Introduction to Gene expression analysis

Katja Nowick
Freie Universität Berlin,
Germany

Goals for today

- Analysis of RNA-Seq data
- Identifying differentially expressed genes
- Co-expression networks
- Gene Ontology enrichment analysis

- Using R and R packages
- Exercises

Dataset for the exercise

Social status alters immune regulation and response to infection in macaques

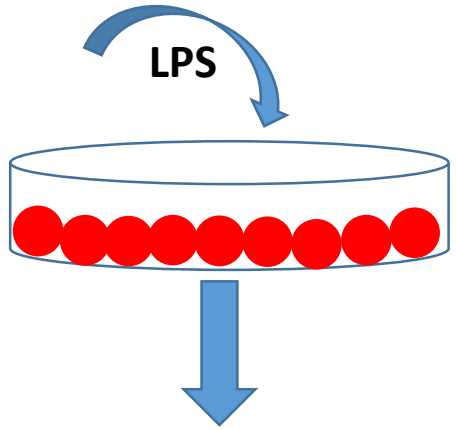
Noah Snyder-Mackler,[1,2*] Joaquín Sanz,[3,4*] Jordan N. Kohn,[5] Jessica F. Brinkworth,[3,6] Shauna Morrow,[1] Amanda O. Shaver,[1†] Jean-Christophe Grenier,[4] Roger Pique-Regi,[7,8] Zachary P. Johnson,[5,9‡] Mark E. Wilson,[5,10] Luis B. Barreiro,[4,11§||] Jenny Tung[1,12,13,14§||]

Snyder-Mackler et al. 2018

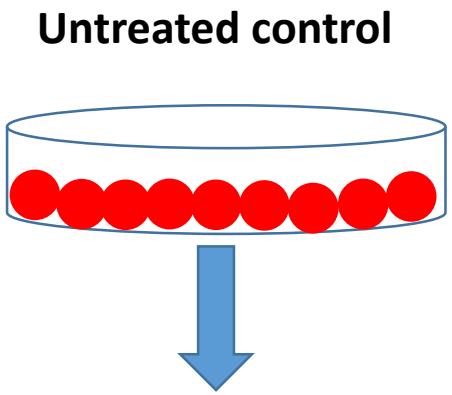
Dataset for the exercise

Social status alters immune regulation and response to infection in macaques

- Lower ranking individuals suffer more stress
- Isolated blood cells from each individual
- Treated cells with LPS to activate the immune system



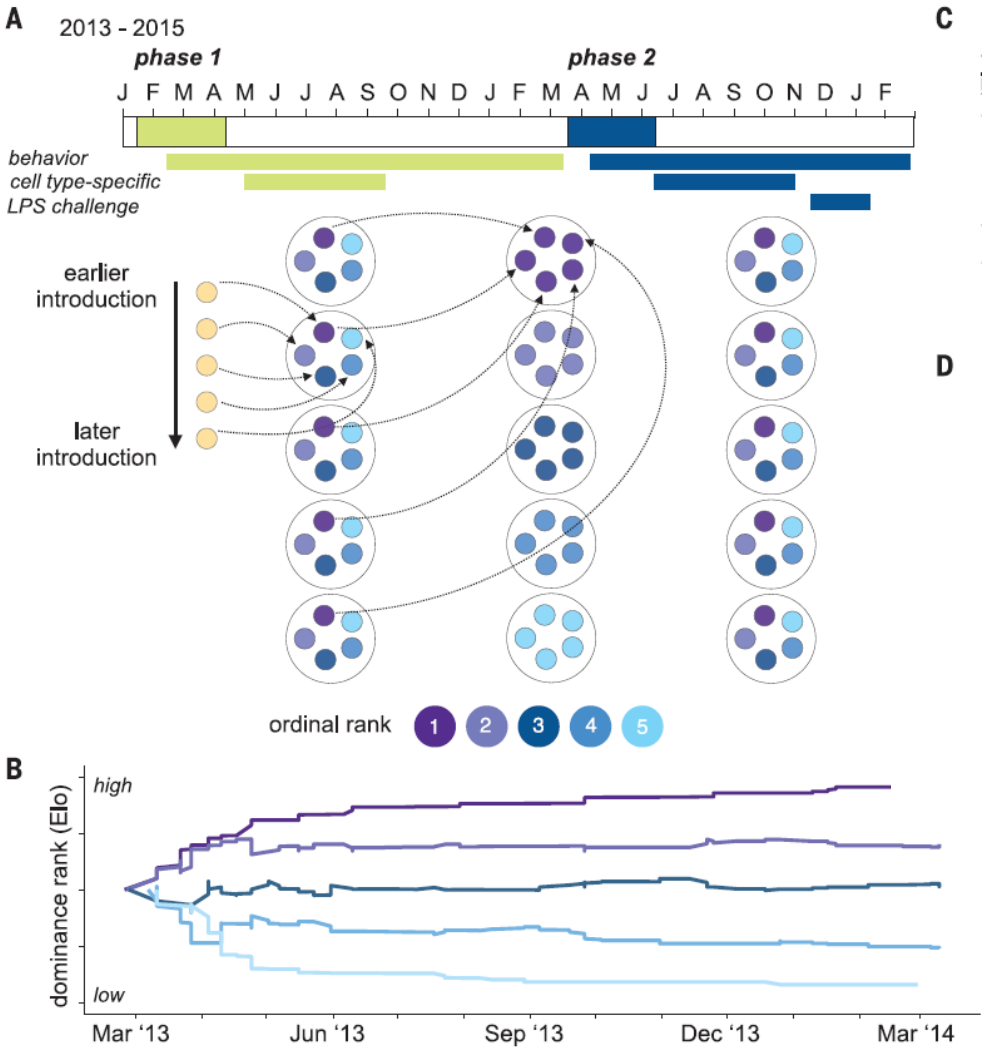
Isolation of RNA for sequencing



Isolation of RNA for sequencing

Snyder-Mackler et al. 2018

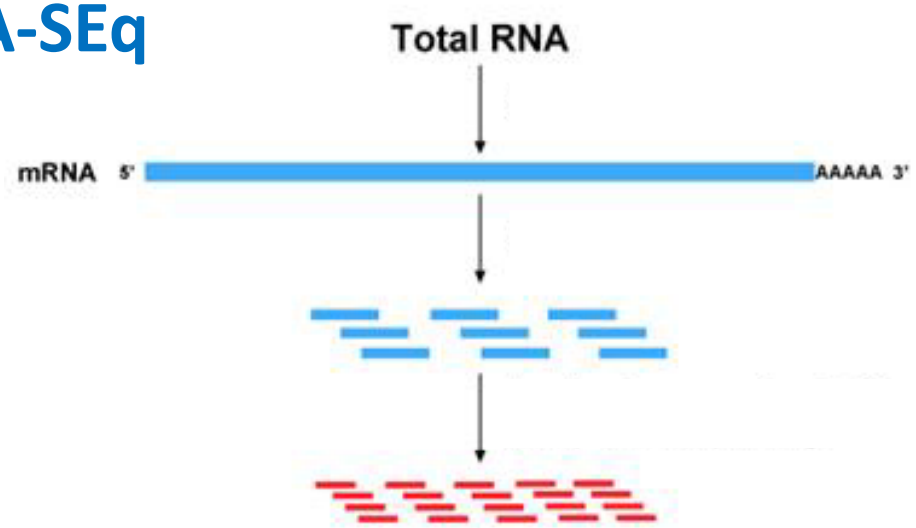
Dataset for the exercise



Snyder-Mackler et al. 2018

Quantification of gene expression

Using RNA-SEq



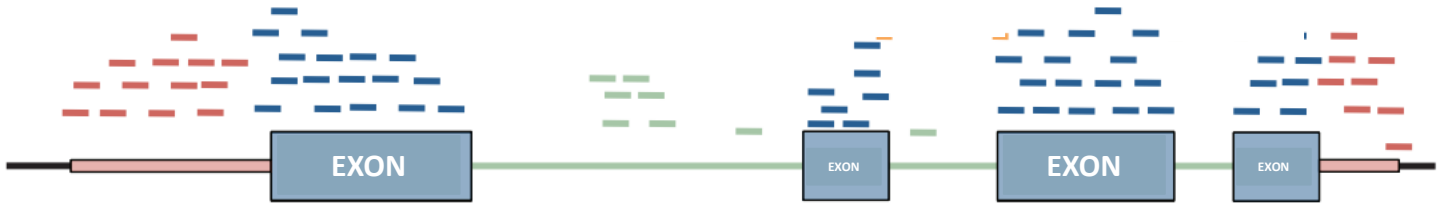
1. Biochemical amplification of all RNAs

2. Fragmentation of all RNAs

3. Biochemical conversion of all RNAs into cDNA

4. Sequencing of all cDNAs
→ Short reads (e.g. 150 nt long)

5. Mapping of reads to a reference genome / genes



6. Count how many reads map to the gene to determine its expression level
Quantification can be done at the level of genes, transcripts, exons

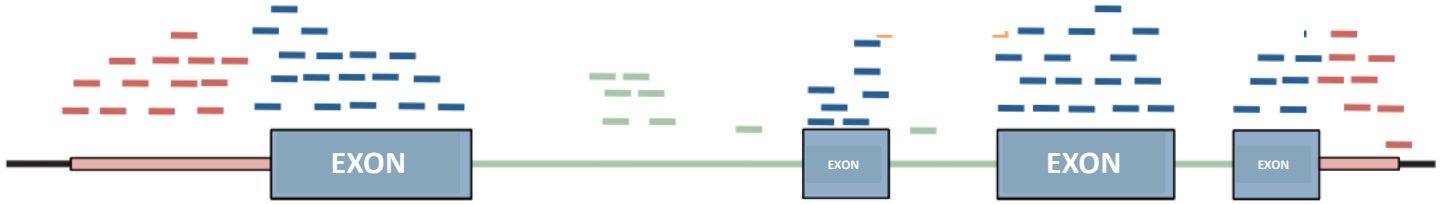
Quantification of gene expression

RNA-Seq
Counts of reads (discrete numbers)
Transcript abundance:
Counts are considered to be linearly related to transcript abundance

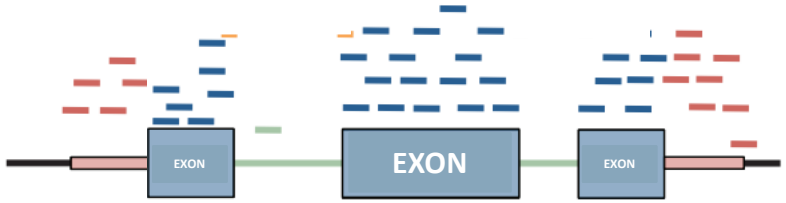
Calculate expression values for genes, transcripts (i.e. different isoforms), or exons? → Differential gene, transcript or exon expression?

Quantification of gene expression

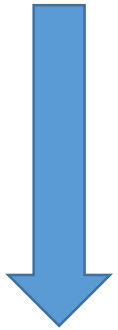
Long gene



Short gene



Has fewer counts



Need to normalize for gene / transcript length!

To normalize or not to normalize …

… for transcript length:

An early method was proposed by Mortazavi et al. in 2008:

RPKM = **R**eads **P**er **K**ilobase of exon model per **M**illion mapped reads

$$\frac{\#\ reads\ overlapping\ exons}{\frac{total\ \#\ mapped\ reads}{10^6} * \frac{total\ length\ of\ all\ exons\ (transcriptome)}{10^3}}$$

Argument against normalization for transcript length in DE analysis:
When comparing the same genes between samples, we expect (and hope) that the biases will affect the same gene in the same way in different samples.
Thus, we do not worry about gene length bias, GC bias and so on, because the biases in effect “cancel out" when making the comparison between samples.

