Multiple Linear Regression

(Chapters 12-13 in
Montgomery, Runger)

12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, 2, \ldots, n$$

In matrix notation this model can be written as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \qquad (12\text{-}6)$$

12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

We wish to find the vector of least squares estimators that minimizes:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$0 = \frac{\partial L}{\partial \beta} = 2X'(y - X\beta)$$

The resulting least squares estimate is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad (12\text{-}7)$$

Analog of $\frac{1}{Var(X)}$

Analog of $Cov(X,Y)$

Multiple Linear Regression Model

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\ \mathbf{X}'\mathbf{y}$$

$$\hat{y} = X\hat{\beta} = \overbrace{X(X'X)^{-1}X'}^{H}y,$$

$$\hat{y} = Hy, \quad \text{and} \quad e = (I - H)y.$$

One can show: $H^2 = H \rightarrow \hat{y}' \cdot e = 0$

12-1: Multiple Linear Regression Models

Estimating σ^2

An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p} \qquad (12\text{-}16)$$

Here $p = K + 1$

R^2 and Adjusted R^2

The **coefficient of multiple determination**

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

The **adjusted R^2** is

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \qquad (12\text{-}23)$$

- The adjusted R^2 statistic penalizes adding terms to the MLR model.
- It can help guard against overfitting (including regressors that are not really useful)

How to know where to stop?

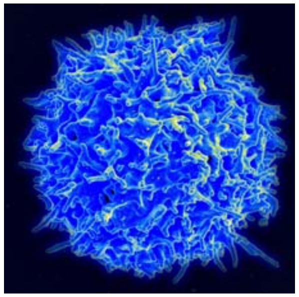
- Adding new variables x_i to MLR watch the adjusted R^2

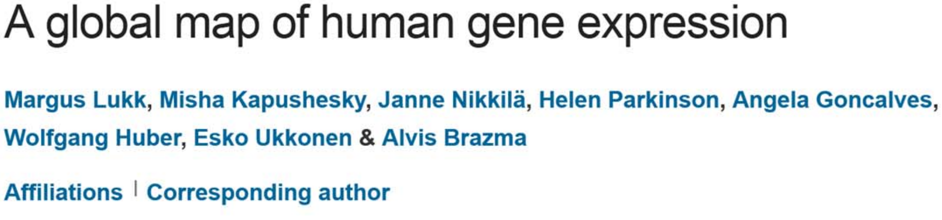
- Once the adjusted R^2 no longer increases = stop

- Now you did the best you can with the data you have

T-cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2011
- All samples are from normal T-cells in different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene’s expression level in a given sample as measured by microarray technology



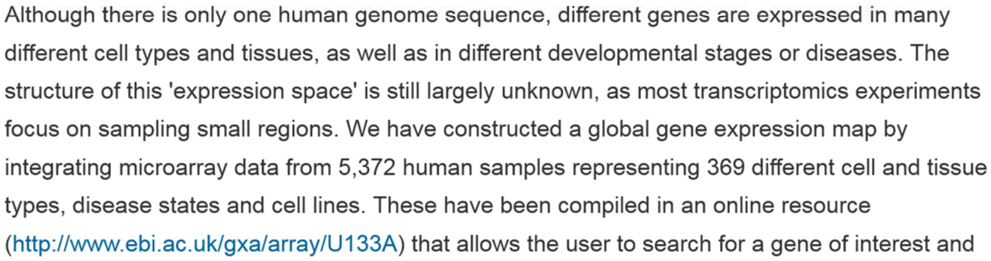


A global map of human gene expression

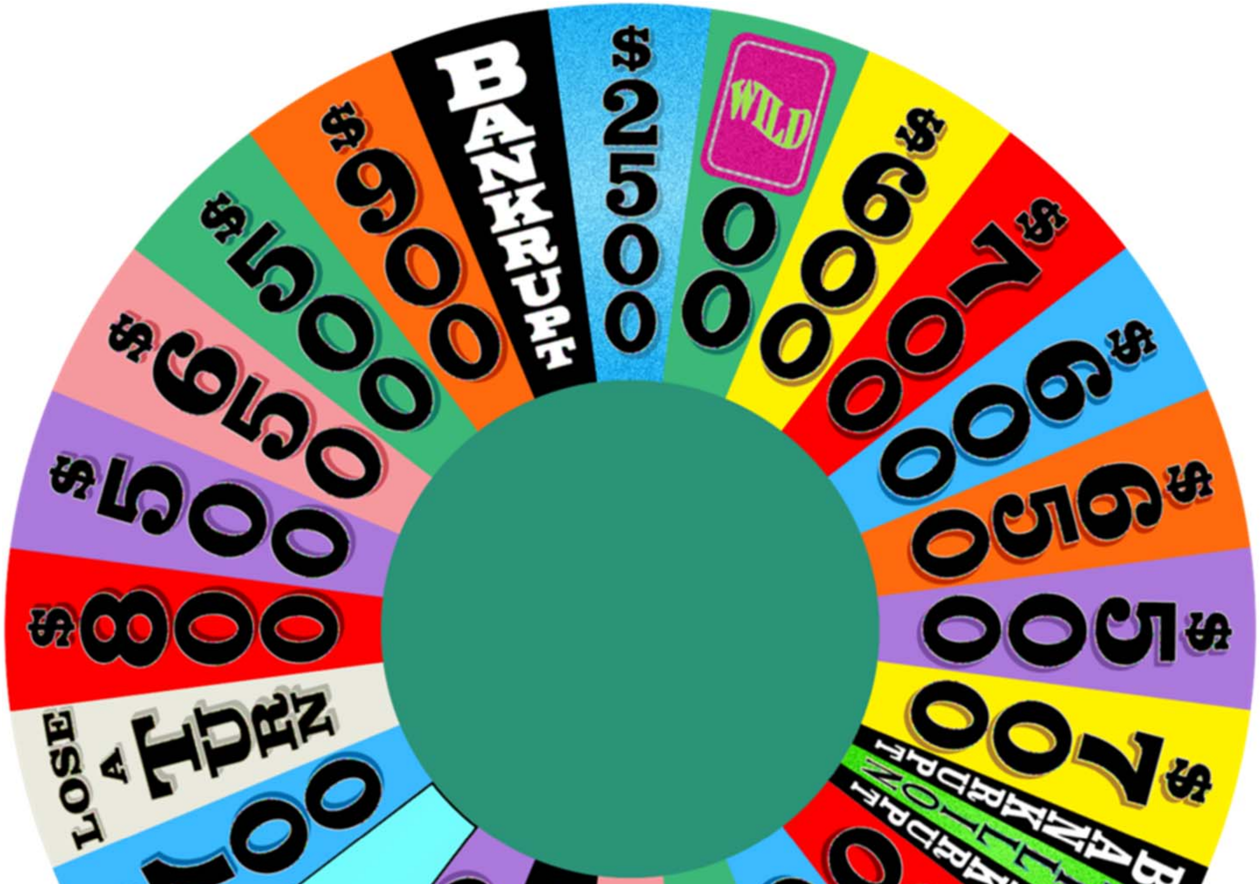
Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

Affiliations | Corresponding author

Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322



Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and



WHEEL OF FORTUNE®

Matlab exercise

- Each group gets one good third gene correlated with their pair and a random gene.
- Compute Multiple Linear Regression (MLR): where y=expression(g1), x1=expression(g2), x2=expression(g3)
- Use lm=fitlm([x1,x2],y)
- How much better did you do with MLR compared to SLR?
- Compute multiple linear regression: where y=expression(g1), x1=expression(g2), x2=expression(g_random)
- How about now? Did random gene work as well as handpicked one?

