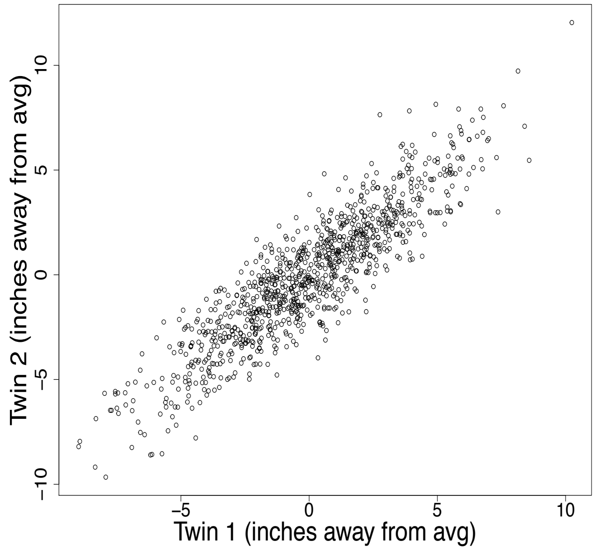
Dimension Reduction

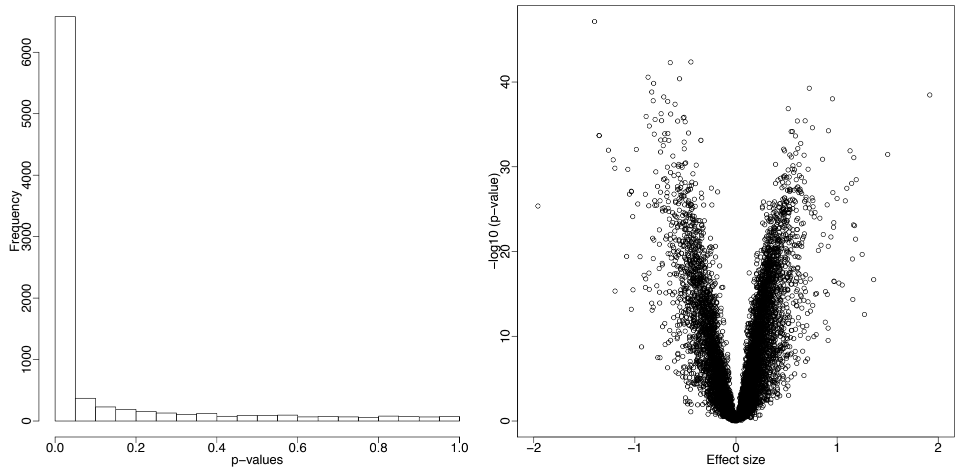
PCA, SVD, MDS, and clustering

Example: height of identical twins



[RI]

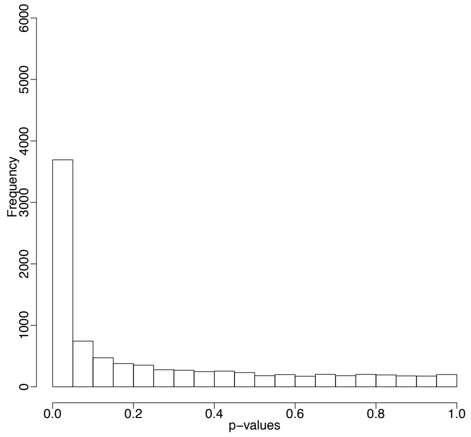
Expression between two ethnic groups

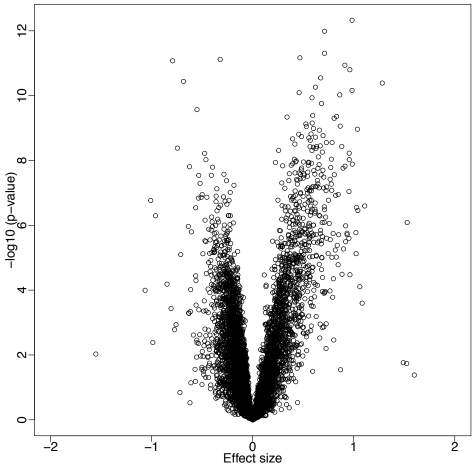


Ethnicity is confounded with year

Year	ASN	CEU
2002	0	32
2003	0	54
2004	0	13
2005	80	3
2006	3	0

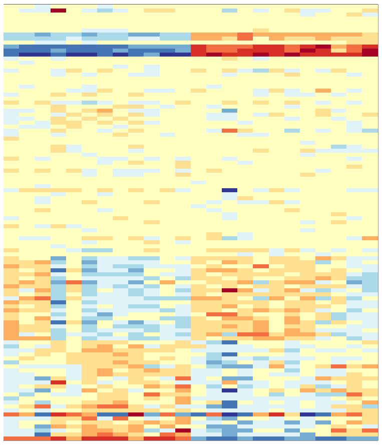
Two batches within ethnic groups





12 males and females, 2 months, 109 genes

	Female	Male
June 2005	3	9
October 2005	9	3



Finding an unknown batch

$$\Delta_i \equiv (Y_{i1},\ldots,Y_{i,n})(1/n_1,\ldots,1/n_1,-1/n_2,\ldots,-1/n_2) = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{i,j} - \frac{1}{n_2}\sum_{j=n_1+1}^{n_1+n_2} Y_{i,j}$$

Find n_1 and n_2 that make this difference large for many genes.

More precisely, maximize:

$$\frac{1}{m}\sum_{i=1}^{m} \Delta_i^2$$

Finding an unknown batch

More generally, let v be any vector with mean 0 and variance 1, find the v that maximizes

$$\sum_{i=1}^{m}\left\{\sum_{j=1}^{n} Y_{i,j} v_j\right\}^2 = (Y_{m\times n} v_{n\times 1})'(Y_{m\times n} v_{n\times 1})$$

The v that maximizes this variance is called the first principal component direction or **eigenvector**, and

$$Y_{m\times n} v_{n\times 1}$$

is the **first principal component.**

Principal components

We can remove the variability explained by v, and find the vector v_2 that maximizes the variability in these residuals.

By continuing this process we end up with n eigenvectors:

$$v_{n\times n} = (v_1 \ldots v_n)$$

Singular value decomposition (SVD)

SVD is a powerful mathematical approach that permits us to compute matrices U, D and V such that

$$Y_{m\times n} = U_{m\times n} D_{n\times n} V'_{n\times n}$$

and V are the eigenvectors.

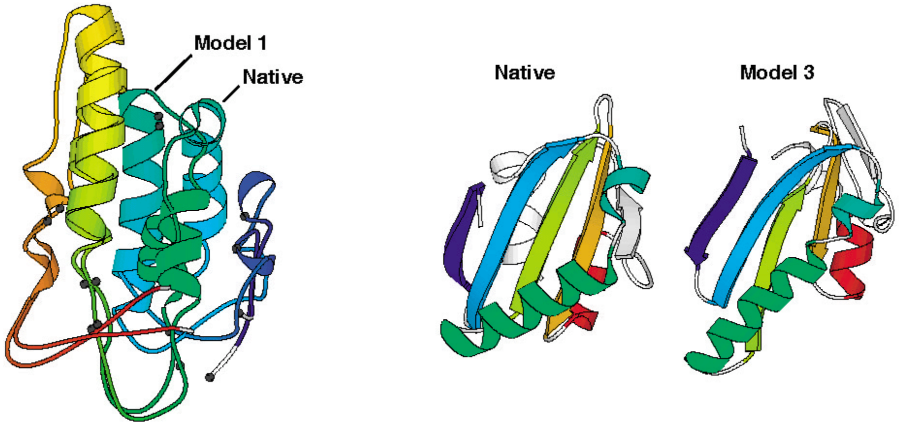
U and V are both orthogonal matrices and D is diagonal.

U orthogonal means that the columns of U are such that

$$U_i' U_i = 1 \quad \text{and} \quad U_i' U_j = 0$$

In other words, the sample standard deviation of each column is 1 and the sample correlation of any two columns is 0.