BUT, when comparing different species:
Gene/transcript lengths, and size of the transcriptomes are different.
Hence we need to normalize for transcript length and mappability.

Mortazavi et al. 2008

To normalize or not to normalize …

… for library size (total number of reads):

Imagine, a gene has the same number of counts in two samples. But the library size was twice as high in the first sample. Then we would conclude that the gene was higher expressed in the second sample.
In other word: if a non-differentially expressed gene has twice as many counts
in one sample than in another, the size factor for this sample should be twice as large as the one for the other sample.
We will see normalization for library size using the R libraries ***DESeq*** and ***edgeR***

… for transcriptome size:

If comparing two tissues with transcriptomes of different sizes, i.e. when there are noticeably more transcripts expressed in one tissues than the other.

Statistics for differential gene expression

Currently no consensus on the best statistics for normalization and analysis of DE

Dillies et al. 2013 compared 7 commonly used normalization methods for RNA-Seq:
Total Count, Upper Quartile, Median, Quantile, DESeq, TMM (edgeR), and RPKM

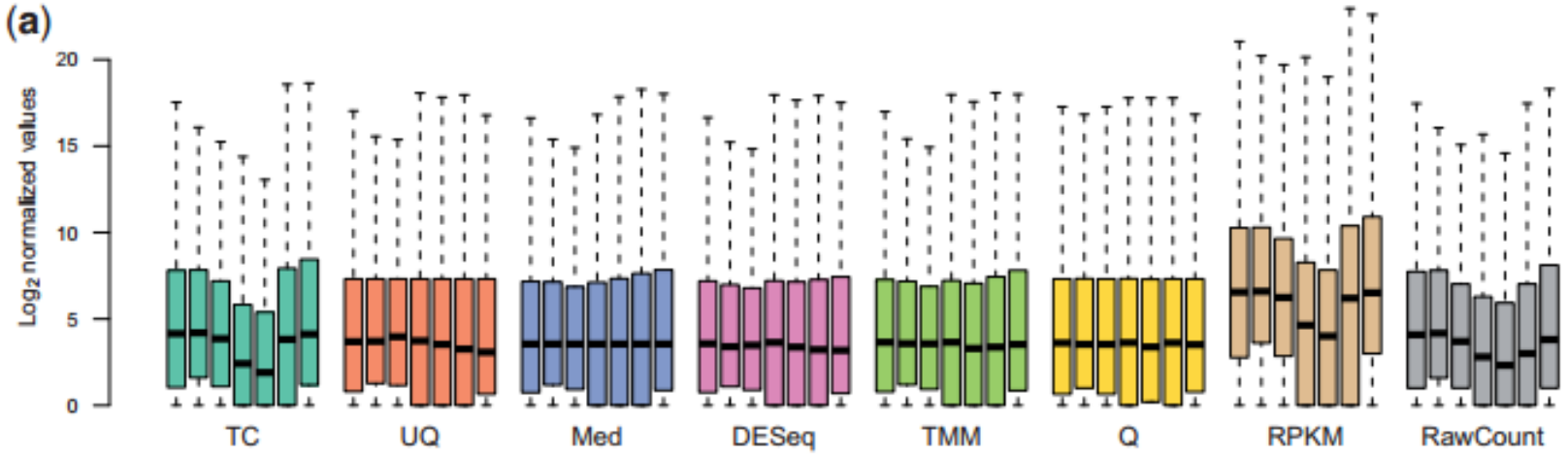
The bad message: all methods come up with different results

	TC	UQ	Med	DESeq	TMM	Q	RPKM	RC
TC	548	547	547	543	547	543	399	175
UQ		**1213**	**1195**	**1160**	**1172**	**1054**	416	184
Med			**1218**	**1147**	**1160**	**1043**	416	183
DESeq				**1249**	**1169**	**1058**	413	184
TMM					**1190**	**1051**	516	184
Q						**1092**	414	184
RPKM							417	149
RC								184

Counts along the diagonal indicate the number of DE genes per method (i.e. 548 DE genes for the TC method, etc.), while counts off the diagonal indicate the number of DE genes in common per pair of methods (i.e. 547 DE genes in common between TC and UQ). Numbers in bold correspond to pairs of methods with very similar lists of DE genes.

Dillies et al. 2013

Statistics for differential gene expression

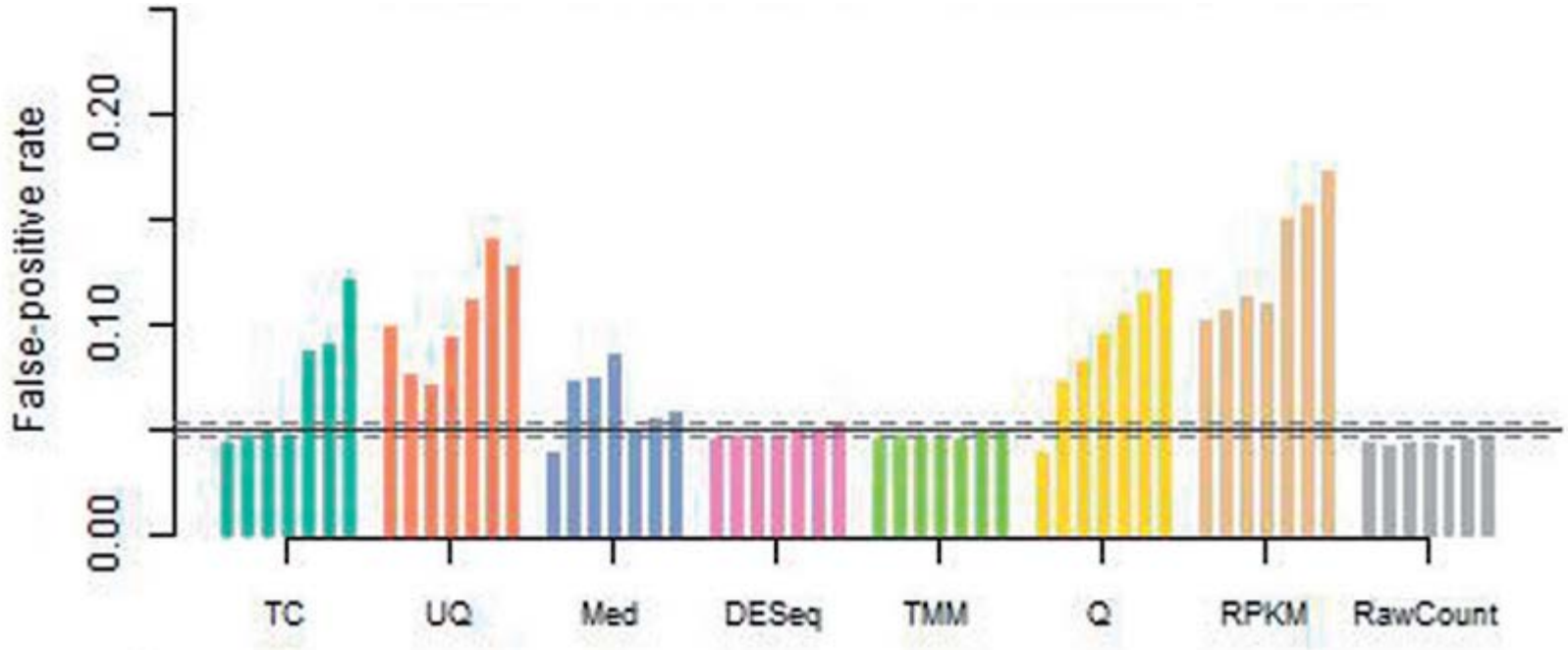


Most genes are assumed to be not DE

After normalization RPKM still looks very much like raw counts ☹

Dillies et al. 2013

Statistics for differential gene expression



False positive rate was calculated from simulated datasets, as an average over datasets with different proportions of DE genes
DESeq and edgR (TMM) were overall the best

Dillies et al. 2013

Statistics for differential gene expression

DESeq

edgeR

Both assume that most genes are not DE
i.e. when comparing across samples, most genes should have similar read counts
→ ratio of ~1
If this is not the case, a correction factor needs to be introduced

- Analyzes the **median** of the ratios across all samples
- If median is not 1 the **read counts** for that sample are corrected so that it becomes 1
- Normalized read counts are calculated from raw read counts divided by the correction factor
- Done by *estimateSizeFactors()*

- Picks one sample as the **reference**
- Analyzes the weighted **mean** of the ratios of each sample to that reference
- If mean is not 1 the **library size** for that sample is corrected so that it becomes 1
- Normalized read counts are recalculated with the corrected library size
- Done by *calcNormFactors()*

We will mainly focus on how to use these packages (not on the statistical details)

Why use logged data?

Makes up- and down-regulation mathematically equivalent

- Ratios are not symmetric around 1
  - Average of 1/10 + 10 is about 5
- Logs of ratios are symmetric around 0
  - Average of log(1/10) and log(10) is 0
  - Remember log(Y/X) = log(Y)-log(X)

Can help equalize variances of different expression levels

Tutorial on Gene Expression and Network Analysis

Let’s get started with importing data and DESeq2

Replication and Multiple Testing

A word on technical replication…

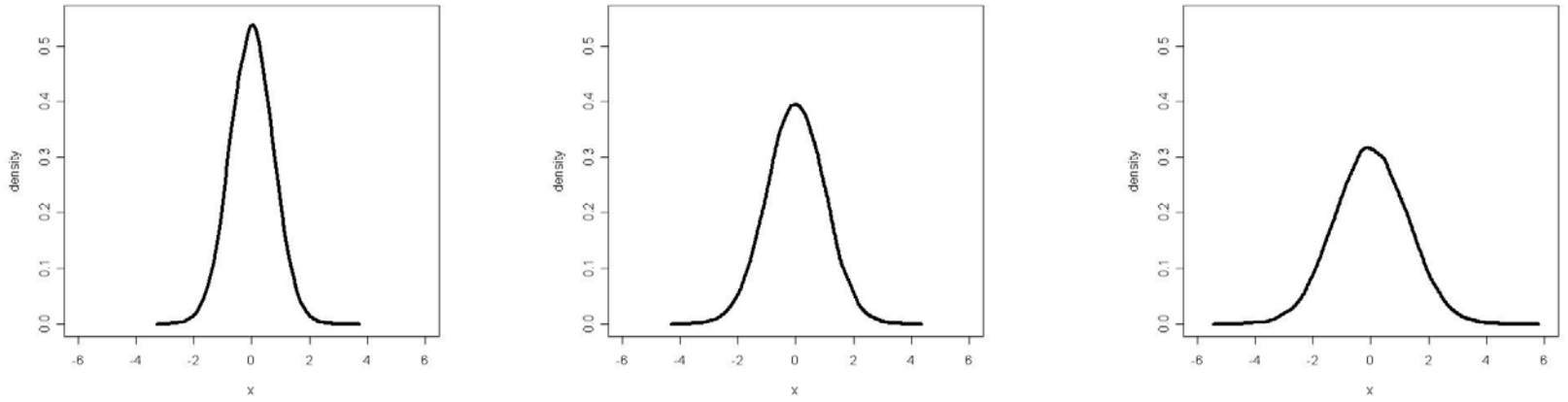
Generally have limited amount of replicates due to cost of the experiments or availability of samples

Technical replication is seen by many statisticians as a waste of time and resources because they do not substantially increase your power to detect differences...