Pairs to correlate

2907	2881	extra: 2629,	random 2445
1994	188	extra: 547,	random 2718
2274	1597	extra: 1994,	random 381
2982	1353	extra: 2303,	random 2741

Multiple linear regression

- load expression_table.mat
- % Single variable regression
- g1=2907; g2=2881;
- y=exp_t(g1,:)'; x=exp_t(g2,:)'
- figure; plot(x,y,'ko')
- lm=fitlm(x,y)
- y_fit=lm.Fitted;
- hold on;
- plot(x,lm.Fitted,'r-');

- %Multiple regression
- g1=2907; g2=2881; g3=2629;
- y=exp_t(g1,:)'; x=[exp_t(g2,:)', exp_t(g3,:)'];
- figure; plot(x(:,1),y,'ko');
- %figure; plot3(x(:,1),x(:,2),y,'ko');
- lm=fitlm(x,y)
- y_fit=lm.Fitted;
- hold on; plot(x(:,1),y_fit,'rd');

Matlab exercise

- Each group gets the third gene correlated with their pair and a random gene.
- Compute multiple linear regression: where y=expression(g1), x1=expression(g2), x2=expression(g3)
- How much better did you do with MLR compared to SLR?
- Show MLR data to 3D scatter plot: plot3(x1,x2,y) and MLR predictions on 2D plot: plot(x1,y)
- Compute multiple linear regression: where y=expression(g1), x1=expression(g2), x2=expression(g_random)
- How about now? Did random gene work as well as handpicked one?

Multiple linear regression

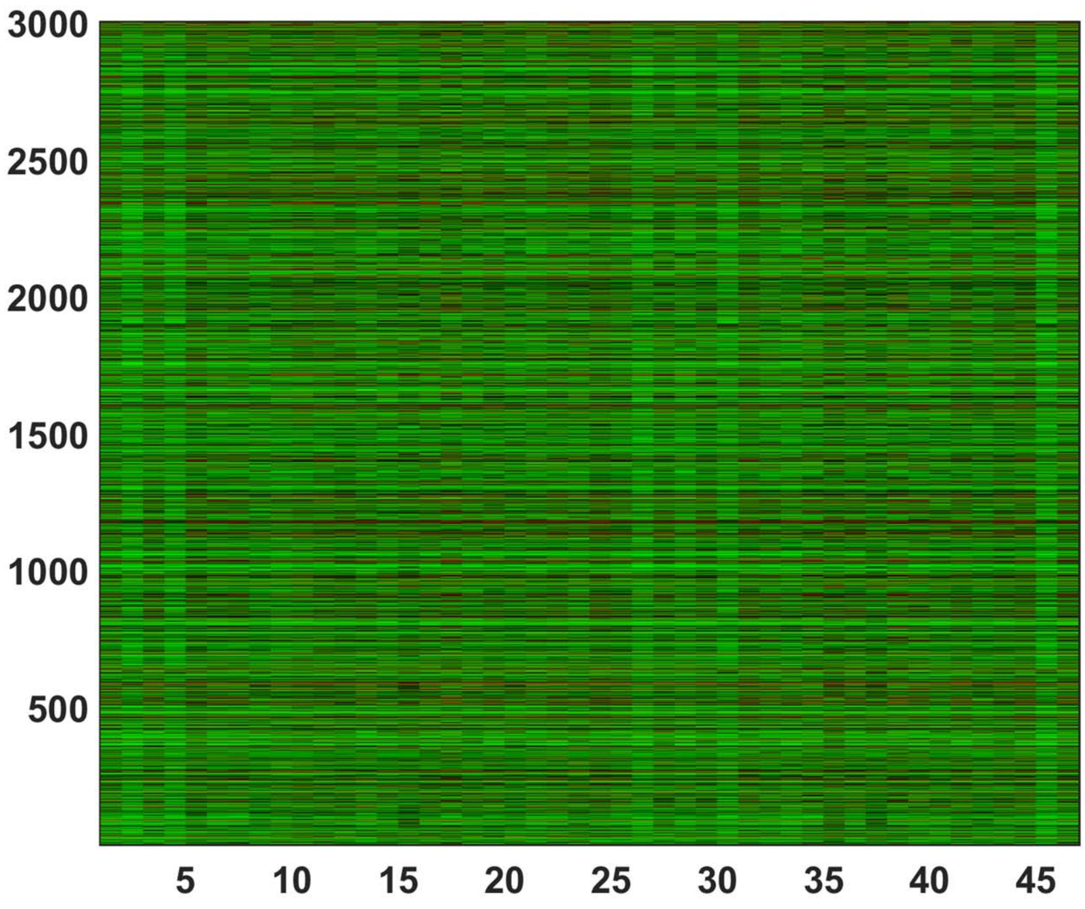
```
load expression_table.mat
% Single variable regression
g1=2907; g2=2881;
y=exp_t(g1,:)'; x=exp_t(g2,:)'
figure; plot(x,y,'ko')
lm=fitlm(x,y)
y_fit=lm.Fitted;
hold on;
plot(x,lm.Fitted,'r-');

%Multiple regression
g1=2907; g2=2881; g3=2629;
y=exp_t(g1,:)'; x=[exp_t(g2,:)', exp_t(g3,:)'];
figure; plot(x(:,1),y,'ko');
%figure; plot3(x(:,1),x(:,2),y,'ko');
lm=fitlm(x,y)
y_fit=lm.Fitted;
hold on; plot(x(:,1),y_fit,'rd');
```

Clustering analysis of gene expression data

Chapter 11 in
Jonathan Pevsner,
Bioinformatics and Functional Genomics,
3rd edition
(Chapter 9 in 2nd edition)

How to interpret the expression data if you still have many genes and many samples?



