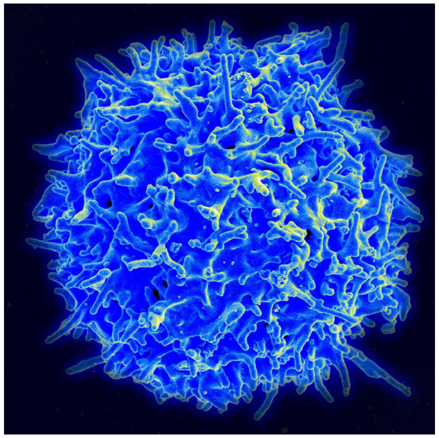
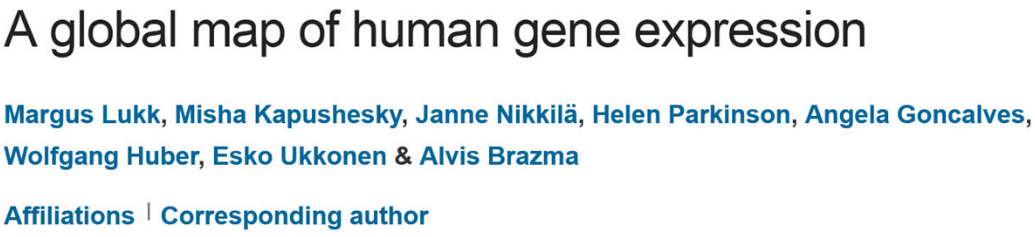
Clustering analysis of gene expression data

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

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



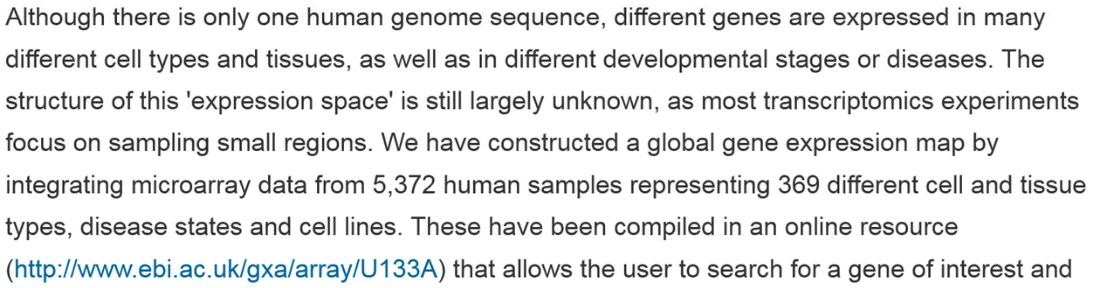


A global map of human gene expression

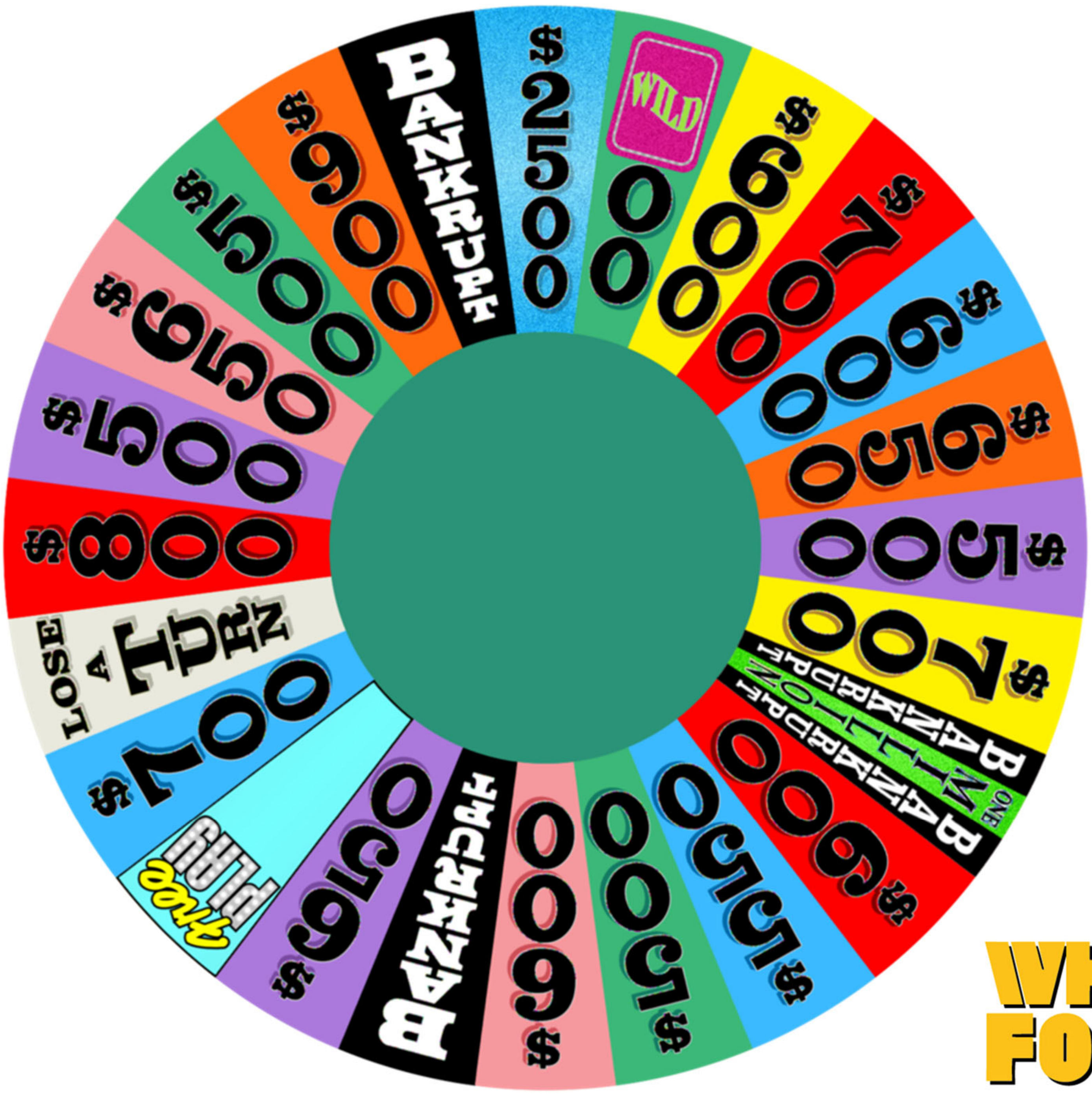
Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen & Alvis Brazma