Biological replicates: include the technical variation and give information on natural/biological variation to determine what amplitude of difference might be biologically relevant

Variation in mRNA levels

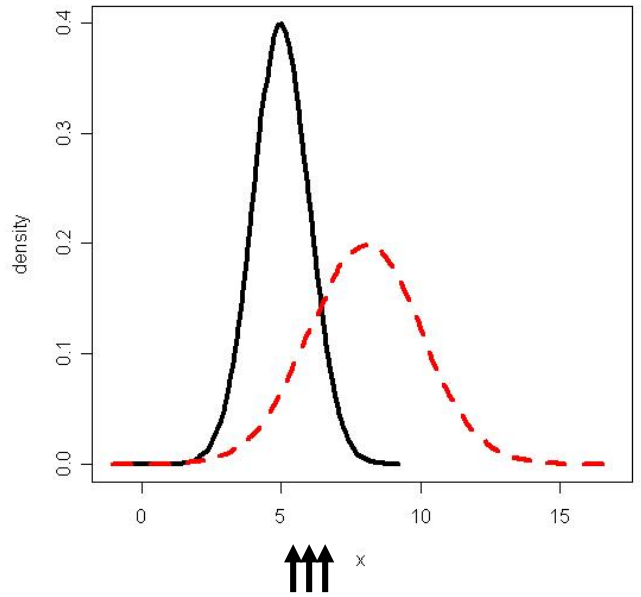
- “Real” differences among samples, even for genetically identical samples (e.g., clones, cell lines).
- “Real” differences within a sample over time.
- So for a “population” of samples, there is a “real” underlying distribution of values



Note: distributions are not necessarily normal or symmetrical!

Population Inference

- Independent samples (replicates) are necessary to make inferences about the underlying population distribution (mean and variance).
- The more replicates, the better you can estimate the distribution!

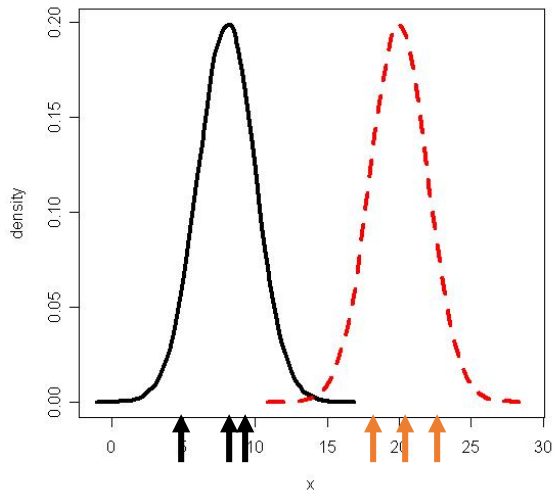


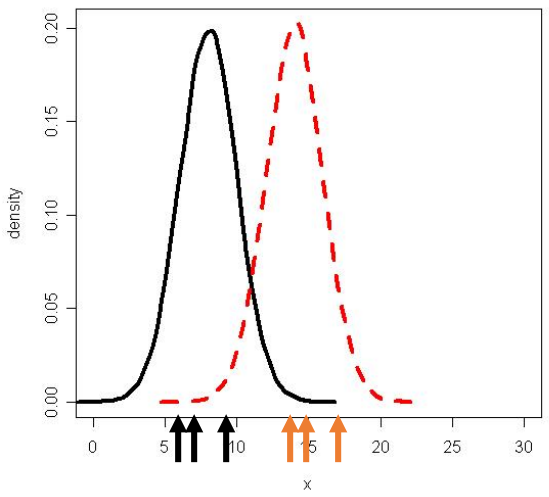
From which distribution did these three samples come?

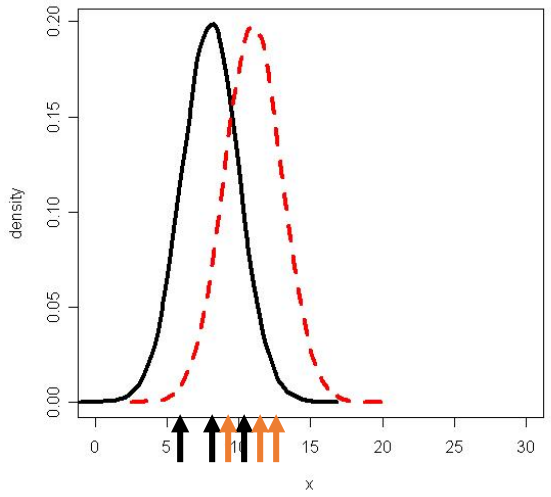
5.0, 5.8, 6.4

Comparing Populations

- How can you tell when two (or more) distributions are different?



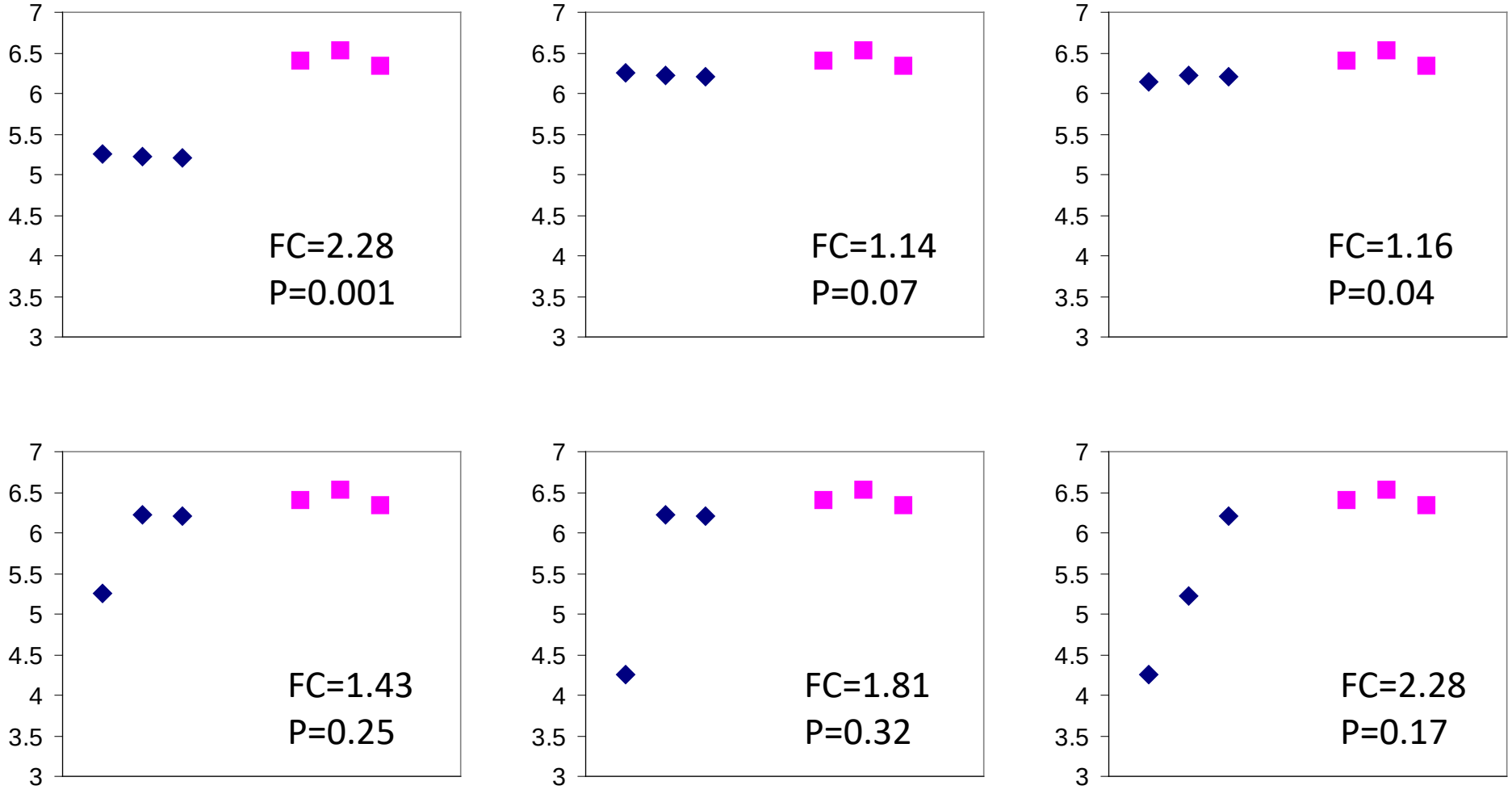




Statistics based on probabilities!

Typically t-test or ANOVA

What is different?



Be aware of effect of outliers or biological variation on FC!

T-test is a common way to identify differentially expressed genes

Other sources of variability (besides factor effects)

- Tissue contamination
- RNA degradation
- Amplification efficiency
- Reverse transcription efficiency
- Hybridization efficiency and specificity
- Clone ID and mapping
- PCR yield & contamination
- Spotting efficiency
- DNA-support binding
- Other array manufacturing related issues
- Image segmentation
- Signal quantification
- “background” correction methods

From Wolfgang Huber

http://bioconductor.org/workshops/2002/Heidelberg02/qcnorm.pdf

Multiple Testing Problem

- Most common statistical tests were good in the times before large scale datasets
- P=0.01 means if you were repeating the same test 100x you would expect the same outcome only ones (1 in 100)
- P<0.01 means you would not expect the same outcome even if you repeated the test 100 x
- Problem: With RNA-Seq, you are testing thousands of genes at the same time
- → If you had 10,000 genes and none were different, you would expect 100 to have p<0.01 by chance alone

Multiple Testing Solutions (1)

Adjust p-values to reflect multiple testing:

- **Bonferroni correction:**
  - Multiply p-values by number of tests done
  - Very conservative; greatly increases false negative rate
- **Benjamini & Hochberg correction:**
  - Less conservative, willing to accept some number of false positives in order to decrease false negative rate

Calculate False Discovery Rate:

- Try multiple p-value cut-offs and use what you are willing to accept

Multiple Testing Solutions (2)

Non-specific filtering to remove genes before analysis

- Remove non-expressed genes (e.g. genes not expressed in all samples)
- This typically reduces the number of tests by 50-75% genes

Multiple Testing Solutions (3)

a priori specification of a set of genes that are likely to be differentially expressed

- E.g. from results from other experiments or other prior knowledge
- Instead of correcting for all genes, this reduces the multiple testing correction to just that number of a priori defined genes
- But don’t define your expectation after analyzing your current dataset or change your expectation after your previous one didn’t work

Tutorial on Gene Expression and Network Analysis

Let’s continue with edgeR and DESeq2

So, I have 400 significant genes … What do I do now?