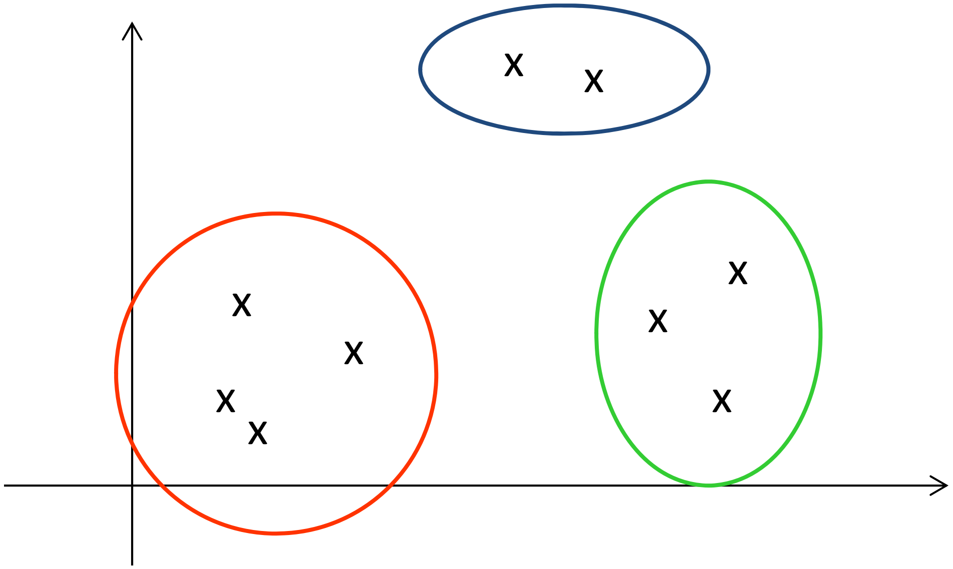
Clustering to the rescue!

Clustering is a part of Machine Learning

- ***Supervised Learning:***
  A machine learning technique whereby a system uses a set of training examples to learn how to correctly perform a task
  Example: a sample of cancer expression profiles each **annotated** with cancer type/tissue.
  Goal: predict cancer type based on expression pattern
- ***Unsupervised Learning (including clustering):***
  In machine learning, unsupervised learning is a class of problems in which one seeks to determine how the data are organized. One only has unlabeled examples.
  Example: a sample of breast cancer expression profiles.
  Goal: Identify several different (yet unknown) subtypes with potentially different treatment

What is clustering?

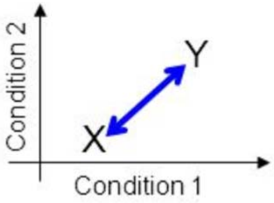
- The goal of clustering is to
  - group data points that are close (or **similar**) to each other
  - Usually we need to identify such groups (or clusters) in an **unsupervised** manner
  - Sometimes we take into account **prior information** (Bayesian methods)
- Need to define distance d_{ij} between objects i and j
- In our case objects could be either genes or samples
- Easy in 2 dimensions but hard in 3000 dimensions
- Need to somehow reduce dimensionality



How to define distance?

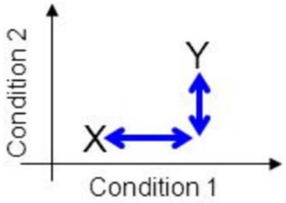
- Euclidean distance:
  - Most commonly used distance
  - Sphere shaped cluster
  - Corresponds to the geometric distance into the multidimensional space

$$d(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$



- City Block (Manhattan) distance:
  - Sum of differences across dimensions
  - Less sensitive to outliers
  - Diamond shaped clusters

$$d(X,Y) = \sum_i |x_i - y_i|$$



The Canberra distance metric is calculated in R by

$$\sum \left(\frac{|x_i - y_i|}{|x_i + y_i|} \right).$$

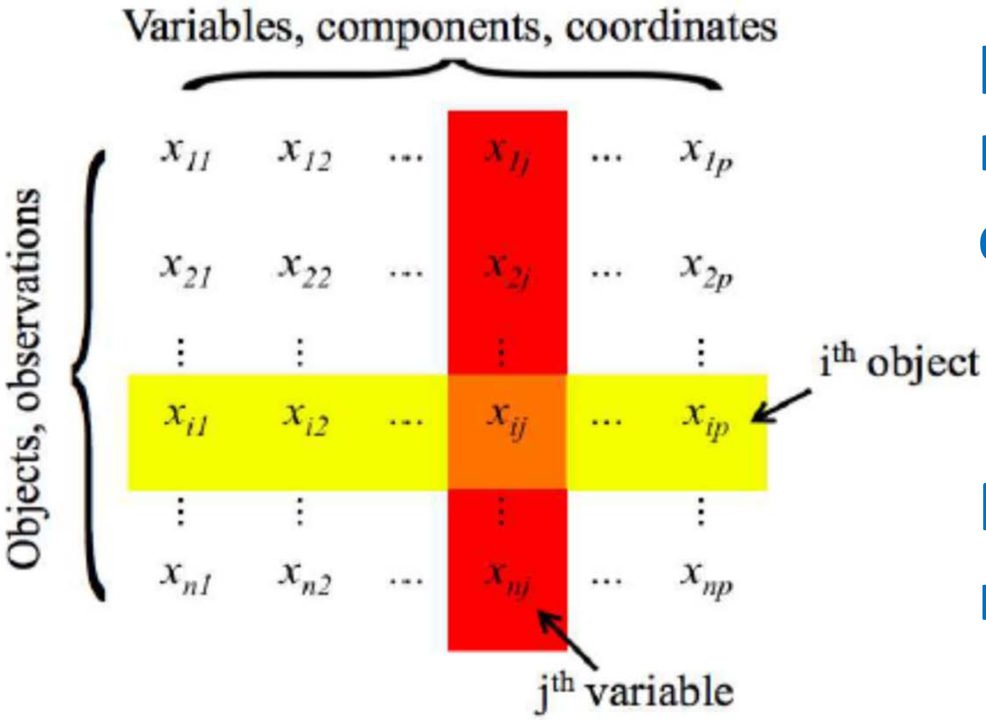
Correlation coefficient distance

$$d(X,Y) = 1 - \rho(X,Y) = 1 - \frac{Cov(X,Y)}{\sqrt{(Var(X) \cdot Var(Y)}}$$

Reminder:
Principal Component Analysis

Multivariable statistics and
Principal Component Analysis (PCA)