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®

Correlated pairs

plausible biological connection based on short description

g1=1994; g2=188; group 1

g1=2872; g2=1269; group 2

g1=1321; g2=10; group 3

g1= 886; g2=819; group 4

g1=2138; g2=1364; group 5

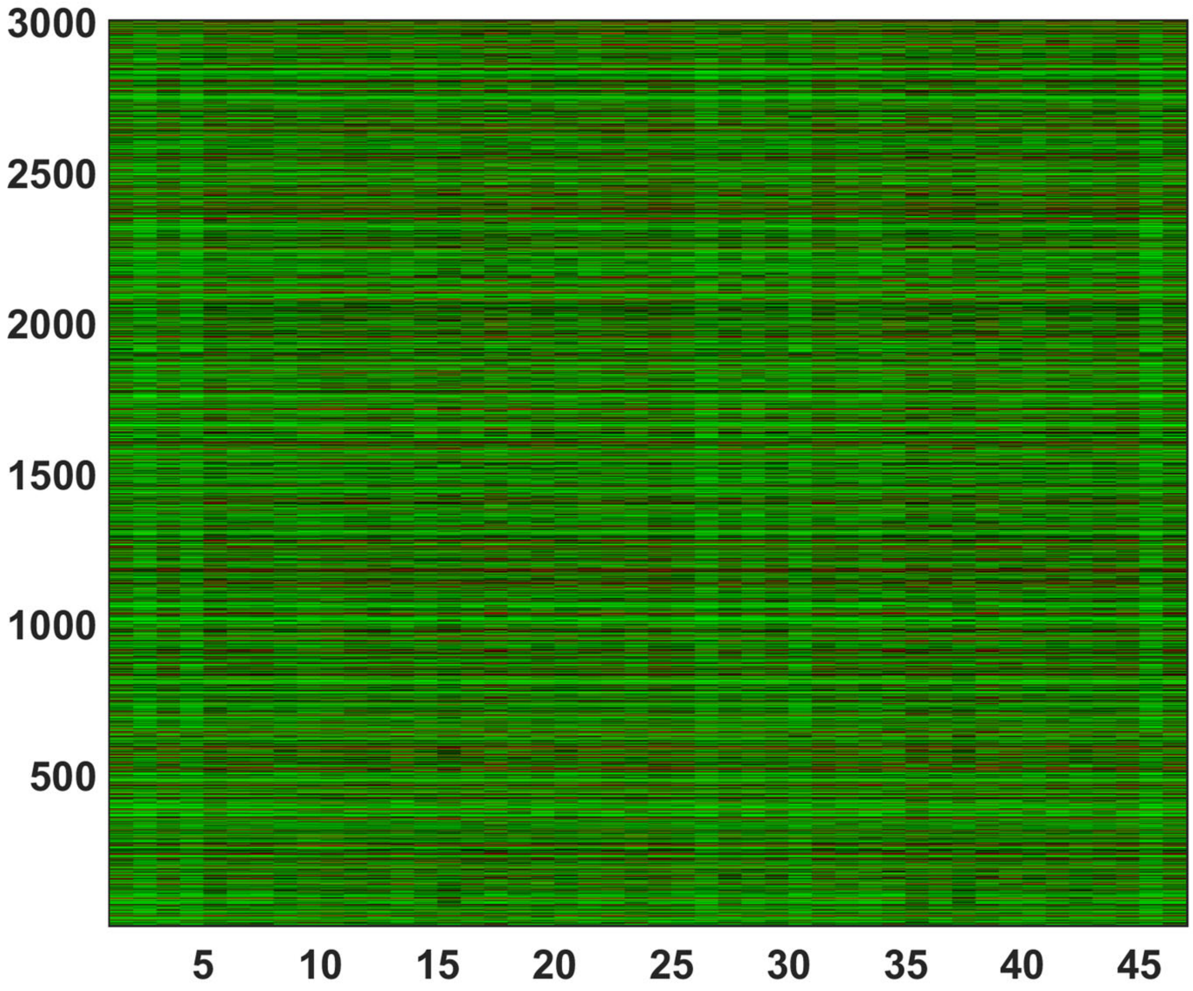
no obvious biological common function

```
g1=1+floor(rand.*3000); g2=1+floor(rand.*3000);
disp([g1, g2])
```

Matlab exercise

- Every group works with g0=2907; g1=1527; g2=2629; g3=2881; g4=1144; g5=1066;
- Compute Multiple Linear Regression (MLR), where y=exp_t (g0);
  x1= exp_t (g1); x2= exp_t (g2);
- How much better the MLR did compared to the Single Linear Regression (SLR)?
- Continue increasing the number of genes in x until R_adj starts to decrease

How to find the entire groups of mutually correlated genes if you have many genes and many samples?



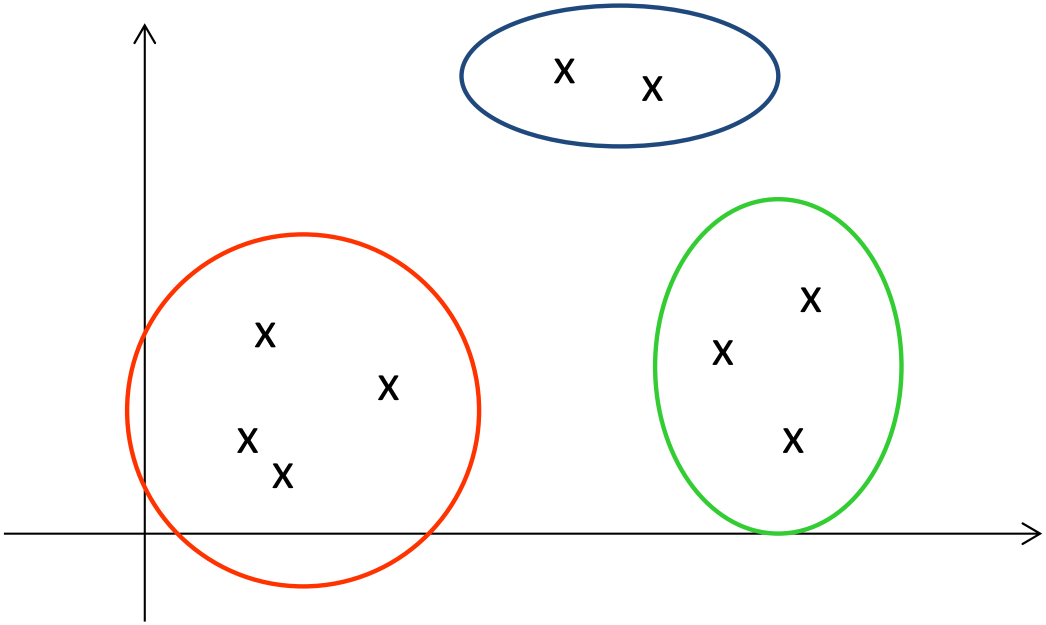
Clustering to the rescue!

