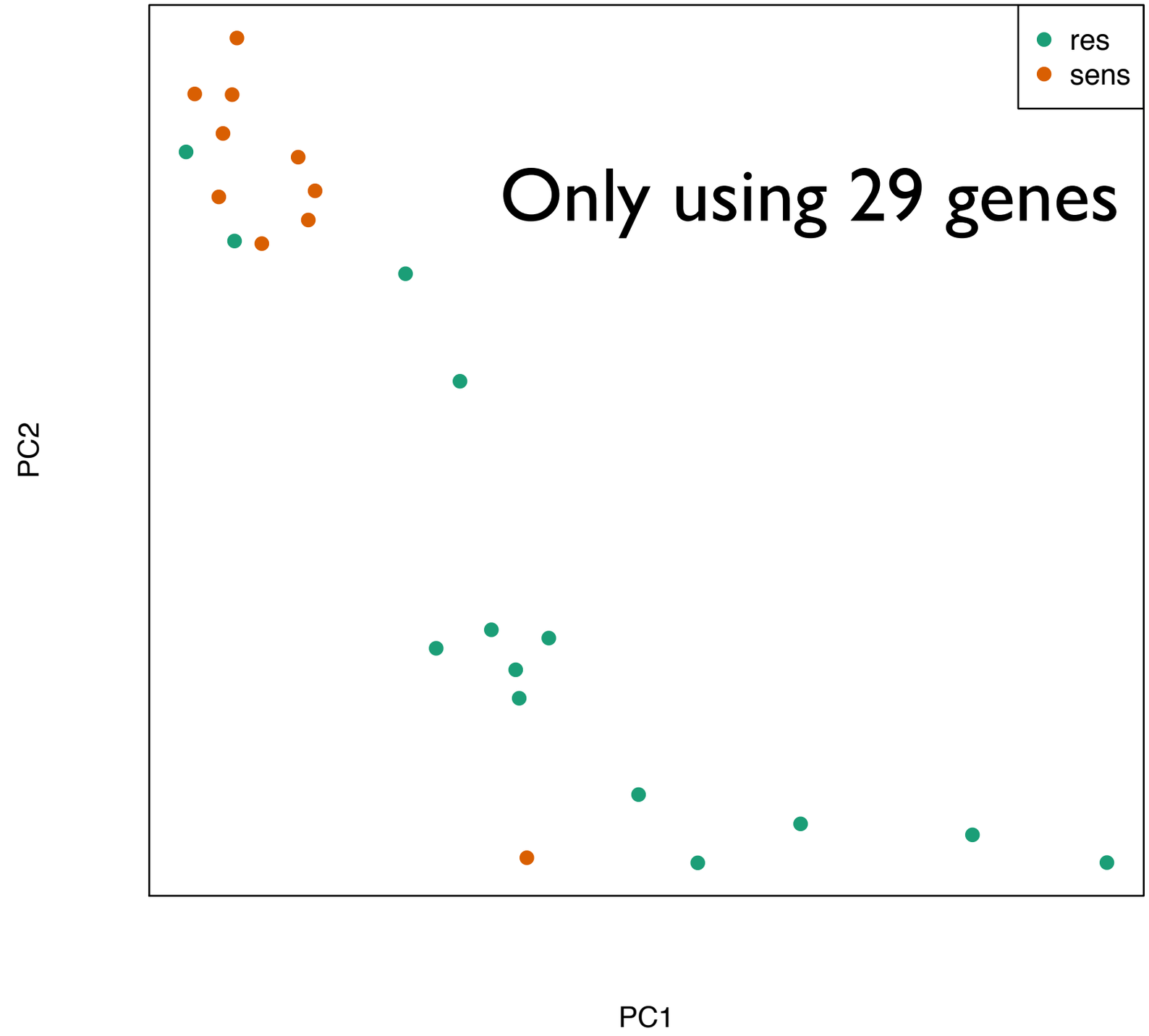
$$\begin{aligned} \max_{u} \quad & X_1' X_1 = v' X' X v \\ \text{s.t.} \quad & \textstyle\sum_{j=1}^{p} |v_j| \leq t \\ & v'v = 1 \end{aligned}$$