- A table of n observations in which p variables were measured



$$\begin{matrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{matrix}$$

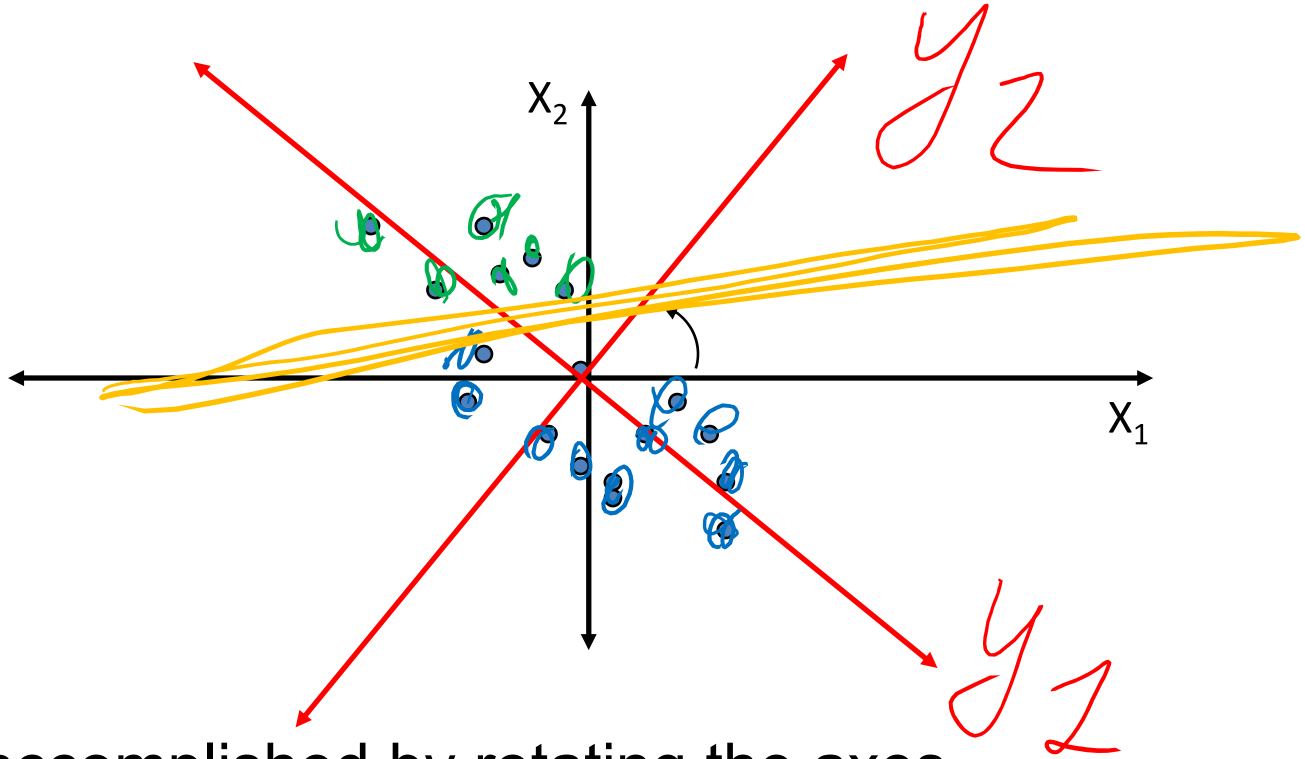
p x p symmetric matrix R of corr. coefficients

$$r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

PCA: Diagonalize matrix R

Trick: Rotate Coordinate Axes

Suppose we have a population measured on p random variables $X_1,\ldots,X_p$. Note that these random variables represent the p-axes of the Cartesian coordinate system in which the population resides. Our goal is to develop a new set of p axes (linear combinations of the original p axes) in the directions of greatest variability:



This is accomplished by rotating the axes.

Adapted from slides by Prof. S. Narasimhan, “Computer Vision” course at CMU

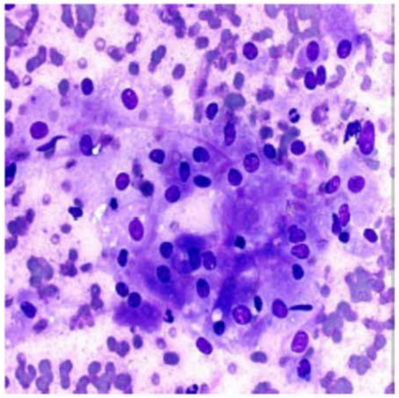
Principle Component Analysis (PCA)

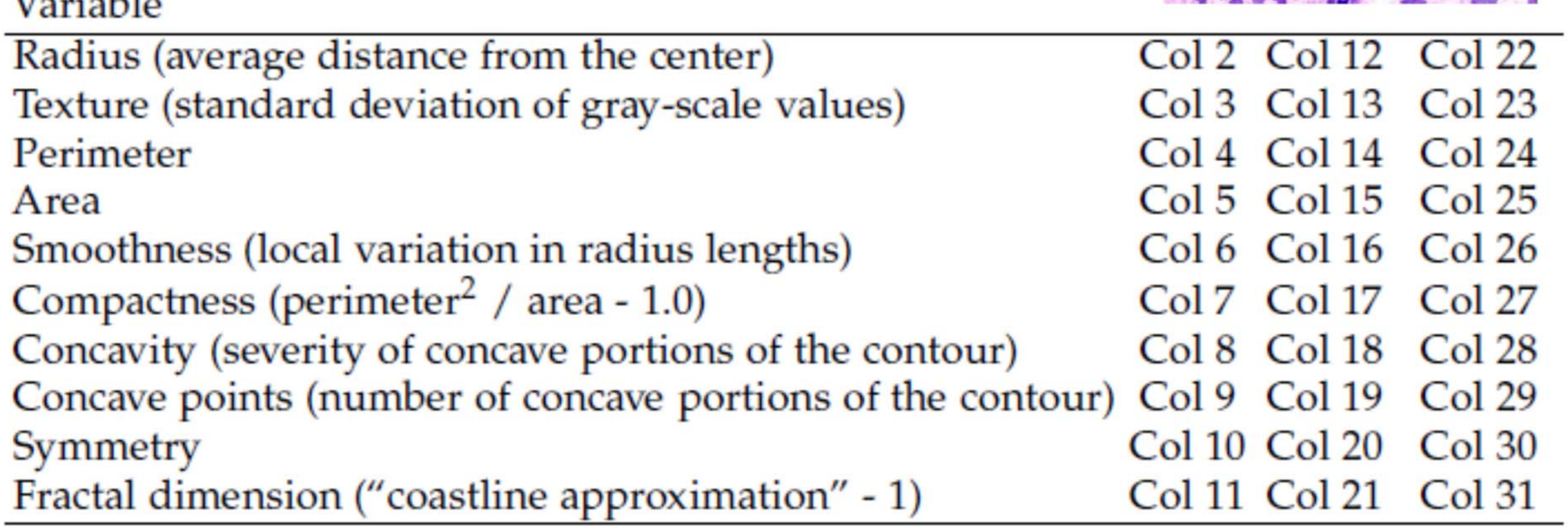
- p x p symmetric matrix *R* of corr. coefficients $r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$
- $R=n^{-1}Z'^{*}Z$ is a “square” of the matrix Z of standardized r.v.: $z_{\alpha k} = \frac{x_{\alpha k}-\mu_k}{\sigma_k}$ → all eigenvalues of R are non-negative
- Diagonal elements=1 → *tr(R)=p*
- Can be diagonalized:
  *R=V*D*V’* where *D* is the diagonal matrix
- d(1,1) –largest eig. value, d(p,p) – the smallest one
- The meaning of V(i,k) – contribution of the data type i to the k-th eigenvector
- tr(D)=p, the largest eigenvalue d(1,1) absorbs a fraction =d(1,1)/p of all correlations can be ~100%
- Scores: Y=Z*V: n x p matrix. Meaning of $Y(\alpha,k)$ – participation of the sample # α in the *k*-th eigenvector

Back to clustering

Let’s work with real cancer data!