RMSD from SVD



PMID 11835488

Principal components from SVD

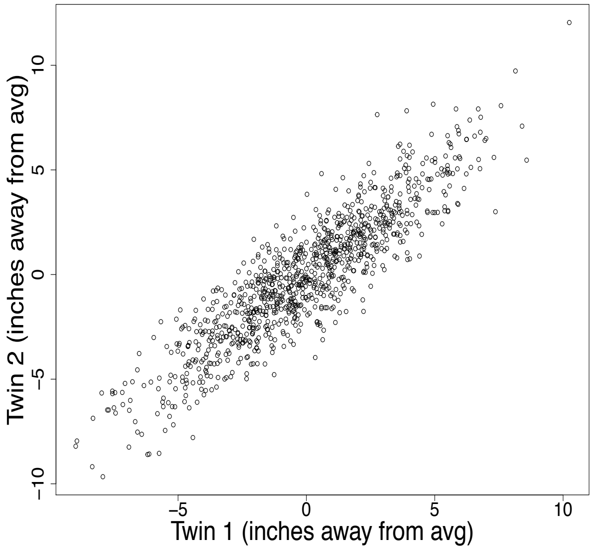
Notice that we can get the principal components from U and D

$$Y_{m\times n}V_{n\times n} = U_{m\times n}D_{n\times n}$$

and the variance from D:

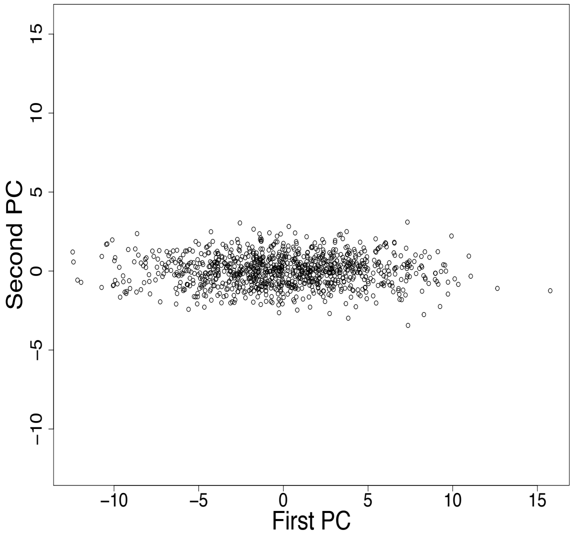
$$(Y_{m\times n}V_{n\times n})'(Y_{m\times n}V_{n\times n}) = DU'UD = D^2_{n\times n}$$

Example: height of identical twins



[RI]

Example: principal components



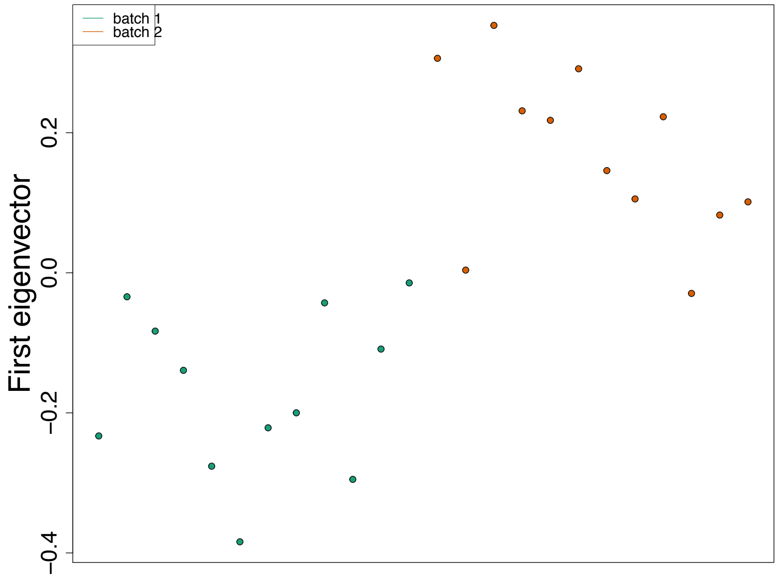
[RI]

Example: eigenvectors and SDs

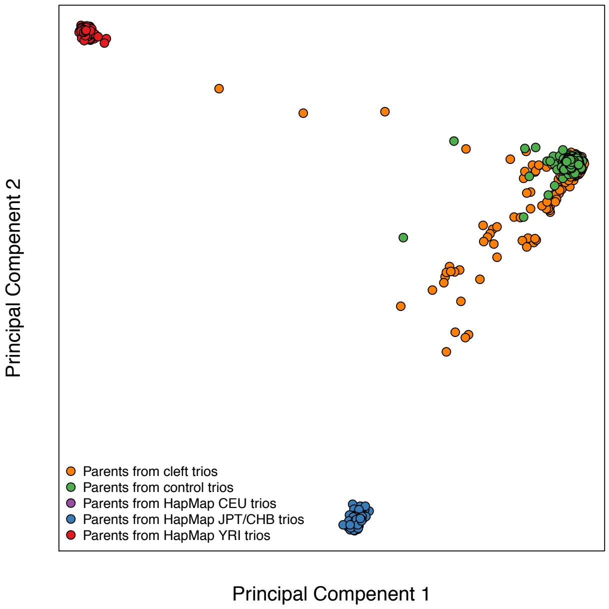
$$V_{2\times 2} = \begin{pmatrix} 0.706 & 0.708 \\ 0.708 & -0.706 \end{pmatrix} \approx \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$\frac{1}{\sqrt{m}} D_{2\times 2} = \begin{pmatrix} 4.3 & 0 \\ 0 & 1 \end{pmatrix}$$

