Gene expression analysis
: Review and more

Sung Wook Chi
Division of Life Sciences, Korea University

What we will learn today

Gene expression analysis (Review & More)

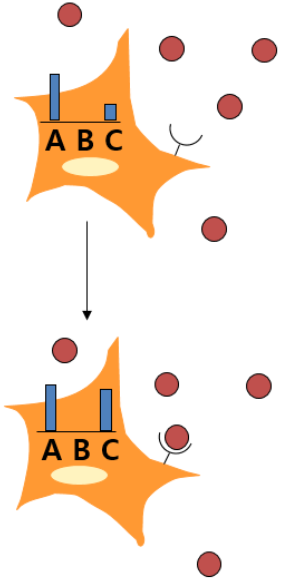
- Northern Blot, qPCR
- Microarray / RNA-Seq
- Gene expression data analysis
- Preprocessing
- Normalization
- Normalization issues for gene amplification
- Differentially Expressed Genes (DEG)
- Clustering

- Ribo-Seq (Ribosome footprinting, Ribosome profiling)

Gene expression analysis

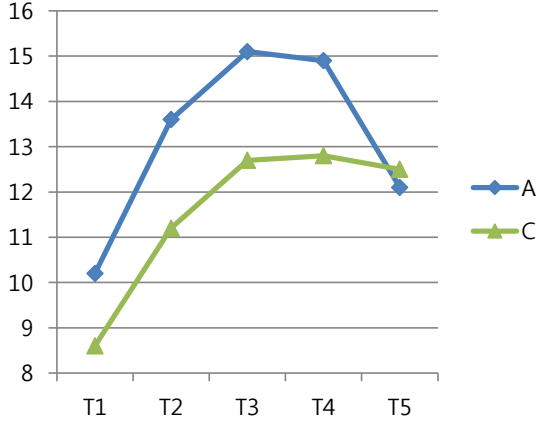
Differentially expression gene (DEG) analysis

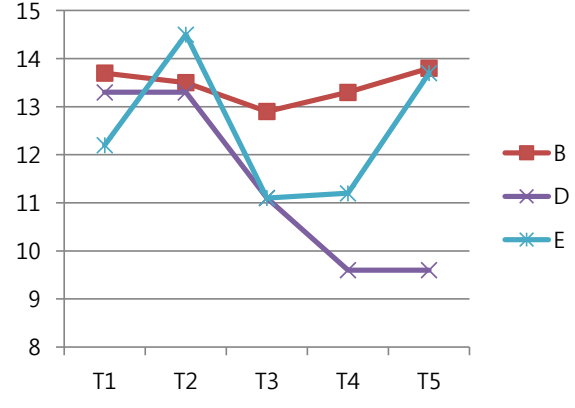
- What genes are (or are not) expressed?
  - In different cells
  - Under different external conditions
  - In different disease states
- How much does their expression change?



Gene expression profiling (Temporal expression)

- Does the change in expression correlate with other observed parameters (time series)?

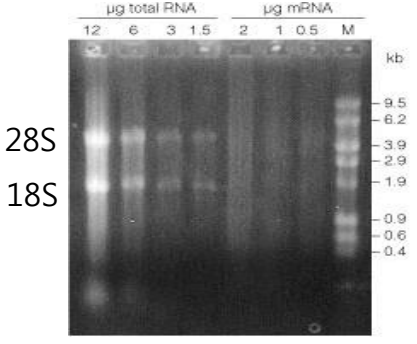




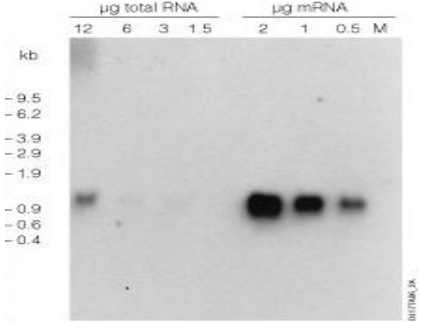
Measuring gene expression (transcripts)

Northern Blot

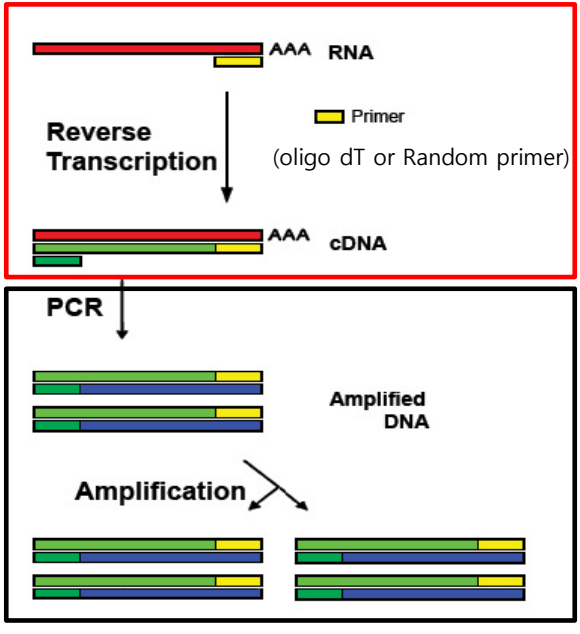
1) RNA extraction
2) Electrophoresis

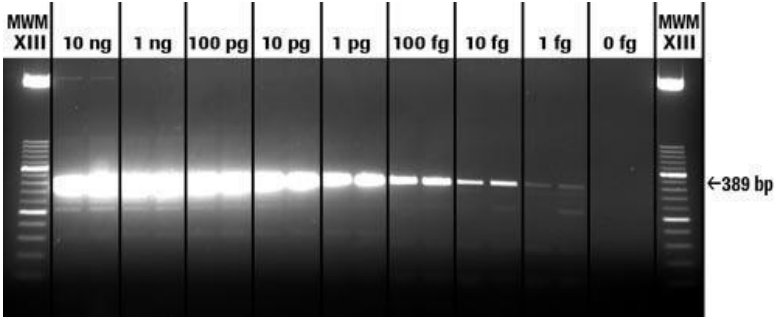


3) Membrane trasfer
4) Probe hybridization
5) Visualization



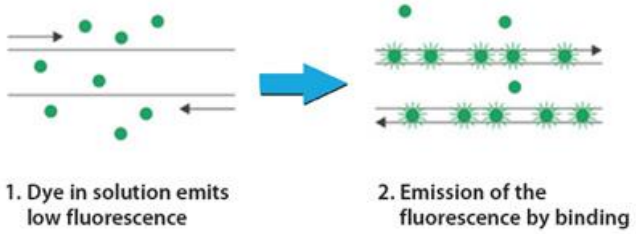
Reverse Transcription (RT) PCR

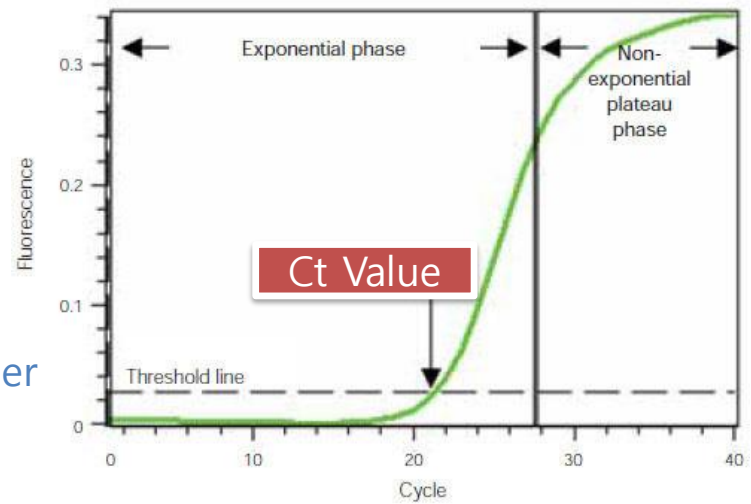




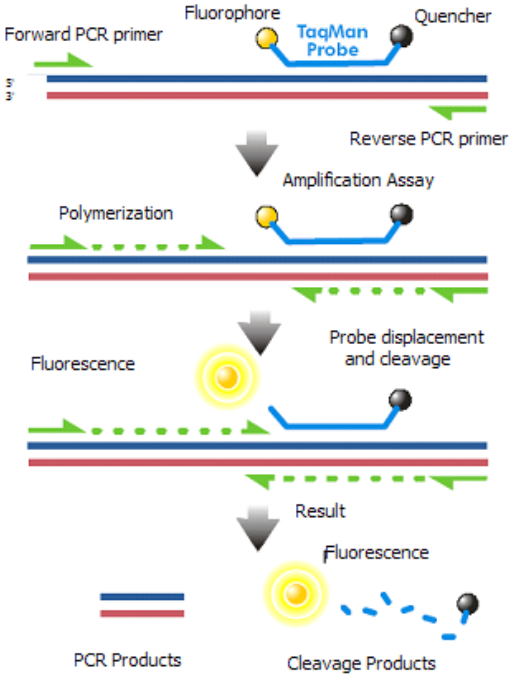
Quantitative PCR (qPCR, real-time PCR)

1) SyBR green : Double strand binding

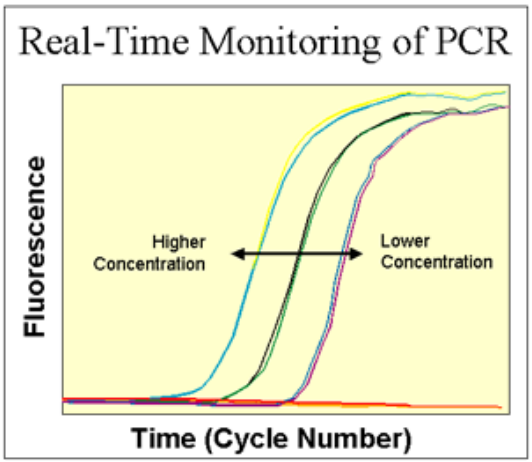




2)TaqMan Probe : Fluorophore & Quencher



The **Ct (cycle threshold)** is defined as the number of cycles required for the fluorescent signal to cross the threshold



S1 vs. S2

ΔCt =Ct (sample) – Ct (control)

ΔΔCt = ΔCt (S1) – ΔCt (S2)

