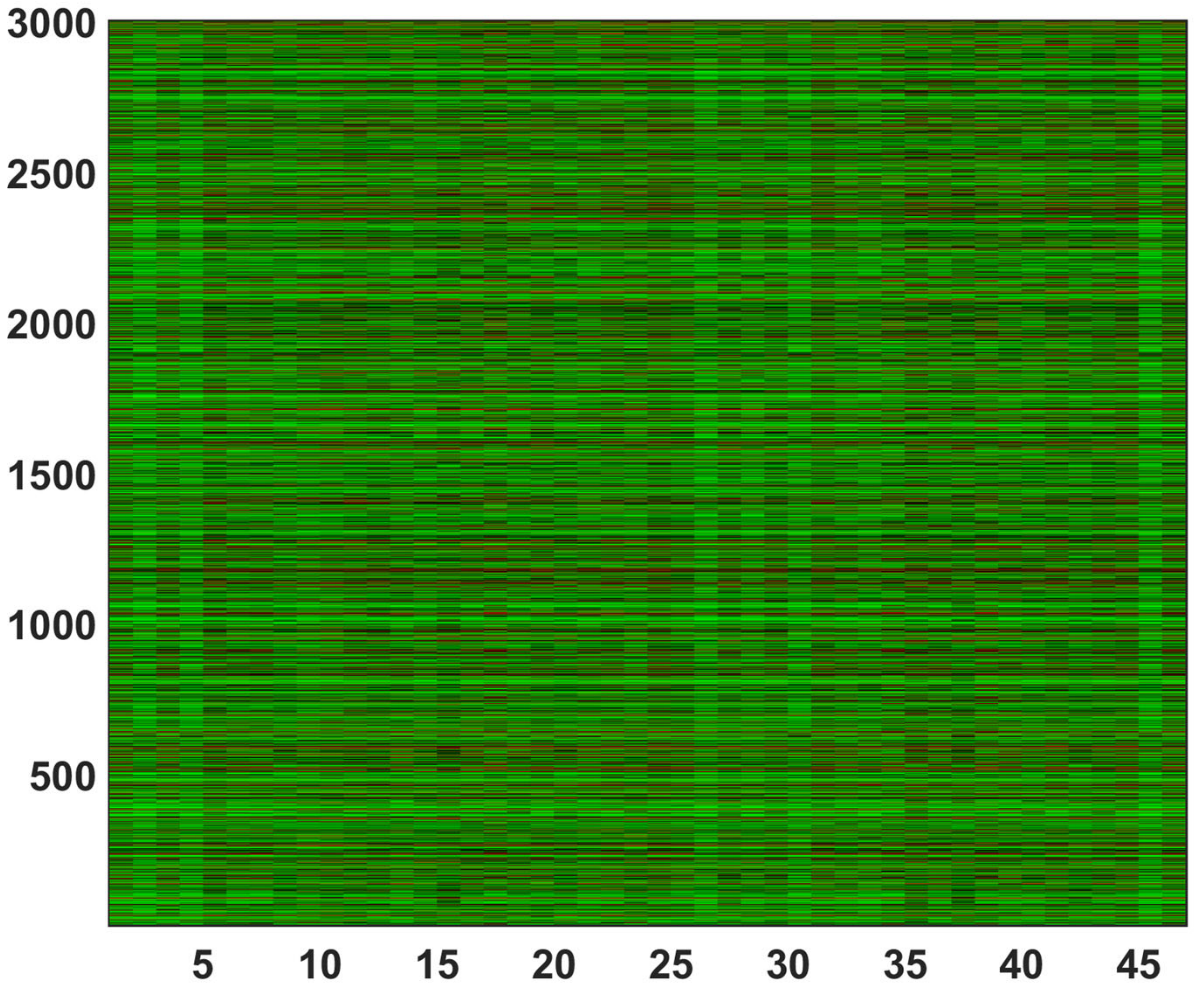
Clustering and network analysis of gene expression data

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

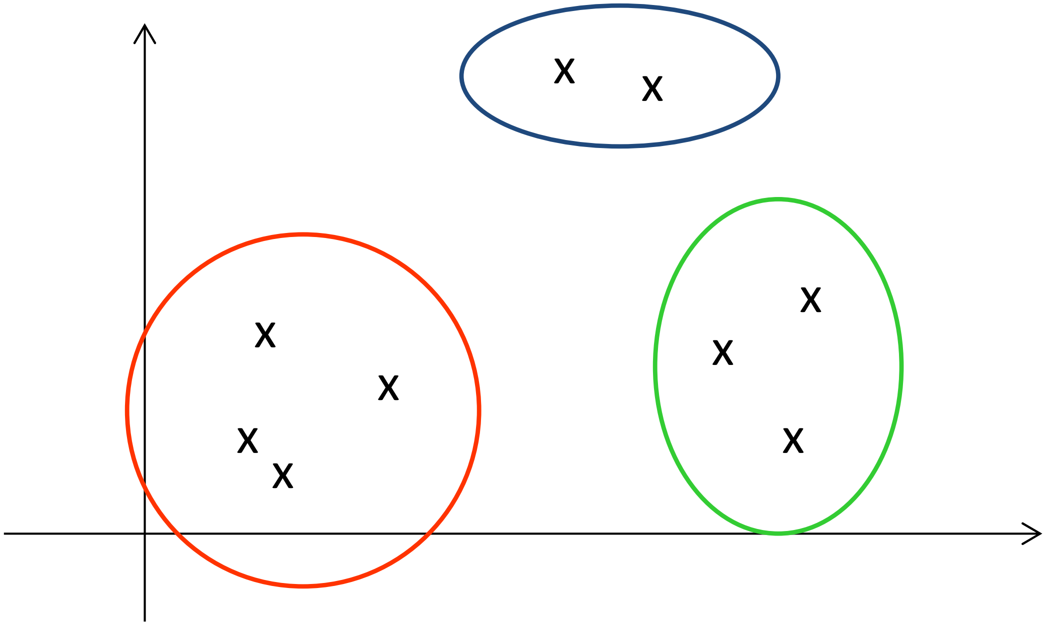
How to find the entire groups of mutually correlated genes if you have many genes and many samples?



Clustering to the rescue!

What is clustering?

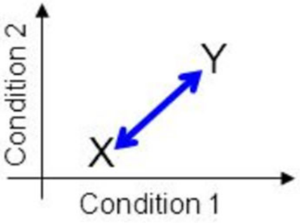
- The goal of clustering is to
  - group data points that are close (or **similar**) to each other
  - Usually, one needs to identify such groups (or clusters) in an **unsupervised** manner
  - Sometimes one takes into account **prior information** (Bayesian methods)
- Need to define some distance d_{ij} 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

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

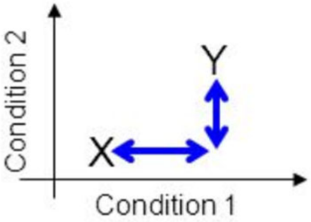


The Canberra distance metric is calculated in R by

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

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

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



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

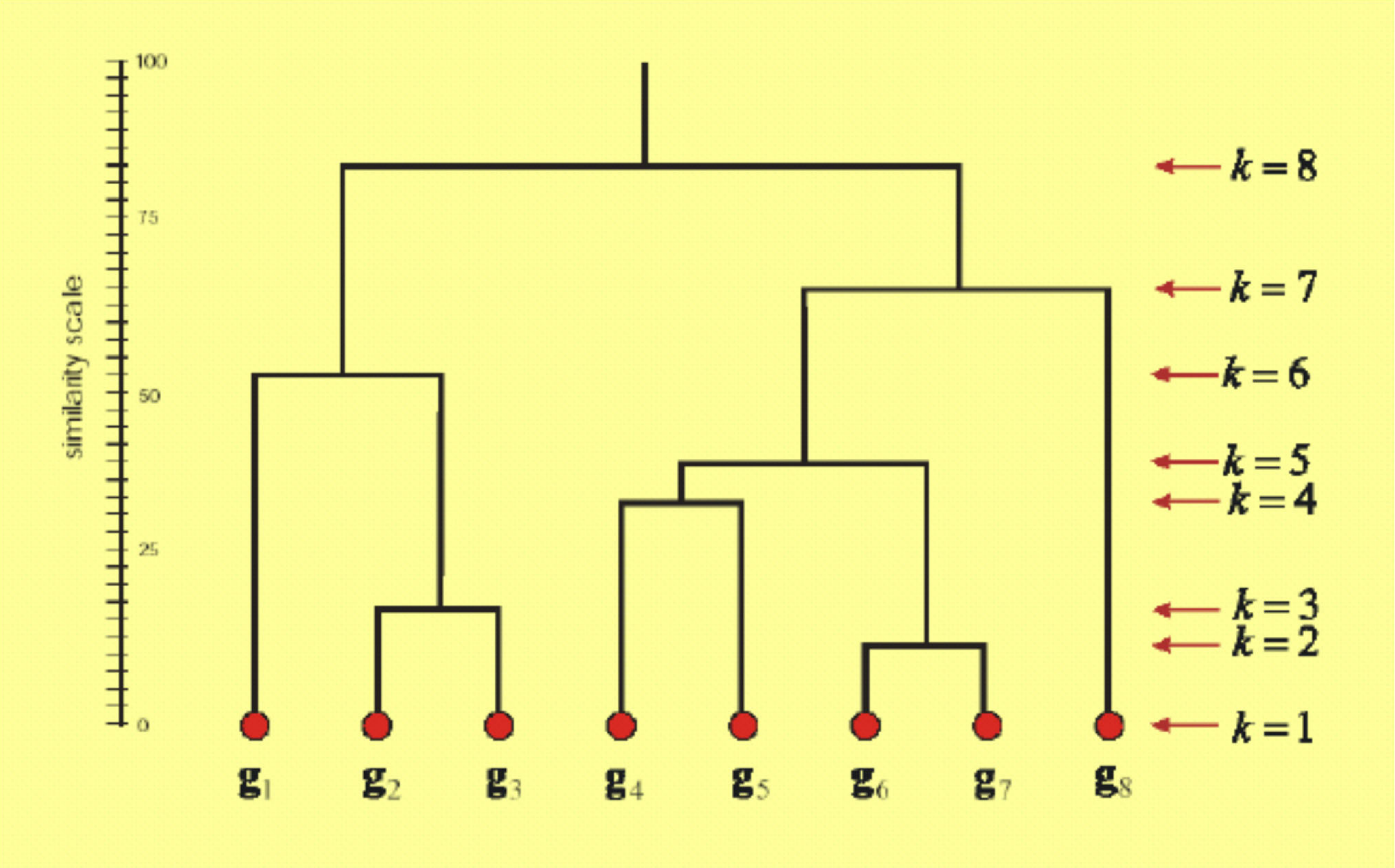
- Hierarchical if one doesn’t know in advance the # of clusters
  - Agglomerative: start with N clusters and gradually merge them into 1 cluster
  - Divisive: start with 1 cluster and gradually break it up into N clusters
- Non-hierarchical algorithms
  - K-means clustering:
    - Iteratively apply the 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