Clustering is a part of Machine Learning

- ***Supervised Learning:***
  A machine learning technique whereby a system uses a set of human-labelled training examples to learn how to correctly perform a task
  Example: a sample of cancer expression profiles each **annotated** with cancer type
  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 treatments

What is clustering?

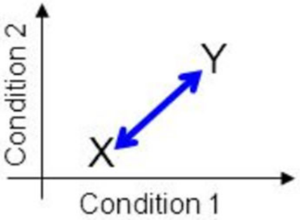
- The goal of clustering is to
  - group data points that are close (or **similar**) to each other
  - Usually, one needs to identify such groups (or clusters) in an **unsupervised** manner
  - Sometimes one takes into account **prior information** (Bayesian methods)
- Need to define some distance d_{ij} between objects i and j
- Clustering is easy in 2 dimensions but hard in 3000 dimensions -> need to somehow reduce dimensionality



How to define the distance?

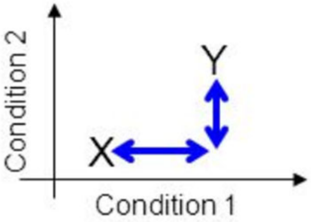
- Euclidean distance:
  - Most commonly used distance
  - Sphere shaped cluster
  - Corresponds to the geometric distance into the multidimensional space

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



- 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

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

Common types of clustering algorithms

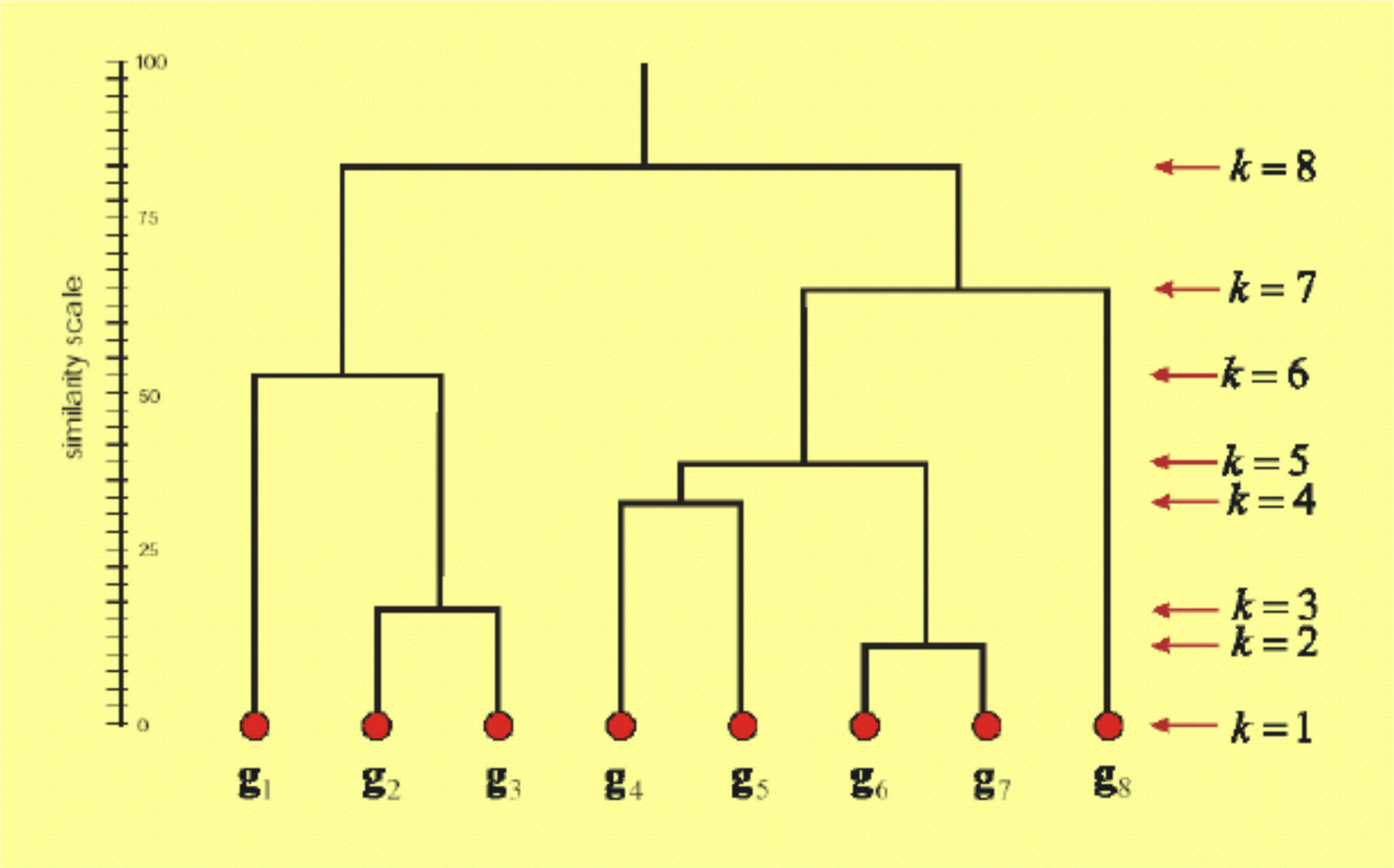
- Hierarchical if one doesn’t know in advance the # of clusters
  - Agglomerative: start with N clusters and gradually merge them into 1 cluster
  - Divisive: start with 1 cluster and gradually break it up into N clusters
- Non-hierarchical algorithms
  - K-means clustering:
    - <u>Iteratively</u>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