Gene expression example

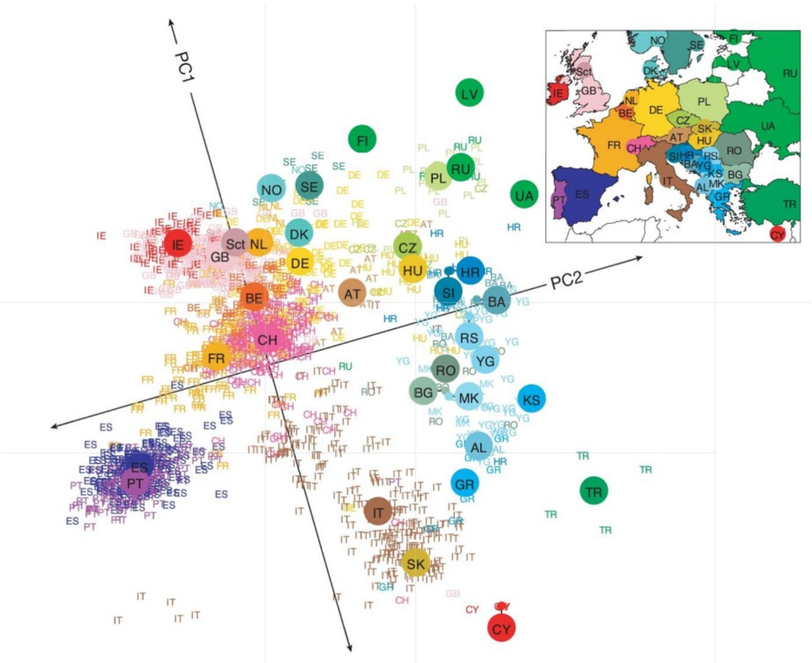


Genetic heterogeneity



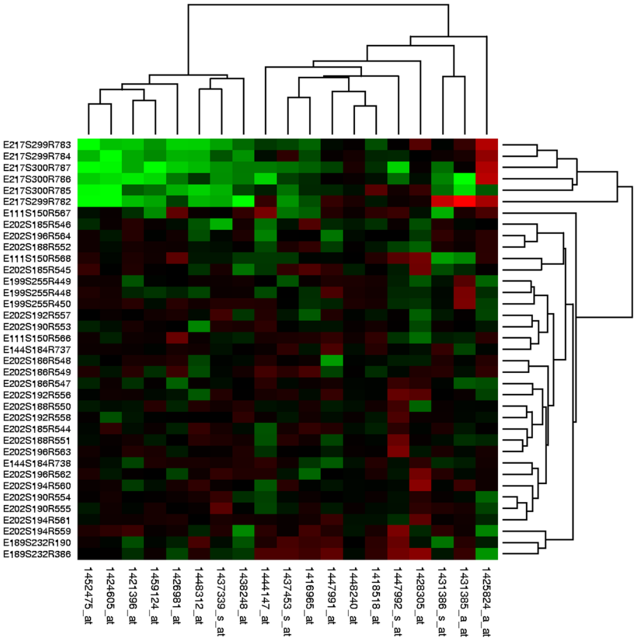
PMID 24528994

Genetic heterogeneity



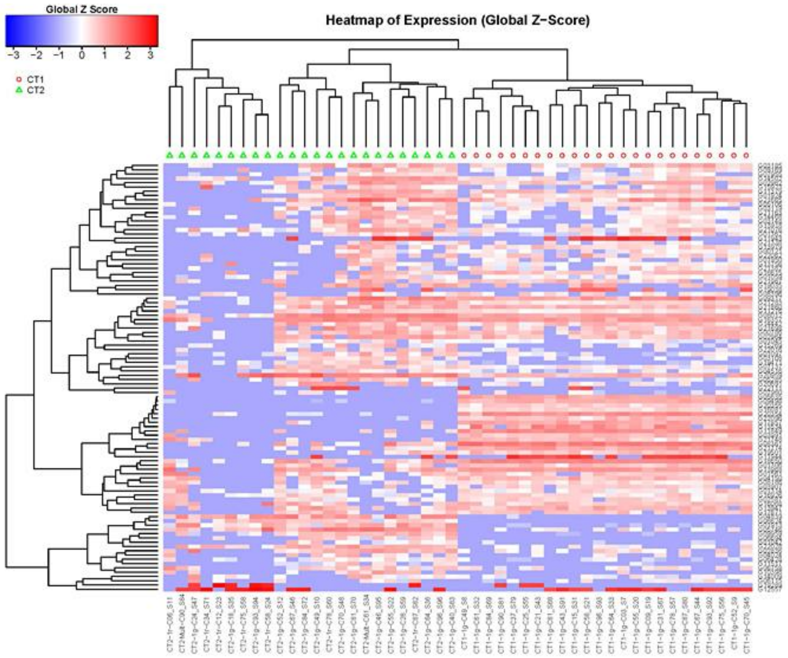
PMID 18758442

A heatmap



commons.wikimedia.org/wiki/File:Heatmap.png

Another heatmap

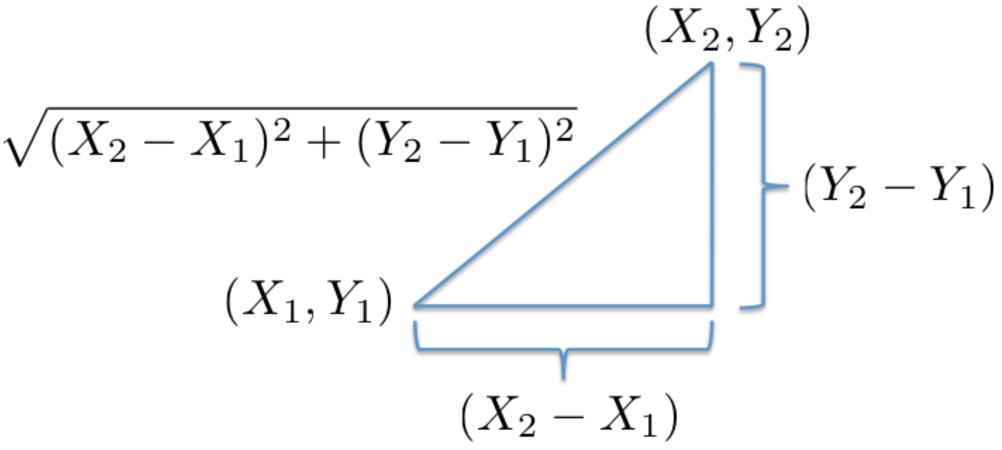


www.fluidigm.com

Distance

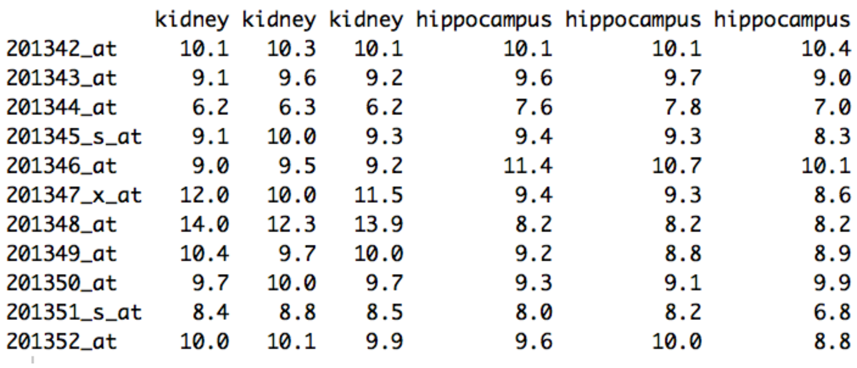
- Clustering organizes things that are “close” into groups.
- What does it mean for two genes to be close?
- What does it mean for two samples to be close?
- Once we know this, how do we define groups?

Distance in two dimension



Gene expression

Subset of a 22,215 x 189 gene expression table:



```
              kidney kidney kidney hippocampus hippocampus hippocampus
201342_at       10.1   10.3   10.1        10.1        10.1        10.4
201343_at        9.1    9.6    9.2         9.6         9.7         9.0
201344_at        6.2    6.3    6.2         7.6         7.8         7.0
201345_s_at      9.1   10.0    9.3         9.4         9.3         8.3
201346_at        9.0    9.5    9.2        11.4        10.7        10.1
201347_x_at     12.0   10.0   11.5         9.4         9.3         8.6
201348_at       14.0   12.3   13.9         8.2         8.2         8.2
201349_at       10.4    9.7   10.0         9.2         8.8         8.9
201350_at        9.7   10.0    9.7         9.3         9.1         9.9
201351_s_at      8.4    8.8    8.5         8.0         8.2         6.8
201352_at       10.0   10.1    9.9         9.6        10.0         8.8
```

Distances

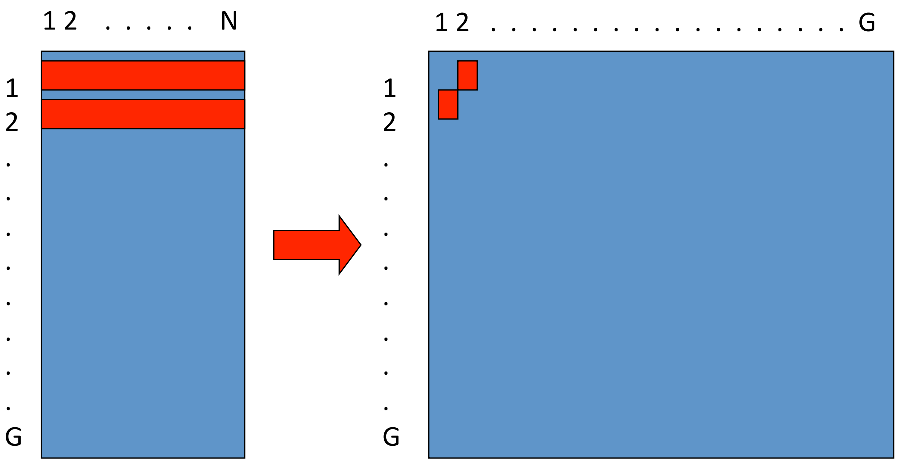
There are 17,776 pairs of samples for which we can compute a distance:

$$d(j,k) = \sqrt{\sum_{i=1}^{22{,}215} (X_{i,j} - X_{i,k})^2}$$

There are 246,742,005 pairs of genes for which we can compute a distance:

$$d(h,i) = \sqrt{\sum_{j=1}^{N} (X_{h,j} - X_{i,j})^2}$$

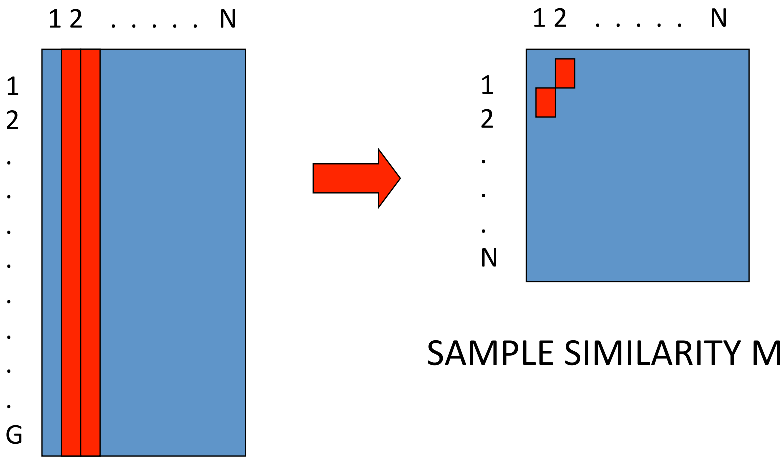
The similarity / distance matrices



DATA MATRIX

GENE SIMILARITY MATRIX

The similarity / distance matrices



SAMPLE SIMILARITY MATRIX

DATA MATRIX

Multidimensional scaling

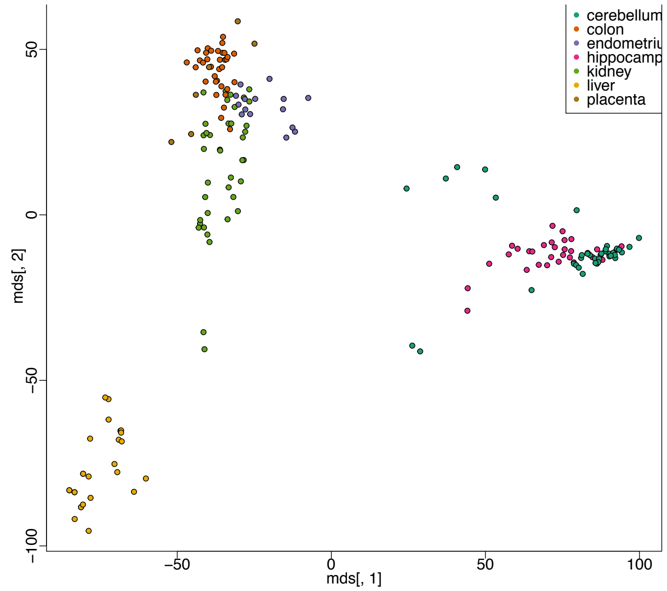
We can find a linear transformation for the data

$$Z = AX$$

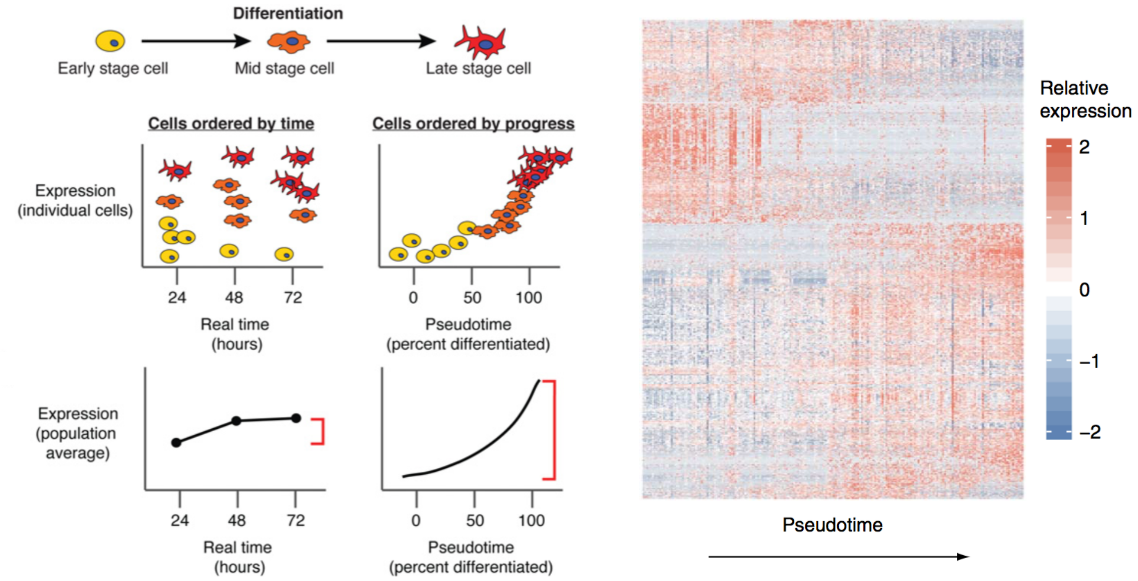
such that

$$\sqrt{\sum_{i=1}^{22{,}215} (X_{i,j} - X_{i,k})^2} \approx \sqrt{(Z_{1,j} - Z_{1,k})^2 + (Z_{2,j} - Z_{2,k})^2}$$

Multidimensional scaling

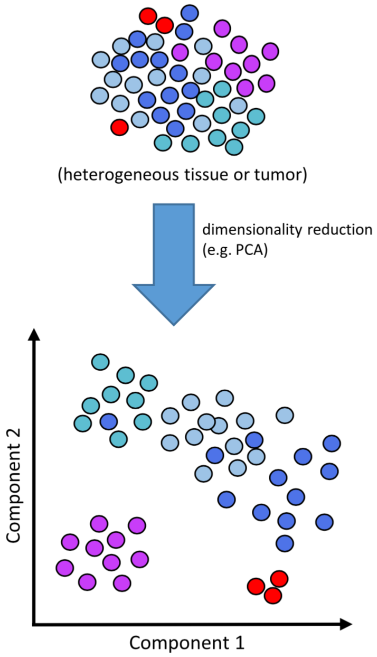


Single cell RNAseq



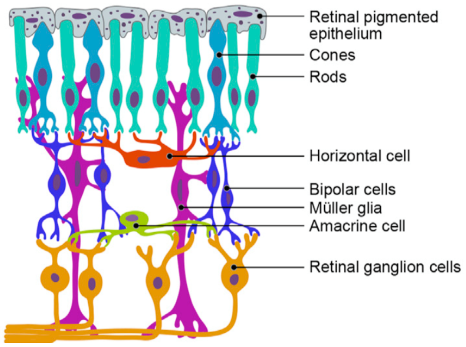
PMIDs 24658644, 26430159

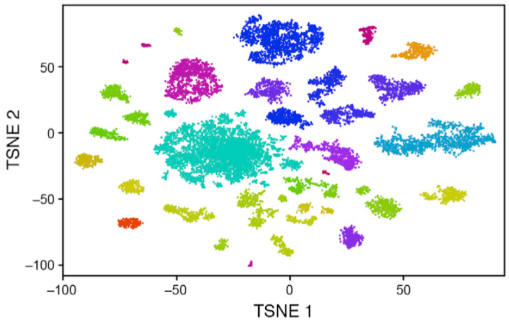
Single cell RNAseq



PMID 26949524

Single cell RNAseq



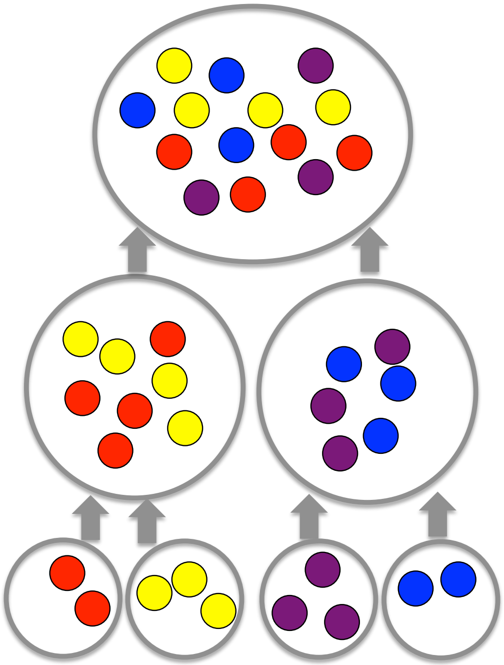


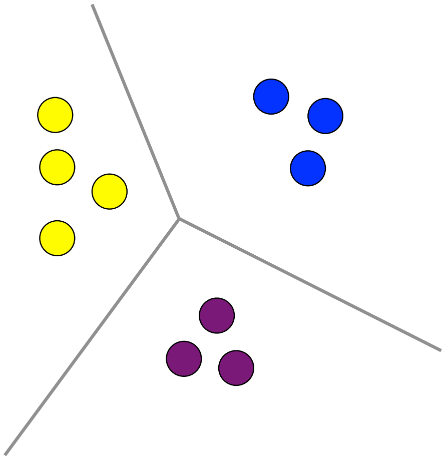
PMID 26000488

Clustering

Hierarchical

Partitioning (K-means)