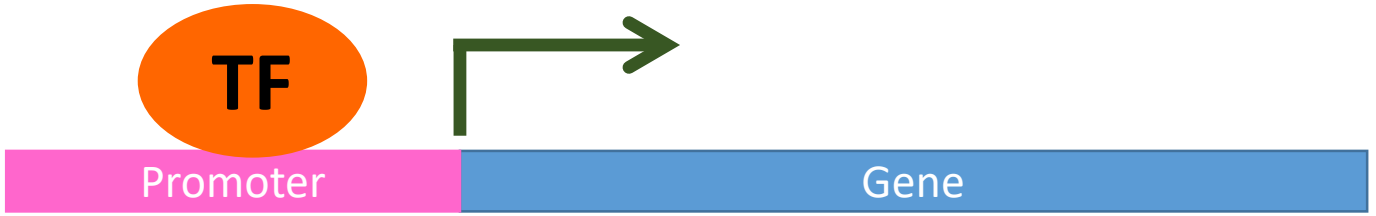
- Production of significant gene lists really represents the *beginning* of the analysis process

  - What are the genes?
  - Are any functional categories over-represented?
  - Can I infer/build a networks or pathway from my data?
  - Which changes are specific to a cell line, species …
  - How do they overlap with other/similar experiments? …

Networks

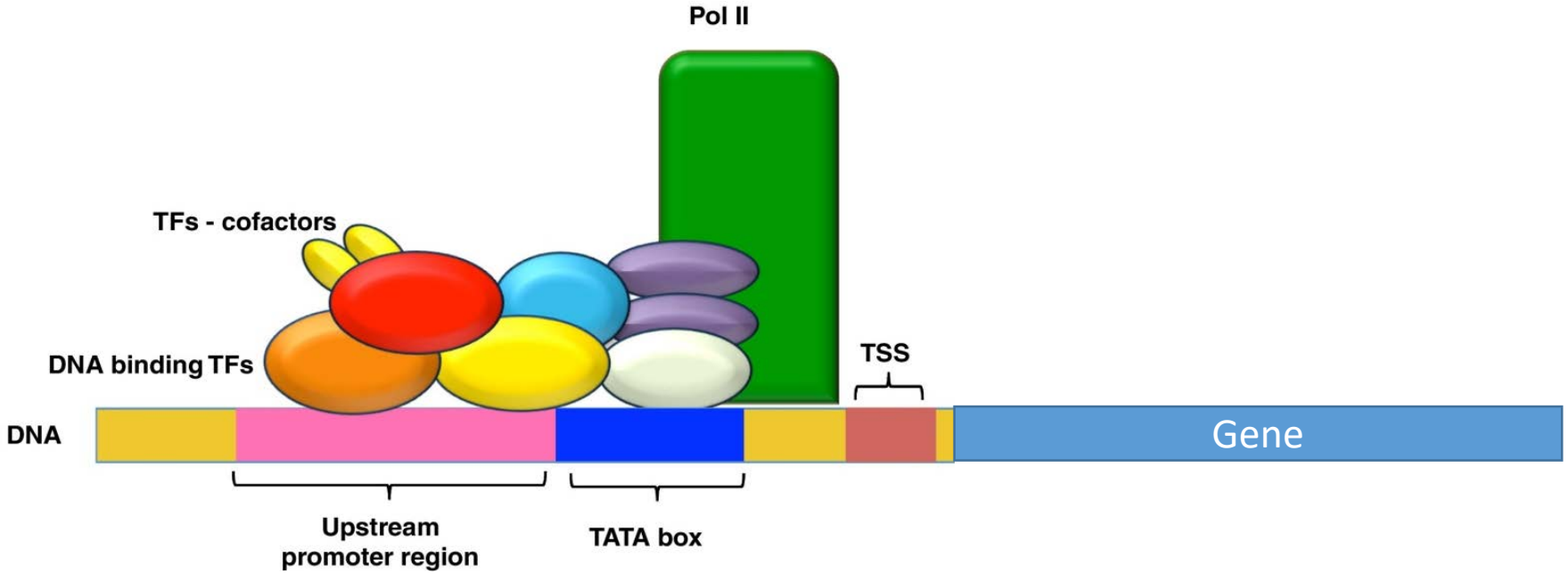
TFs regulate expression of other genes

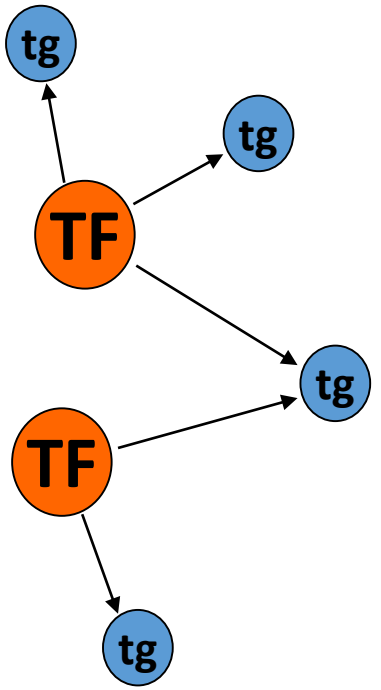
Transcription factors bind to DNA to regulate their target genes



TFs regulate expression of other genes

Transcription factors bind to DNA to regulate their target genes
And co-factors bind to transcription factors

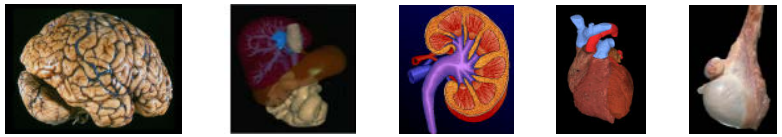




Many TFs have to come together to start/stop transcription of a target → form networks

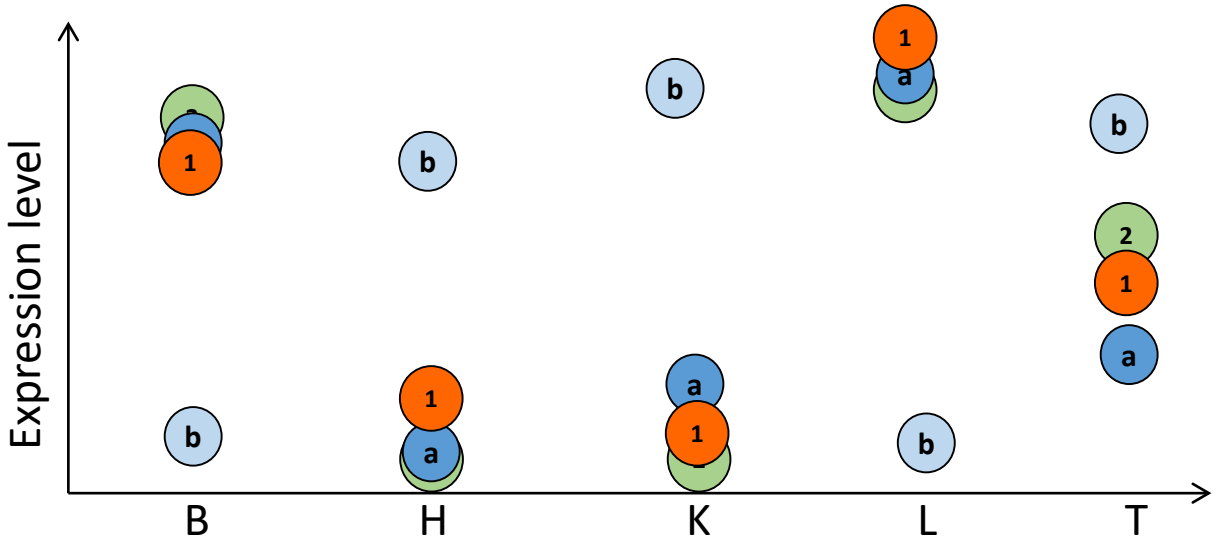
Correlation of gene expression patterns

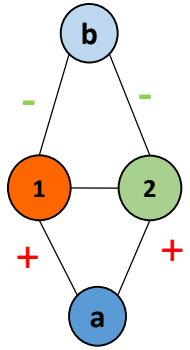
Expression profiles



Co-expression patterns

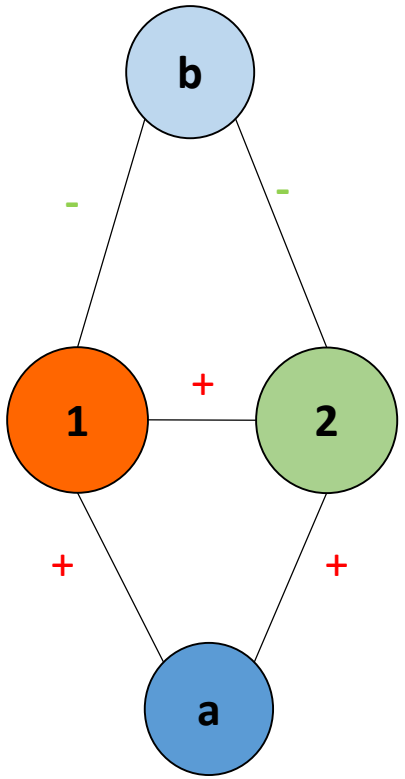
Significant Spearman correlations





→ Co-expression network

Co-expression networks

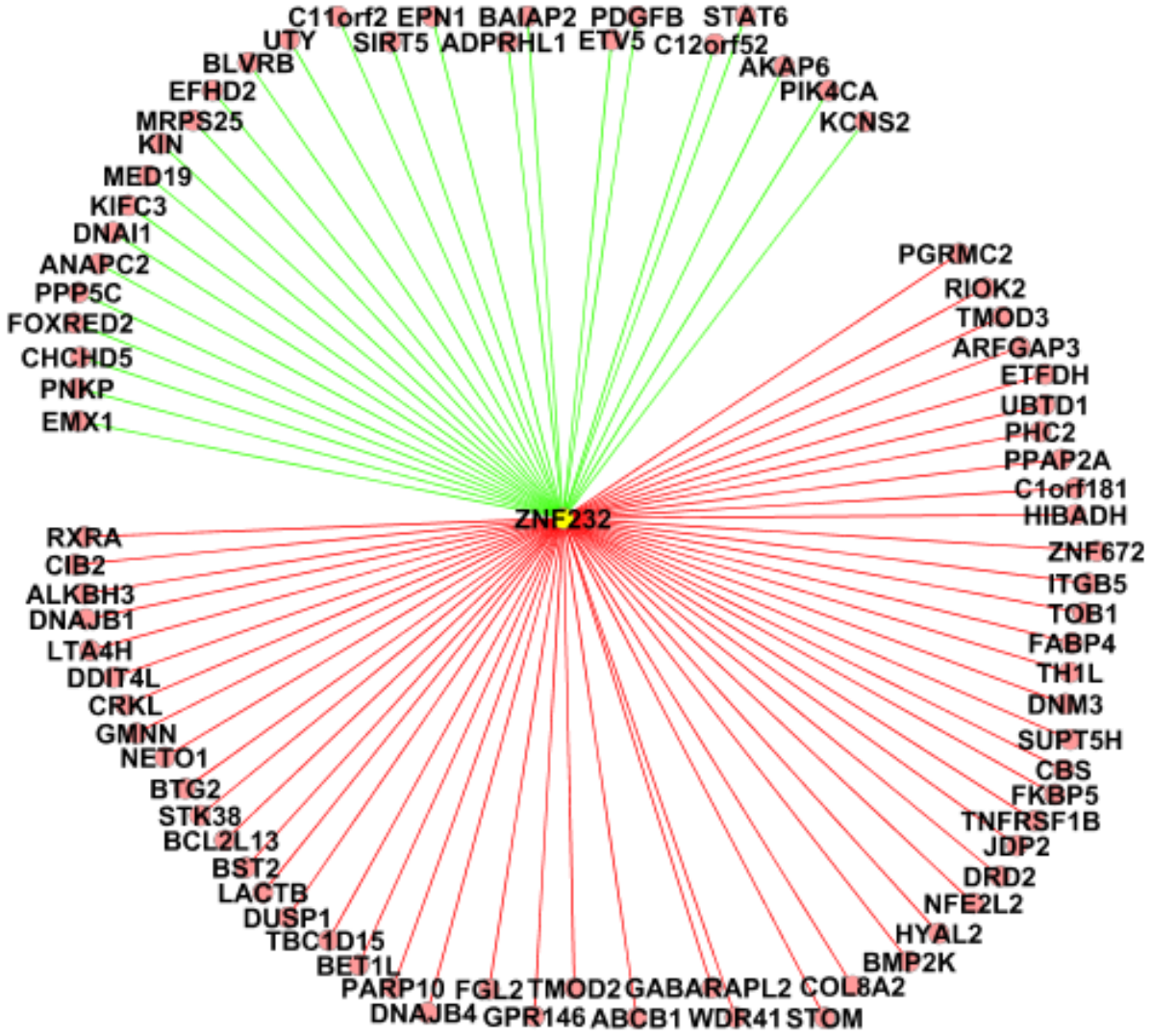


Nodes = genes/proteins
TFs and targets

Links = relationship between the genes
correlated in expression
positive or negative
undirected!