Hierarchical clustering

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

250 genes on chromosome 21

25 samples

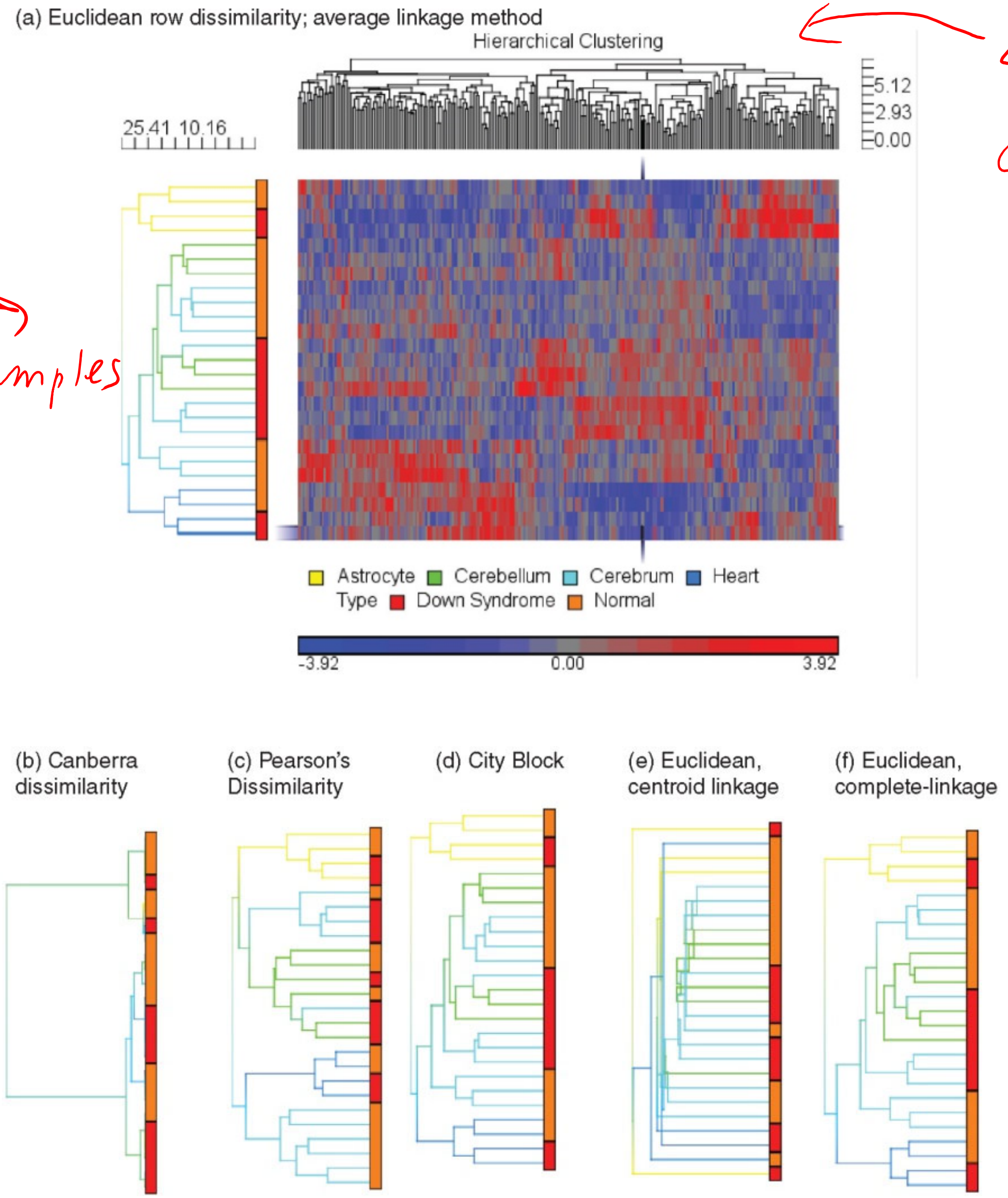
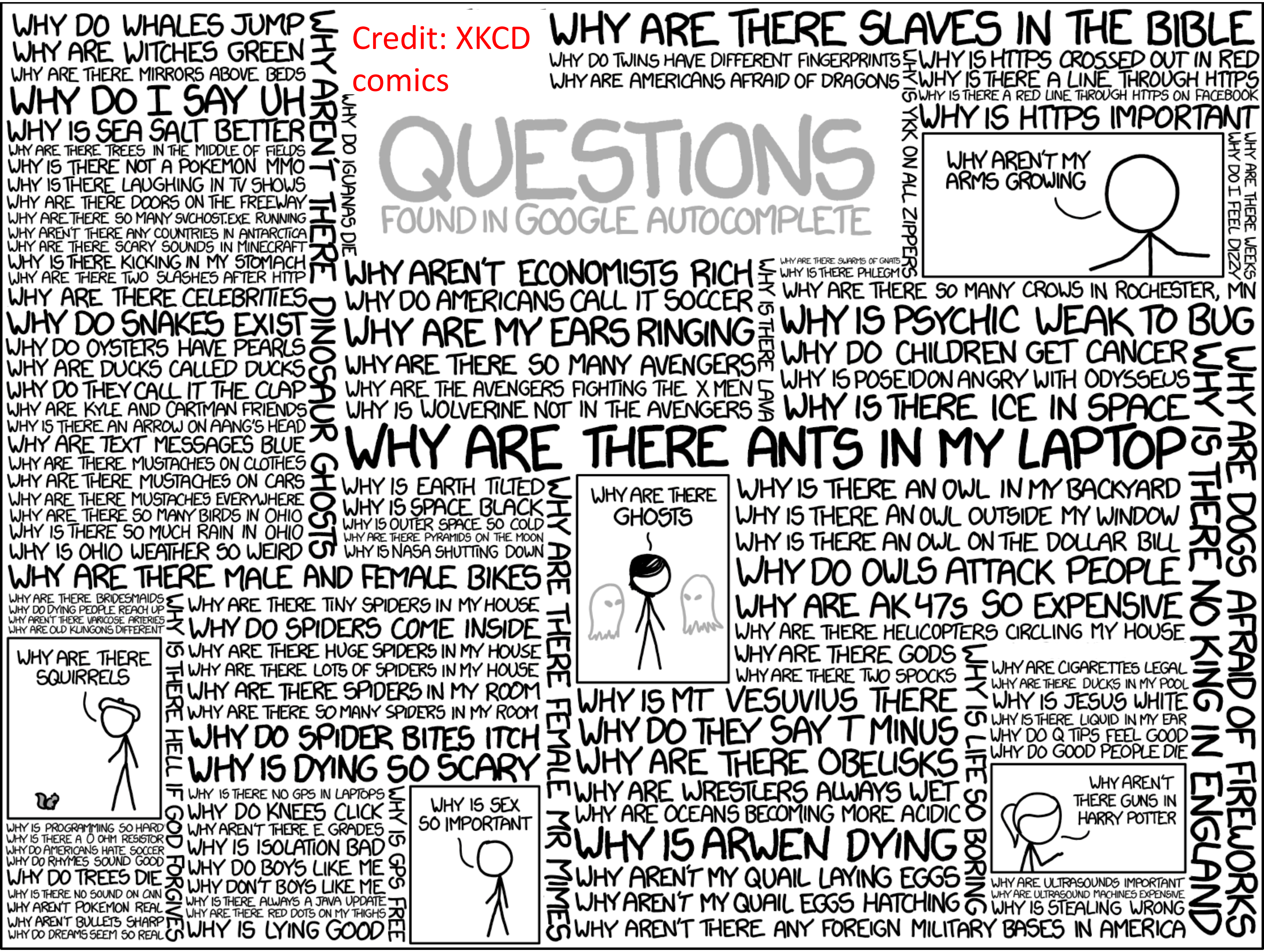