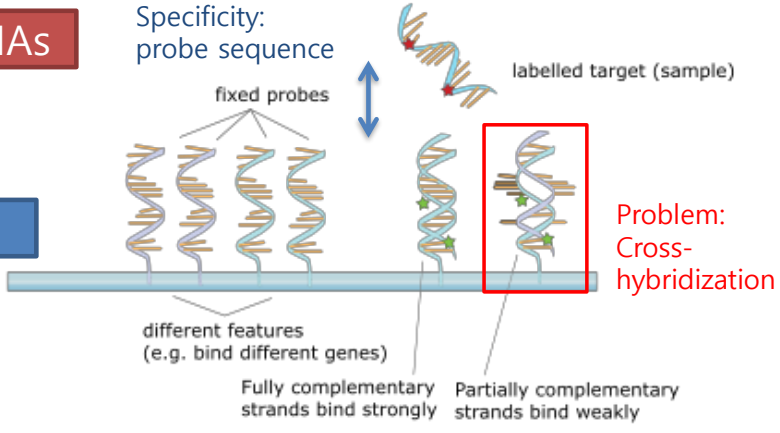
$2^{-\Delta\Delta Ct}$

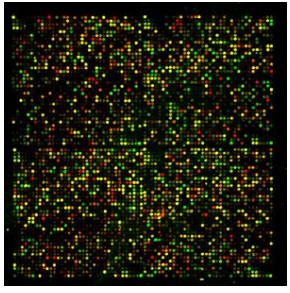
Relative expression

Microarrays

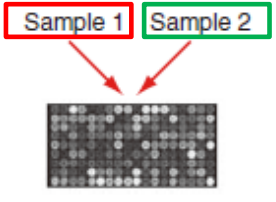
Labelled mRNAs

Fixed probes

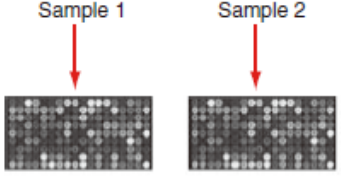


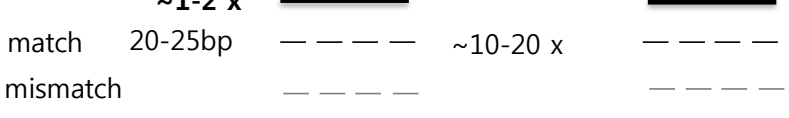


1) cDNA microarray (two channel)



2) Oligonucleotide microarray : Affymetrix (one channel)

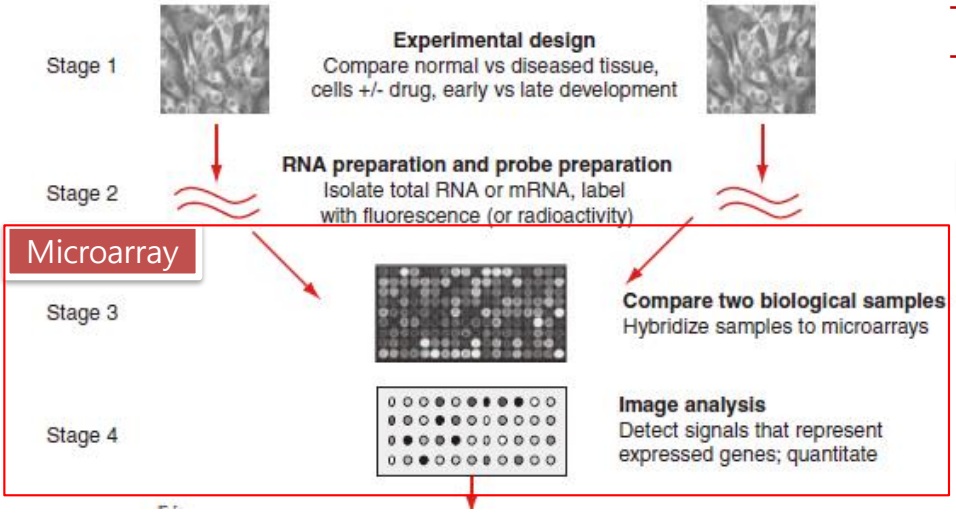


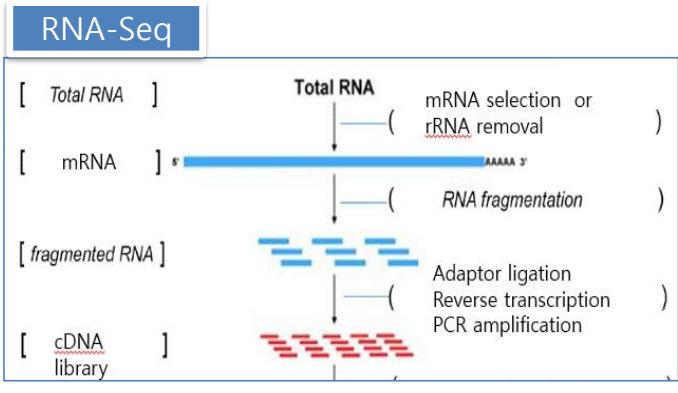


Unique in genome
Nonoverlapping

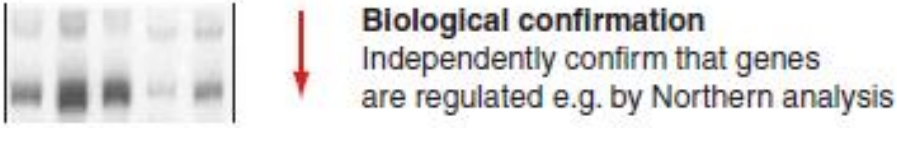
Global Gene Expression Analysis

- Microarray
- RNA-Seq





Data analysis: Identify significantly regulated genes (e.g. using scatter plots)
Identify co-regulated genes (e.g. cluster analysis); classify samples



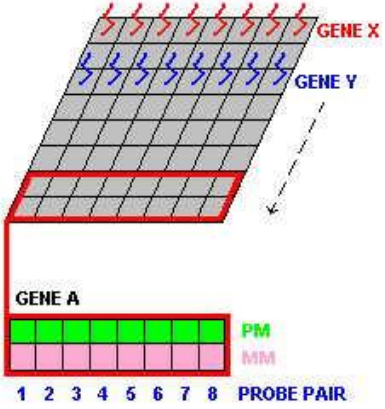
Measurement (Microarray) : ratio, log ratio

cDNA array – ratio, log ratio

$$T_i = \frac{R_i}{G_i} \text{ OR } \log \text{ratio} = \log_2 \frac{R_i}{G_i}$$

- *Filter out near zero-value probes (~10,000 expressed transcripts)*

Oligonucleotide array (Affymetrix array)



$$\text{Difference}_{\text{probe pair}} = PM - MM$$

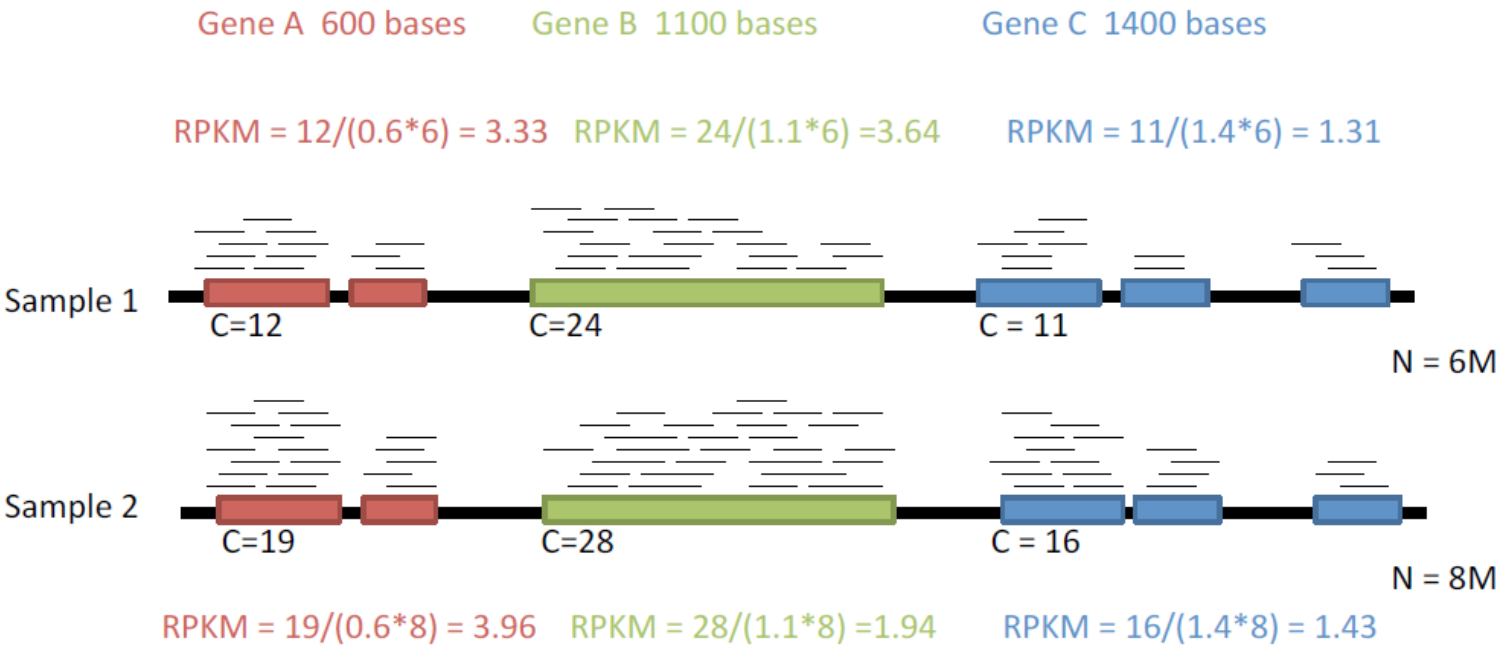
$$\text{Average Difference}_{\text{probe set}} = \sum_{i=1}^{n} \frac{(PM_i - MM_i)}{n}$$

MAS, RMA,GCRMA

Probe intensity -> Log (Normalized probe intensity)

- *Filter out probes with absent calls (~10,000 expressed transcripts)*

Measurement (RNA-Seq) : RPKM (FPKM)