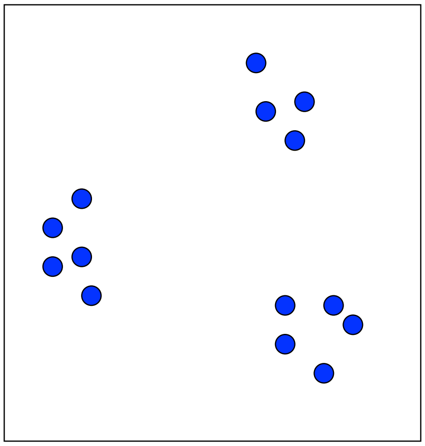




[140.688]

K-means

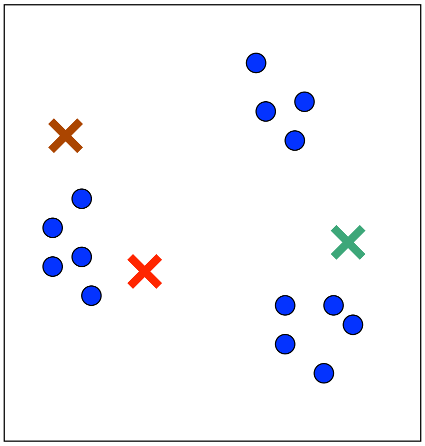
- We start with some data.
- For example:
  - We are showing expression for two samples for 14 genes.
  - We are showing expression for two genes for 14 samples.
- This is simplifaction.



Iteration = 0

K-means

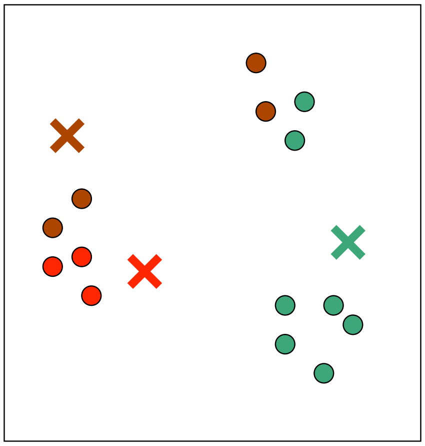
- Choose K *centroids.*
- These are starting values that the user picks.
- There are some data driven ways to do it.



Iteration = 0

K-means

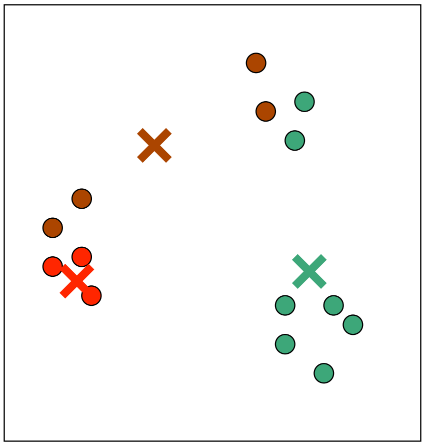
- Make first *partition* by finding the closest centroid for each point.
- This is where distance is used.



Iteration = 1

K-means

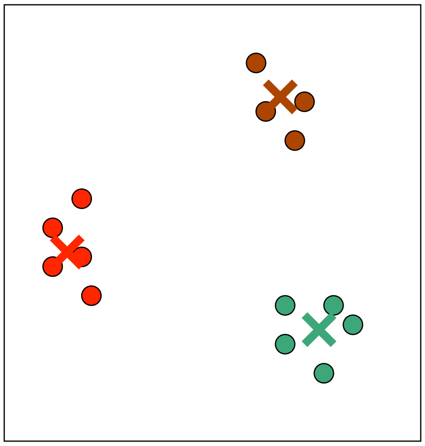
- Now re-compute the centroids by taking the *middle* of each cluster.



Iteration = 2

K-means

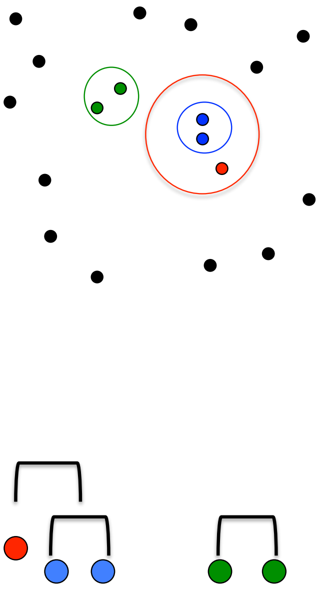
- Repeat until the centroids stop moving or until you get tired of waiting.



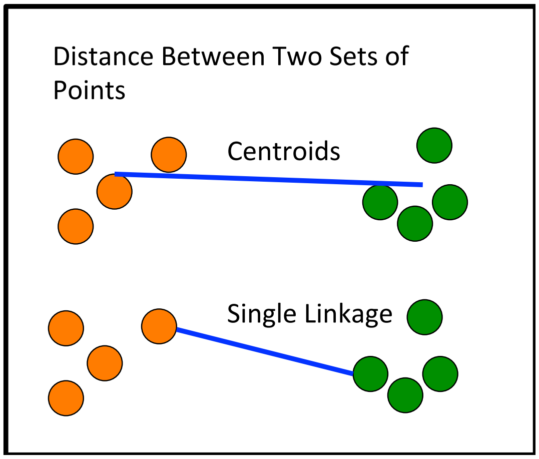
Iteration = 3

[140.688]

Hierarchical clustering algorithm



1. Say every point is its own cluster
2. Merge “closest” points
3. Repeat



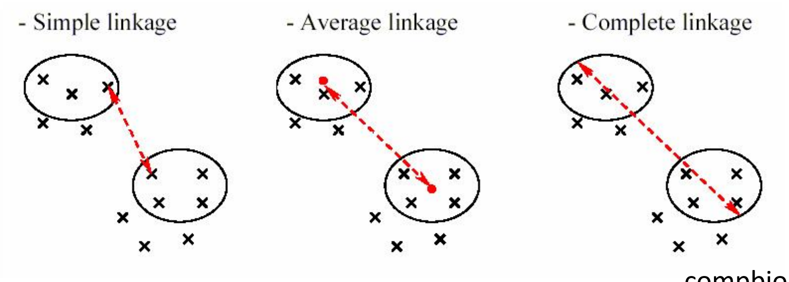
Distance Between Two Sets of Points

Linkage

Single linkage defines the distance between clusters as the distance between the closest two points. Single linkage can lead to a lot of singleton clusters, and to clusters that look stringlike in high dimensions.

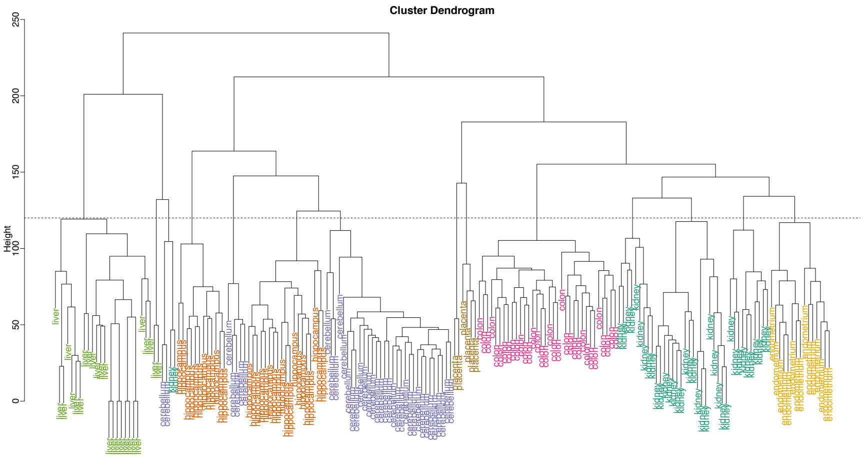
Complete linkage defines the distance between clusters as the distance between the farthest two points. Complete linkage tends to lead to more compact spherical structures.

Average linkage is the average of all the pairwise distances between points in the two clusters. Average linkage is between single and complete linkage in terms of the type of clusters it outputs.