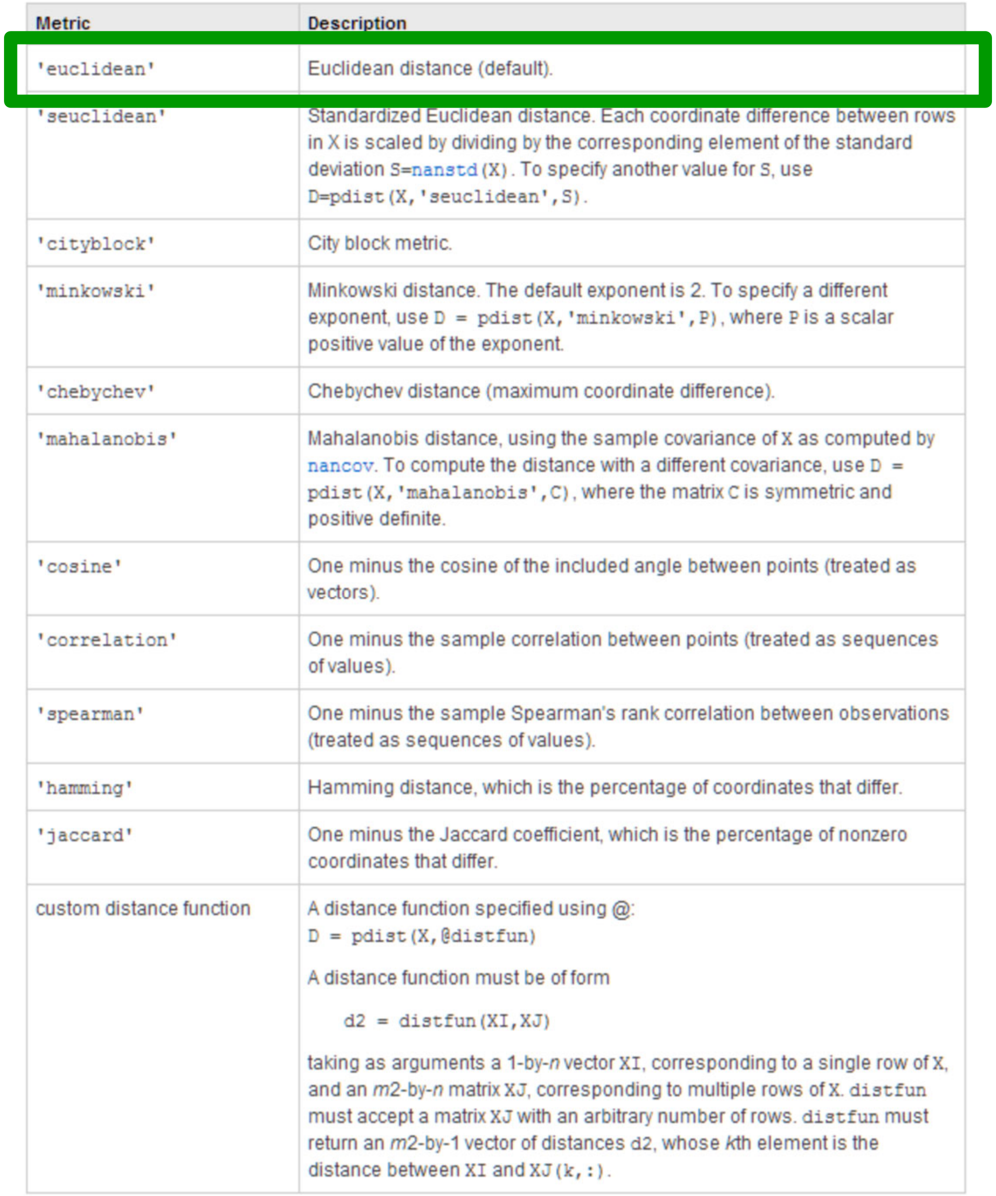
FIGURE 11.16 Hierarchical clustering of 250 chromosome 21 transcripts in 25 samples using Partek software. (a) Hierarchical clustering of microarray data using the default settings of Euclidean dissimilarity for rows (samples) and columns (transcripts). Colors correspond to expression intensity values.