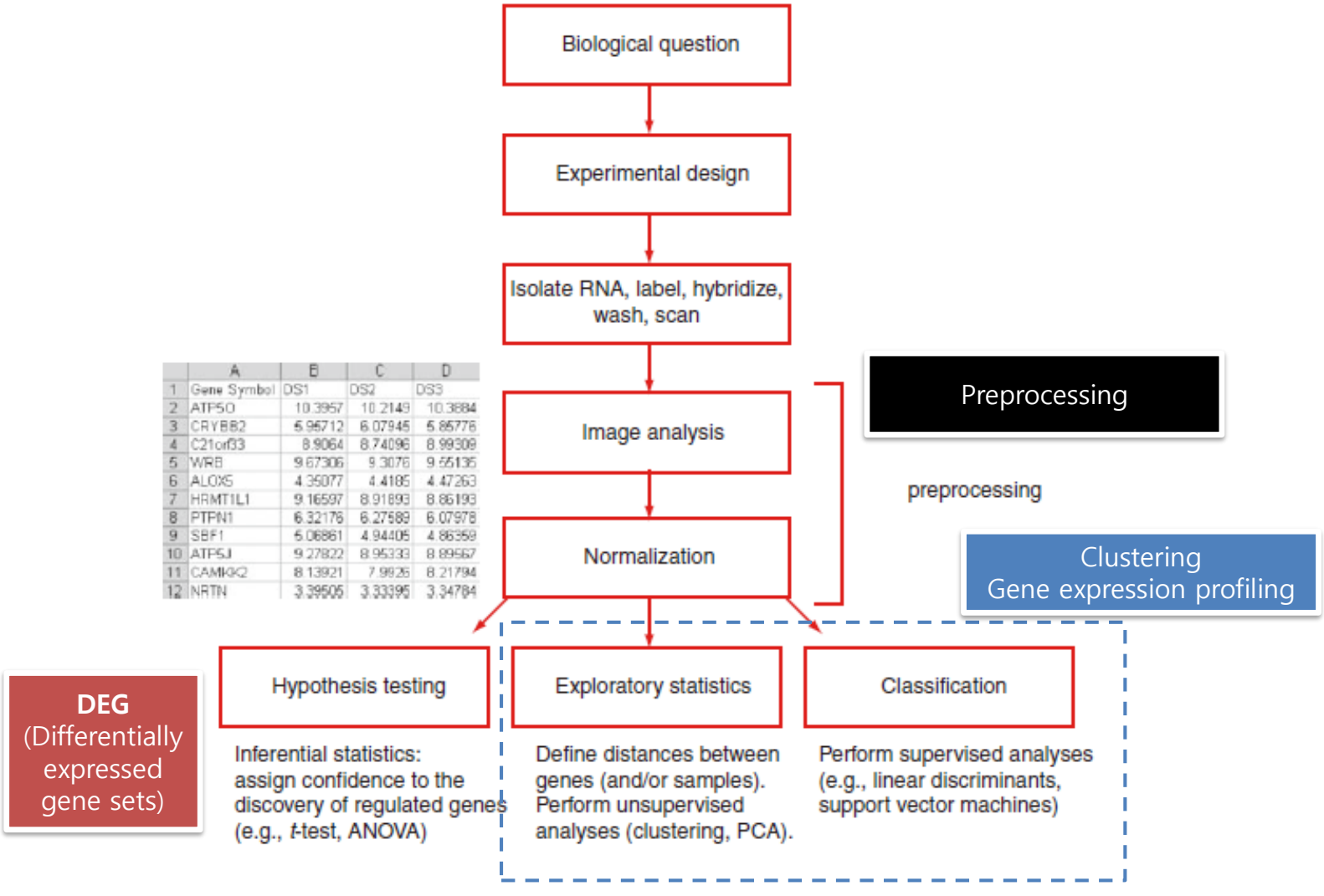


- **RPKM**
  - Reads Per Kilobase of exon model per Million mapped reads

$$R = \frac{10^9 C}{NL}$$

C : the number of mappable reads that fell onto the gene's exons
N : the total number of mappable reads in the experiment
L : the sum of the exons in base pairs.

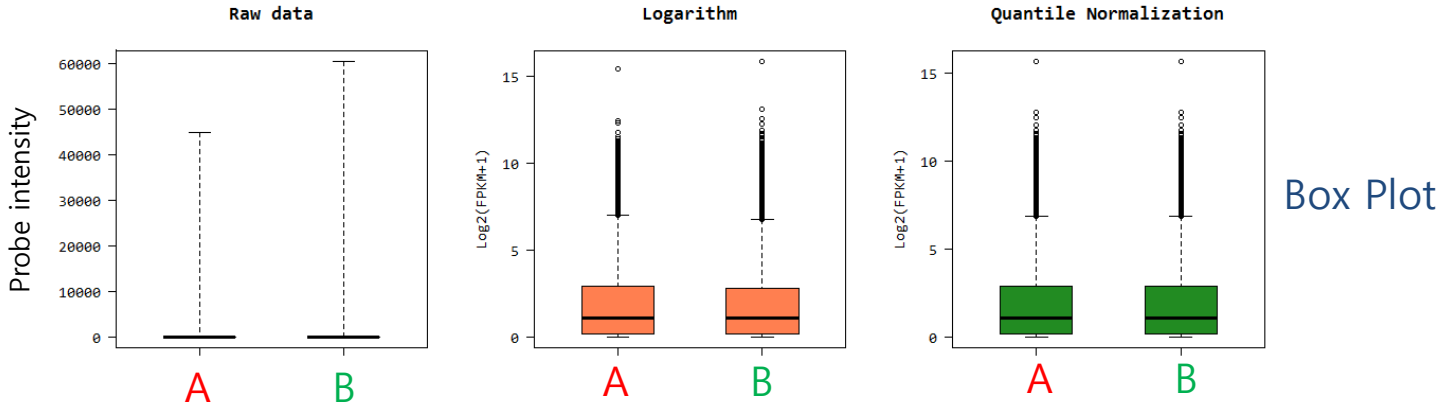
Gene Expression analysis



	A	B	C	D
1	Gene Symbol	DS1	DS2	DS3
2	ATP5O	10.3957	10.2149	10.3884
3	CRYBB2	5.95712	6.07945	5.85776
4	C21orf33	8.9064	8.74096	8.99309
5	WRB	9.67306	9.3076	9.55135
6	ALOX5	4.35077	4.4185	4.47263
7	HRMT1L1	9.16597	8.91893	8.86193
8	PTPN1	6.32176	6.27589	6.07978
9	SBF1	5.06861	4.94405	4.86359
10	ATP5J	9.27822	8.95333	8.89567
11	CAMK2	8.13921	7.9926	8.21794
12	NRTN	3.39505	3.33395	3.34784

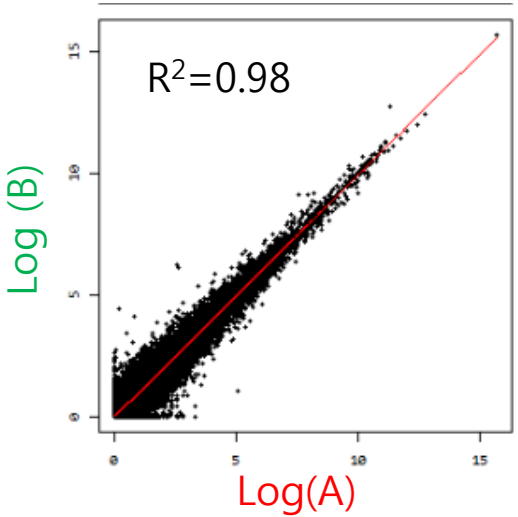
Pre-processing

Raw data (probe intensity) – filtering – Logarithm - Normalization



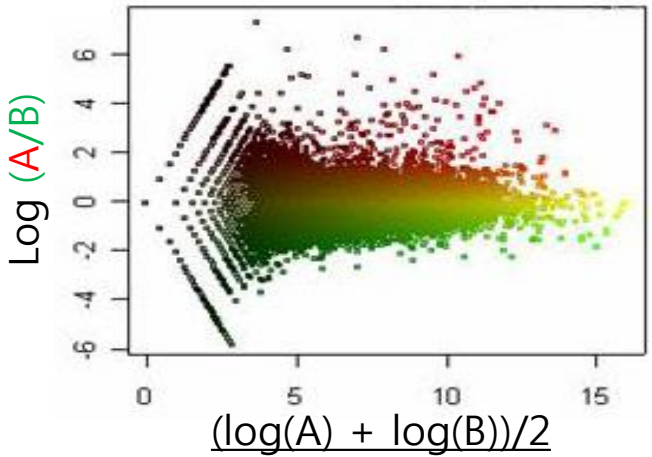
Box Plot

Scatter Plot



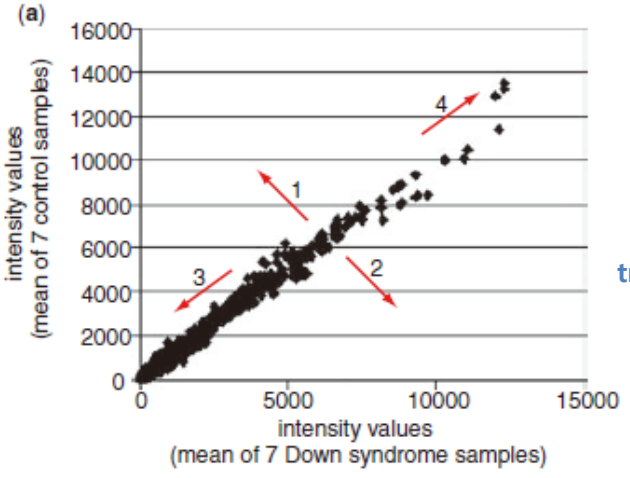
Reproducibility

MA plot



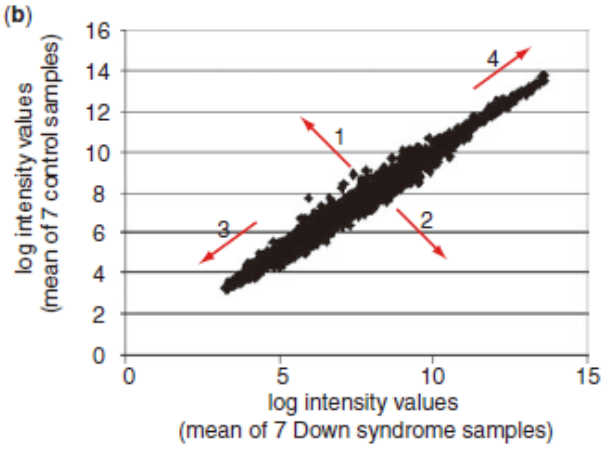
Bias to probe intensity

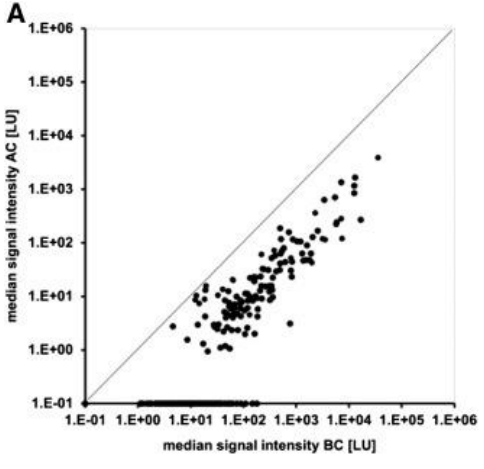
Filtering, Log transformation, & normalization