Co-expressed genes of a TF

- Potential target genes
- Potential interaction partners

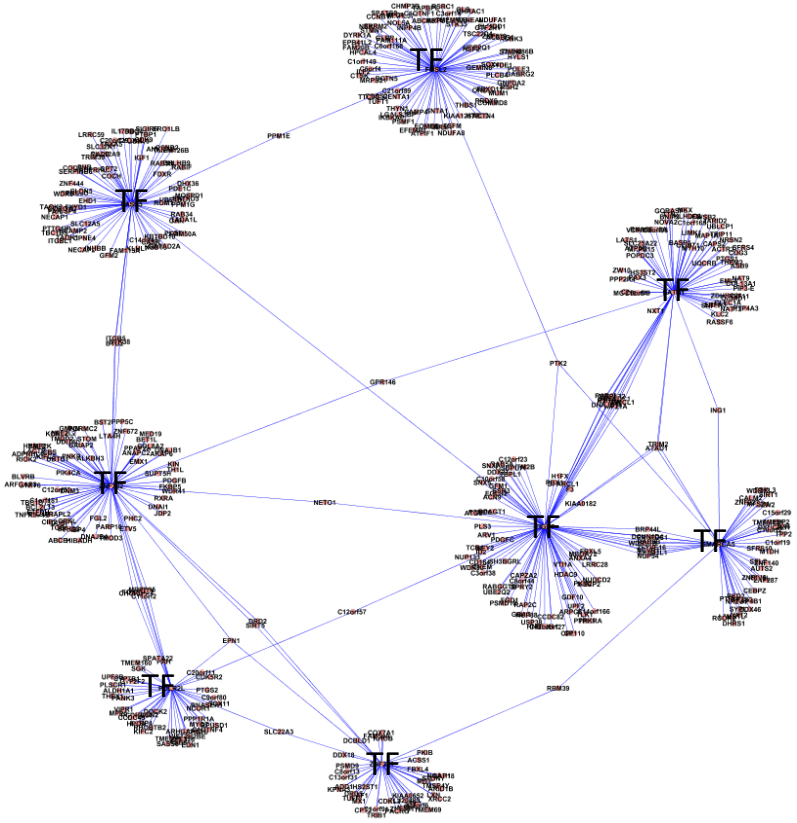


Prediction of targets and interaction partners

Genes correlated with the 24 TFs (Spearman Rank correlation)

Assumption: Genes that are expressed together, function in the same molecular pathways

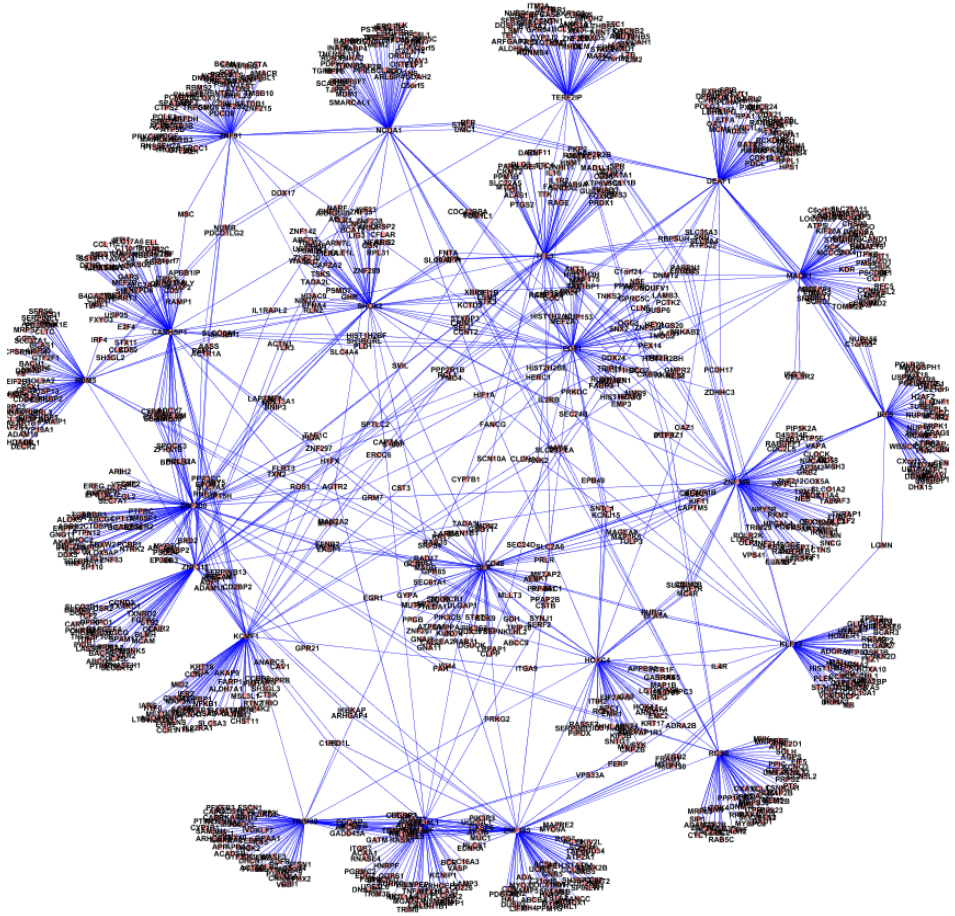
- Putative target genes
- Putative interaction partners



(Simplified representation of TFs and their co-expressed genes)

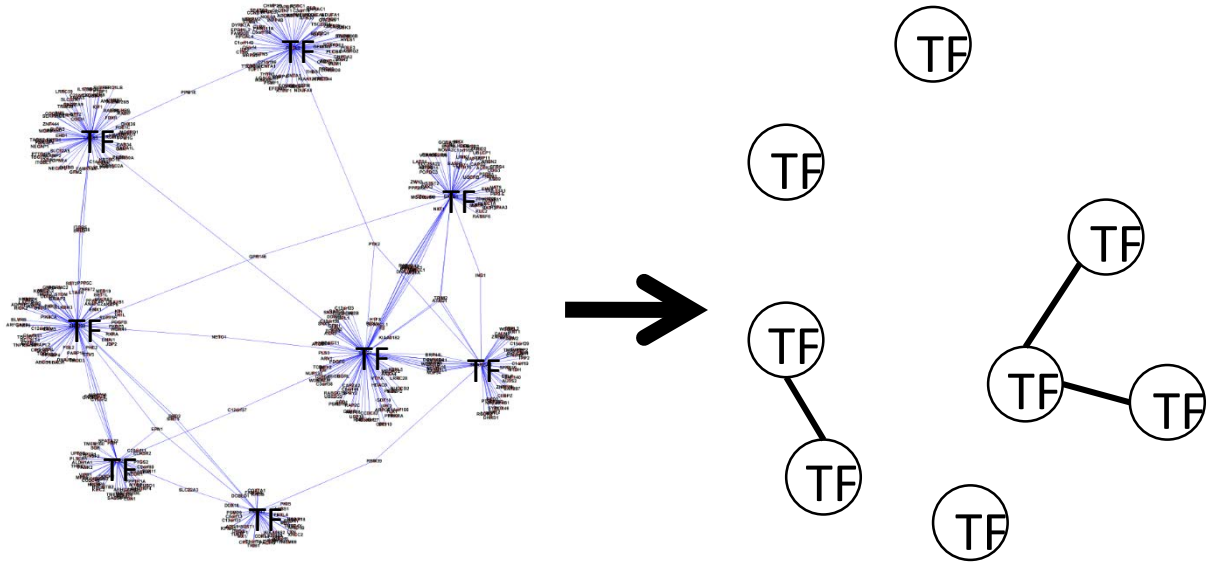
TFs Form Interaction Networks

→ Representation of TFs and their co-expressed genes



→ Many TFs share correlated genes

Weighted Topological Overlap (wTO) network for capturing TF interactions

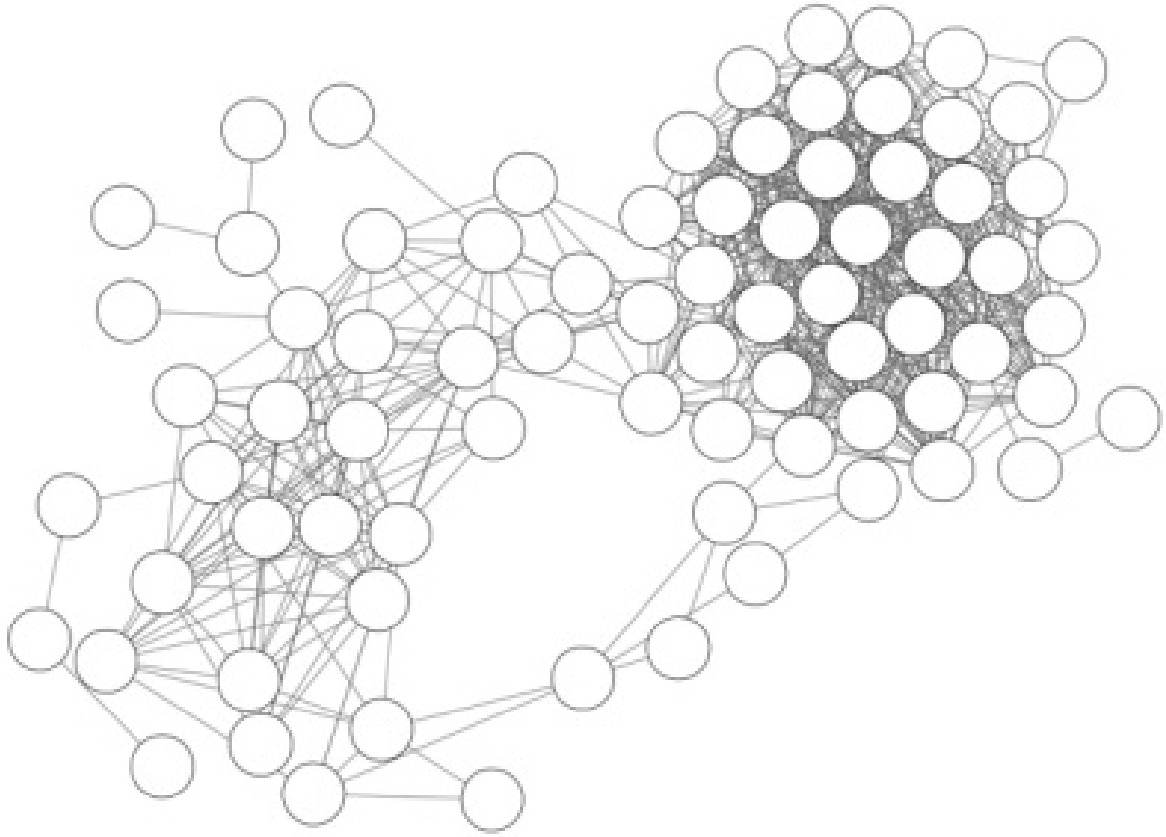


$$\omega_{ij} = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(K_i, K_j) + 1 - |a_{ij}|'}$$

i, j	TFs, nodes in the network
u	genes correlated with the TFs
a_{ij}	rho of the Spearman rank correlation between expression values of TFs *i* and *j*
a_{iu}	rho of the Spearman rank correlation between expression values of TF *i* and gene *u*
K_i	connectivity of TF *i*, $\sum_i a_{ij}$