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 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 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 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 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
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 ARE THERE WEEKS
WHY DO I FEEL DIZZY
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 FEMALE MR MIMES
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 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 THERE NO KING IN ENGLAND
WHY ARE DOGS AFRAID OF FIREWORKS

Credit: XKCD comics

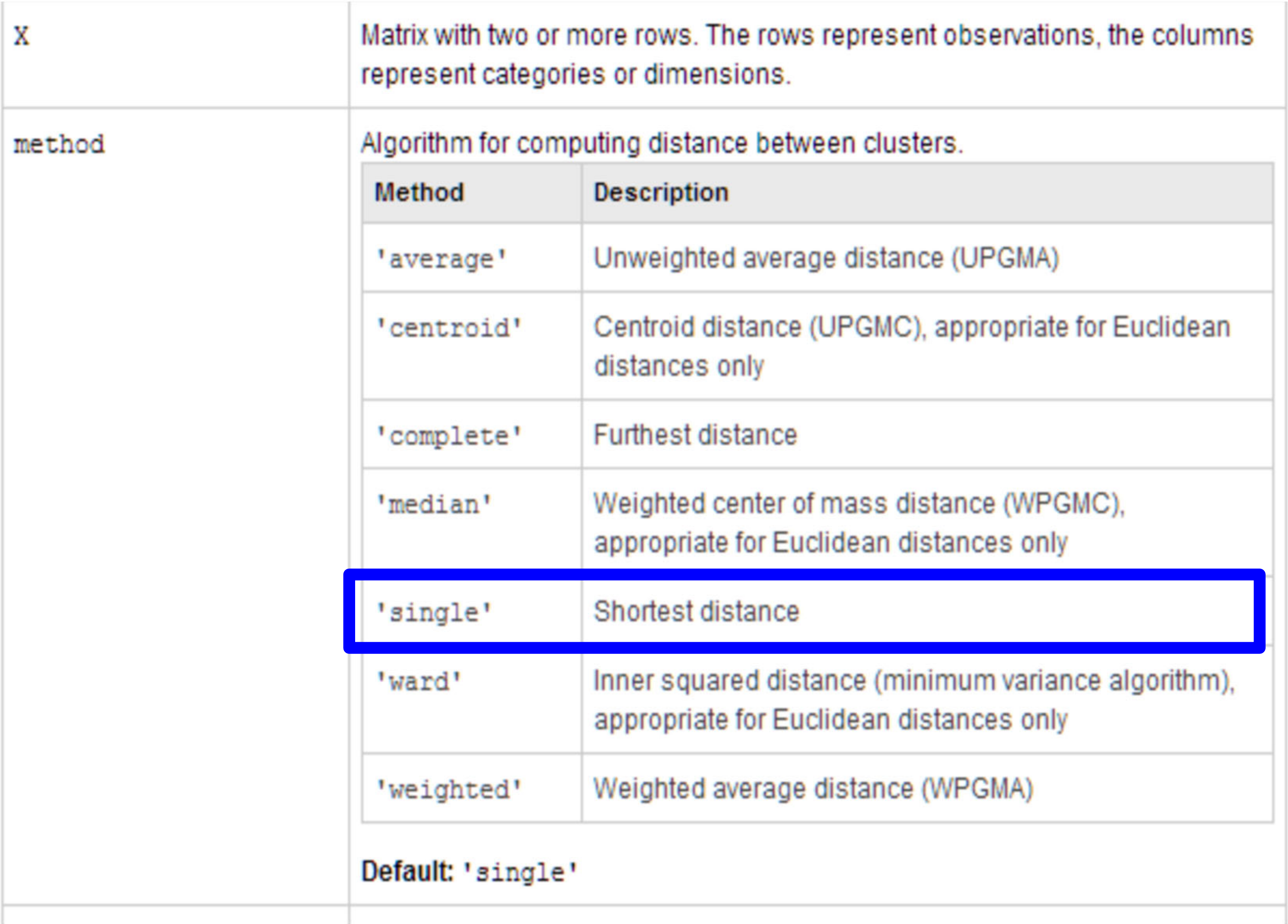
Clustering
Matlab demo

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 *m*2-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 *m*2-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’,…)



<table>
<tr><td>X</td><td>Matrix with two or more rows. The rows represent observations, the columns represent categories or dimensions.</td></tr>
<tr><td>method</td><td>Algorithm for computing distance between clusters.
<table>
<tr><th>Method</th><th>Description</th></tr>
<tr><td>'average'</td><td>Unweighted average distance (UPGMA)</td></tr>
<tr><td>'centroid'</td><td>Centroid distance (UPGMC), appropriate for Euclidean distances only</td></tr>
<tr><td>'complete'</td><td>Furthest distance</td></tr>
<tr><td>'median'</td><td>Weighted center of mass distance (WPGMC), appropriate for Euclidean distances only</td></tr>
<tr><td>'single'</td><td>Shortest distance</td></tr>
<tr><td>'ward'</td><td>Inner squared distance (minimum variance algorithm), appropriate for Euclidean distances only</td></tr>
<tr><td>'weighted'</td><td>Weighted average distance (WPGMA)</td></tr>
</table>
<b>Default:</b> 'single'</td></tr>
</table>