- Data from Wolberg, Street, and Mangasarian (1994)
- Fine-needle aspirates = biopsy for breast cancer
- Black dots – cell nuclei. Irregular shapes/sizes may mean cancer
- 212 cancer patients and 357 healthy (column 1)
- 30 other properties (see table)





Variable			
Radius (average distance from the center)	Col 2	Col 12	Col 22
Texture (standard deviation of gray-scale values)	Col 3	Col 13	Col 23
Perimeter	Col 4	Col 14	Col 24
Area	Col 5	Col 15	Col 25
Smoothness (local variation in radius lengths)	Col 6	Col 16	Col 26
Compactness (perimeter^2 / area - 1.0)	Col 7	Col 17	Col 27
Concavity (severity of concave portions of the contour)	Col 8	Col 18	Col 28
Concave points (number of concave portions of the contour)	Col 9	Col 19	Col 29
Symmetry	Col 10	Col 20	Col 30
Fractal dimension (“coastline approximation” - 1)	Col 11	Col 21	Col 31

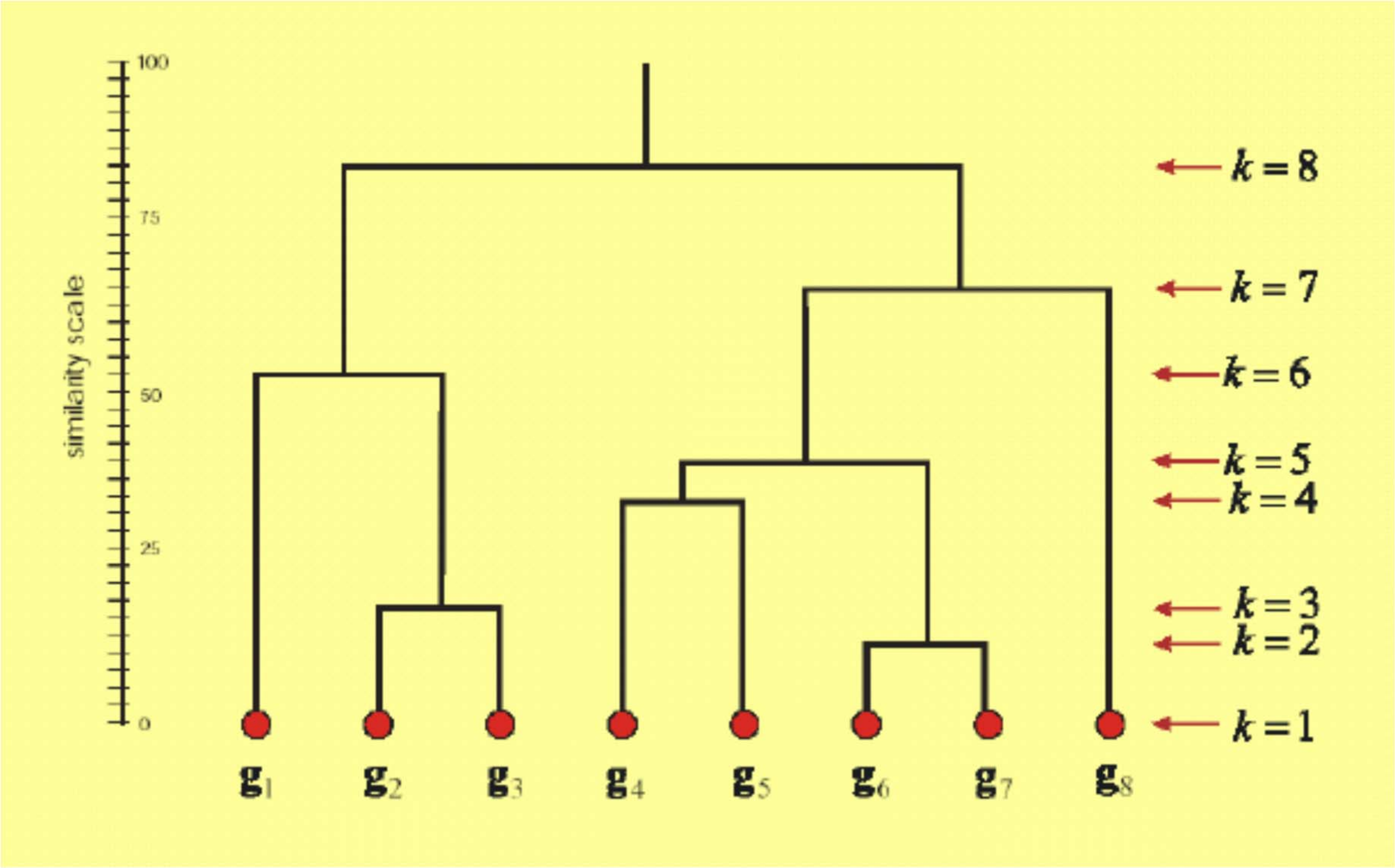
Common types of clustering algorithms

- Hierarchical if don’t know in advance # of clusters
  - Agglomerative: start: N clusters, merge into 1 cluster
  - Divisive: start with 1 cluster and breaks it up into N
- Non-hierarchical algorithms
  - Principal Component Analysis (PCA)
    - plot pairs of top eigenvectors of the covariance matrix Cov(X_i, X_j) and uses visual information to group
  - K-means clustering:
    - Iteratively apply the following two steps:
    - Calculate the centroid (center of mass) of each cluster
    - Assign each to the cluster to the nearest centroi

UPGMA algorithm

- Hierarchical agglomerative clustering algorithm
- **UPGMA** = Unweighted Pair Group Method with Arithmetic mean
- Iterative algorithm:
- Start with a pair with the smallest d(X,Y)
- Cluster these two together and replace it with their arithmetic mean (X+Y)/2
- Recalculate all distances to this new “cluster node”
- Repeat until all nodes are merged