Filtering

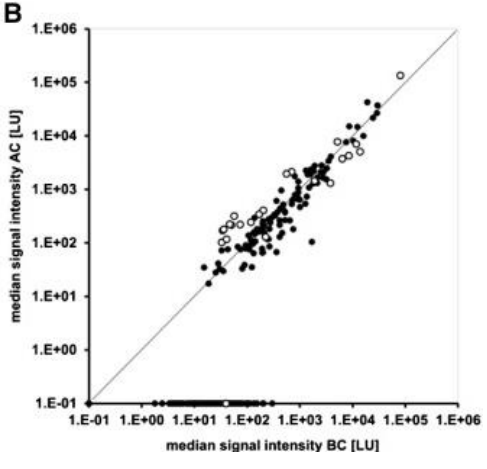
Log transformation





Median normalization

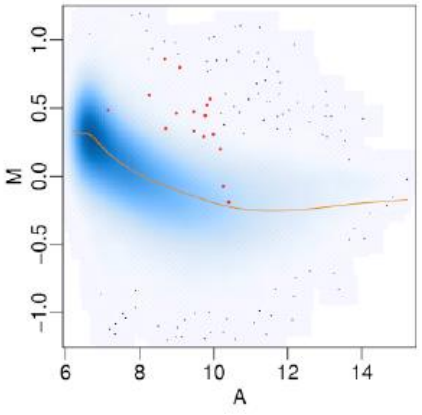
Global normalization

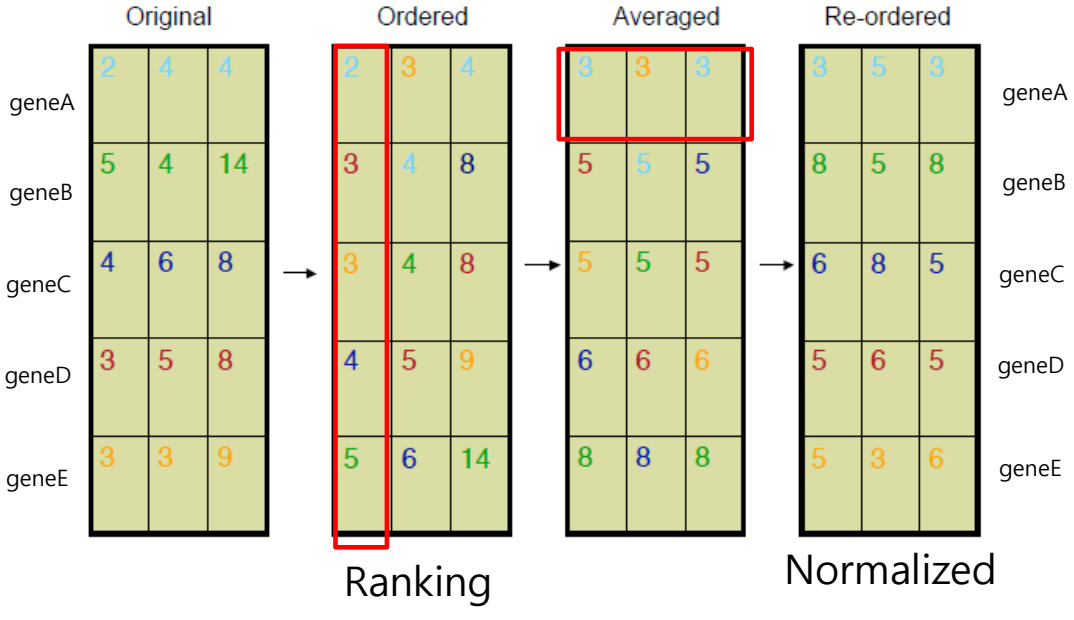


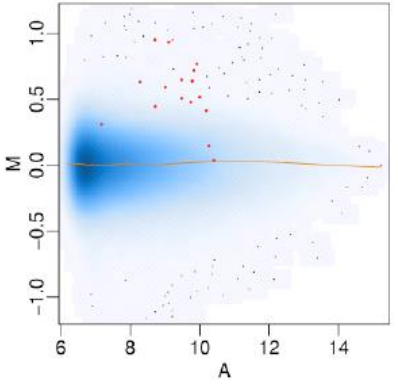
Rank quantile normalization

Rank quantile normalization

- All these non-linear methods perform similarly
- Basic idea:
  - order value in each array
  - take average across probes
  - Substitute probe intensity with average
  - Put in original order

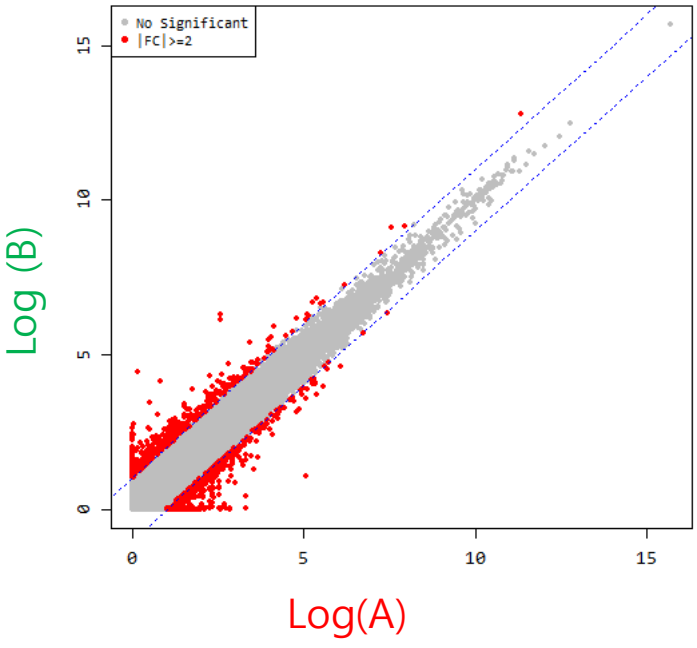




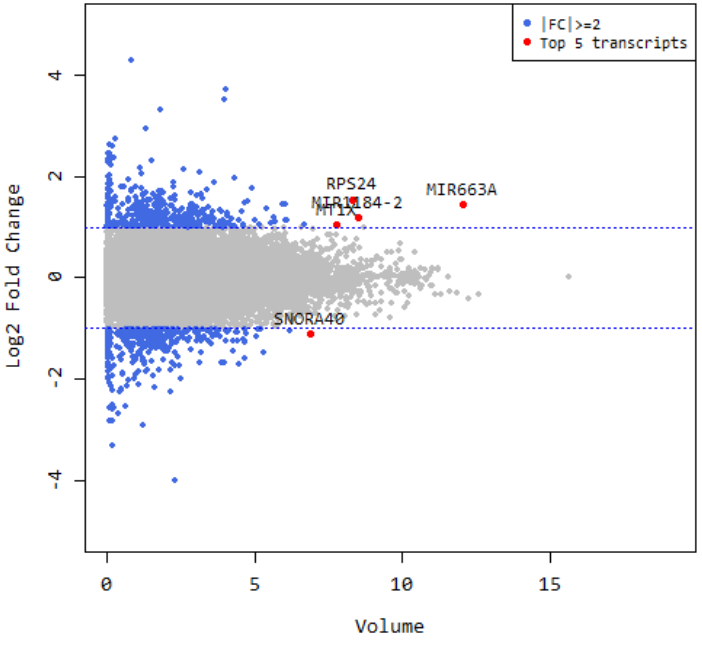


Differential expression analysis

Using cutoff (fold change)

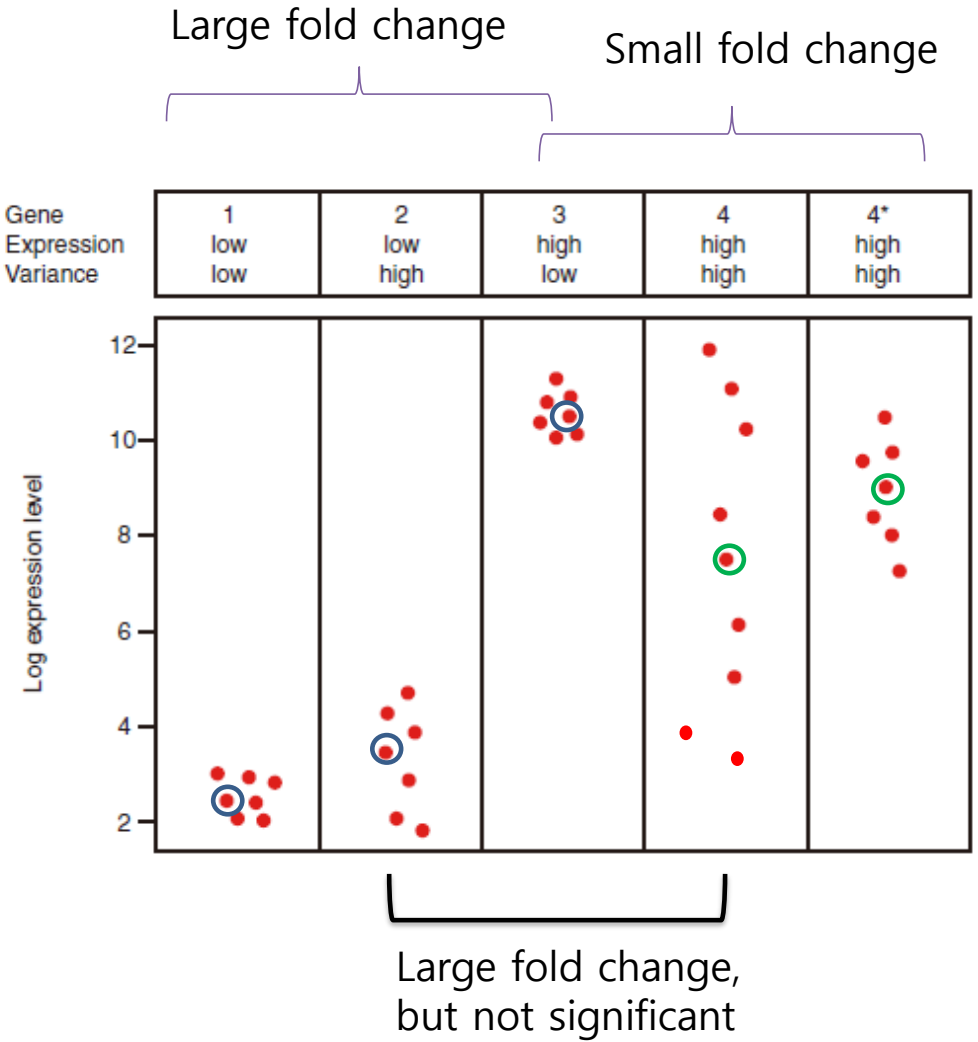


|Fold Change| >=2



|Fold Change| >=2

Problem of using fold change for DEG



3 vs 4

3 vs 4*

Similar fold change

But

3 vs 4* is more significant

Statistical test is required

Combining p-values and fold changes

- What is important biologically? : How significant is the difference?
  : How large is the difference?
- Both amount can be used to identify genes
- What cutoffs to use?
- How many genes should be selected?
- Where are your positive controls?