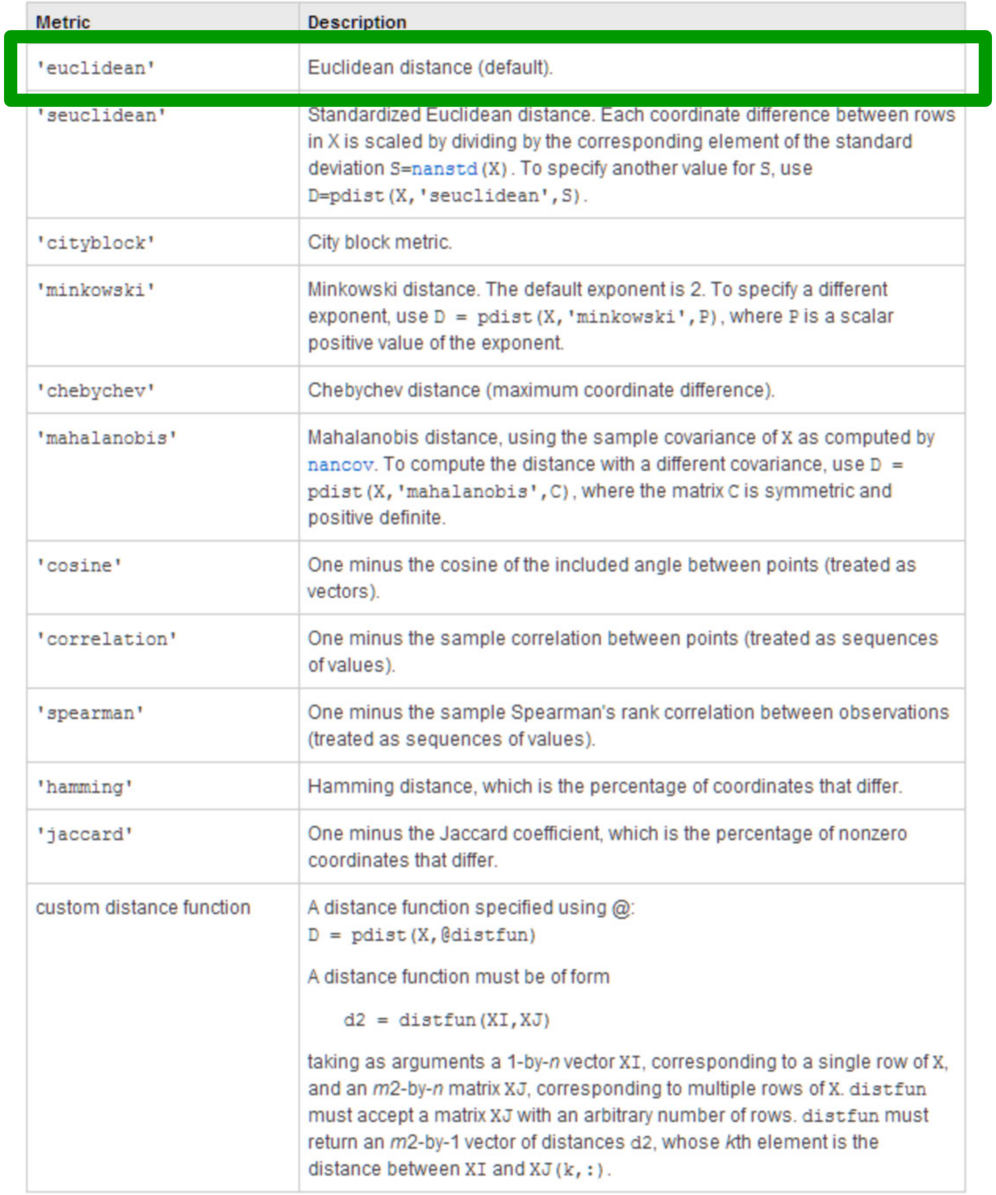
- Hierarchical agglomerative clustering algorithm
- **UPGMA** = Unweighted Pair Group Method with Arithmetic mean
- Iterative algorithm:
- Start with a pair with the smallest d(X,Y)
- Cluster these two together and replace it with their arithmetic mean (X+Y)/2
- Recalculate all distances to this new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

Output of UPGMA algorithm



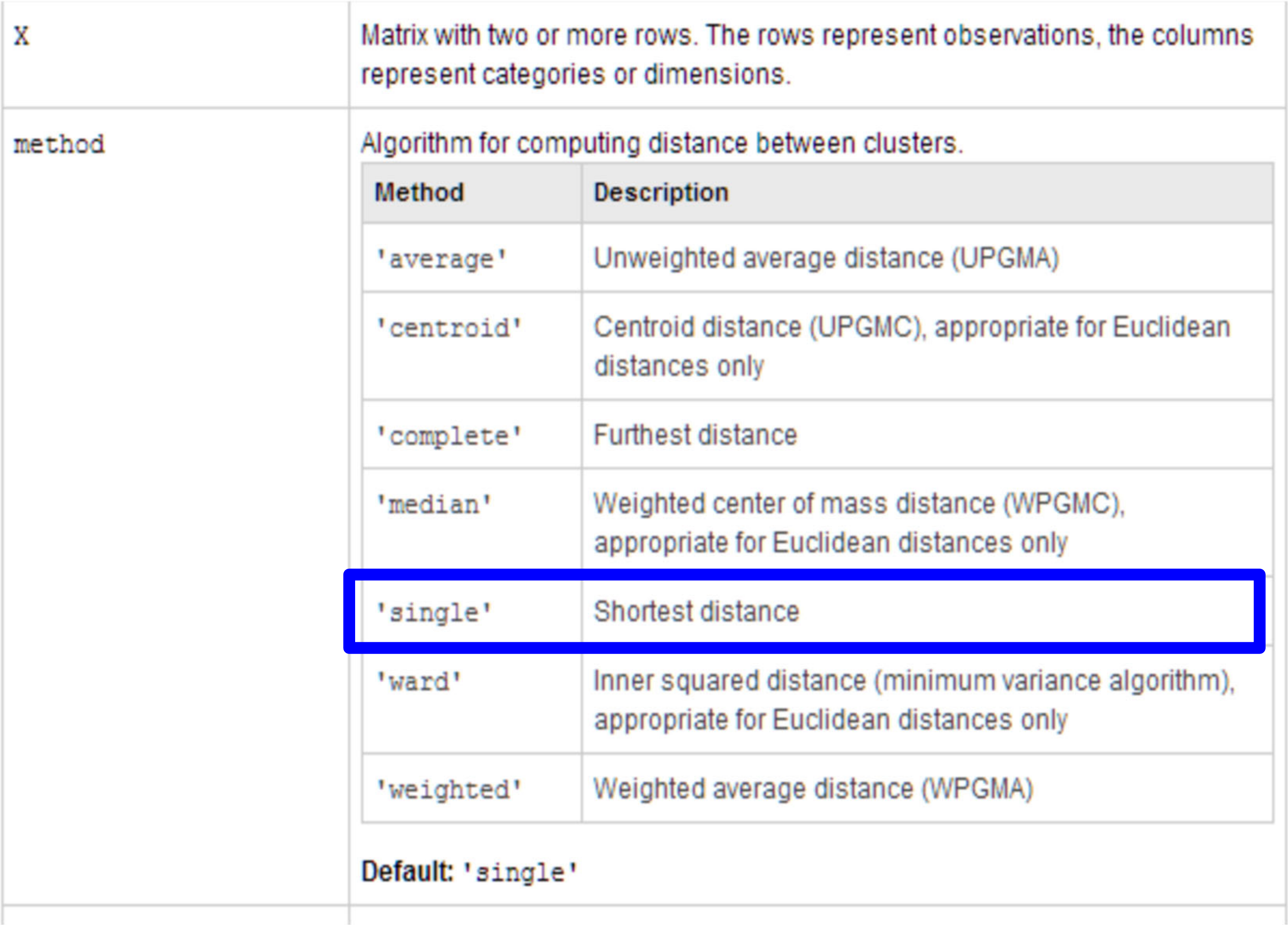
Clustering in Matlab

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’,…)



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)

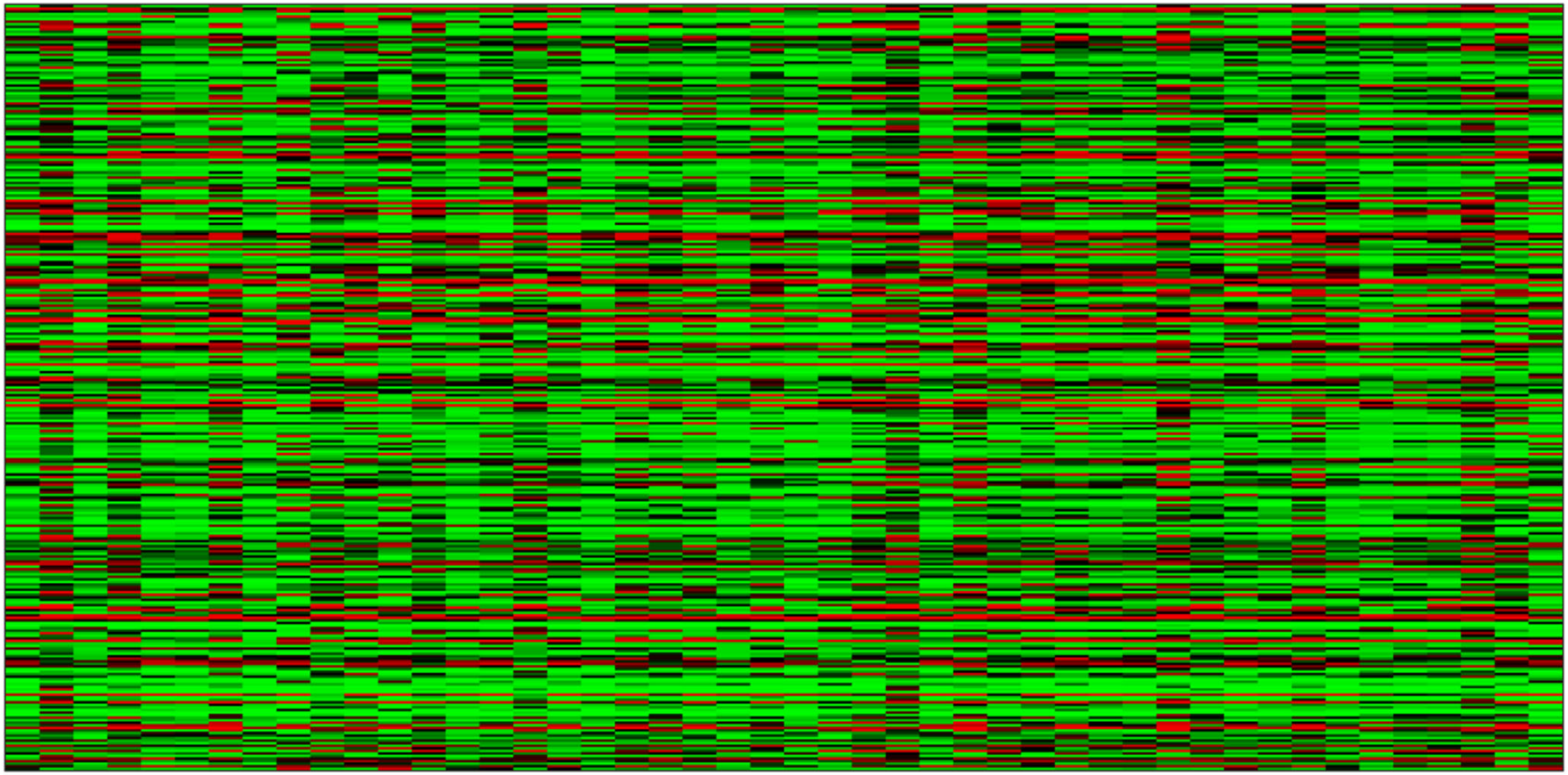
Default: 'single'