Output of UPGMA algorithm



UPGMA algorithm

25 samples

250 genes on chromosome 21

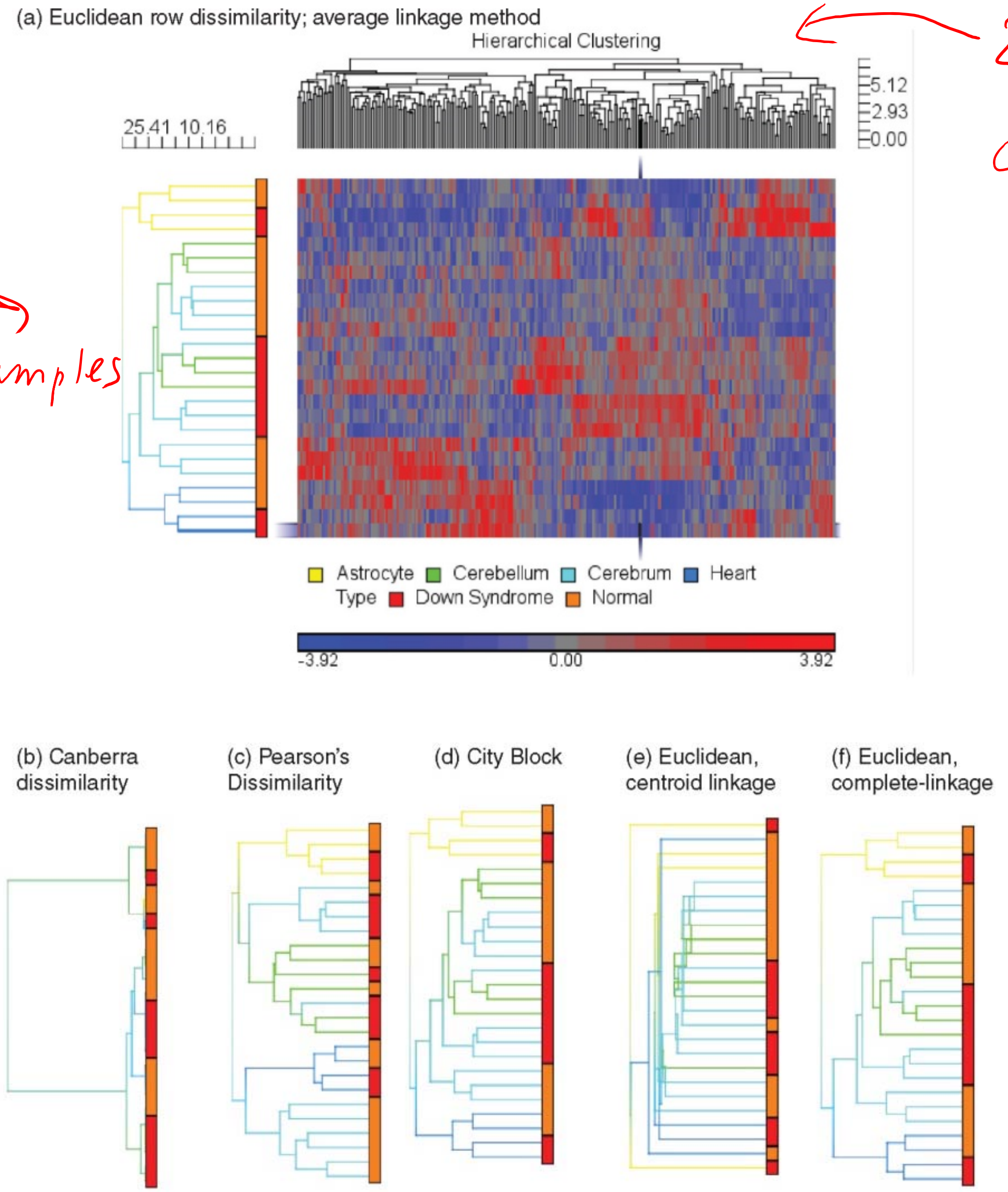
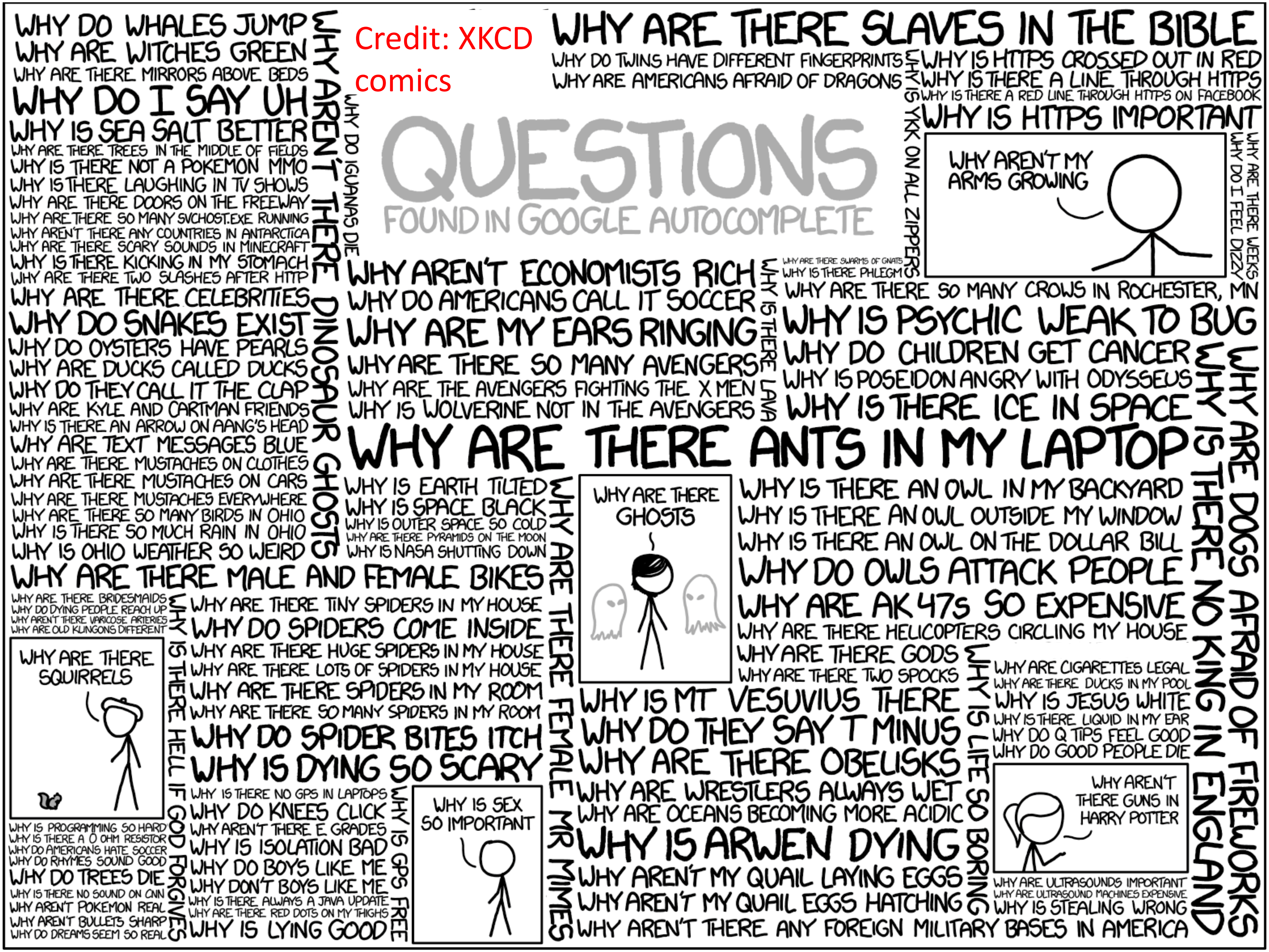


FIGURE 11.16 Hierarchical clustering of 250 chromosome 21 transcripts in 25 samples using Partek software. (a) Hierarchical clustering of microarray data using the default settings of Euclidean dissimilarity for rows (samples) and columns (transcripts). Colors correspond to expression intensity values.