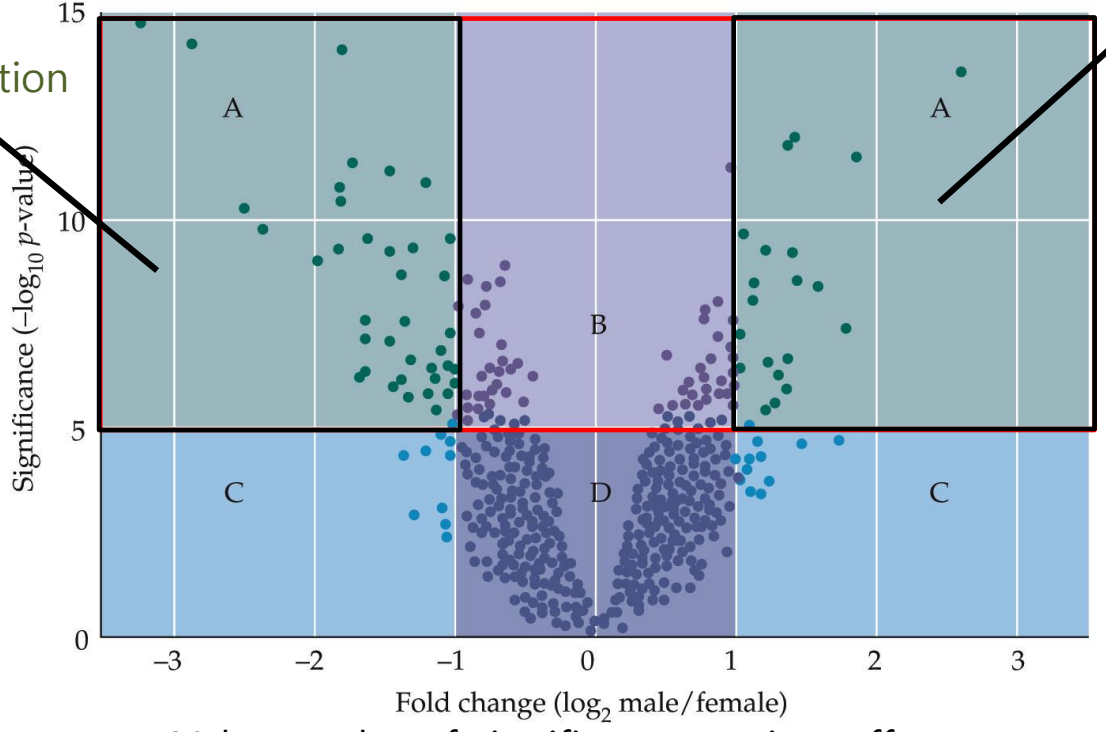
>2 fold downregulation

P<0.00001

>2 fold upregulation

P<0.00001

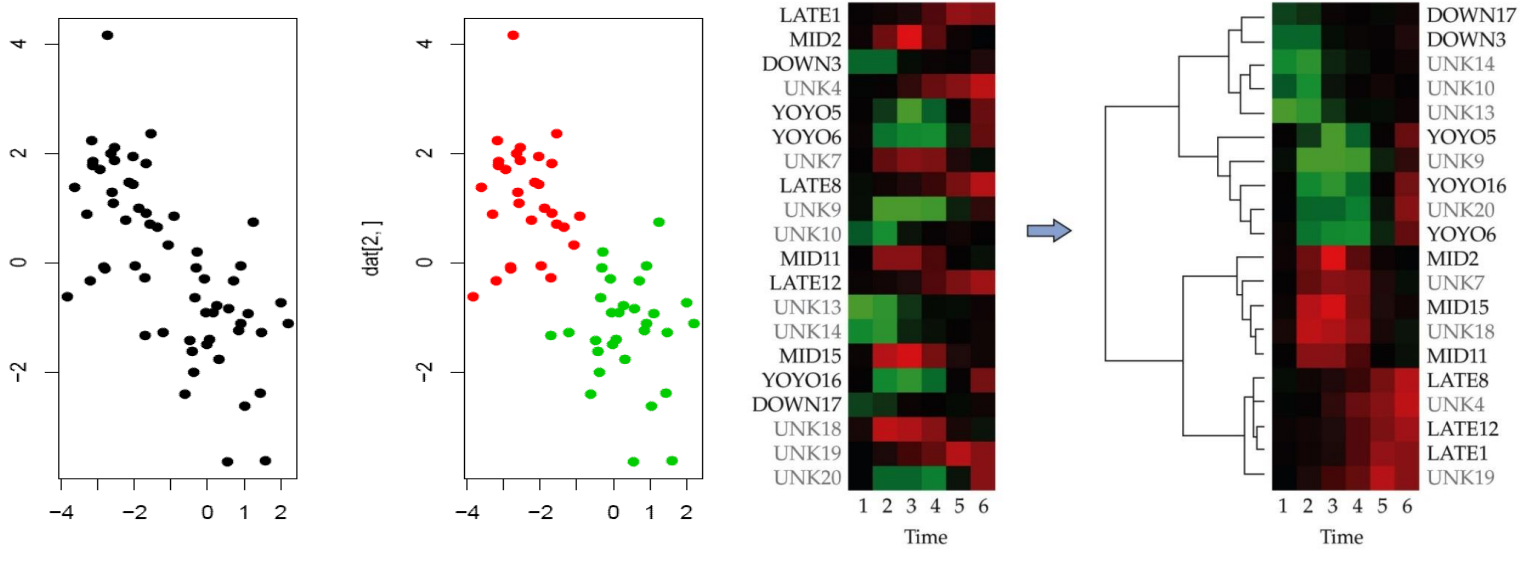
P<0.00001



Volcano plot of significance against effect

Clustering analysis

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in so me sense or another) to each other than to those in other groups (clusters).



- Unsupervised Analysis
- Supervised Analysis: classification rules

1. **Distance measurement**
2. **Linkage method**
3. **Clustering method**

Clustering analysis: Distance & Linkage

Distance a measure of similarity between genes.

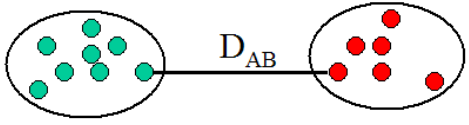
	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Gene A	x_{1A}	x_{2A}	x_{3A}	x_{4A}	x_{5A}	x_{6A}
Gene B	x_{1B}	x_{2B}	x_{3B}	x_{4B}	x_{5B}	x_{6B}

Some distances: (MeV provides 11 metrics)

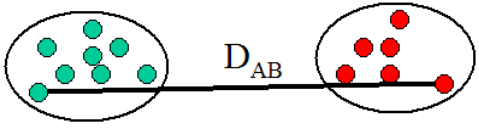
1. Euclidean: $\sqrt{\sum_{i=1}^{6}(x_{iA} - x_{iB})^2}$
2. Manhattan: $\sum_{i=1}^{6}|x_{iA} - x_{iB}|$
3. Pearson correlation

Linkage

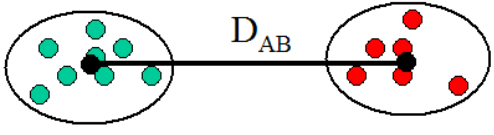
Single-linkage (minimum)



Complete-linkage (maximum)



Average-linkage



Linkage method

Hierarchical clustering : serially making clusters with genes with minimum distances

An example of a hierarchical clustering using single linkage algorithm.

	T1	T2	T3	T4	T5
Gene A	10.2	13.6	15.1	14.9	12.1
Gene B	13.7	13.5	12.9	13.3	13.8
Gene C	8.6	11.2	12.7	12.8	12.5
Gene D	13.3	13.3	11.1	9.6	9.6
Gene E	12.2	14.5	11.1	11.2	13.7

Hierarchical Clustering

Manhattan: $\sum_{i=1}^{6} |x_{iA} - x_{iB}|$

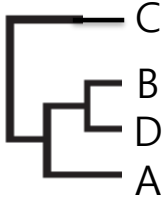
Distance matrix

	B	C	D
A	2.3	4.8	4.3
B		2.5	2
C			3.7

|13.6-13.5|+|15.1-12.9| = 2.3

Single-linkage (minimum)

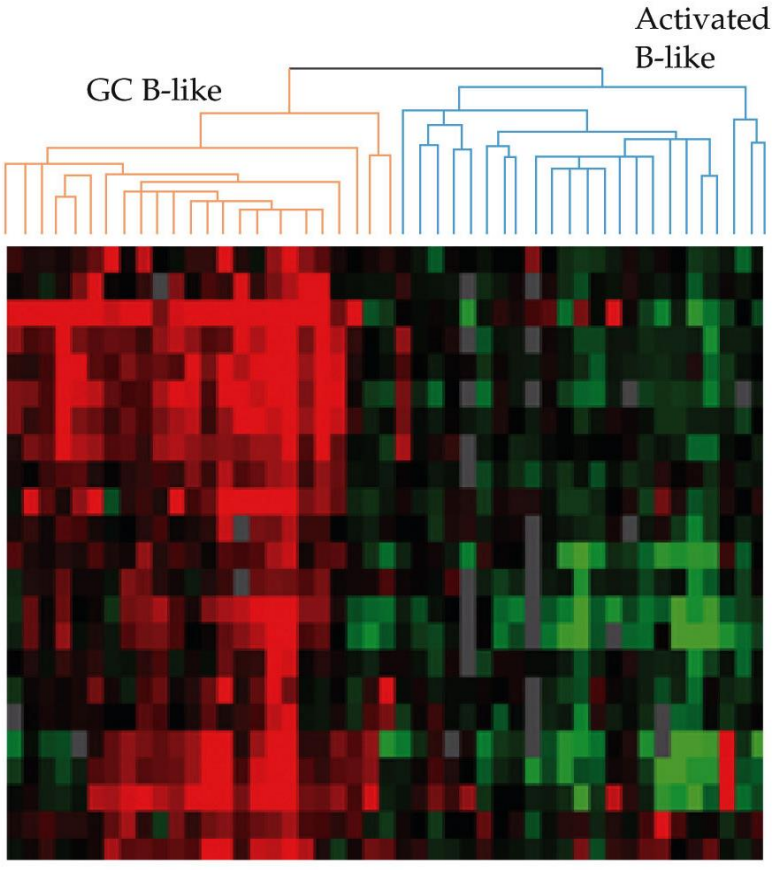
	B-D	C
A	2.3	4.8
C	2.5	



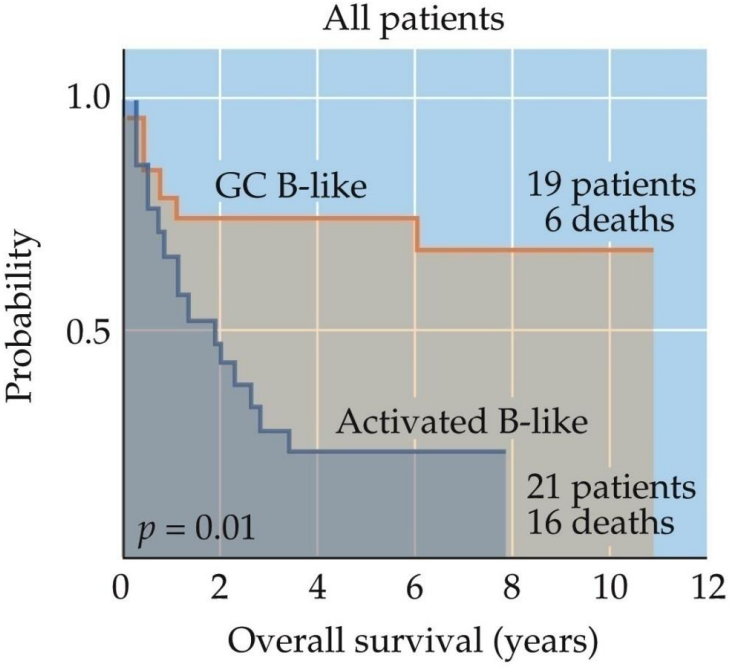
	T2	T3
C	11.2	12.7
B	13.5	12.9
D	13.3	11.1
A	13.6	15.1

Clustering analysis : Molecular pharmacology of cancers

Clustering of distinct cancer types, such as **diffuse large B-cell lymphomas (DLBCL),** uncovers the existence of novel molecular subtypes (**GC B-like** and **Activated B-like**) that may be predictive of survival probability



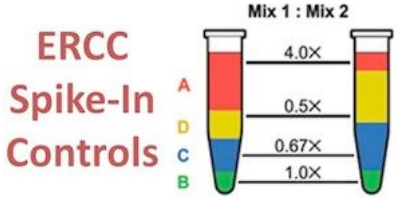
Survival curve

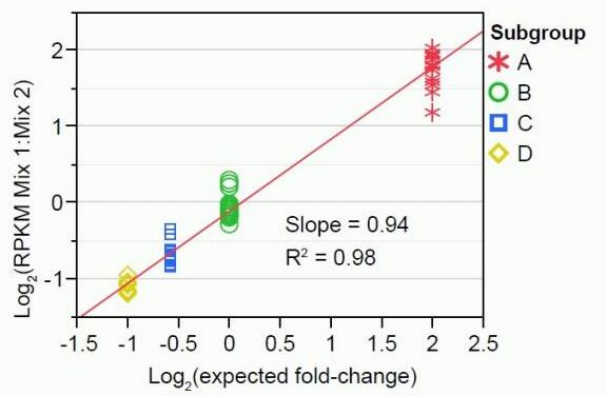


Kaplan-Meier plots

Issue of normalization

- Which genes to use for normalization
  - **Housekeeping genes**
  - **Spiked controls**.