Clustering group exercise

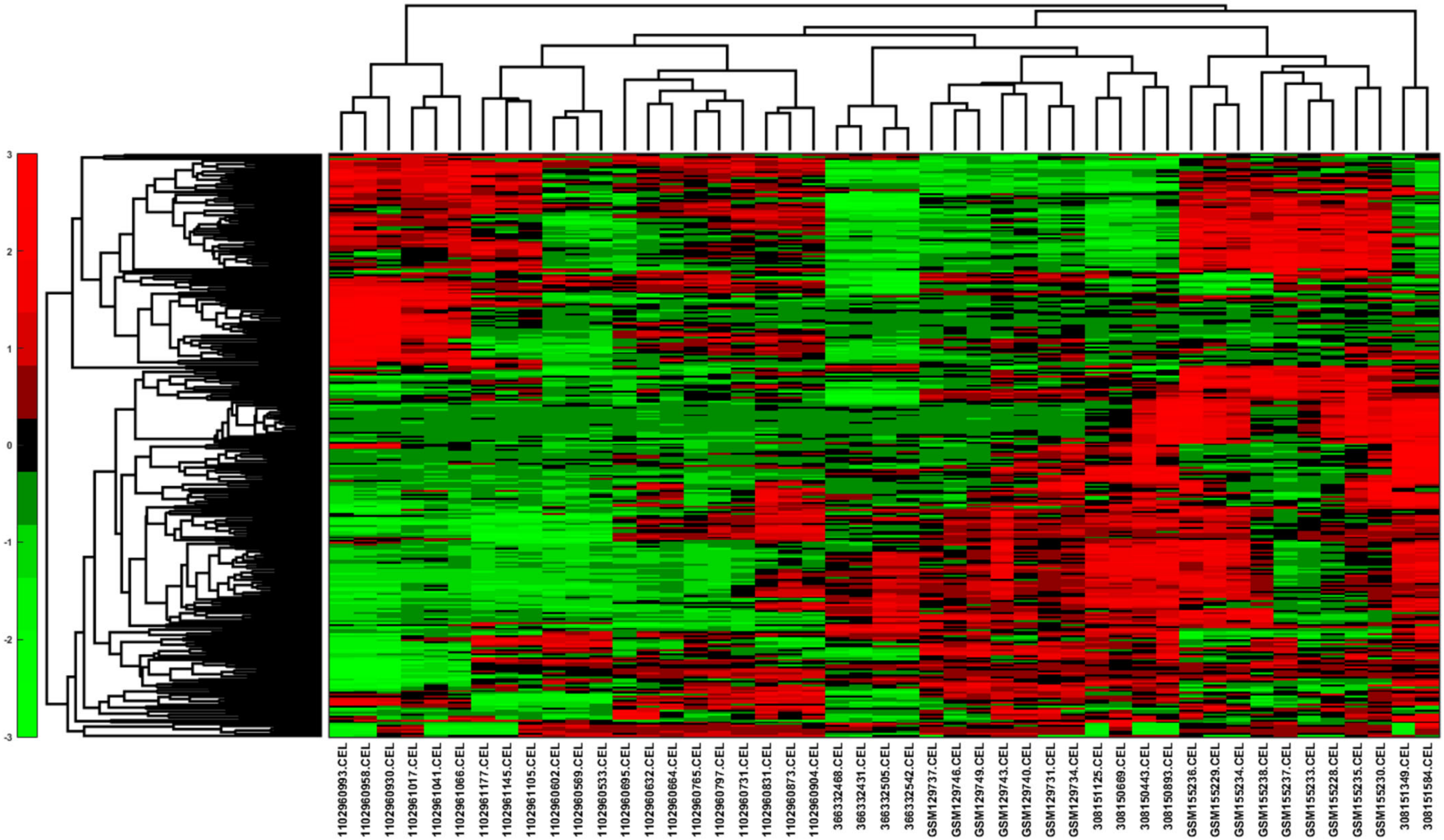
- Each group will analyze a cluster of genes identified in the T cell expression table
- Analyze the table of top 100 genes **by variance** in 47 samples
- Cluster them using:
  - Group 1: UPGMA = ‘linkage’, ‘average’, ‘RowPDistValue’, ’euclidean’,
  - Group 2: ‘linkage’, ‘single’, ‘RowPDistValue’, ’cityblock’,
  - 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)

```
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,'ColumnLab
els',array_names,'linkage',
'average','RowPDist','euclidean');
set(CGobj1,'RowLabels',gene_names1,'ColumnLab
els',array_names,'linkage',
'average','RowPDist','correlation');
```

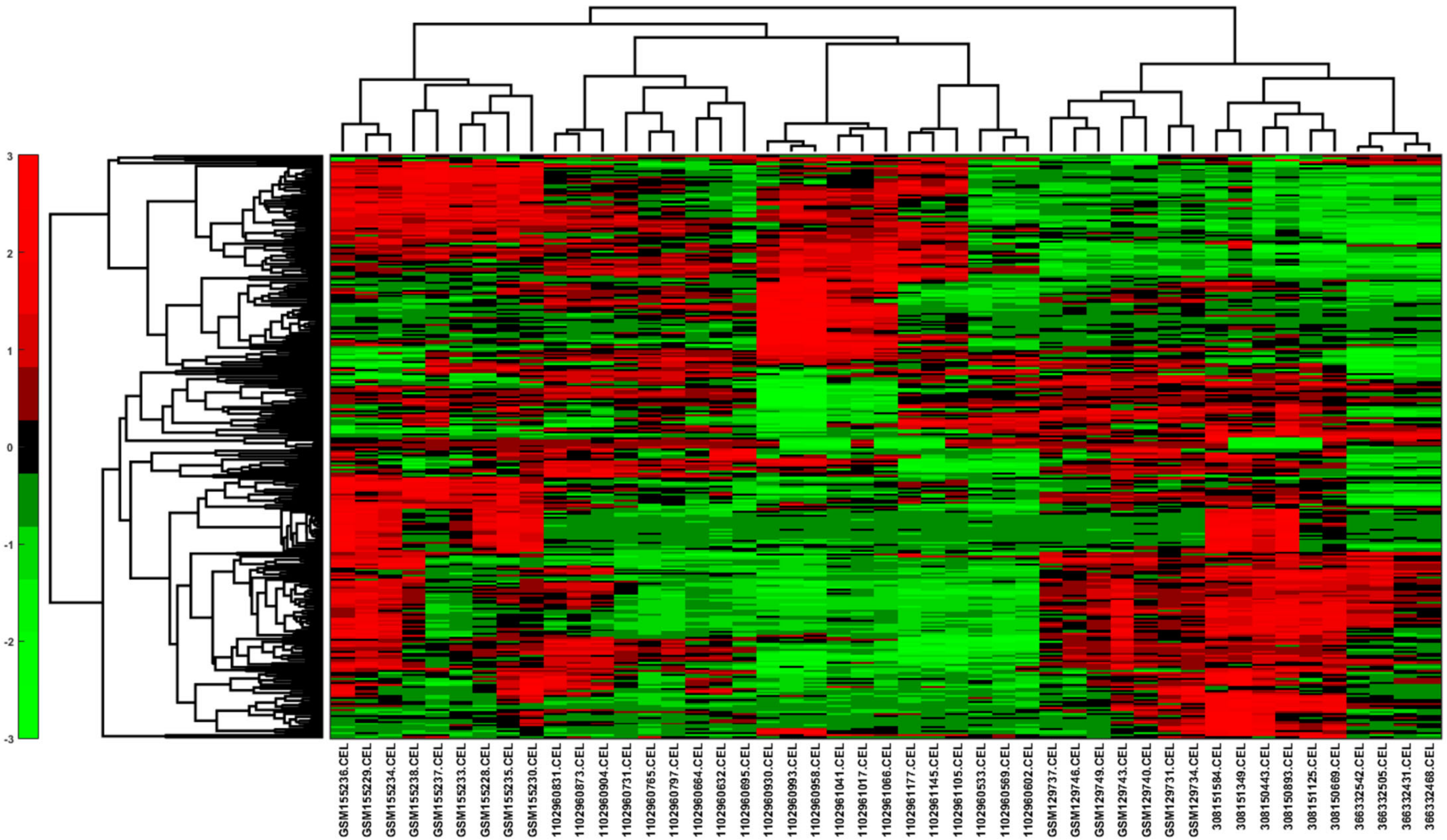
Before clustering



UPGMA hierarchical clustering, Euclidian distance



UPGMA hierarchical clustering, correlation distance



Search for shared biological functions

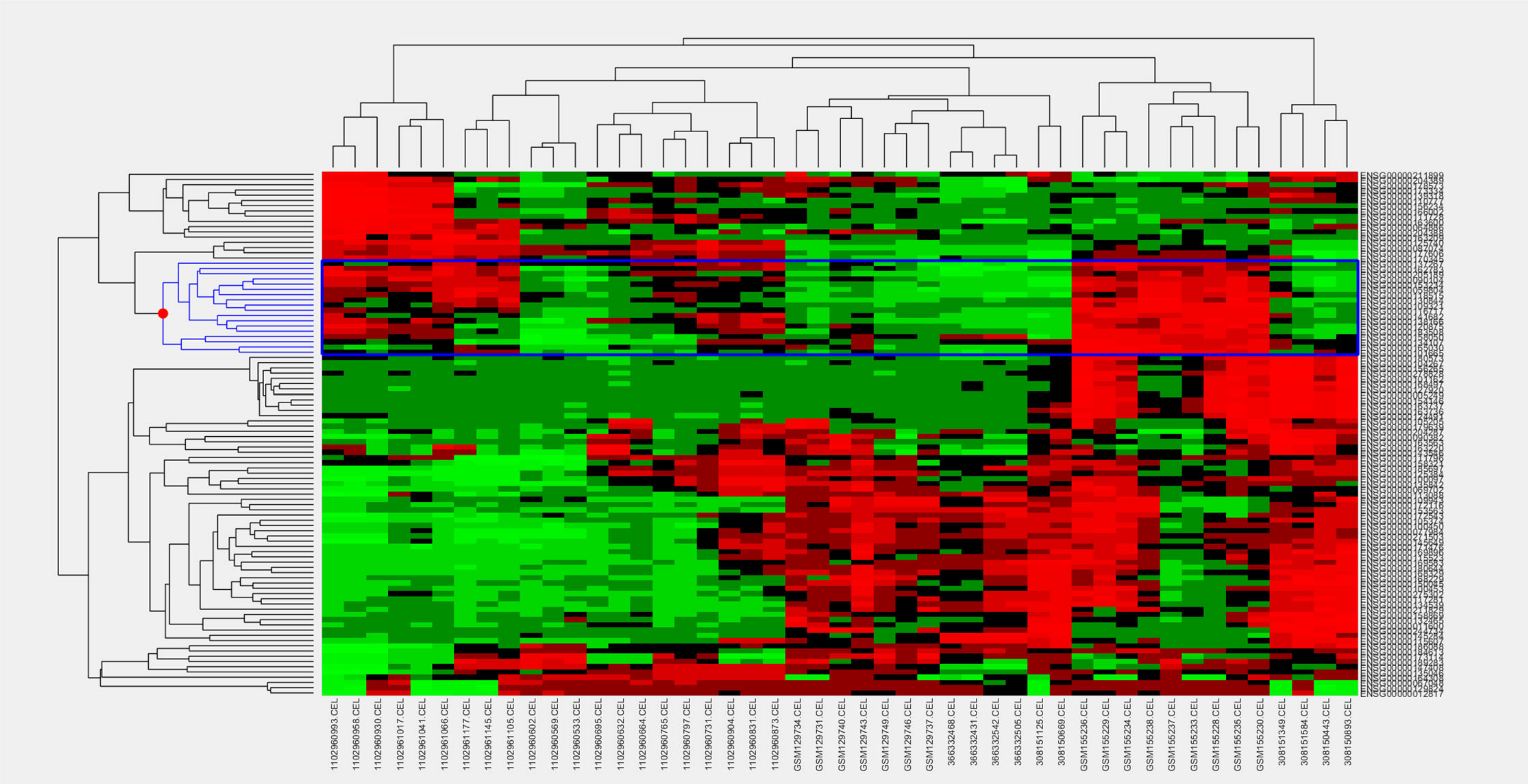
- copy the list of displayed genes
- go to "Start Analysis" on https://david.ncifcrf.gov/tools.jsp
- Paste genes from gene list displayed by Matlab into the box in the left panel of the website
- select ENSEMBL_GENE_ID and “gene list” radio button
- Click "Functional Annotation Clustering“
- Select groups in “Annotation Summary Results” which have many genes from your list. Definitely select “PUBMED_ID” and interaction databases like “Biogrid”
- First look at "Functional Annotation Chart" rectangular button below to display all overrepresented terms. Sort by “Benjamini” correction for multiple hypotheses testing
- Select "Functional Annotation Clustering" rectangular button below to display annotation results for gene list broken into multiple groups (clusters) each with related biological functions
- Write down the # of genes in the cluster and the top functions in two most interesting clusters