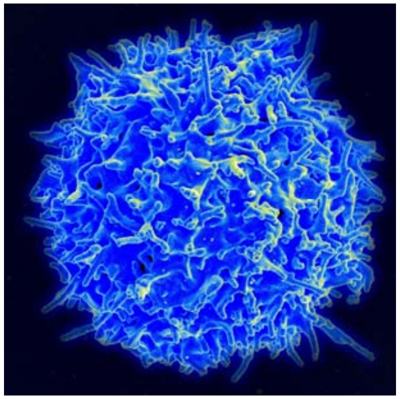

QUESTIONS
FOUND IN GOOGLE AUTOCOMPLETE
WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD
WHY AREN'T THERE DINOSAUR GHOSTS
WHY DO IGUANAS DIE
WHY ARE THERE SLAVES IN THE BIBLE
WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS
WHY IS YKK ON ALL ZIPPERS
WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT
WHY AREN'T MY ARMS GROWING
WHY DO I FEEL DIZZY
WHY ARE THERE WEEKS
WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS
WHY IS THERE LAVA
WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE
WHY ARE THERE ANTS IN MY LAPTOP
WHY IS EARTH TILTED
WHY IS SPACE BLACK
WHY IS OUTER SPACE SO COLD
WHY ARE THERE PYRAMIDS ON THE MOON
WHY IS NASA SHUTTING DOWN
WHY ARE THERE GHOSTS
WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS
WHY IS THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS
WHY ARE THERE MALE AND FEMALE BIKES
WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KLINGONS DIFFERENT
WHY ARE THERE SQUIRRELS
WHY IS THERE HELL IF GOD FORGIVES
WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY
WHY ARE THERE FEMALE MR MIMES
WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA
WHY IS LIFE SO BORING
WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE
WHY AREN'T THERE GUNS IN HARRY POTTER
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL
WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD
WHY IS GPS FREE
WHY IS SEX SO IMPORTANT

Credit: XKCD comics

Matlab demo

Human T cell expression data

- The matrix contains 47 expression samples from Lukk et al, Nature Biotechnology 2010
- All samples are from T cells in different individuals
- Only the top 3000 genes with the largest variability were used
- The value is log2 of gene's expression level in a given sample as measured by the microarray technology