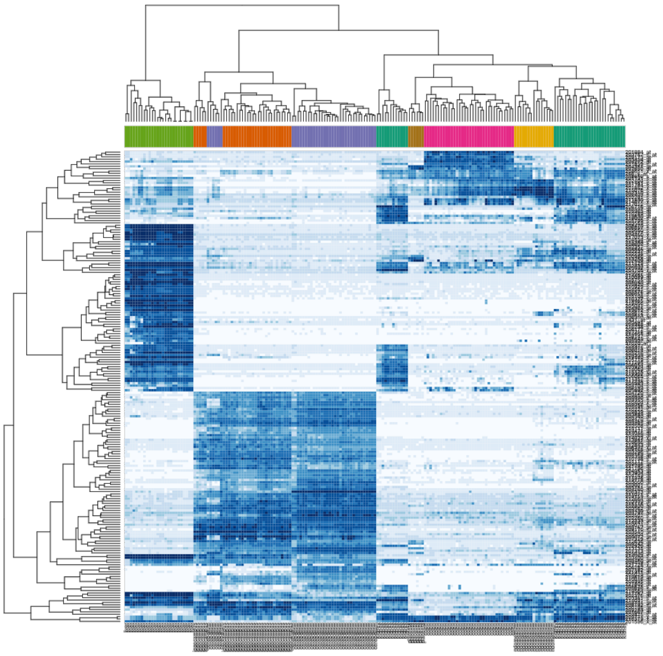


compbio.pbworks.com

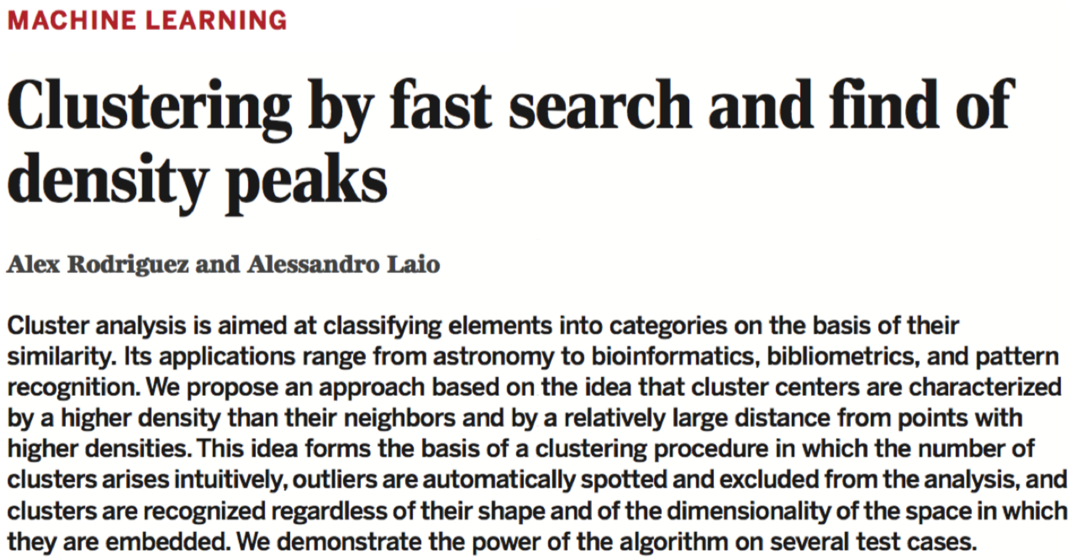
A dendogram



A heatmap



[RI]



MACHINE LEARNING

Clustering by fast search and find of density peaks

Alex Rodriguez and Alessandro Laio

Cluster analysis is aimed at classifying elements into categories on the basis of their similarity. Its applications range from astronomy to bioinformatics, bibliometrics, and pattern recognition. We propose an approach based on the idea that cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities. This idea forms the basis of a clustering procedure in which the number of clusters arises intuitively, outliers are automatically spotted and excluded from the analysis, and clusters are recognized regardless of their shape and of the dimensionality of the space in which they are embedded. We demonstrate the power of the algorithm on several test cases.

PMID 24970081

The algorithm has its basis in the assumptions that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density. For each data point i, we compute two quantities: its local density ρ_i and its distance δ_i from points of higher density. Both these quantities depend only on the distances d_{ij} between data points, which are assumed to satisfy the triangular inequality. The local density ρ_i of data point i is defined as

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \qquad (1)$$

where $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise, and d_c is a cutoff distance. Basically, ρ_i is equal to the number of points that are closer than d_c to point i. The algorithm is sensitive only to the relative magnitude of ρ_i in different points, implying that, for large data sets, the results of the analysis are robust with respect to the choice of d_c.

δ_i is measured by computing the minimum distance between the point i and any other point with higher density:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \qquad (2)$$

For the point with highest density, we conventionally take $\delta_i = \max_j(d_{ij})$. Note that δ_i is much larger than the typical nearest neighbor distance only for points that are local or global maxima in the density. Thus, cluster centers are recognized as points for which the value of δ_i is anomalously large.

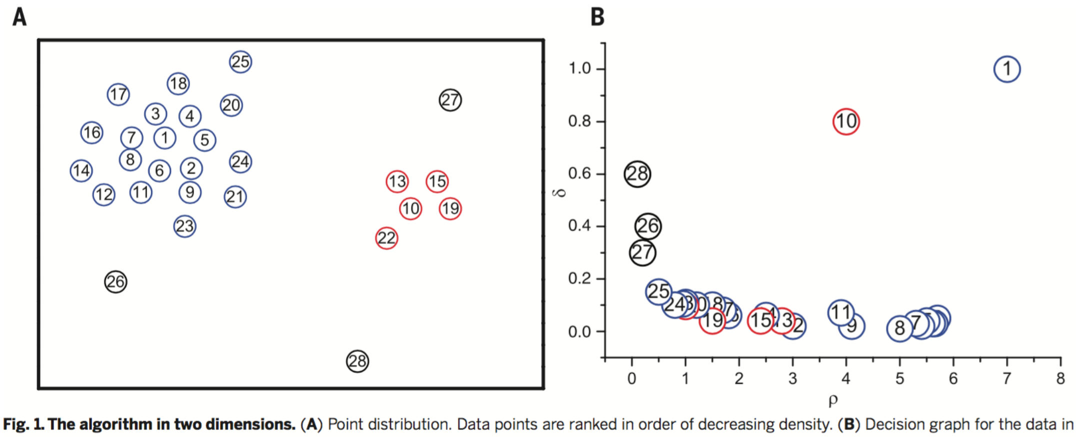


Fig. 1. The algorithm in two dimensions. (**A**) Point distribution. Data points are ranked in order of decreasing density. (**B**) Decision graph for the data in (A). Different colors correspond to different clusters.