Sparse PCA via L1 regularization

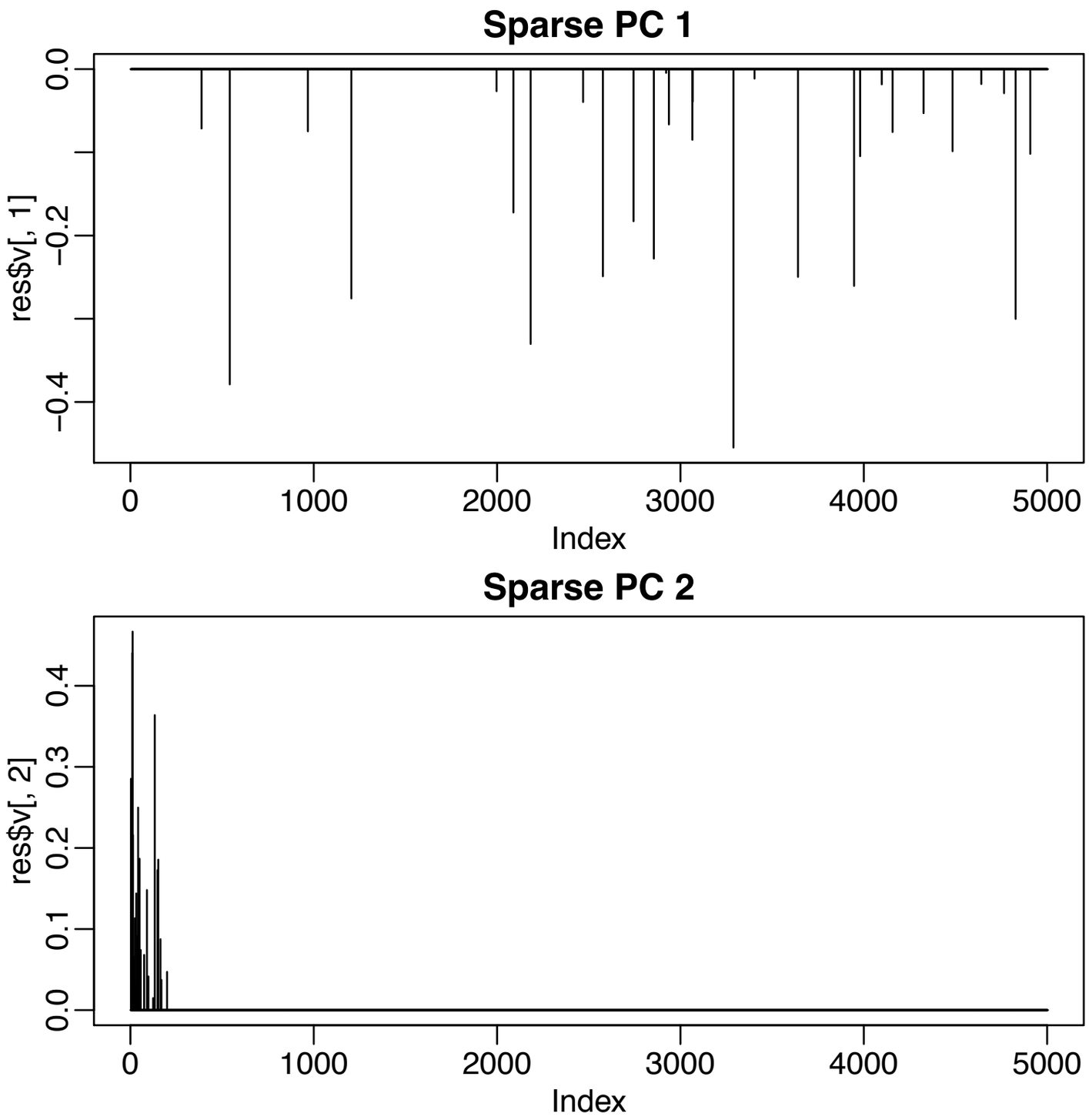
- That's too hard, but this is where thinking about approximation helps

$$\begin{aligned} \min_{V_q} \quad & \textstyle\sum_{i=1}^{N} \|(x_i - \bar{x}) - V_q V_q'(x_i - \bar{x})\|^2 \\ \text{s.t.} \quad & \|v_j\|_1 \leq t,\ j = 1, \ldots, q \\ & V_q' V_q = I \end{aligned}$$

Results on dataset



Interpreting PCA



Other methods

- "Direct" sparse PCA: constrain the number of genes directly
- Multidimensional Scaling (MDS): model distances between objects explicitly
- Non-negative matrix factorization: non-negativity constraints on approximation
  - has a nice "parts" interpretation
- In general, probabilistic PCA (likelihood instead of least-squares)