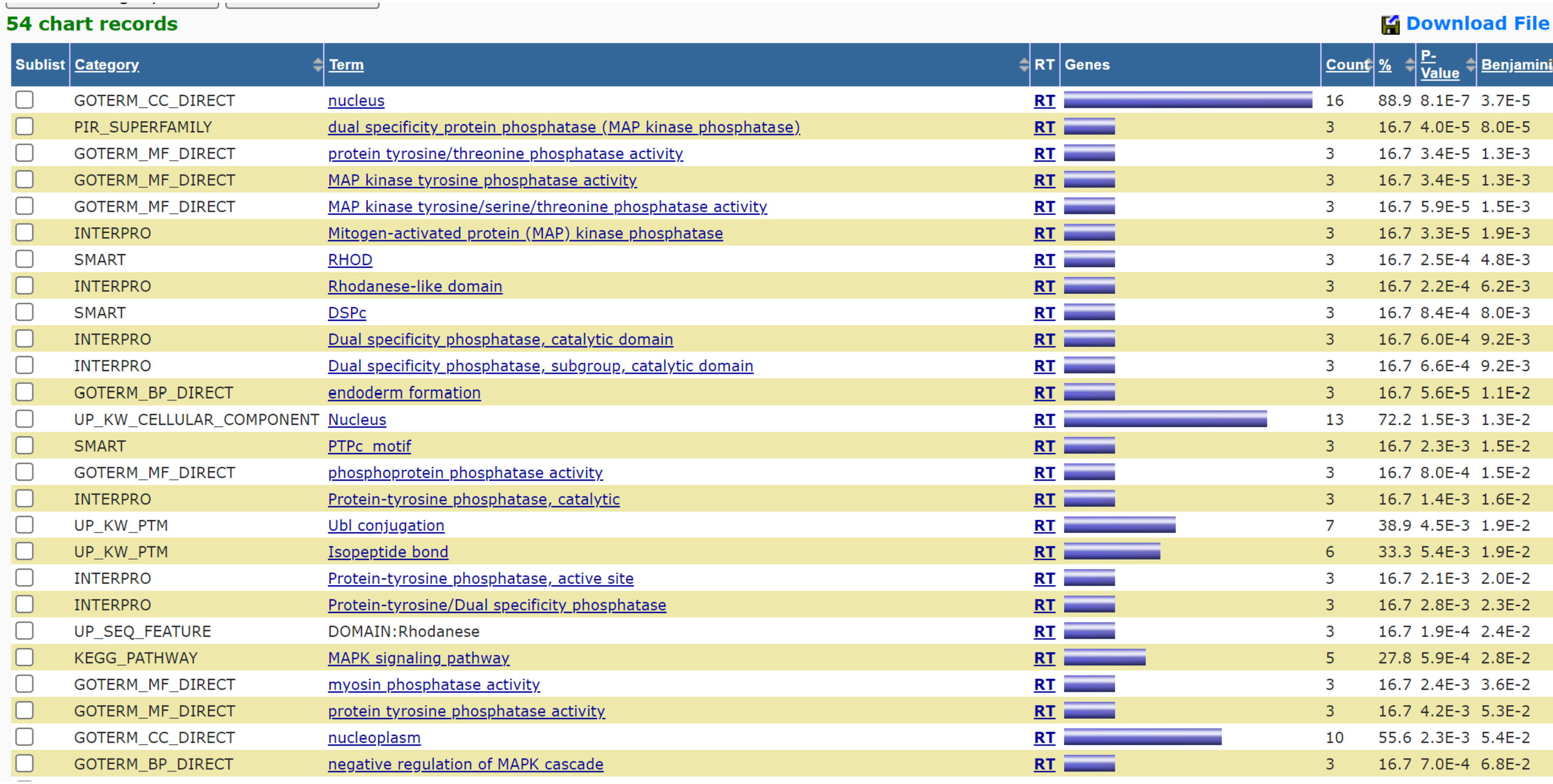
```
%%%
%Which biological functions are
overrepresented in different clusters?
%1) Pick a cluster:
%2) Select a node on the tree of rows,
%3) Right click
%4) Choose “export group info” into
the workspace
%5) Name it gene_list
%Run the following two Matlab
commands to display genes
g1=gene_list.RowNodeNames;
for m=1:length(g1);
disp(g1{m});
end;
```

% select ENSEMBL_GENE_ID and “gene list” radio button
% Click "Functional Annotation Clustering“
% Select groups in “Annotation Summary Results”
% which have many genes from your list.
% Definitely select “PUBMED_ID” and
% interaction databases like “Biogrid”
% First look at "Functional Annotation Chart" rectangular button below
% to display all overrepresented terms.
% Sort by “Benjamini” correction for multiple hypotheses testing
% Select "Functional Annotation Clustering" rectangular button below
% to display annotation results for gene list broken into multiple groups
% (clusters) each with related biological functions
% Write down the # of genes in the cluster and the top functions
% in two most interesting clusters

Using options:

‘linkage’, ‘average’, ‘RowPDistValue’, ’euclidean’,





54 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	GOTERM_CC_DIRECT	nucleus	RT		16	88.9	8.1E-7	3.7E-5
□	PIR_SUPERFAMILY	dual specificity protein phosphatase (MAP kinase phosphatase)	RT		3	16.7	4.0E-5	8.0E-5
□	GOTERM_MF_DIRECT	protein tyrosine/threonine phosphatase activity	RT		3	16.7	3.4E-5	1.3E-3
□	GOTERM_MF_DIRECT	MAP kinase tyrosine phosphatase activity	RT		3	16.7	3.4E-5	1.3E-3
□	GOTERM_MF_DIRECT	MAP kinase tyrosine/serine/threonine phosphatase activity	RT		3	16.7	5.9E-5	1.5E-3
□	INTERPRO	Mitogen-activated protein (MAP) kinase phosphatase	RT		3	16.7	3.3E-5	1.9E-3
□	SMART	RHOD	RT		3	16.7	2.5E-4	4.8E-3
□	INTERPRO	Rhodanese-like domain	RT		3	16.7	2.2E-4	6.2E-3
□	SMART	DSPc	RT		3	16.7	8.4E-4	8.0E-3
□	INTERPRO	Dual specificity phosphatase, catalytic domain	RT		3	16.7	6.0E-4	9.2E-3
□	INTERPRO	Dual specificity phosphatase, subgroup, catalytic domain	RT		3	16.7	6.6E-4	9.2E-3
□	GOTERM_BP_DIRECT	endoderm formation	RT		3	16.7	5.6E-5	1.1E-2
□	UP_KW_CELLULAR_COMPONENT	Nucleus	RT		13	72.2	1.5E-3	1.3E-2
□	SMART	PTPc_motif	RT		3	16.7	2.3E-3	1.5E-2
□	GOTERM_MF_DIRECT	phosphoprotein phosphatase activity	RT		3	16.7	8.0E-4	1.5E-2
□	INTERPRO	Protein-tyrosine phosphatase, catalytic	RT		3	16.7	1.4E-3	1.6E-2
□	UP_KW_PTM	Ubl conjugation	RT		7	38.9	4.5E-3	1.9E-2
□	UP_KW_PTM	Isopeptide bond	RT		6	33.3	5.4E-3	1.9E-2
□	INTERPRO	Protein-tyrosine phosphatase, active site	RT		3	16.7	2.1E-3	2.0E-2
□	INTERPRO	Protein-tyrosine/Dual specificity phosphatase	RT		3	16.7	2.8E-3	2.3E-2
□	UP_SEQ_FEATURE	DOMAIN:Rhodanese	RT		3	16.7	1.9E-4	2.4E-2
□	KEGG_PATHWAY	MAPK signaling pathway	RT		5	27.8	5.9E-4	2.8E-2
□	GOTERM_MF_DIRECT	myosin phosphatase activity	RT		3	16.7	2.4E-3	3.6E-2
□	GOTERM_MF_DIRECT	protein tyrosine phosphatase activity	RT		3	16.7	4.2E-3	5.3E-2
□	GOTERM_CC_DIRECT	nucleoplasm	RT		10	55.6	2.3E-3	5.4E-2
□	GOTERM_BP_DIRECT	negative regulation of MAPK cascade	RT		3	16.7	7.0E-4	6.8E-2

Gene list being analyzed

Clustering options and stringency

score for the group based on the EASE scores of each term members. The higher, the more enriched.

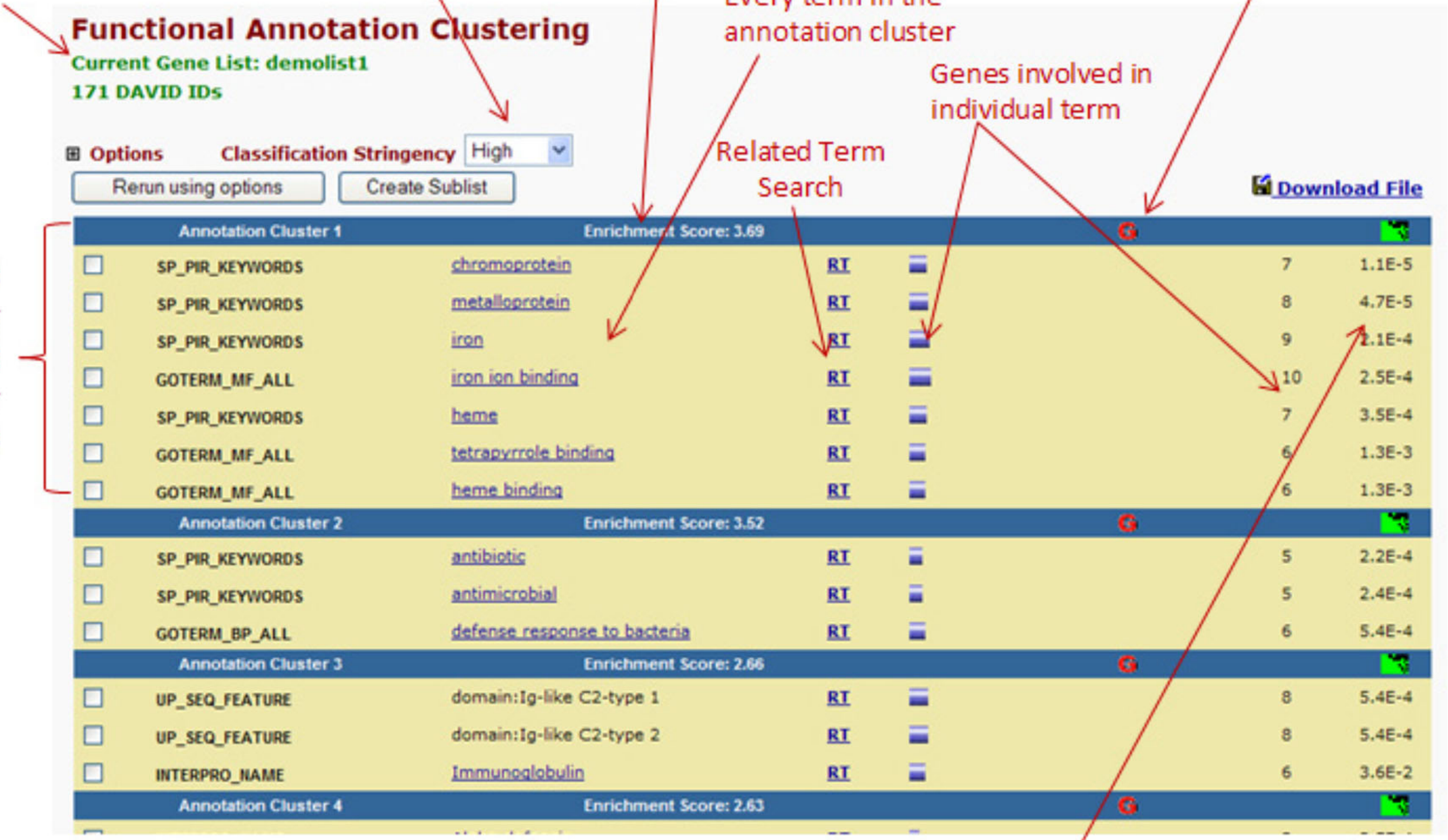
Every term in the annotation cluster

Related Term Search

Genes involved in individual term

ALL genes involved in this annotation cluster

A group of terms having similar biological meaning due to sharing similar gene members