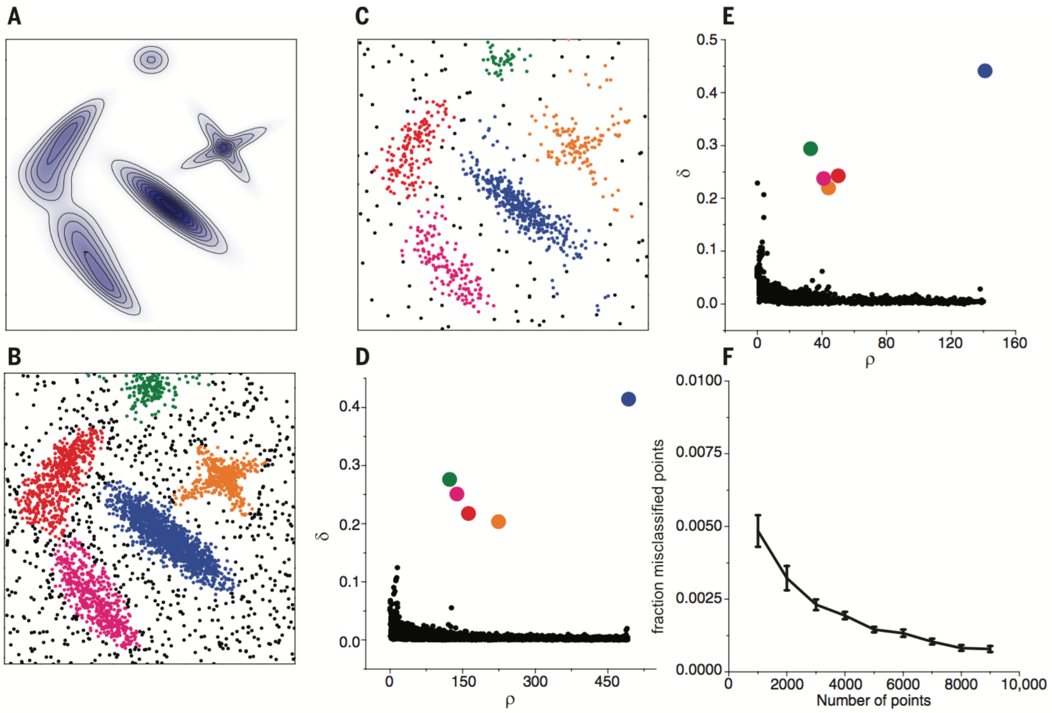


Fig. 2. Results for synthetic point distributions. (**A**) The probability distribution from which point distributions are drawn. The regions with lowest intensity correspond to a background uniform probability of 20%. (**B** and **C**) Point distributions for samples of 4000 and 1000 points, respectively. Points are colored according to the cluster to which they are assigned. Black points belong to the cluster halos. (**D** and **E**) The corresponding decision graphs, with the centers colored by cluster. (**F**) The fraction of points assigned to the incorrect cluster as a function of the sample dimension. Error bars indicate the standard error of the mean.

PMID 24970081

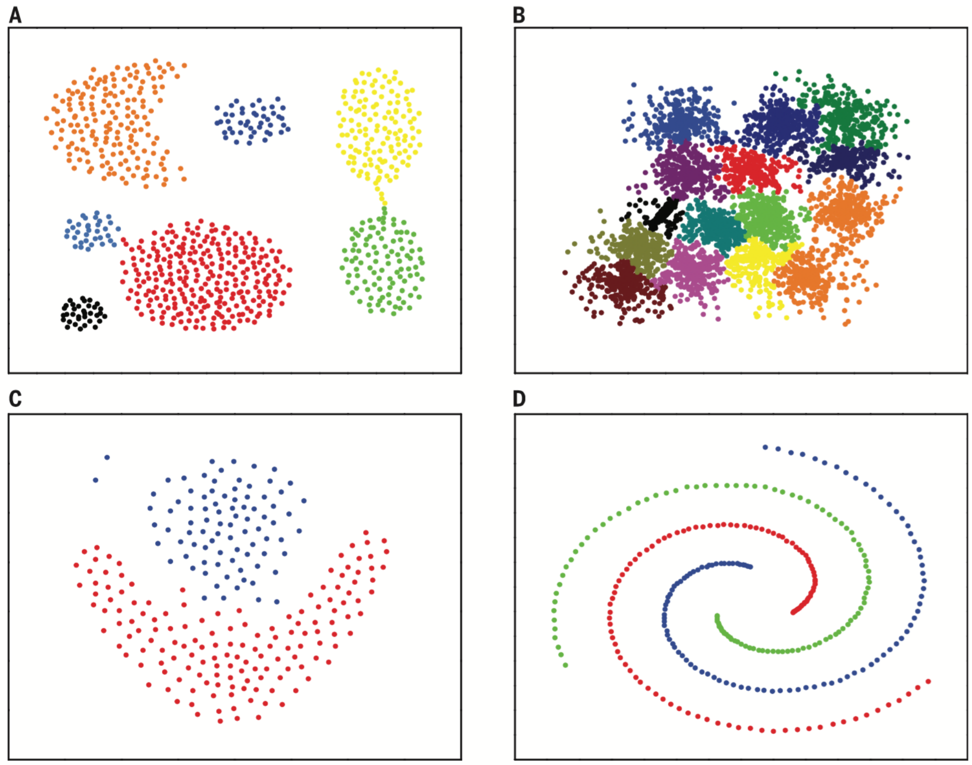


Fig. 3. Results for test cases in the literature. Synthetic point distributions from (*12*) (**A**), (*13*) (**B**), (*14*) (**C**), and (*15*) (**D**).

PMID 24970081

Result 1

The distance is equivalent to the correlation when the data are standardized.

$$\frac{1}{M}\sum_{i=1}^{M}\left(\frac{X_i-\bar{X}}{s_X}-\frac{Y_i-\bar{Y}}{s_Y}\right)^2=$$

$$\frac{1}{M}\sum_{i=1}^{M}\left(\frac{X_i-\bar{X}}{s_X}\right)^2+\frac{1}{M}\sum_{i=1}^{M}\left(\frac{Y_i-\bar{Y}}{s_Y}\right)^2-\frac{2}{M}\sum_{i=1}^{M}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)=$$

$$2(1-r)$$

Result 2

The difference in the averages can drive the distance.

$$\frac{1}{M}\sum_{i=1}^{M}(X_i-Y_i)^2=\frac{1}{M}\sum_{i=1}^{M}\left\{(X_i-\bar{X})-(Y_i-\bar{Y})+(\bar{X}-\bar{Y})\right\}^2$$

$$=\frac{1}{M}\sum_{i=1}^{M}\left\{(X_i-\bar{X})-(Y_i-\bar{Y})\right\}^2+2(\bar{X}-\bar{Y})\frac{1}{M}\sum_{i=1}^{M}\left\{(X_i-\bar{X})-(Y_i-\bar{Y})\right\}+\frac{1}{M}\sum_{i=1}^{M}(\bar{X}-\bar{Y})^2$$

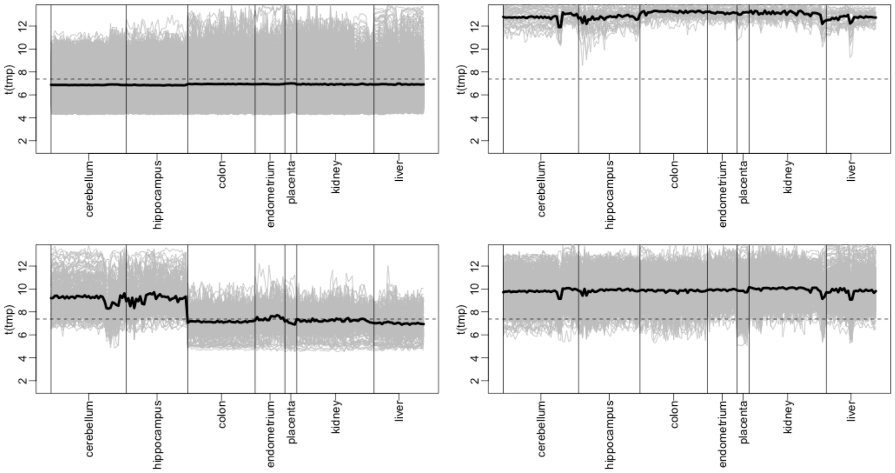
$$=2(1-r)+\frac{1}{M}\sum_{i=1}^{M}(\bar{X}-\bar{Y})^2$$

$$=2(1-r)+(\bar{X}-\bar{Y})^2$$

(assuming $\frac{1}{M}\sum_{i=1}^{M}(X_i-\bar{X})^2=1$)