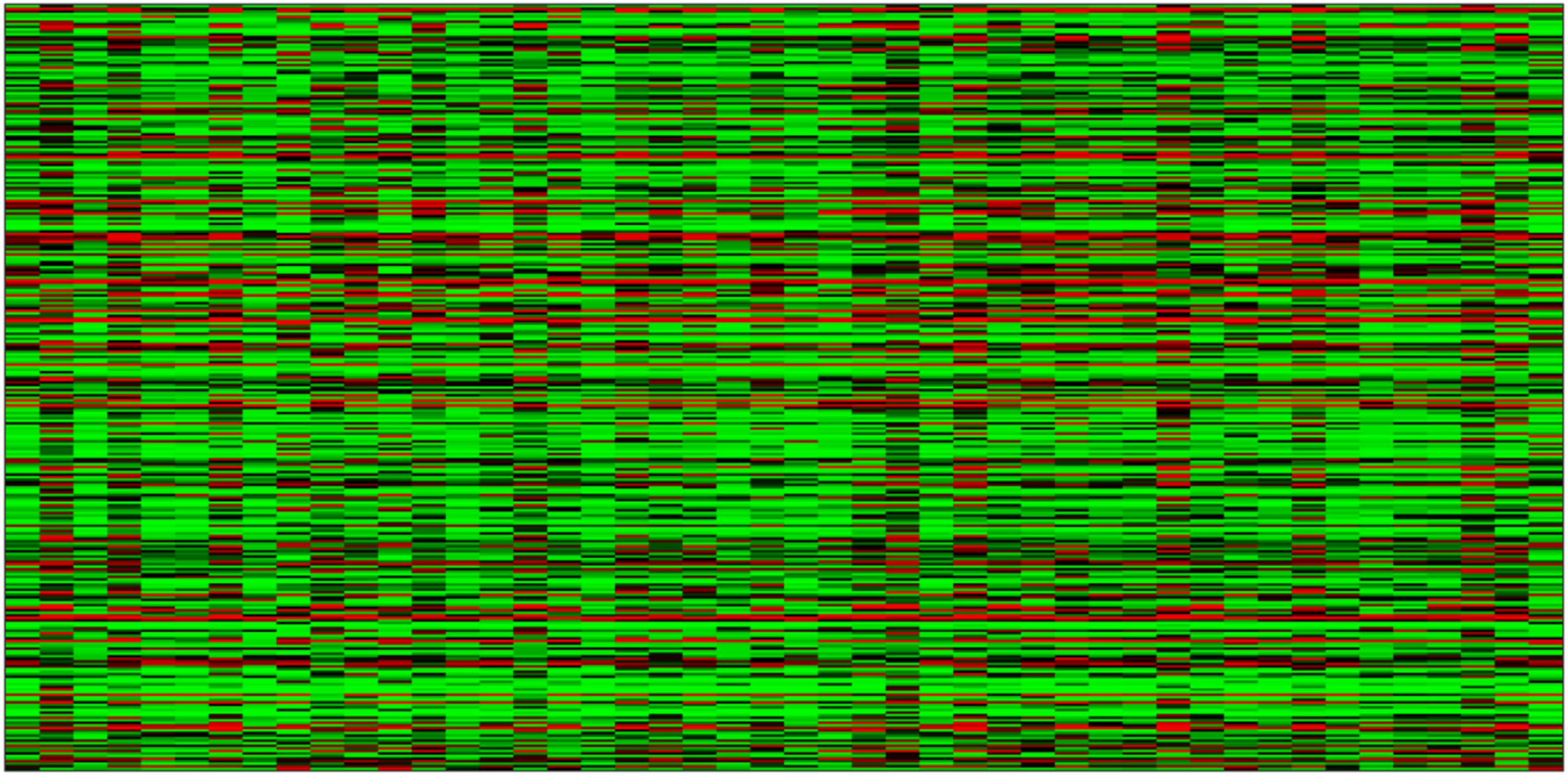
Clustering group exercise

- Each group will analyze a cluster of genes identified in the T cell expression table
- Analyze the table of top 100 genes **by variance** in 47 samples
- Cluster them using:
  - Group 1: UPGMA = ‘linkage’, ‘average’, ‘RowPDistValue’, ’euclidean’,
  - Group 2: ‘linkage’, ‘single’, ‘RowPDistValue’, ’cityblock’,
  - Group 3: ‘linkage’, ‘average’, ‘RowPDistValue’, ’correlation’,
  - Group 4: UPGMA = ‘linkage’, ‘single’, ‘RowPDistValue’, ’euclidean’,
  - Group 5: UPGMA = ‘linkage’, ‘weighted’, ‘RowPDistValue’, ’correlation’,
- Use clustergram(…, 'Standardize','Row', ‘linkage’, ***as specified for your group,*** ‘RowPDistValue’ ***as specified for your group,*** 'RowLabels',gene_names1,'ColumnLabels‘, array_names)