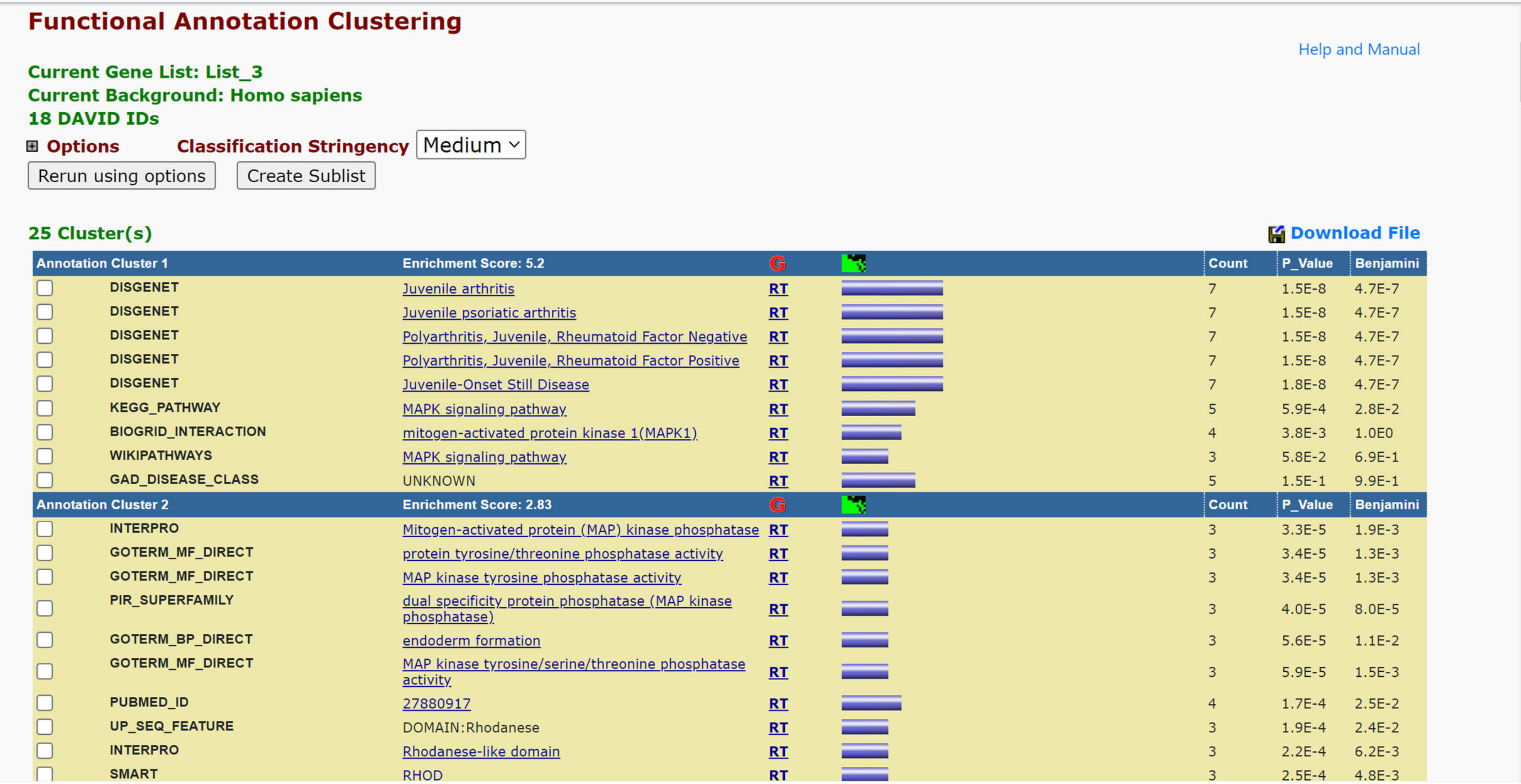
Functional Annotation Clustering

Current Gene List: demolist1
171 DAVID IDs

⊞ Options   Classification Stringency: High

Annotation Cluster 1		**Enrichment Score: 3.69**				
□	SP_PIR_KEYWORDS	chromoprotein	RT	▬	7	1.1E-5
□	SP_PIR_KEYWORDS	metalloprotein	RT	▬	8	4.7E-5
□	SP_PIR_KEYWORDS	iron	RT	▬	9	2.1E-4
□	GOTERM_MF_ALL	iron ion binding	RT	▬	10	2.5E-4
□	SP_PIR_KEYWORDS	heme	RT	▬	7	3.5E-4
□	GOTERM_MF_ALL	tetrapyrrole binding	RT	▬	6	1.3E-3
□	GOTERM_MF_ALL	heme binding	RT	▬	6	1.3E-3
Annotation Cluster 2		**Enrichment Score: 3.52**				
□	SP_PIR_KEYWORDS	antibiotic	RT	▬	5	2.2E-4
□	SP_PIR_KEYWORDS	antimicrobial	RT	▬	5	2.4E-4
□	GOTERM_BP_ALL	defense response to bacteria	RT	▬	6	5.4E-4
Annotation Cluster 3		**Enrichment Score: 2.66**				
□	UP_SEQ_FEATURE	domain:Ig-like C2-type 1	RT	▬	8	5.4E-4
□	UP_SEQ_FEATURE	domain:Ig-like C2-type 2	RT	▬	8	5.4E-4
□	INTERPRO_NAME	Immunoglobulin	RT	▬	6	3.6E-2
Annotation Cluster 4		**Enrichment Score: 2.63**				

EASE Score, the modified Fisher Exact P-Value. They are identical to that in the Chart Report. The smaller, the more enriched.



Functional Annotation Clustering

Help and Manual

Current Gene List: List_3
Current Background: Homo sapiens
18 DAVID IDs

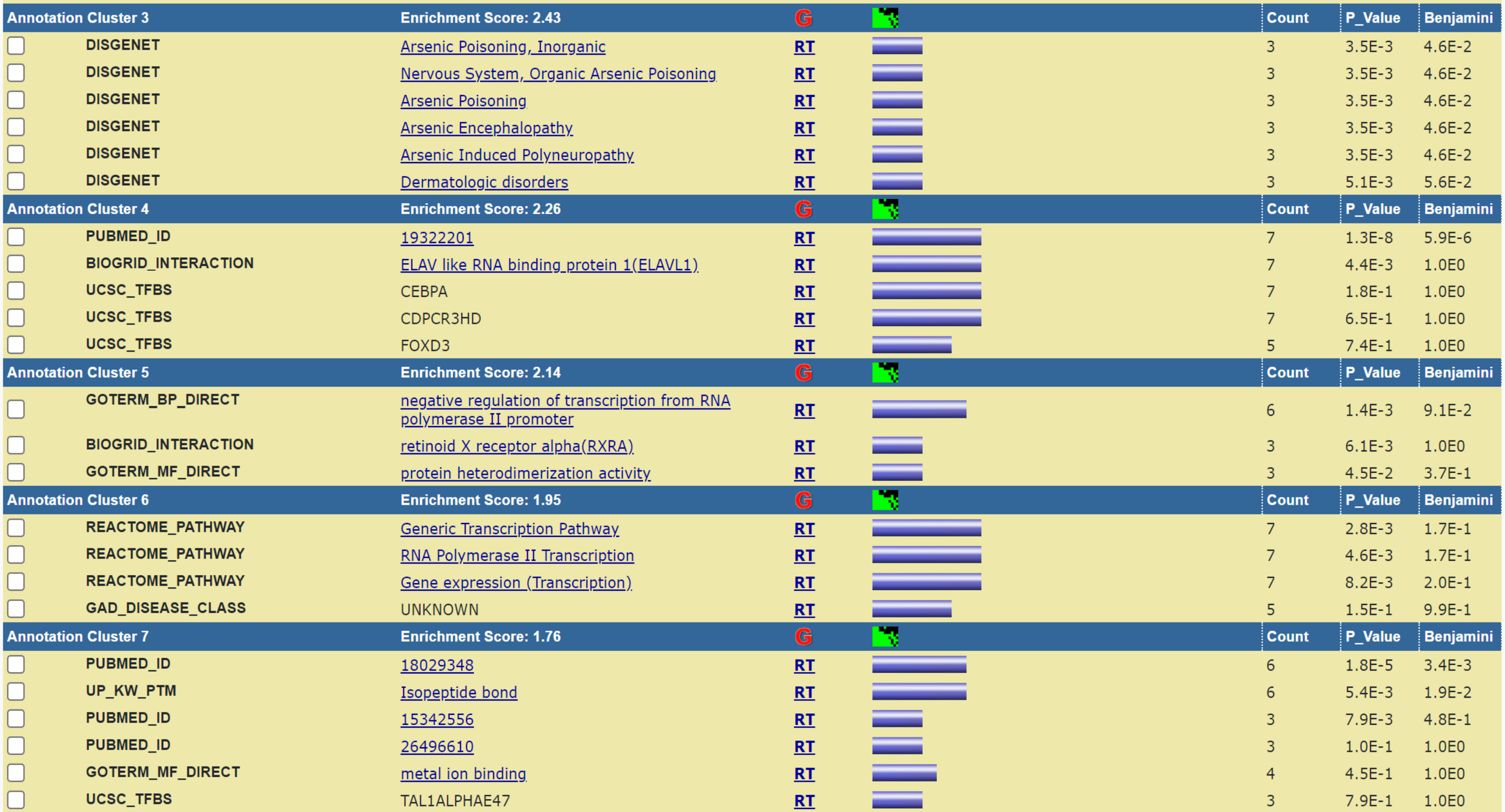
⊞ **Options** **Classification Stringency** Medium

Rerun using options | Create Sublist

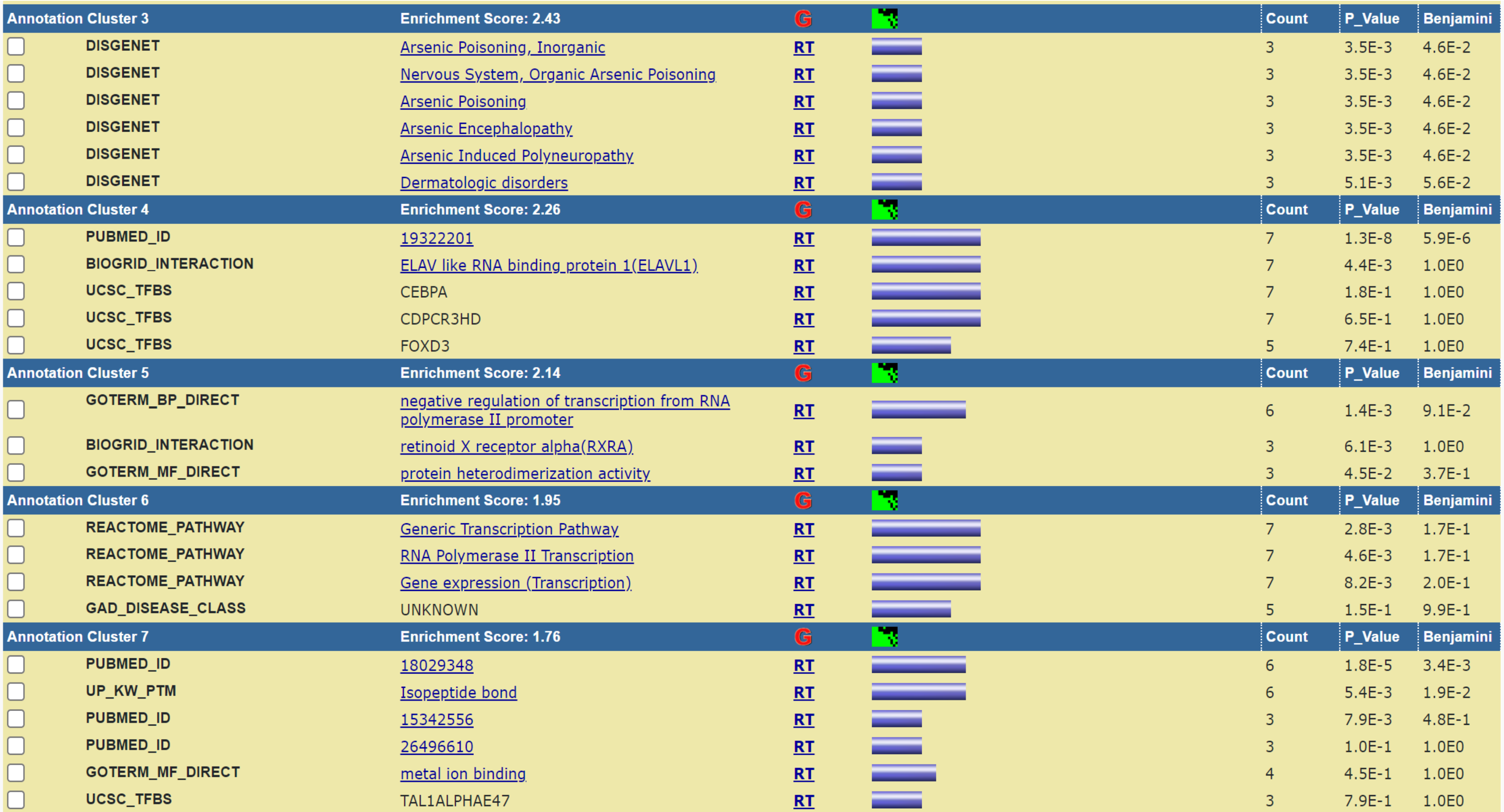
25 Cluster(s)