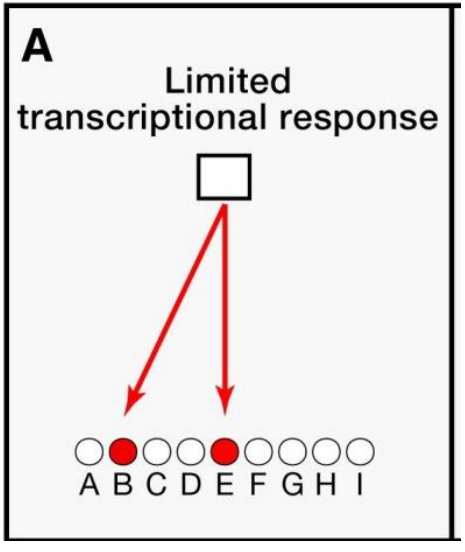
- **Using all genes**
  - Simplest approach – use all adequately expressed genes for normalization. The assumption is that the majority of genes on the array are housekeeping genes and the proportion of over expressed genes is similar to that of the under expressed genes.

Not changed

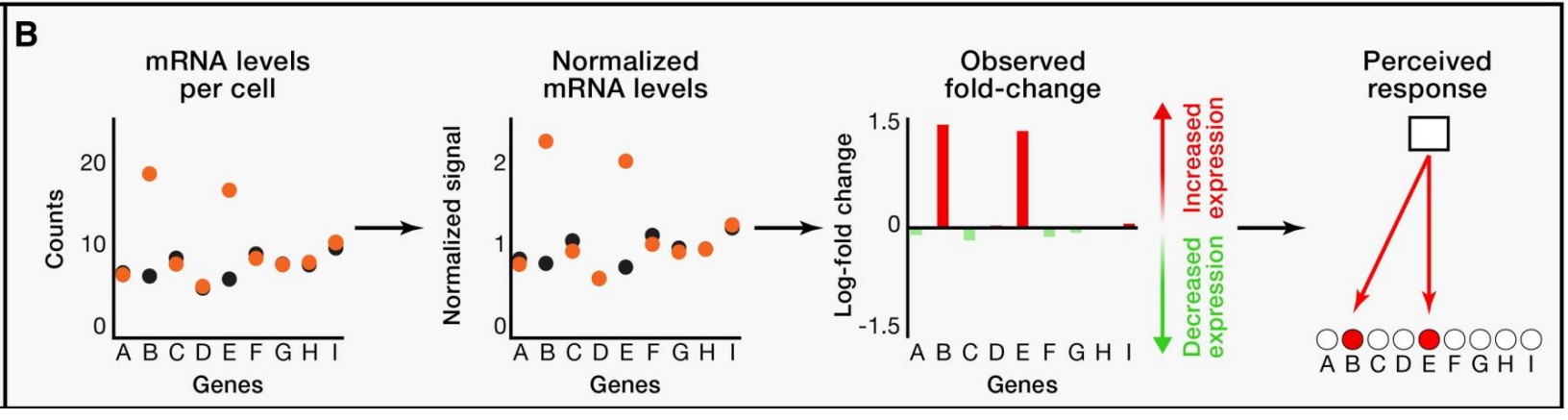
  - If the genes one the chip are specially selected, then this method will not work.

Normalization and Interpretation of Expression Data

Schematic representation of pattern of change in gene expression when levels of total RNA in the two cells are similar

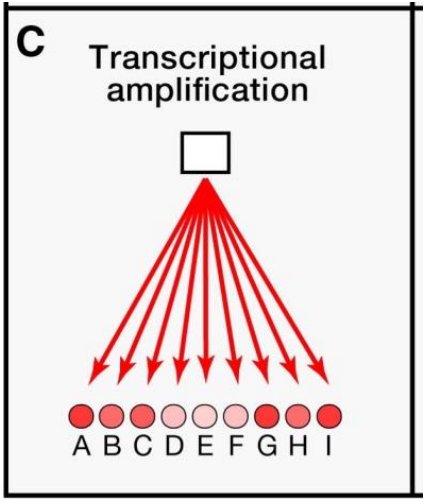


microarray normalization when the overall levels of mRNA per cell are not changing in two conditions



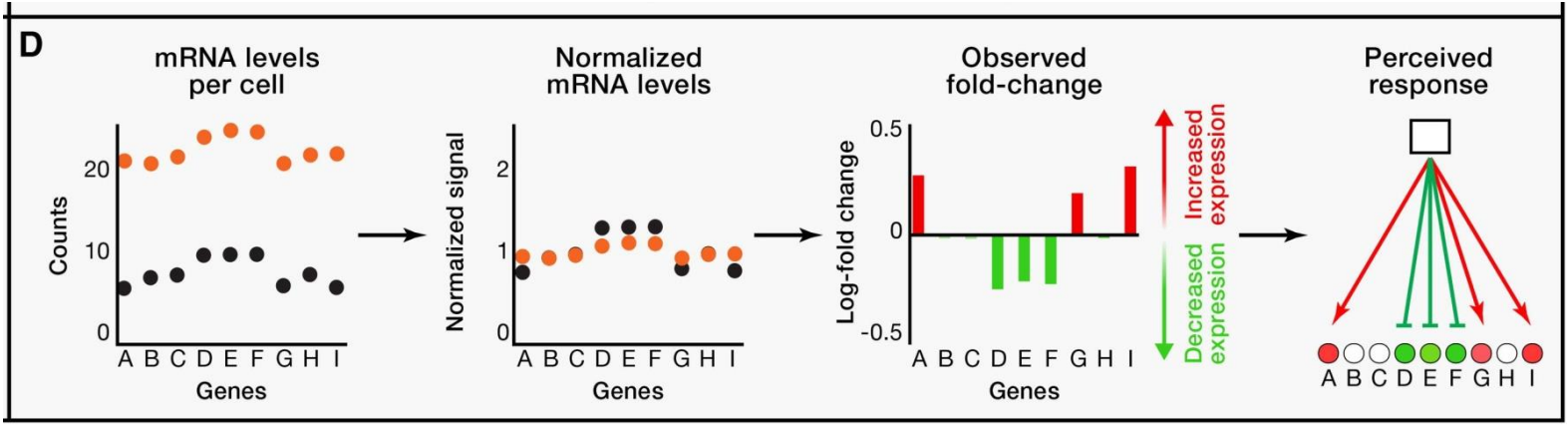
Normalization and Interpretation of Expression Data

Schematic representation of pattern of change in gene expression when levels of total RNA in the two cells is different such as in transcriptional amplification, where most genes are expressed at higher levels.

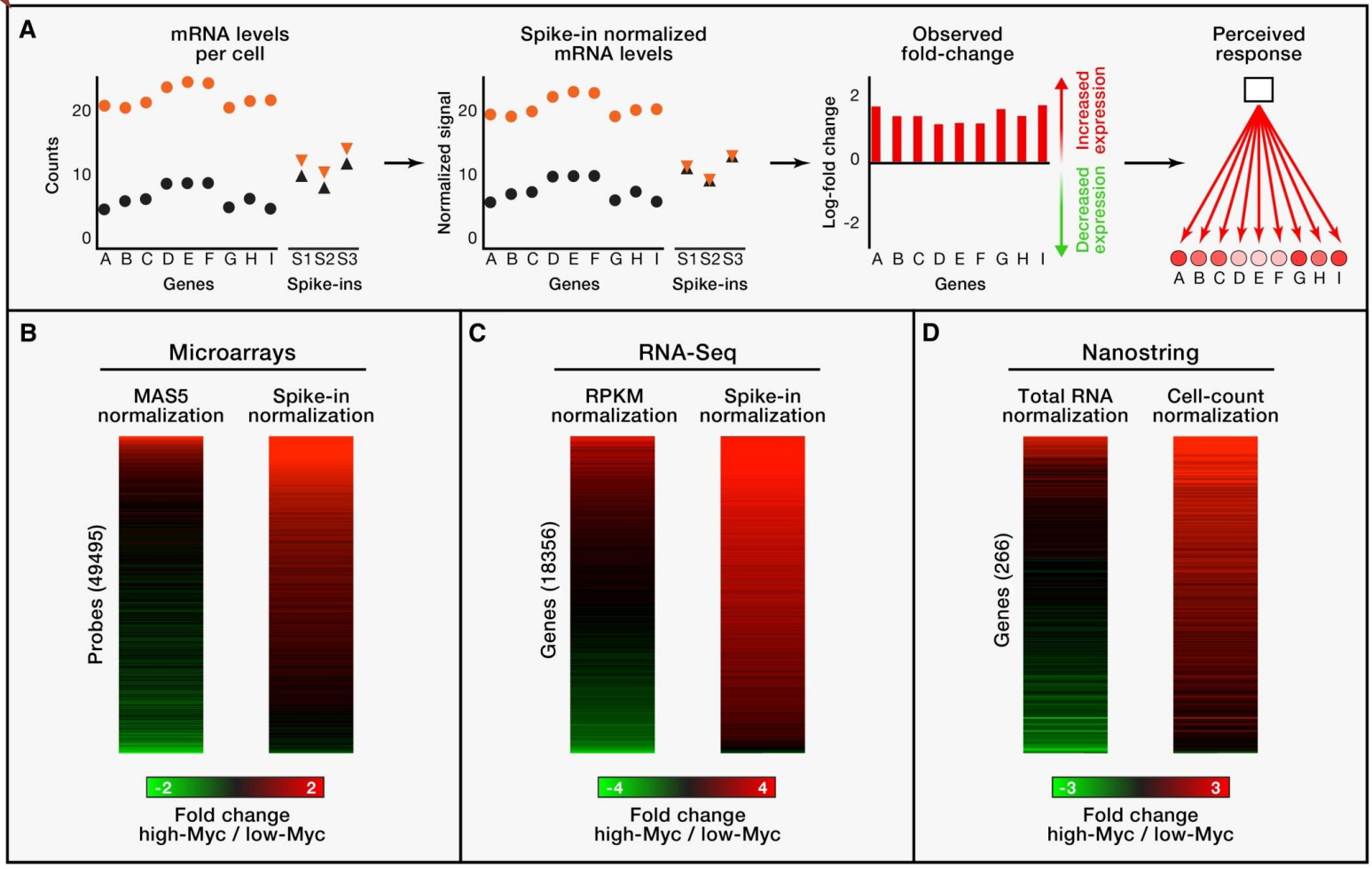


"Transcriptional amplification"
: *high level of c-Myc produces two to three times more total RNA.*

microarray normalization when the overall levels of mRNA per cell are increased in one condition compared to another



Spike-In Controls, Normalized to Cell Number, Enable Accurate Interpretation of Transcriptional Changes



Cell 2012 151, 476-482DOI: (10.1016/j.cell.2012.10.012)

"Spike-in" or "Cell-count" normalization !