Nowick et al. PNAS 2009

wTO Network Representation



Nodes = TFs

links = commonality of TFs in correlated genes

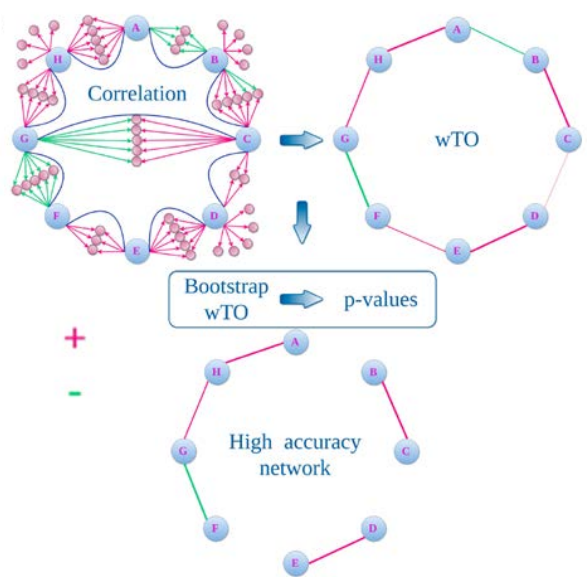
Two R packages: wTO and CoDiNA

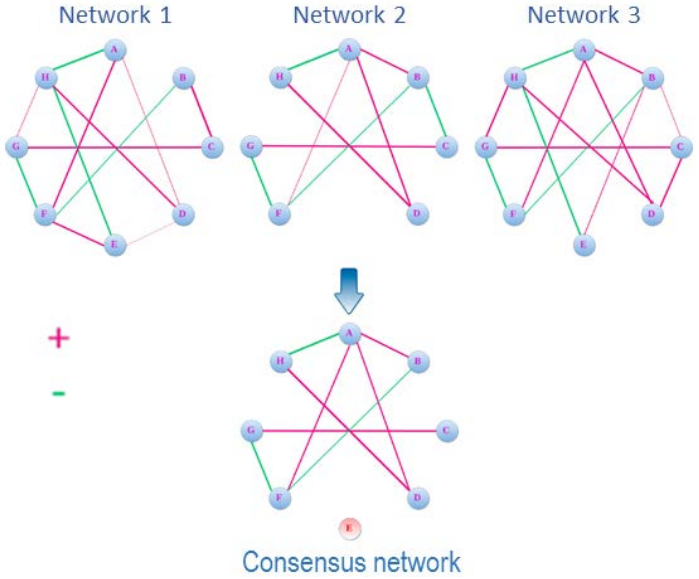
wTO

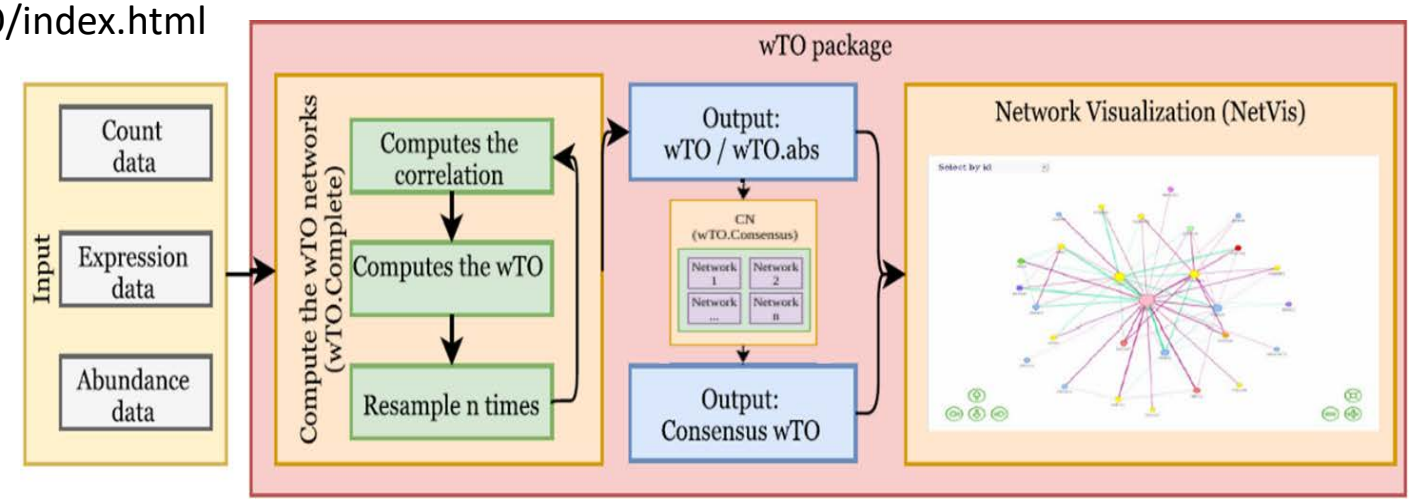
(Weighted Topological Overlap)

In CRAN and published:

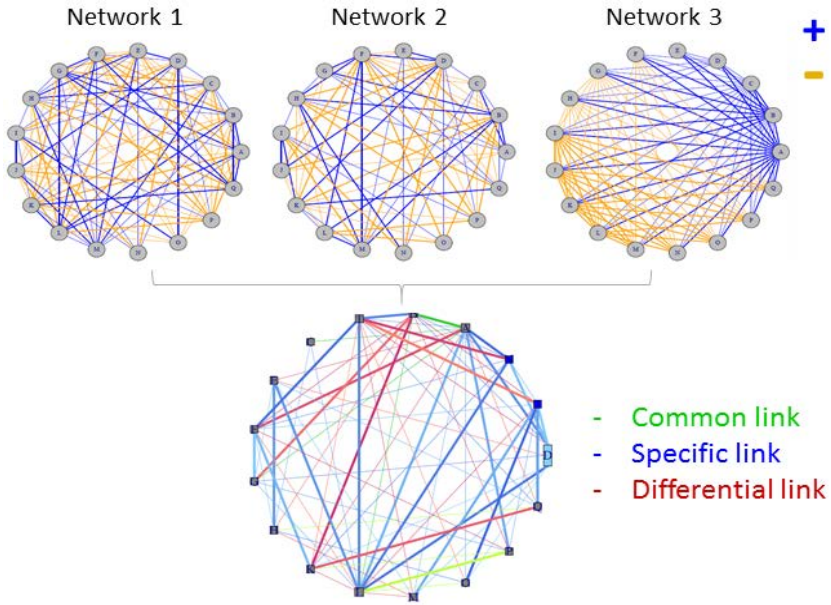
Gysi et al. *BMC Bioinformatics* (2018)
https://cran.r-project.org/web/packages/wTO/index.html







Two R packages: wTO and CoDiNA



CoDiNA

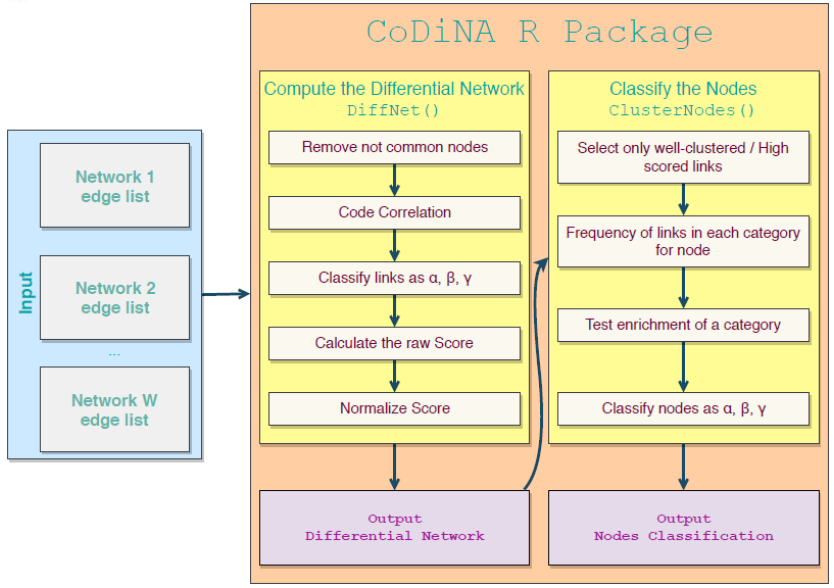
(Co-expression Differential Network Analysis)

In CRAN and arXiv:

Gysi et al. PLoS One (2020)
https://cran.r-project.org/web/packages/CoDiNA/index.html

For as many networks as desired

- Differences between tissues
- Differences between diseases
- Differences between species
- Differences between treatments
- …



Package downloads:
CoDiNA: 15363 times
wTO: 28575 times

Correlations

e.g. between the physical statures of parents and their offspring
between the demand for a product and its price

Correlation does not imply causation!

e.g. one may observe a correlation between an ordinary alarm clock ringing and daybreak, though there is no direct causal relationship between these events

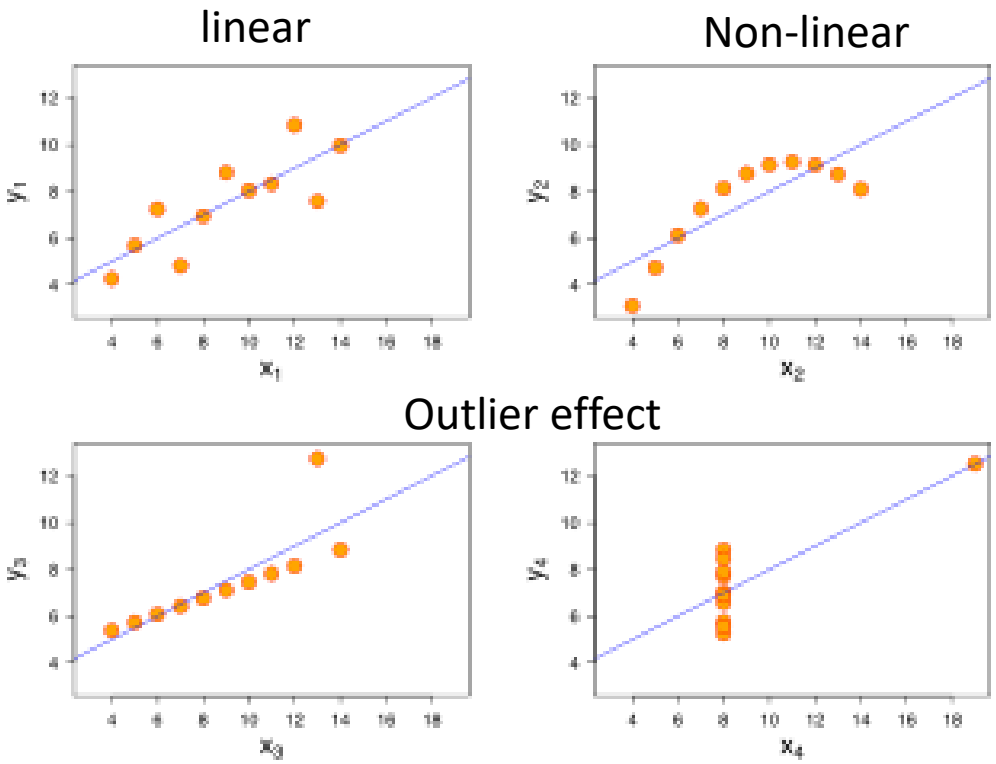
Direction often not clear

e.g. correlation between mood and health in people: Does improved mood lead to improved health, or does good health lead to good mood?

Correlations

Pearson correlation: linear relationship

observations independent from each other



Have a look at your data!

Anscombe's quartet by Francis Anscombe: The four y variables have the same mean (7.5), standard deviation (4.12), correlation (0.816) and regression line (y = 3 + 0.5x).