Download File

	Annotation Cluster 1	Enrichment Score: 5.2	G		Count	P_Value	Benjamini
□	DISGENET	Juvenile arthritis	RT		7	1.5E-8	4.7E-7
□	DISGENET	Juvenile psoriatic arthritis	RT		7	1.5E-8	4.7E-7
□	DISGENET	Polyarthritis, Juvenile, Rheumatoid Factor Negative	RT		7	1.5E-8	4.7E-7
□	DISGENET	Polyarthritis, Juvenile, Rheumatoid Factor Positive	RT		7	1.5E-8	4.7E-7
□	DISGENET	Juvenile-Onset Still Disease	RT		7	1.8E-8	4.7E-7
□	KEGG_PATHWAY	MAPK signaling pathway	RT		5	5.9E-4	2.8E-2
□	BIOGRID_INTERACTION	mitogen-activated protein kinase 1(MAPK1)	RT		4	3.8E-3	1.0E0
□	WIKIPATHWAYS	MAPK signaling pathway	RT		3	5.8E-2	6.9E-1
□	GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
	Annotation Cluster 2	**Enrichment Score: 2.83**	G		**Count**	**P_Value**	**Benjamini**
□	INTERPRO	Mitogen-activated protein (MAP) kinase phosphatase	RT		3	3.3E-5	1.9E-3
□	GOTERM_MF_DIRECT	protein tyrosine/threonine phosphatase activity	RT		3	3.4E-5	1.3E-3
□	GOTERM_MF_DIRECT	MAP kinase tyrosine phosphatase activity	RT		3	3.4E-5	1.3E-3
□	PIR_SUPERFAMILY	dual specificity protein phosphatase (MAP kinase phosphatase)	RT		3	4.0E-5	8.0E-5
□	GOTERM_BP_DIRECT	endoderm formation	RT		3	5.6E-5	1.1E-2
□	GOTERM_MF_DIRECT	MAP kinase tyrosine/serine/threonine phosphatase activity	RT		3	5.9E-5	1.5E-3
□	PUBMED_ID	27880917	RT		4	1.7E-4	2.5E-2
□	UP_SEQ_FEATURE	DOMAIN:Rhodanese	RT		3	1.9E-4	2.4E-2
□	INTERPRO	Rhodanese-like domain	RT		3	2.2E-4	6.2E-3
□	SMART	RHOD	RT		3	2.5E-4	4.8E-3

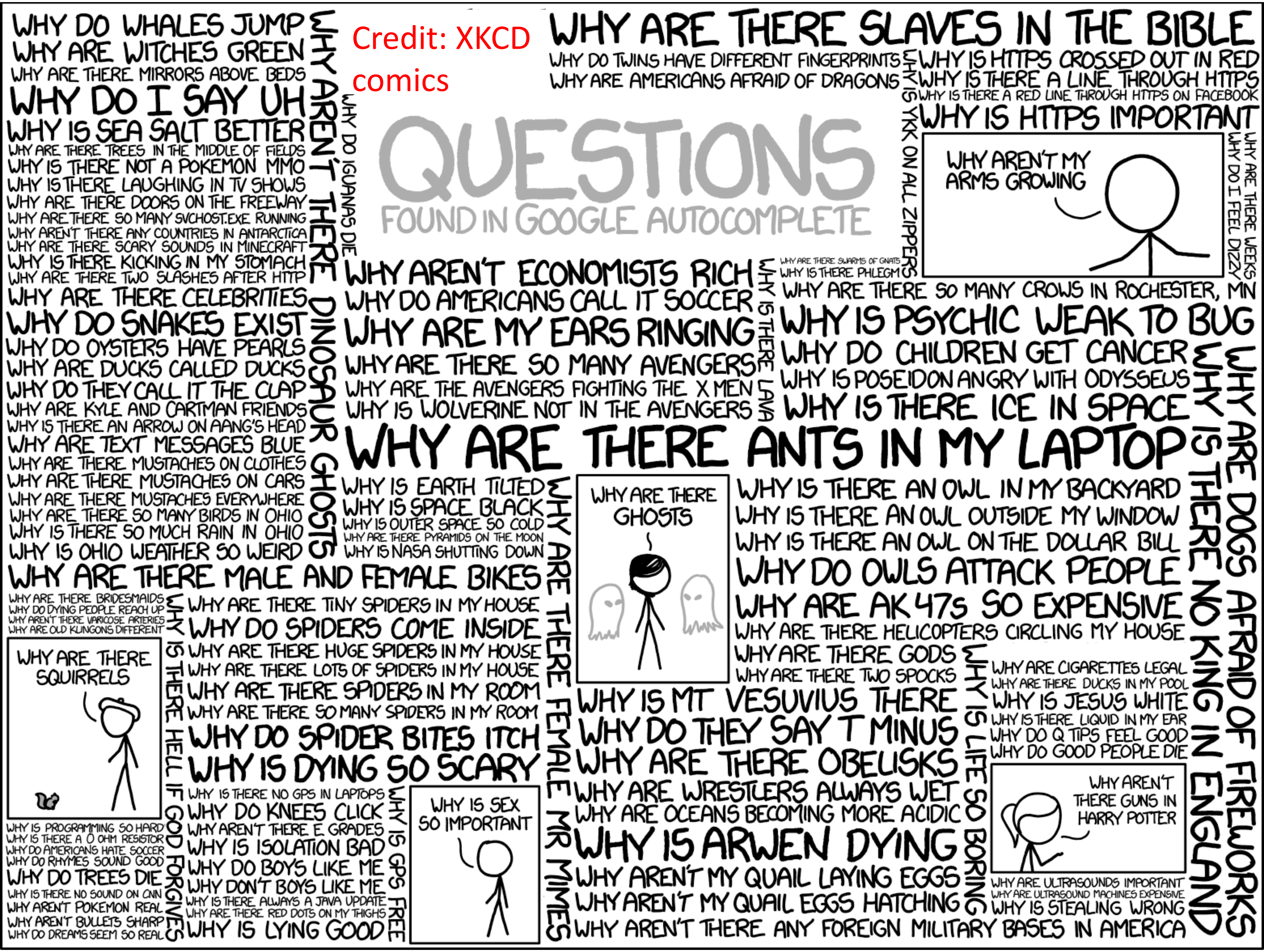


Annotation Cluster 3		Enrichment Score: 2.43	G		Count	P_Value	Benjamini
□	DISGENET	Arsenic Poisoning, Inorganic	RT		3	3.5E-3	4.6E-2
□	DISGENET	Nervous System, Organic Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Encephalopathy	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Induced Polyneuropathy	RT		3	3.5E-3	4.6E-2
□	DISGENET	Dermatologic disorders	RT		3	5.1E-3	5.6E-2
Annotation Cluster 4		Enrichment Score: 2.26	G		Count	P_Value	Benjamini
□	PUBMED_ID	19322201	RT		7	1.3E-8	5.9E-6
□	BIOGRID_INTERACTION	ELAV like RNA binding protein 1(ELAVL1)	RT		7	4.4E-3	1.0E0
□	UCSC_TFBS	CEBPA	RT		7	1.8E-1	1.0E0
□	UCSC_TFBS	CDPCR3HD	RT		7	6.5E-1	1.0E0
□	UCSC_TFBS	FOXD3	RT		5	7.4E-1	1.0E0
Annotation Cluster 5		Enrichment Score: 2.14	G		Count	P_Value	Benjamini
□	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		6	1.4E-3	9.1E-2
□	BIOGRID_INTERACTION	retinoid X receptor alpha(RXRA)	RT		3	6.1E-3	1.0E0
□	GOTERM_MF_DIRECT	protein heterodimerization activity	RT		3	4.5E-2	3.7E-1
Annotation Cluster 6		Enrichment Score: 1.95	G		Count	P_Value	Benjamini
□	REACTOME_PATHWAY	Generic Transcription Pathway	RT		7	2.8E-3	1.7E-1
□	REACTOME_PATHWAY	RNA Polymerase II Transcription	RT		7	4.6E-3	1.7E-1
□	REACTOME_PATHWAY	Gene expression (Transcription)	RT		7	8.2E-3	2.0E-1
□	GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 7		Enrichment Score: 1.76	G		Count	P_Value	Benjamini
□	PUBMED_ID	18029348	RT		6	1.8E-5	3.4E-3
□	UP_KW_PTM	Isopeptide bond	RT		6	5.4E-3	1.9E-2
□	PUBMED_ID	15342556	RT		3	7.9E-3	4.8E-1
□	PUBMED_ID	26496610	RT		3	1.0E-1	1.0E0
□	GOTERM_MF_DIRECT	metal ion binding	RT		4	4.5E-1	1.0E0
□	UCSC_TFBS	TAL1ALPHAE47	RT		3	7.9E-1	1.0E0



Annotation Cluster 3		Enrichment Score: 2.43	G		Count	P_Value	Benjamini
□	DISGENET	Arsenic Poisoning, Inorganic	RT		3	3.5E-3	4.6E-2
□	DISGENET	Nervous System, Organic Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Poisoning	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Encephalopathy	RT		3	3.5E-3	4.6E-2
□	DISGENET	Arsenic Induced Polyneuropathy	RT		3	3.5E-3	4.6E-2
□	DISGENET	Dermatologic disorders	RT		3	5.1E-3	5.6E-2
Annotation Cluster 4		Enrichment Score: 2.26	G		Count	P_Value	Benjamini
□	PUBMED_ID	19322201	RT		7	1.3E-8	5.9E-6
□	BIOGRID_INTERACTION	ELAV like RNA binding protein 1(ELAVL1)	RT		7	4.4E-3	1.0E0
□	UCSC_TFBS	CEBPA	RT		7	1.8E-1	1.0E0
□	UCSC_TFBS	CDPCR3HD	RT		7	6.5E-1	1.0E0
□	UCSC_TFBS	FOXD3	RT		5	7.4E-1	1.0E0
Annotation Cluster 5		Enrichment Score: 2.14	G		Count	P_Value	Benjamini
□	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		6	1.4E-3	9.1E-2
□	BIOGRID_INTERACTION	retinoid X receptor alpha(RXRA)	RT		3	6.1E-3	1.0E0
□	GOTERM_MF_DIRECT	protein heterodimerization activity	RT		3	4.5E-2	3.7E-1
Annotation Cluster 6		Enrichment Score: 1.95	G		Count	P_Value	Benjamini
□	REACTOME_PATHWAY	Generic Transcription Pathway	RT		7	2.8E-3	1.7E-1
□	REACTOME_PATHWAY	RNA Polymerase II Transcription	RT		7	4.6E-3	1.7E-1
□	REACTOME_PATHWAY	Gene expression (Transcription)	RT		7	8.2E-3	2.0E-1
□	GAD_DISEASE_CLASS	UNKNOWN	RT		5	1.5E-1	9.9E-1
Annotation Cluster 7		Enrichment Score: 1.76	G		Count	P_Value	Benjamini
□	PUBMED_ID	18029348	RT		6	1.8E-5	3.4E-3
□	UP_KW_PTM	Isopeptide bond	RT		6	5.4E-3	1.9E-2
□	PUBMED_ID	15342556	RT		3	7.9E-3	4.8E-1
□	PUBMED_ID	26496610	RT		3	1.0E-1	1.0E0
□	GOTERM_MF_DIRECT	metal ion binding	RT		4	4.5E-1	1.0E0
□	UCSC_TFBS	TAL1ALPHAE47	RT		3	7.9E-1	1.0E0

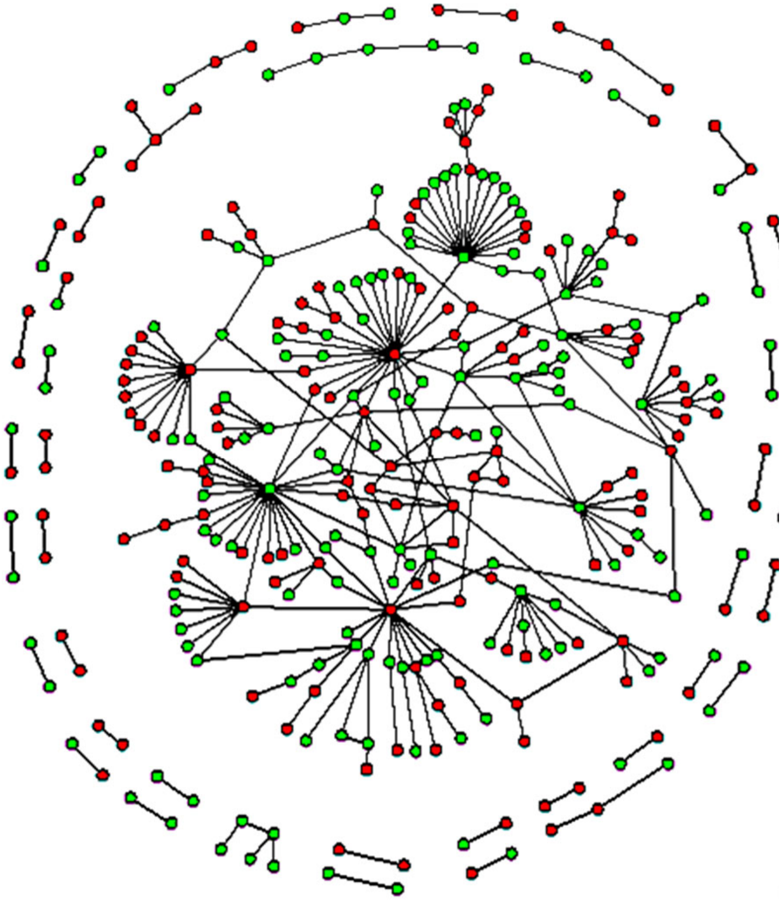
Credit: XKCD comics


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 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 GHOSTS
WHY ARE THERE FEMALE MR MIMES
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 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 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