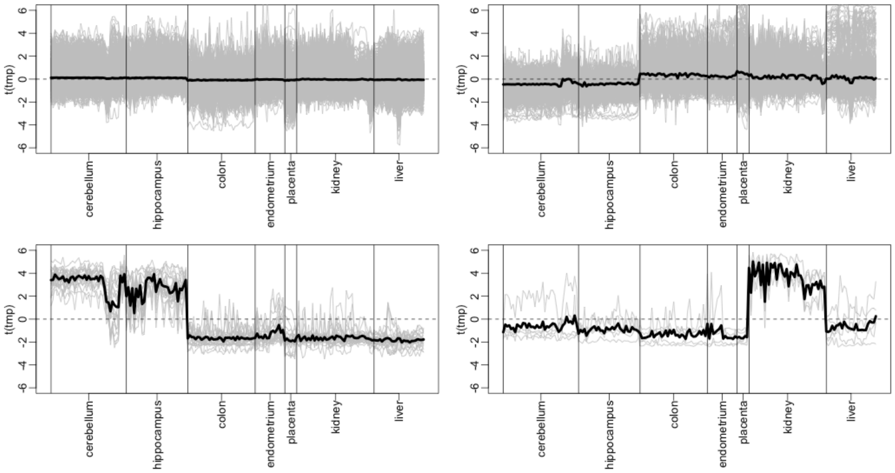
Four gene cluster

no mean removal



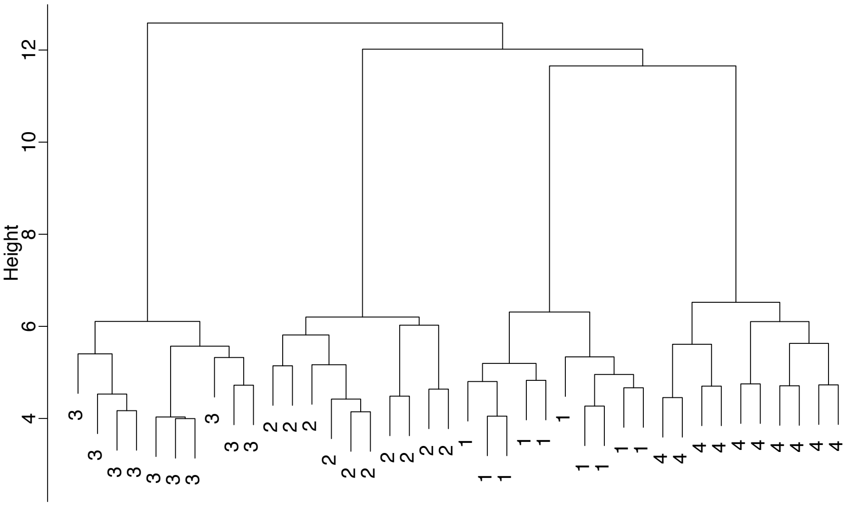
Four gene cluster

after mean removal



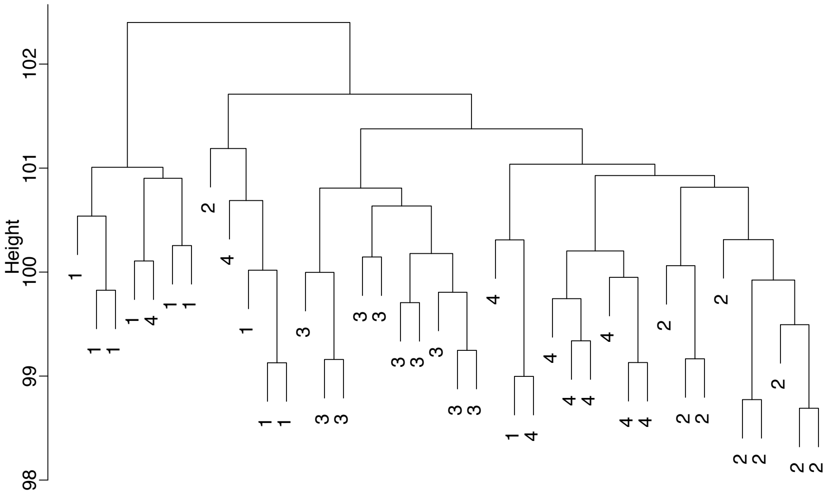
Simulation

only differentially expressed genes

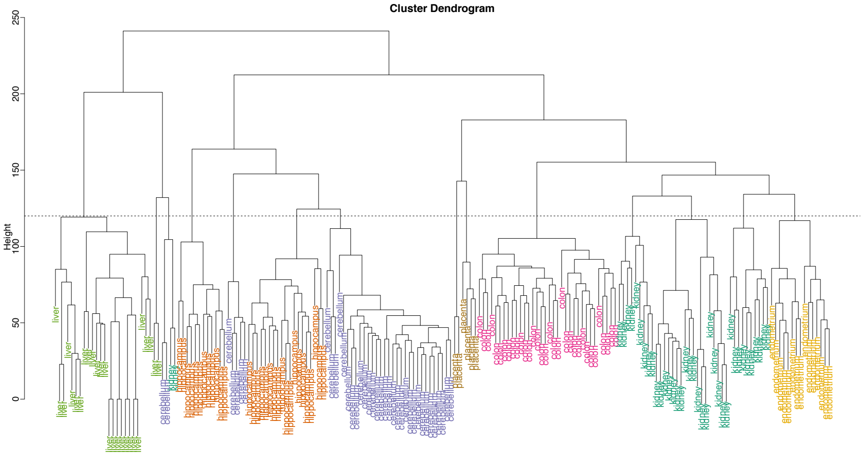


Simulation

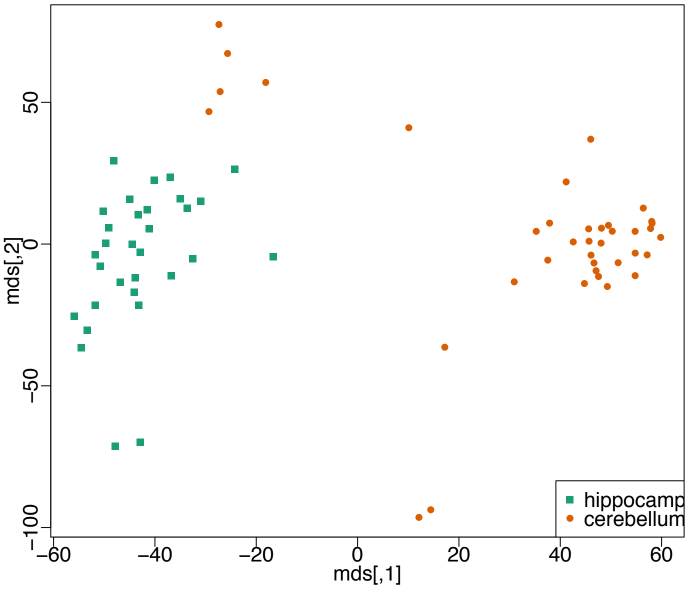
all genes



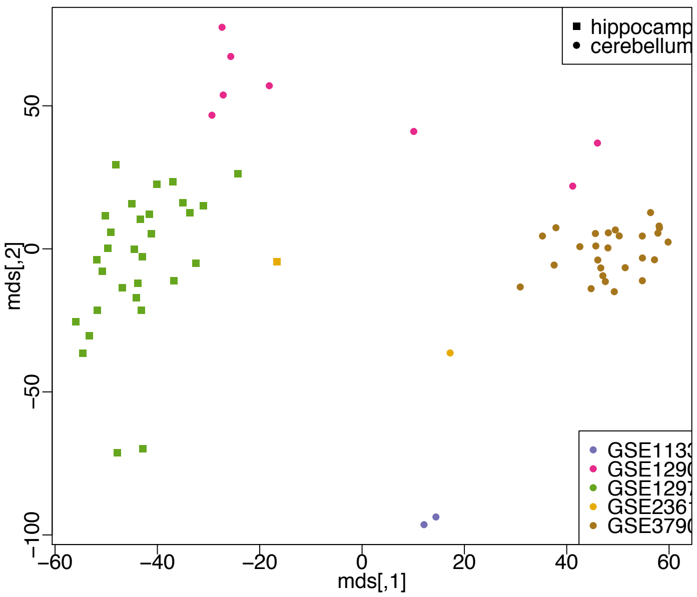
Batch effects



Color represents tissue

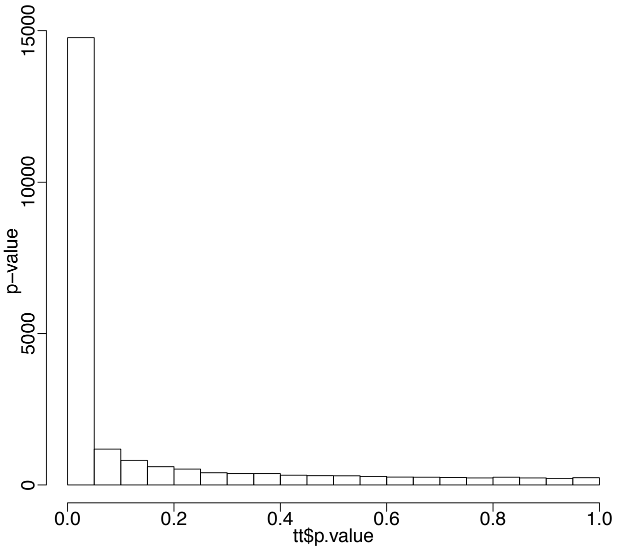


Color represents study



Null distribution of p-values

cerebellum only


p−value
tt$p.value