Correlations

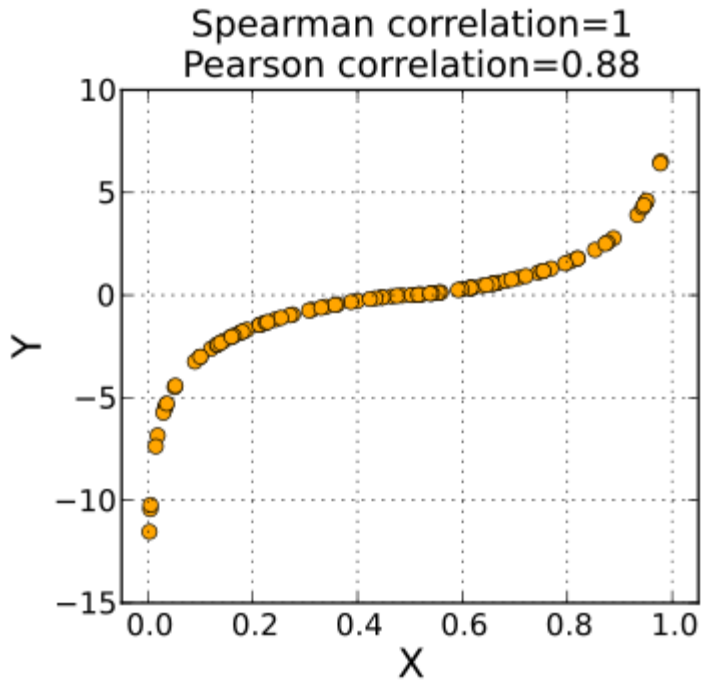
Pearson correlation: linear relationship

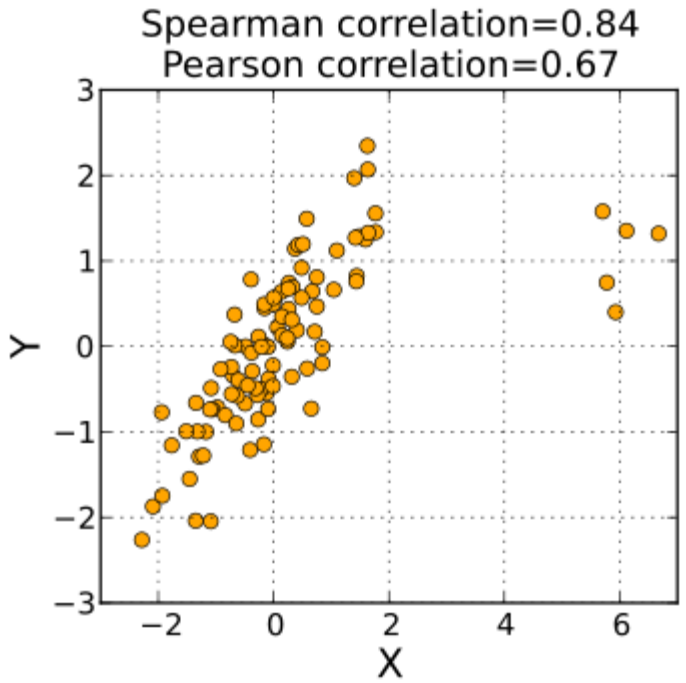
observations independent from each other

Spearman correlation: rank relationship

observations don't need to be independent

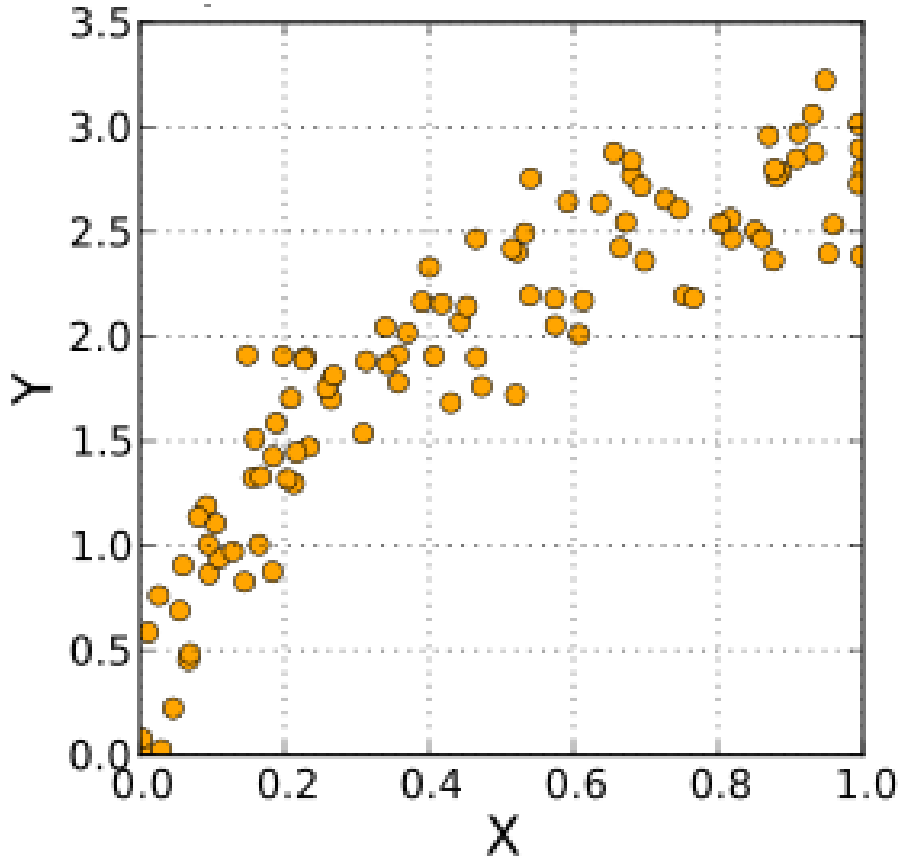
less sensitive to outliers



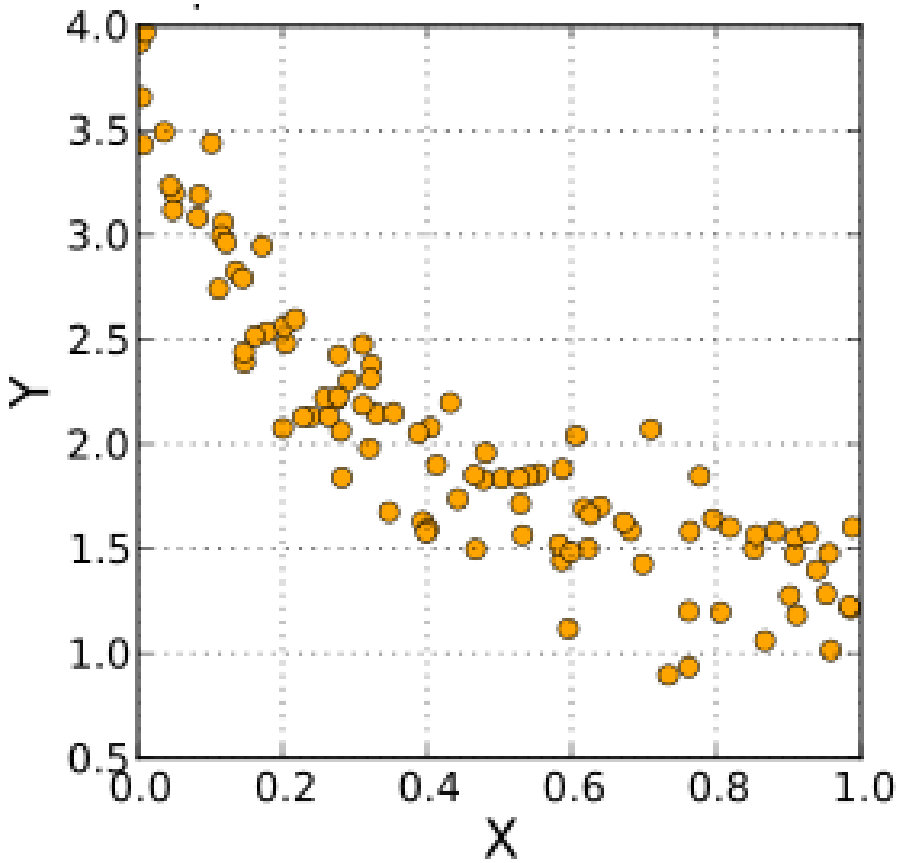


Correlations

Positive correlation



Negative correlation



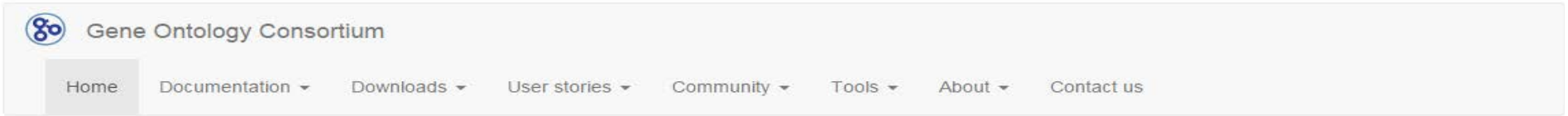
Tutorial on Gene Expression and Network Analysis

Let’s run some network analyses

Gene Ontology Enrichment

Gene Ontology Data base

http://geneontology.org/



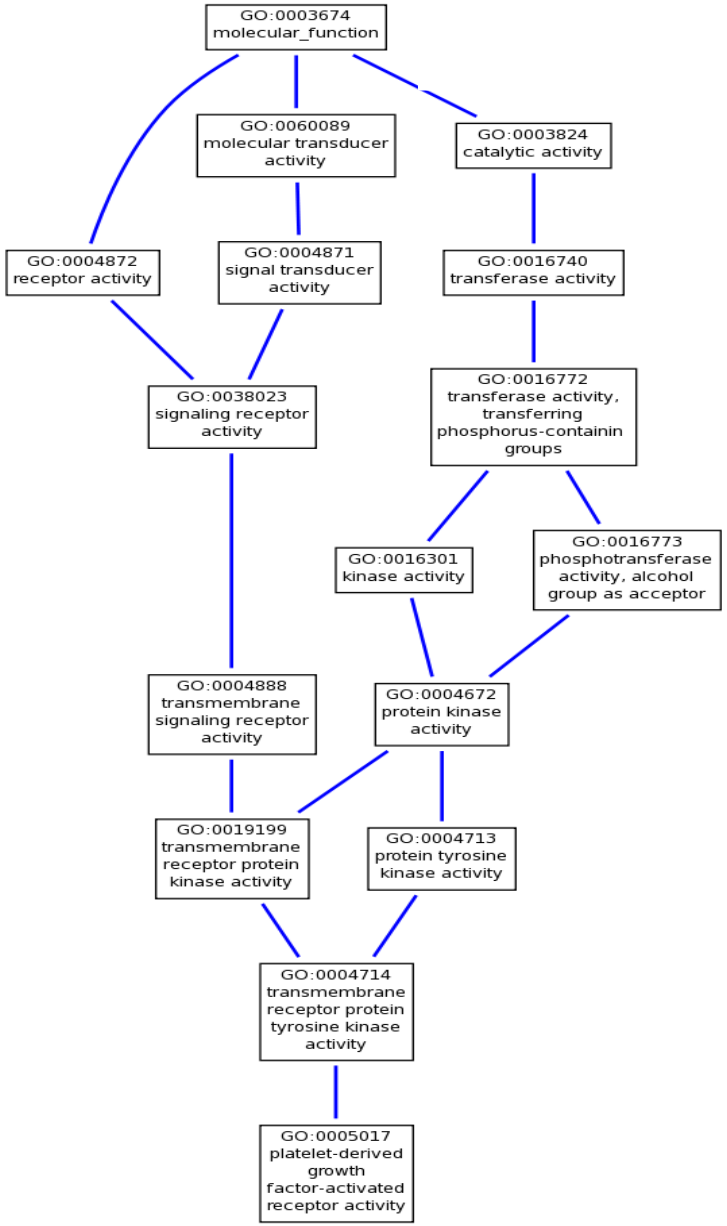
- Gene Ontology Consortium founded in 1988
- Originally three model organisms:
  FlyBase (Drosophila), Saccharomyces Genome Database (SGD) and Mouse Genome Database (MGD)
- Contains information of many data bases, including plants, animals, microorganisms
- Controlled vocabulary
- Genes classified into functional categories = “GO-groups”