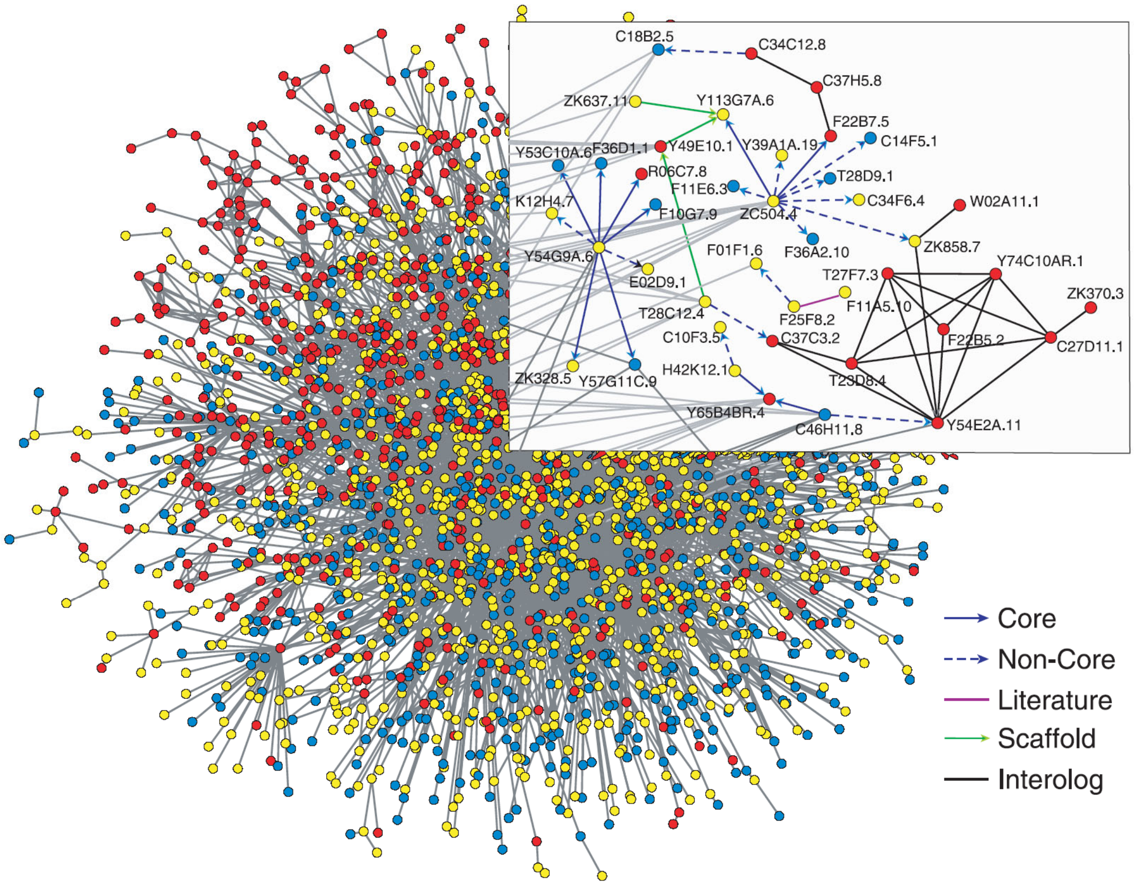
Basic concepts of network analysis

Reminder from the first lecture

Protein-Protein binding
IntAct Database (Dec 2015)
Interactions: 577,297   Proteins: 89,716

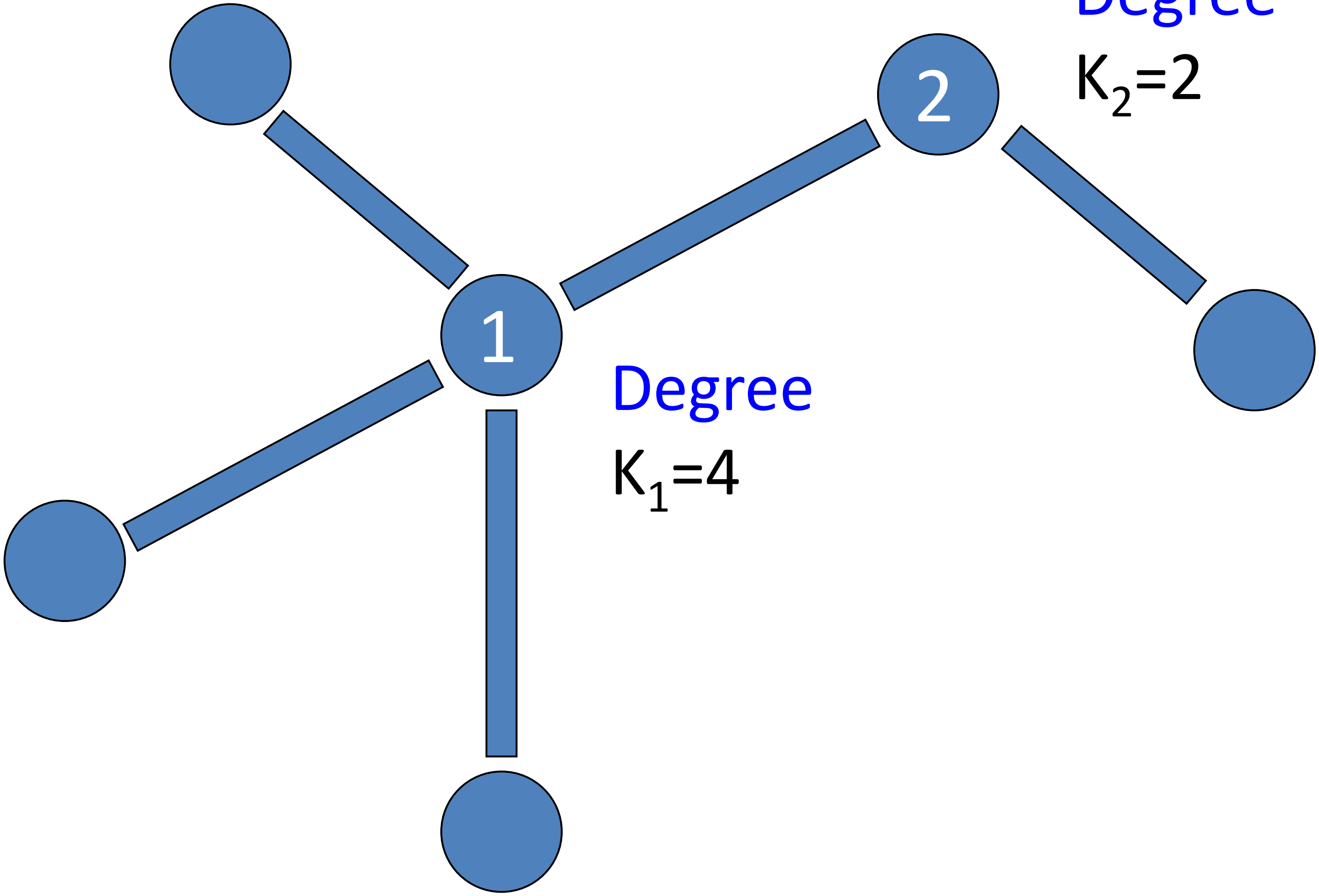


Baker’s yeast *S. cerevisiae* (only nuclear proteins shown)
From S. Maslov, K. Sneppen, Science 2002