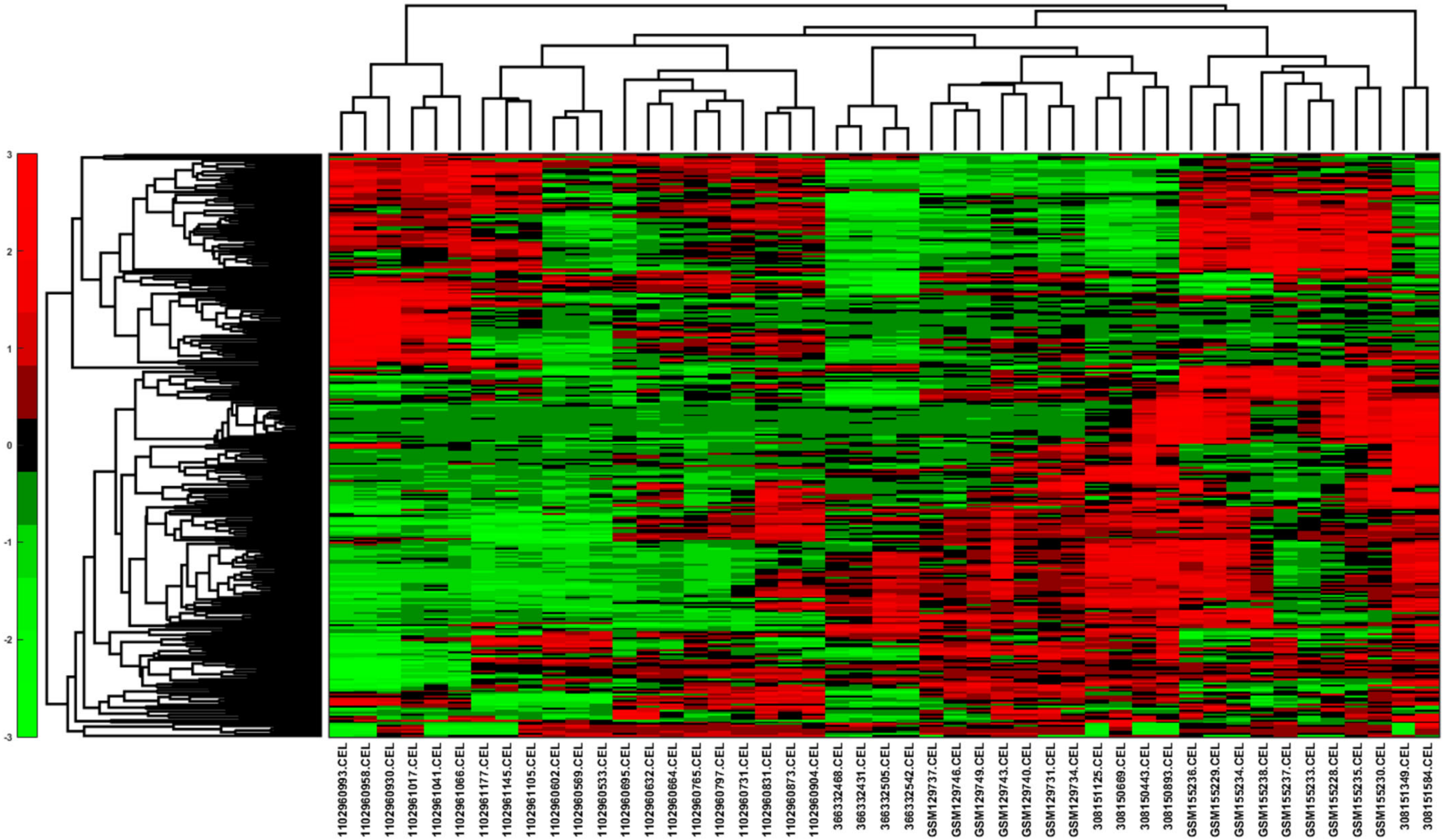
Matlab code

- load expression_table.mat
- gene_variation=std(exp_t')';
- [a,b]=sort(gene_variation,'descend');
- ngenes=100;
- exp_t1=exp_t(b(1:ngenes),:);
- gene_names1=gene_names(b(1:ngenes));
- %%% for group 1
- CGobj1 = clustergram(exp_t1, 'Standardize','Row', 'RowLabels', gene_names1,'ColumnLabels',array_names)
- set(CGobj1,'RowLabels',gene_names1,'ColumnLabels', array_names,'linkage', 'average','RowPDist','euclidean');

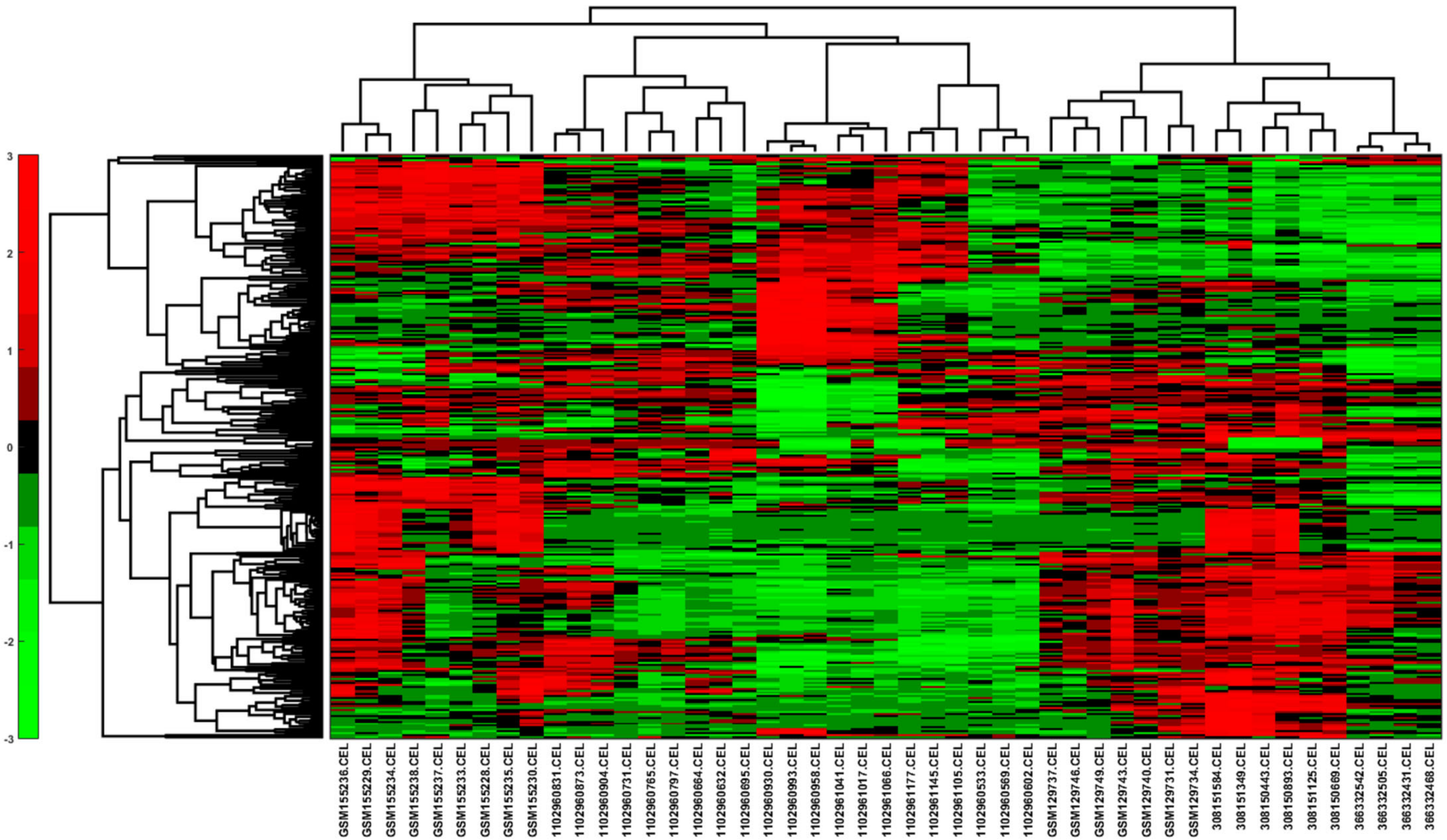
Before clustering



UPGMA hierarchical clustering, Euclidian distance



UPGMA hierarchical clustering, correlation distance


GSM155236.CEL
GSM155229.CEL
GSM155234.CEL
GSM155238.CEL
GSM155237.CEL
GSM155233.CEL
GSM155228.CEL
GSM155235.CEL
GSM155230.CEL
1102960831.CEL
1102960873.CEL
1102960904.CEL
1102960731.CEL
1102960765.CEL
1102960797.CEL
1102960664.CEL
1102960632.CEL
1102960695.CEL
1102960930.CEL
1102960993.CEL
1102960958.CEL
1102961041.CEL
1102961017.CEL
1102961066.CEL
1102961177.CEL
1102961145.CEL
1102961105.CEL
1102960533.CEL
1102960569.CEL
1102960602.CEL
GSM129737.CEL
GSM129746.CEL
GSM129749.CEL
GSM129743.CEL
GSM129740.CEL
GSM129731.CEL
GSM129734.CEL
308151584.CEL
308151349.CEL
308150443.CEL
308150893.CEL
308151125.CEL
308150669.CEL
366332542.CEL
366332505.CEL
366332431.CEL
366332468.CEL