- **3 Taxonomies:** Biological process
  Molecular function
  Cellular component

Ashburner, M. et al. (2000) Gene Ontology: tool for the unification of biology. Nat. Genet., 25, 25-29

Gene Ontology Data base

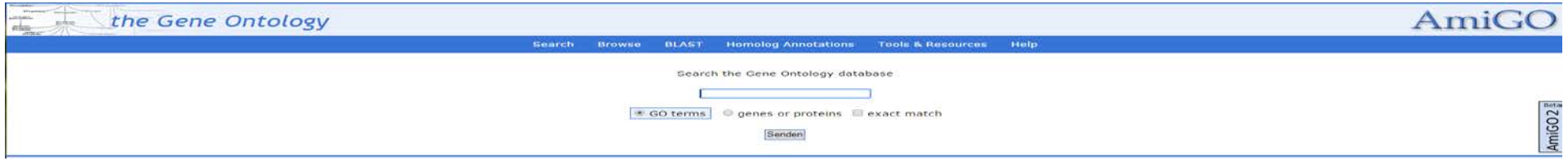
Each taxonomy is organized as **“directed acyclic graph” (DAG):**
i.e. children can have more than one parent



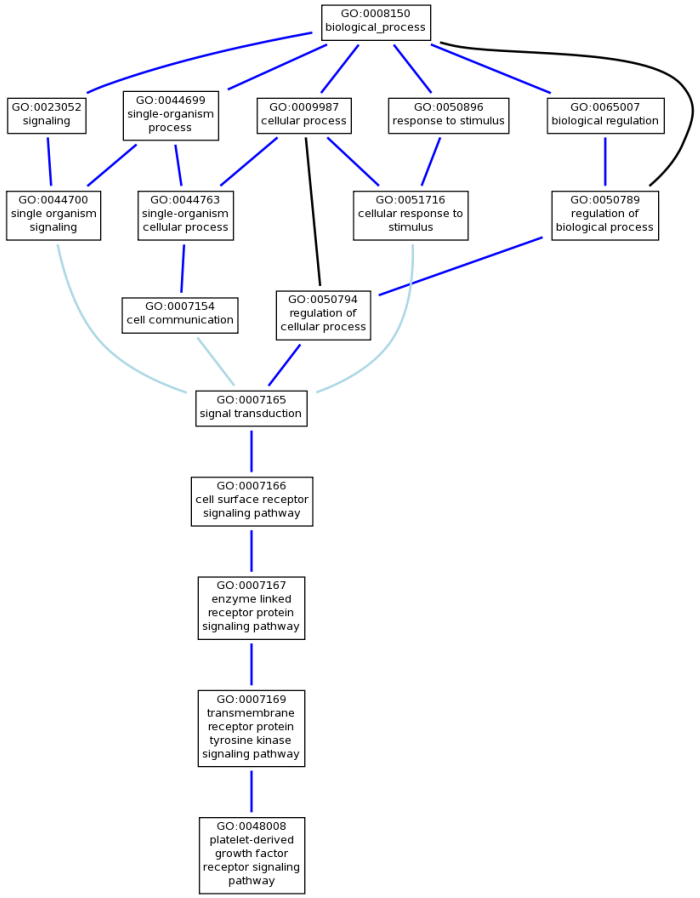
Example: Annotation for PDGF in Molecular function

AmiGO Browser: To search for GO annotations

http://amigo.geneontology.org/amigo

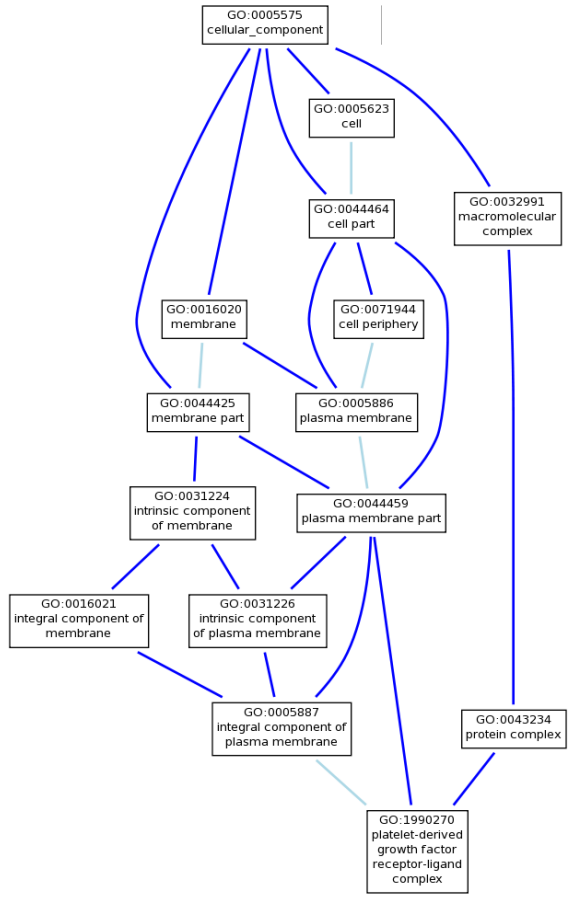


Biological Process



Molecular Function

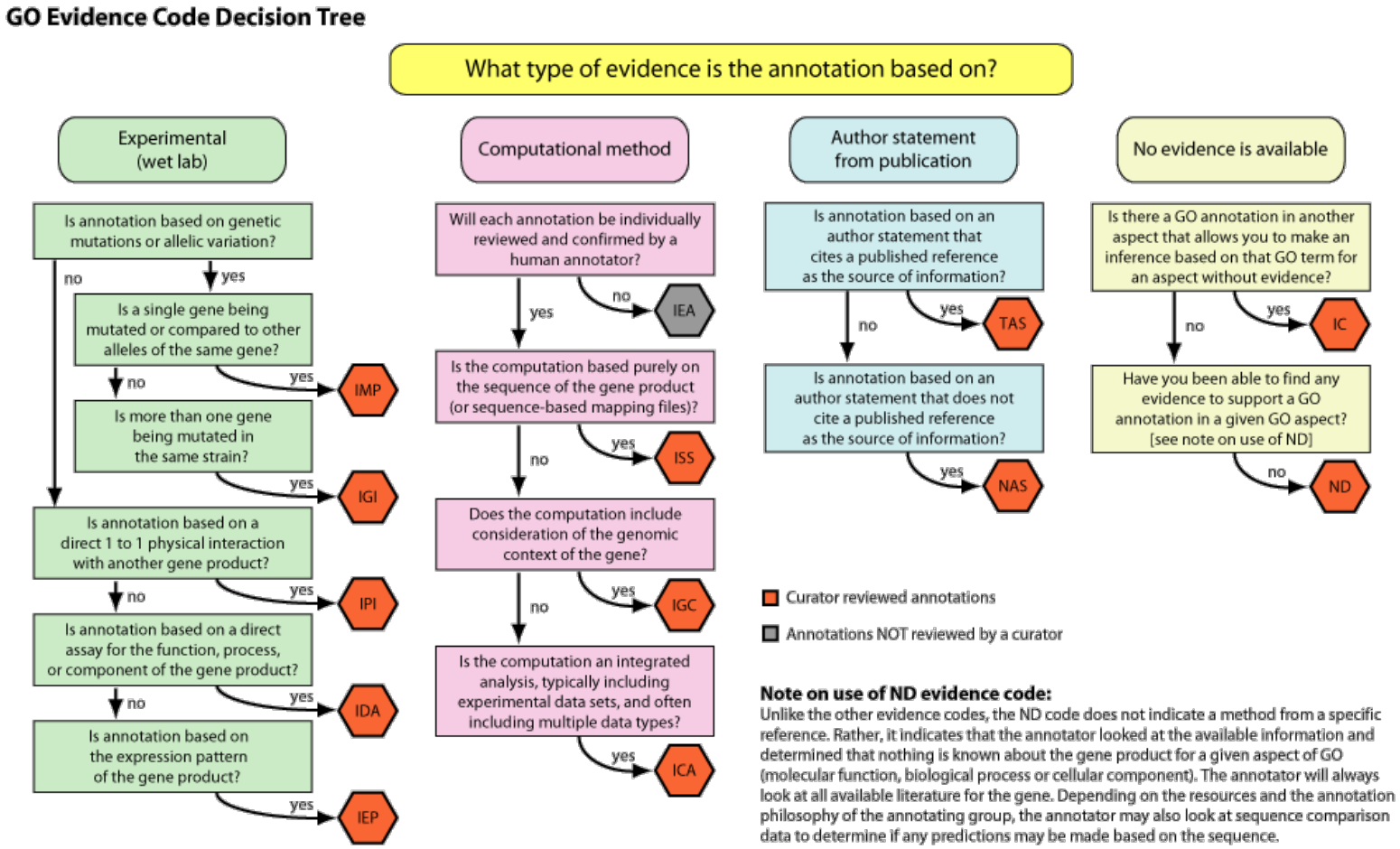
Cellular Component



Example: Annotation for PDGF

Where do the annotations come from?

Curators enter data into GO-Data base
Decide on the **Evidence-Code** in the following way:



GO enrichment tools

There are many tools, some online, some as packages

- Panther: http://geneontology.org
- Gorilla: http://cbl-gorilla.cs.technion.ac.il
- Bingo
- Ontologizer
- func

- topGO
- GOfuncR
- Gostats
- goseq

…

GO enrichment tools

Most use **Fisher's exact test** to test for enrichment

	Men	Women	*Row total*
Studying	1	9	*10*
Not-studying	11	3	*14*
Column total	*12*	*12*	*24*

The question we ask about these data is: knowing that 10 of these 24 teenagers are studiers, and that 12 of the 24 are female, and assuming the null hypothesis that men and women are equally likely to study, what is the probability that these 10 studiers would be so unevenly distributed between the women and the men? If we were to choose 10 of the teenagers at random, what is the probability that 9 or more of them would be among the 12 women, and only 1 or fewer from among the 12 men?
In other words: **Are women enriched among the studiers?**

GO enrichment tools

Most use **Fisher's exact test** to test for enrichment

	DE genes	Not DE genes	Row total
In GO group X			
Not in GO group X			
Column total			

Are the genes among the DE genes enriched for a particular GO group?

Since many GO groups are tested, correction for multiple testing necessary

Tutorial on Gene Expression and Network Analysis

Let’s finish with some GO enrichment analysis

Further exercises

Pick as many exercises as you want to
Feel free to work in groups
Have fun!

1. DESeq2 allows for more visualization and data exploration. Check out the vignette and play around.
2. CoDiNA allows for a comparison of an unlimited number of networks. Make networks for the individuals of each rank (i.e. 5 networks, only from the controls) and see if they differ.
3. We only looked at the wTO of TFs. But it would be interesting to know, with which genes hub TFs of the CoDiNA network are correlated in the LPS and in the NC samples. Pick the three highest ranking hubs, find their correlated genes (i.e. potential target genes) and check for GO enrichment among them.