Worm *C. elegans*
From S. Lee et al , Science 2004

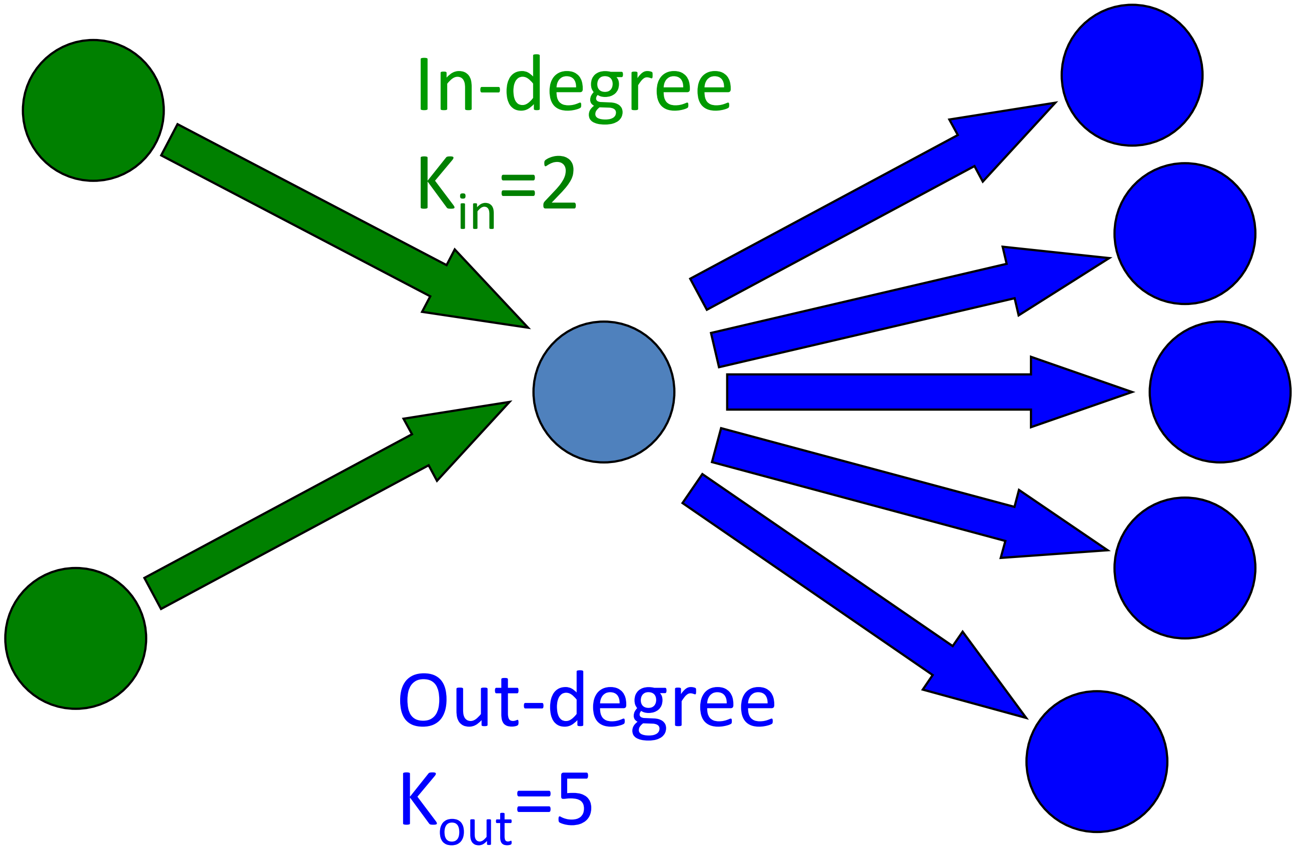
Degree of a node – its # of neighbors



Degree
K_2=2

Degree
K_1=4

Directed networks have in- and out- degrees



In-degree

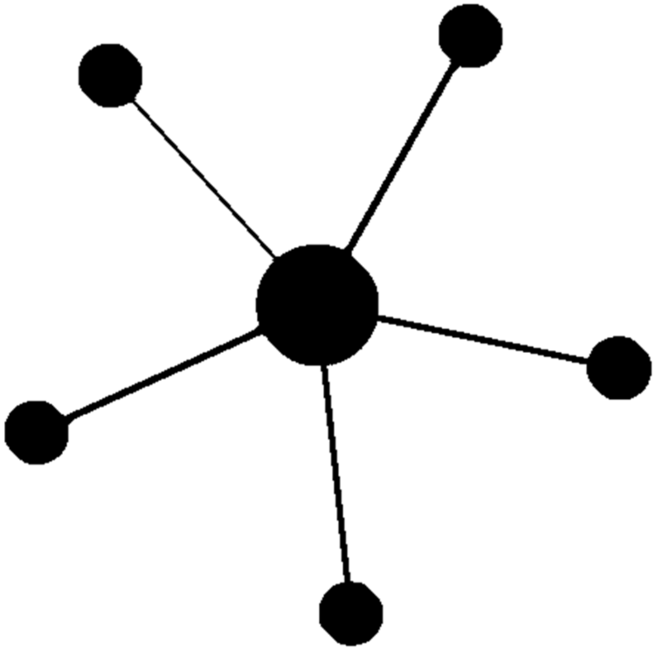
K_{in}=2

Out-degree

K_{out}=5

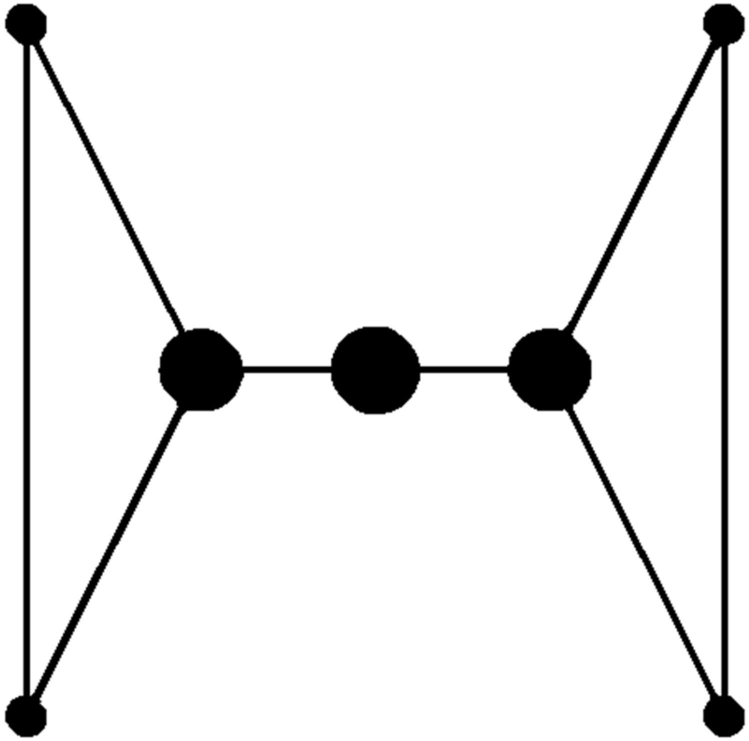
How to find “important” nodes?

- By their degree
- Hubs = important
- Example: Google’s PageRank



How to find “important” nodes?

- By their connectivity
- Connectors = important
- Betweenness-centrality

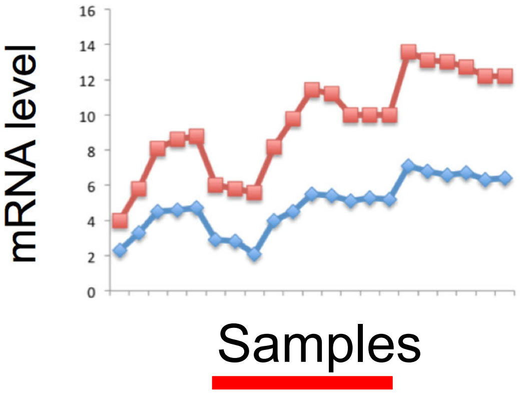


Betweenness centrality: definition

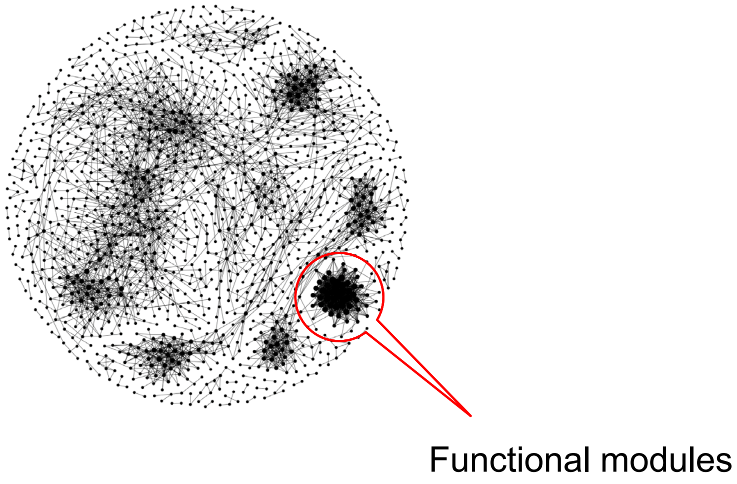
- Take a node i
- There are (N-1)*(N-2)/2 pairs of other nodes
- For each pair find the shortest path on the network
- If more than one shortest path, sample them equally
- Betweenness-centrality C(i) ~ the number of shortest paths going through node i

To analyze correlations in expression for all pairs of genes: Co-expression networks

How to construct a co-expression network?



A co-expression network



- Start with a matrix of log2 of expression levels of N genes in K samples (conditions): for our T-cell data N=3000, K=47

- For each of N(N-1)/2 pairs of genes i and j calculate the correlation coefficient $\rho_{ij}=\sigma_{ij}/\sigma_i\sigma_j$ of gene levels across K samples

- Put a threshold, e.g. $\rho_{ij}>0.85$, or otherwise select the most correlated pairs of genes (~4500 in our case). Now you have a weighted network.

- Identify densely interconnected functional modules in this network.

- Modules can be used to infer unknown functions of genes via “Guilt by Association” principle.

How to install Gephi software for network analysis?

- Install Gephi from: https://gephi.org/users/download/
- One of the common problems with installation is the version of Java on your computer. One possible solution is here: https://github.com/gephi/gephi/issues/1787.

Sometimes after installation Gephi may complain that it cannot find java version 1.8 or higher. In this case you need to go to C:\Program Files\Gephi-0.9.2\etc
Open file gephi.conf using notepad.exe (MS Word does not work!).
Add a line jdkhome="C:\Program Files (x86)\Java\jre1.8.0_231"
(the numbers in …jre1.8.0_231 may be changed to reflect the actual directory where Java is installed on your computer).
If JDK is not installed on your computer, you need to install itfirst from https://www.java.com/en/download/win10.jsp"

Co-expression network analysis exercise

- Start Gephi and open coexpression_network_random_start.gephi
- Run "Layout" → Fruchterman Reingold → Speed 10.0
- Run "Average degree", "Network diameter", "Modularity" in the Statistics tab in the right panel.
- Color nodes by "modularity class":
  Appearance → Nodes → Partition → Palette Icon → Modularity class
- Size nodes first by "degree".
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - If the nodes are too small, select "Min size": 10 and "Max size":80
  - Nodes in large tightly connected clusters have large degree
- Then size nodes by "betweenness-centrality"
  Appearance → Nodes → Ranking → Multiple Circles Icon → Betweenness-centrality
  - Large circles are "coordinator" genes connecting different co-expressed clusters to each other. Potentially biologically interesting

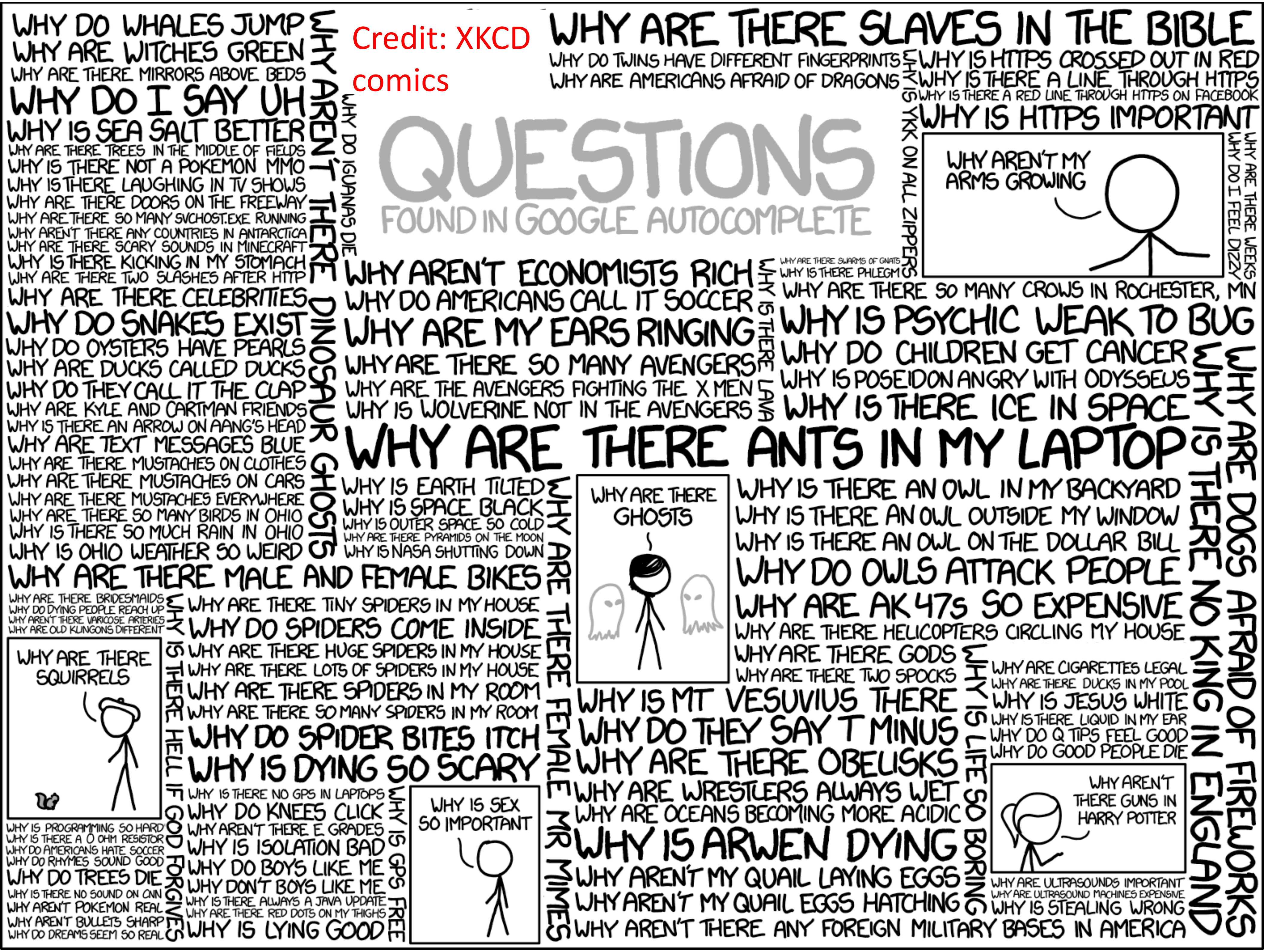
Disease-disease similarity network

- Based on the table summarizing **all current medical knowledge** of genes implicated in diseases:
  - Rows: 516 common human diseases
  - Columns: 25,000 human genes
  - Matrix element $D_{i\alpha}$ =1 if the gene α is known to be involved in the disease i . 0 – otherwise
- Constructed disease-disease similarity network:
  - Weight of the edge - # of shared genes between two diseases
  - Easy to construct: the adjacency matrix A of the network is simply $A=D \bullet D^{+}$

Disease network analysis exercise

- Start Gephi and open disease_disease_random_start.gephi
- Run "Layout" → Fruchterman Reingold → Speed 10.0
  Observe how clusters emerge.
- Run "Average degree", "Network diameter", "Modularity" analysis tools in the right panel.
- Color nodes with medical term: "disorder class"
  Appearance → Nodes → Partition → Palette Icon → Disorder class
- Then color nodes by "modularity class". See how well it agrees with the previous color.
  Appearance → Nodes → Partition → Palette Icon → Modularity class
- Size nodes first by "degree".
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - Which disease has the largest degree?
- Size nodes by "betweenness centrality"
  Appearance → Nodes → Ranking → Multiple Circles Icon → Degree
  - Which diseases have the largest betweenness-centrality?

  These "connector" diseases linking different diseases clusters to each other. They highlight potentially interesting connections between diseases


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 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 GHOSTS
WHY ARE THERE FEMALE MR MIMES
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 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 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

Credit: XKCD comics