Normalization issue

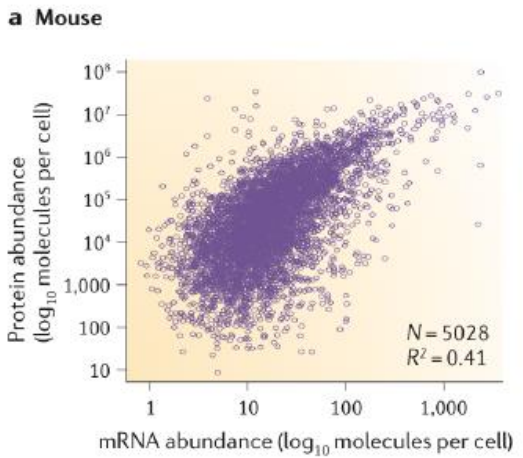
- global gene expression analysis (microarray, RNA-sequencing, and digital quantification) detect a widespread increase in transcripts/cells in cells that experience transcriptional amplification (c-Myc) .

-fail to detect the widespread increase of transcription when inappropriate normalization methods (**conventional global gene normalization**) are used

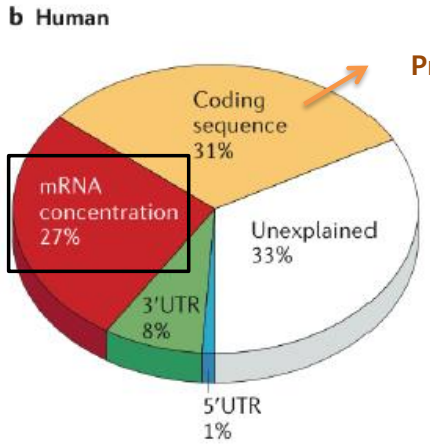
- Instead, they erroneously suggest the interpretation that a similar number of genes show increases and decreases in expression.

Correlation of protein and mRNA levels

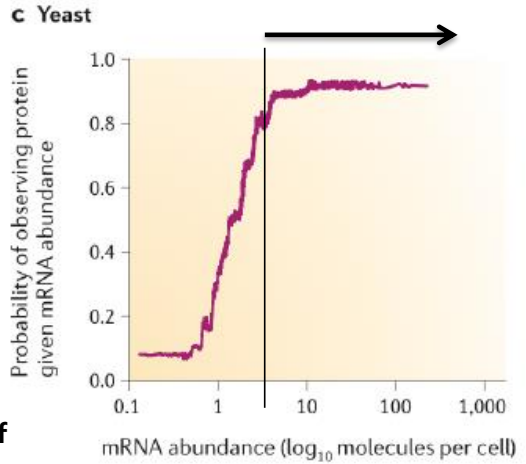
protein



mRNA

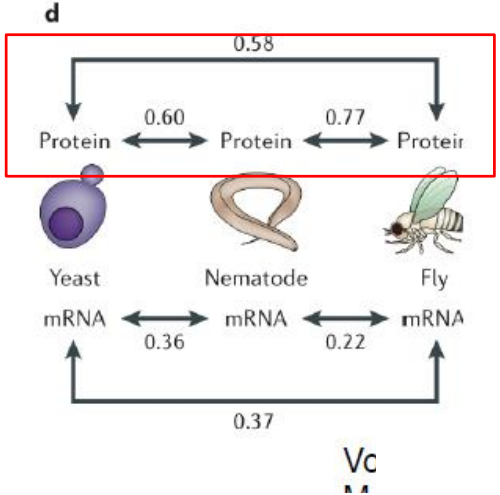


protein



mRNA

(Limitation of proteomics)

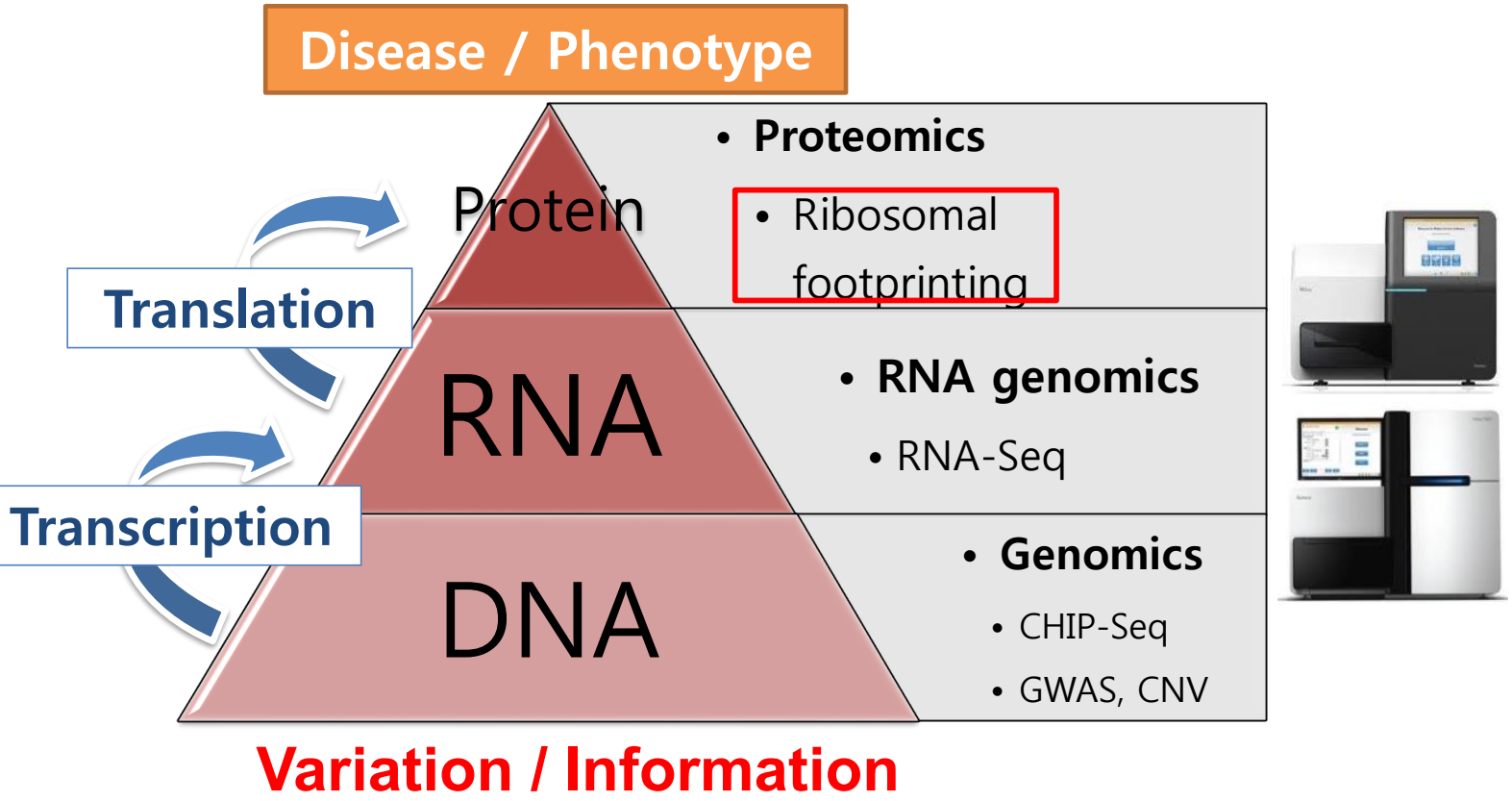


Abundance of proteins are more conserved

Vogel, C. & Marcotte, E.M. *Nat Rev Genet* ***13,*** 227–232 (2012).

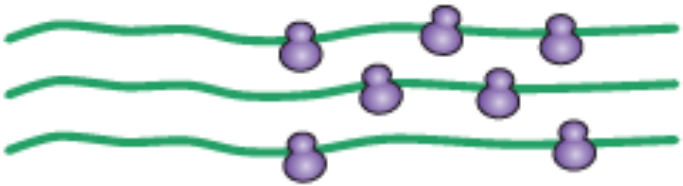
Functional Genomics

: A functional link in the post-genomic era

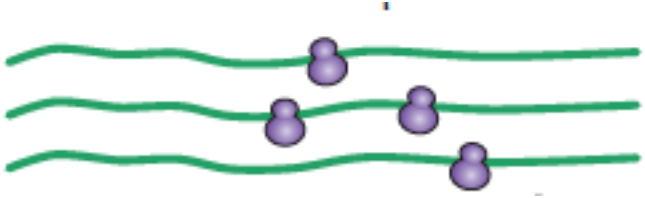


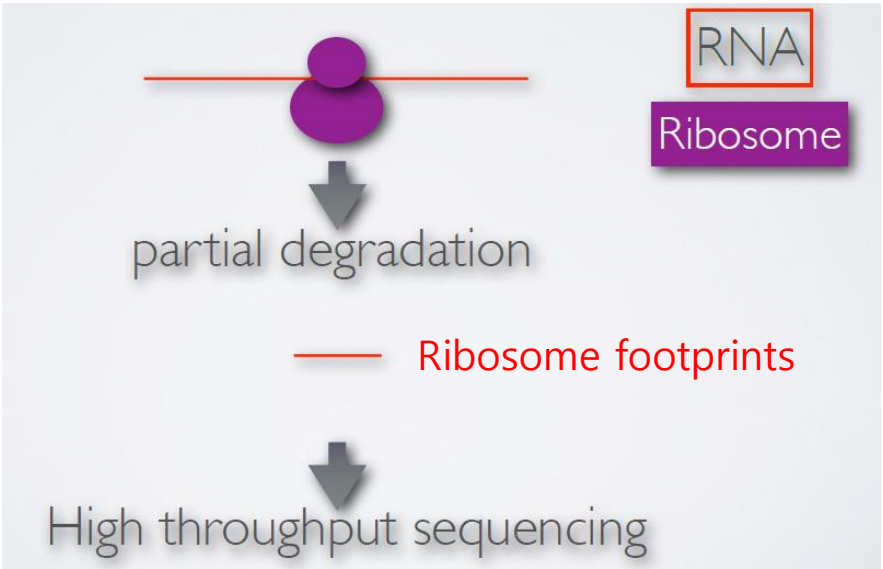
Ribosome Footprinting

Active translation (Polysome)