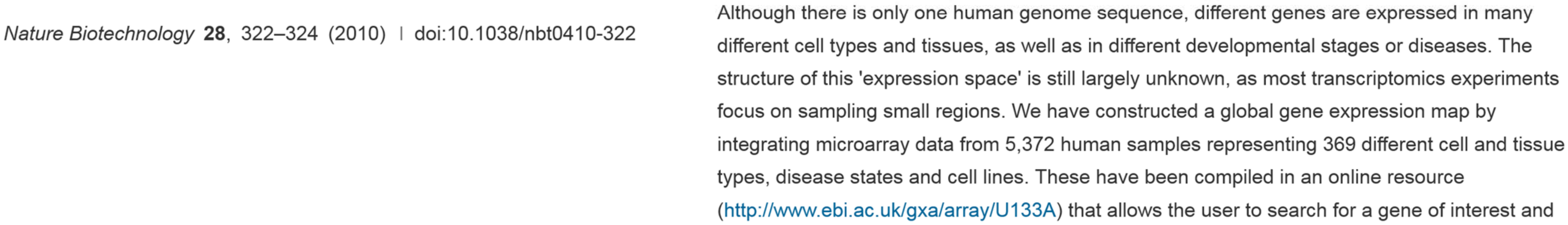


a T cell

A global map of human gene expression

Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

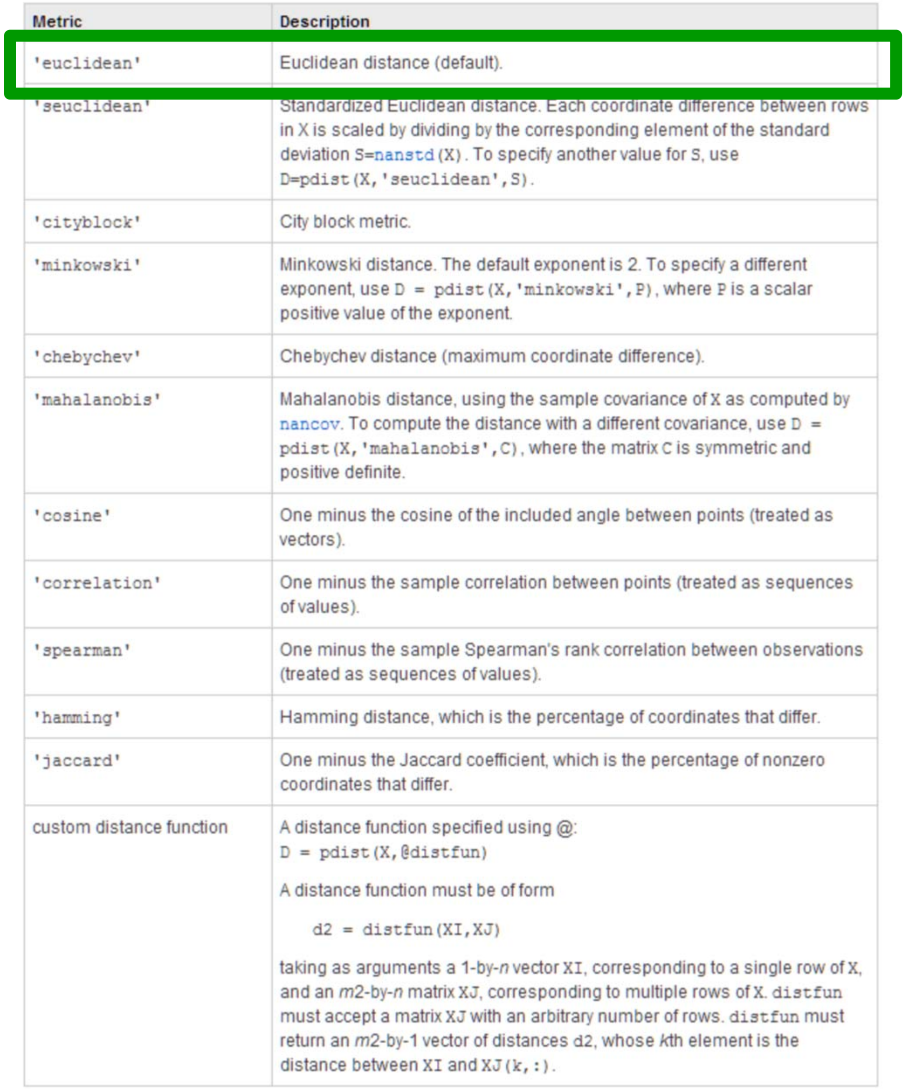
Affiliations | Corresponding author



Nature Biotechnology **28**, 322–324 (2010) | doi:10.1038/nbt0410-322

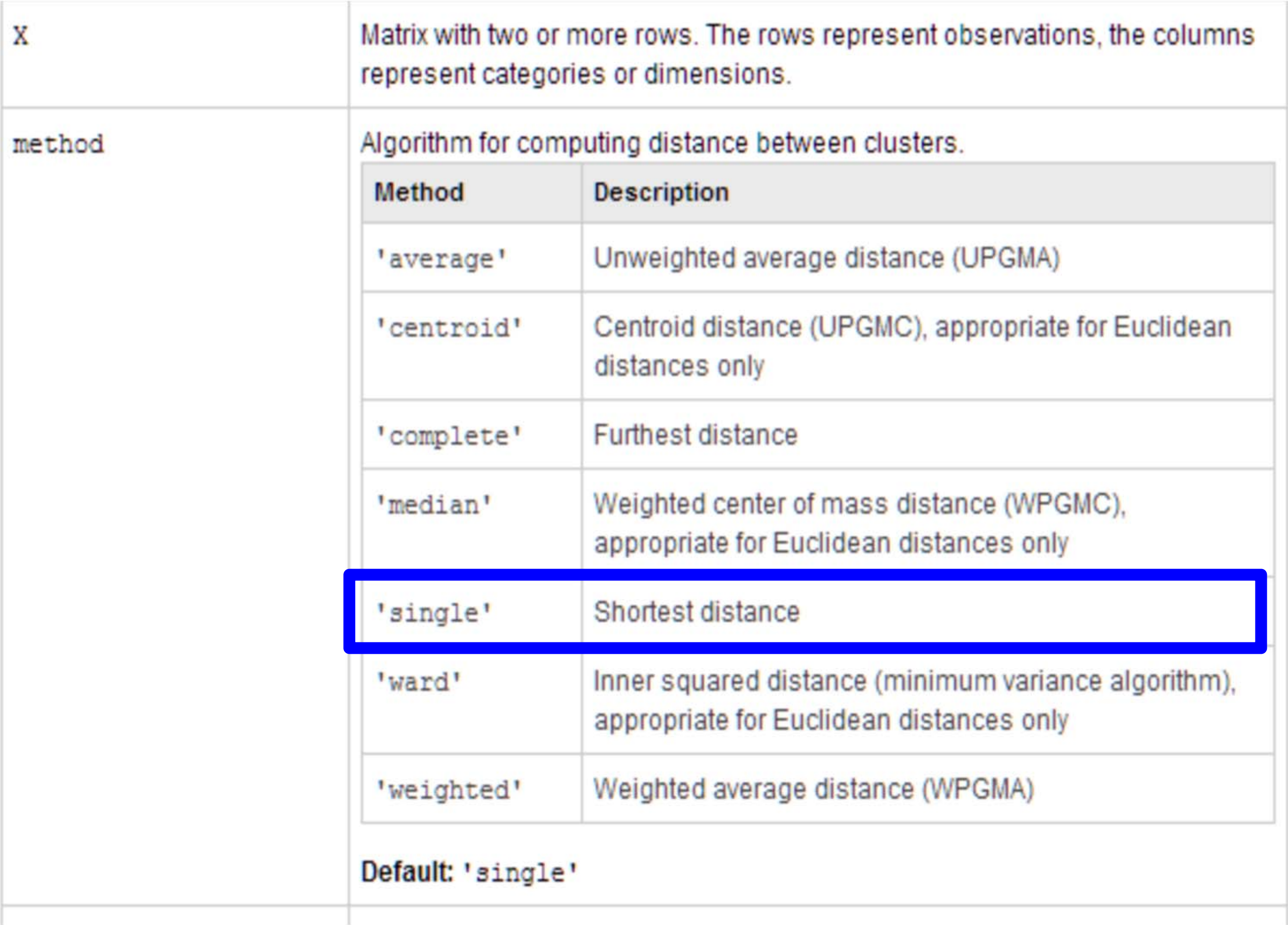
Although there is only one human genome sequence, different genes are expressed in many different cell types and tissues, as well as in different developmental stages or diseases. The structure of this 'expression space' is still largely unknown, as most transcriptomics experiments focus on sampling small regions. We have constructed a global gene expression map by integrating microarray data from 5,372 human samples representing 369 different cell and tissue types, disease states and cell lines. These have been compiled in an online resource (http://www.ebi.ac.uk/gxa/array/U133A) that allows the user to search for a gene of interest and

Choices of distance metrics in clustergram(... 'RowPDistValue' ..., 'ColumnPDistValue' ...,)



Metric	Description
'euclidean'	Euclidean distance (default).
'seuclidean'	Standardized Euclidean distance. Each coordinate difference between rows in X is scaled by dividing by the corresponding element of the standard deviation S=nanstd(X). To specify another value for S, use D=pdist(X,'seuclidean',S).
'cityblock'	City block metric.
'minkowski'	Minkowski distance. The default exponent is 2. To specify a different exponent, use D = pdist(X,'minkowski',P), where P is a scalar positive value of the exponent.
'chebychev'	Chebychev distance (maximum coordinate difference).
'mahalanobis'	Mahalanobis distance, using the sample covariance of X as computed by nancov. To compute the distance with a different covariance, use D = pdist(X,'mahalanobis',C), where the matrix C is symmetric and positive definite.
'cosine'	One minus the cosine of the included angle between points (treated as vectors).
'correlation'	One minus the sample correlation between points (treated as sequences of values).
'spearman'	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).
'hamming'	Hamming distance, which is the percentage of coordinates that differ.
'jaccard'	One minus the Jaccard coefficient, which is the percentage of nonzero coordinates that differ.
custom distance function	A distance function specified using @: D = pdist(X,@distfun) A distance function must be of form d2 = distfun(XI,XJ) taking as arguments a 1-by-*n* vector XI, corresponding to a single row of X, and an *m2*-by-*n* matrix XJ, corresponding to multiple rows of X. distfun must accept a matrix XJ with an arbitrary number of rows. distfun must return an *m2*-by-1 vector of distances d2, whose *k*th element is the distance between XI and XJ(k,:).

Choices of hierarchical clustering algorithm in clustergram(…’linkage’,…)



X	Matrix with two or more rows. The rows represent observations, the columns represent categories or dimensions.
method	Algorithm for computing distance between clusters.

Method	Description
'average'	Unweighted average distance (UPGMA)
'centroid'	Centroid distance (UPGMC), appropriate for Euclidean distances only
'complete'	Furthest distance
'median'	Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only
'single'	Shortest distance
'ward'	Inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only
'weighted'	Weighted average distance (WPGMA)

Default: 'single'