Less translation (monosome)



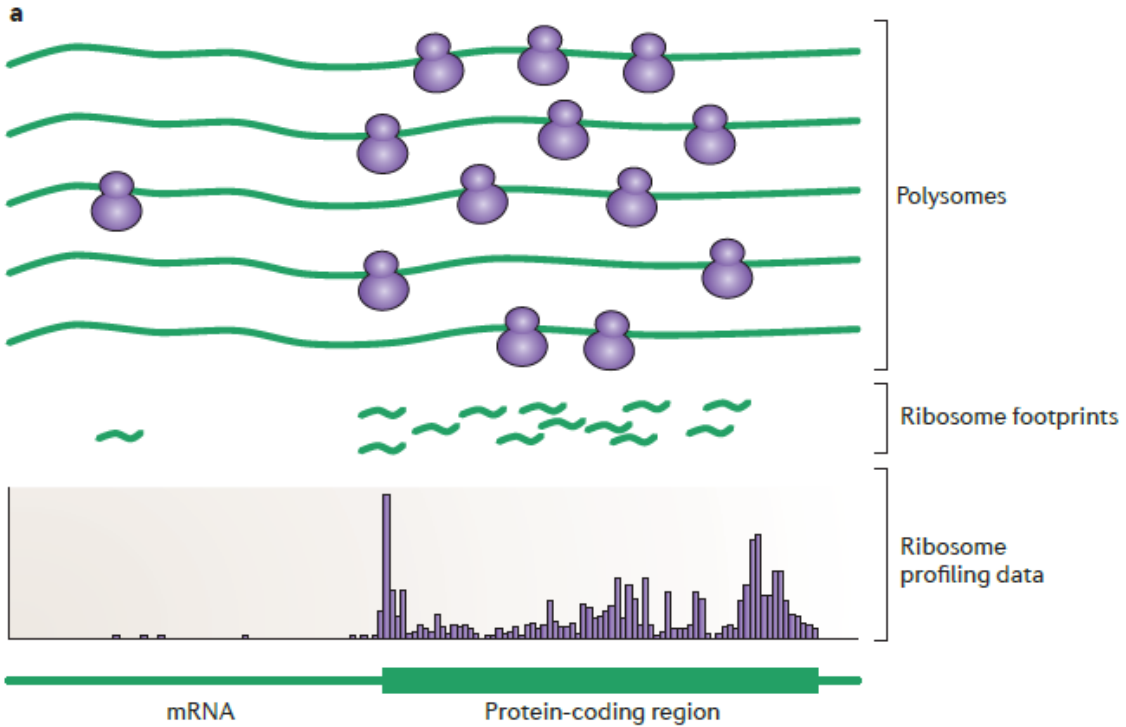


- "Freeze" ribosomes in the RNA with cycloheximide
- Digest mRNA-protein complex (mRNP) with Rnase
- Isolate ribosomes with the protected mRNA fragment (Centrifugation)
- NGS

Ribo-Seq (Ribosome profiling)

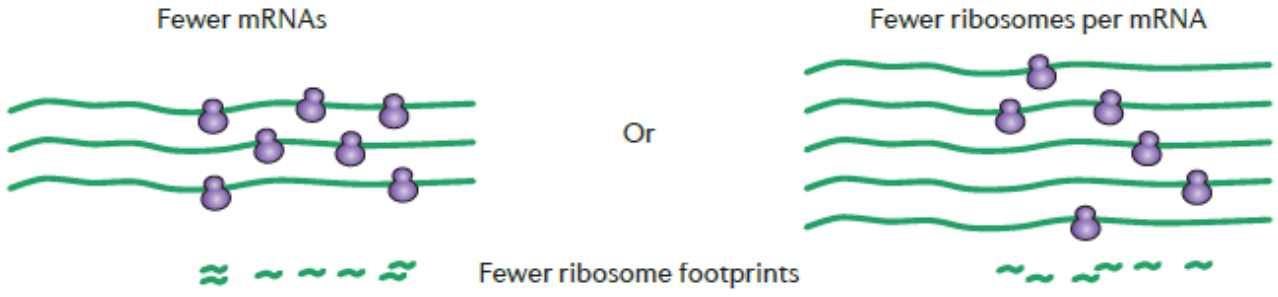
Science, 2009

Ribo-Seq: Analysis of ribosome occupancy data.

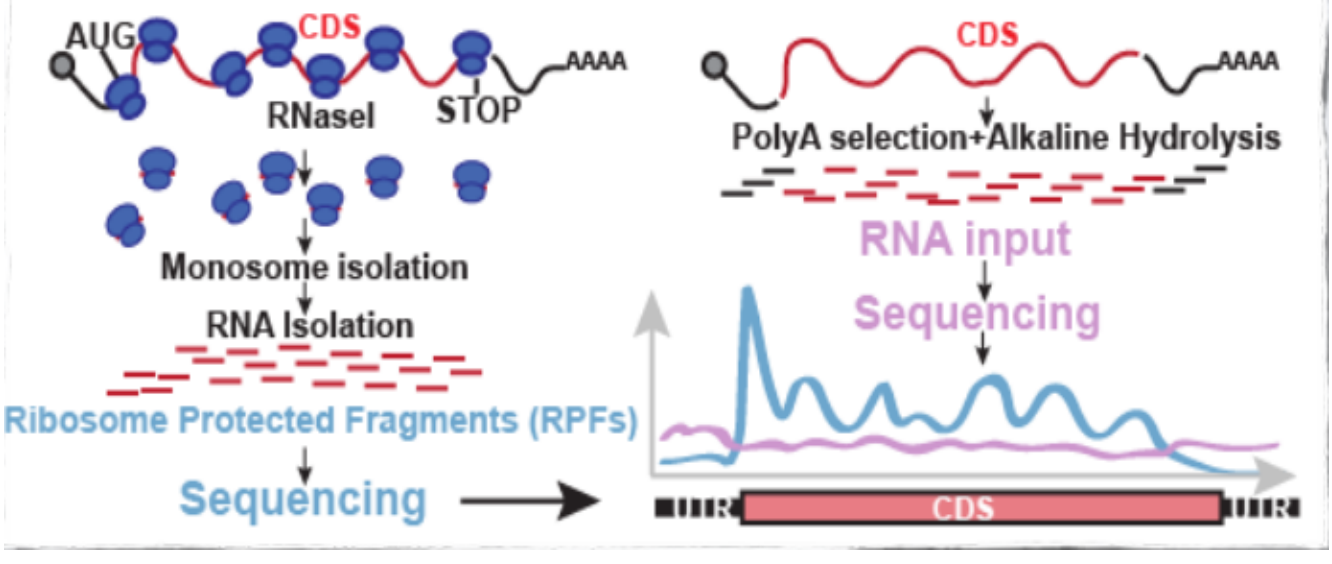


-Genome annotation: AUG, STOP, micro-peptides, non-coding genes
-Measure translational regulation
-Genome-wide visualization of translation

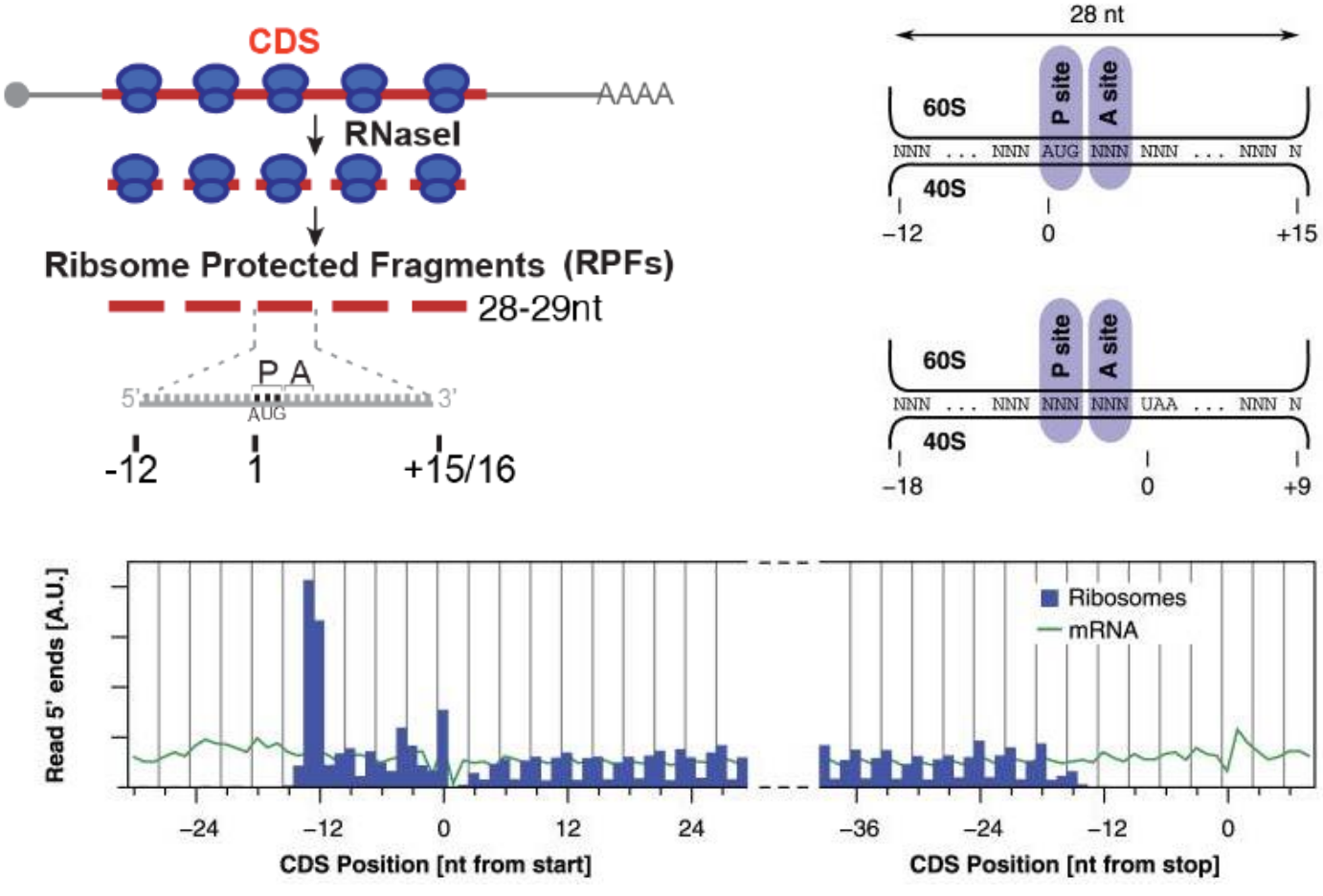
Ribo-Seq analysis: together with RNA-Seq



Should be performed with RNA-Seq



Ribo-Seq: Subcodon resolution of translation



-Ribosome protected fragments map to the Coding sequence
-Input RNA map to the whole transcript

Ribo-Seq data analysis

Ribosome pausing