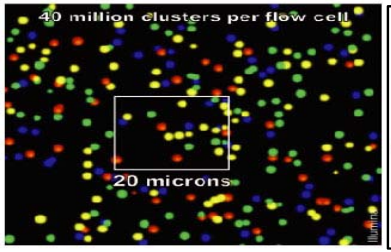
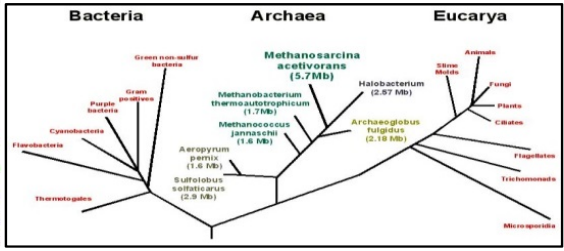
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



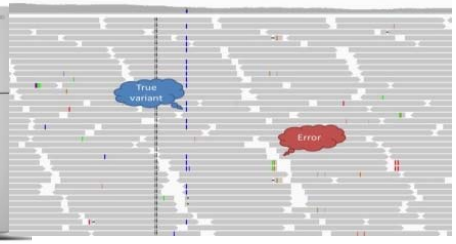


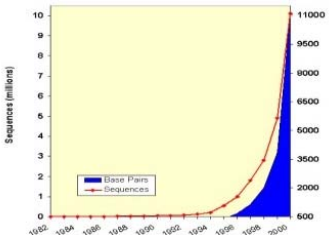
Transcriptome Analysis with noncoding RNAs

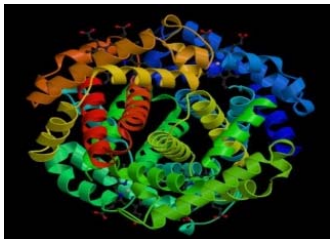
北京大学生物信息学中心 高歌

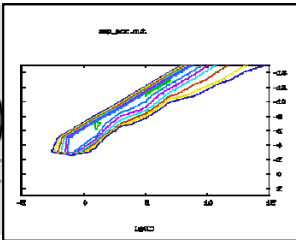
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

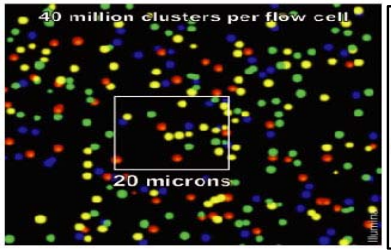


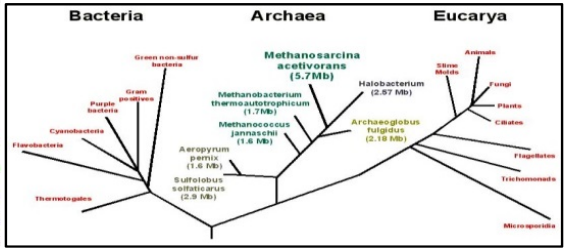






TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



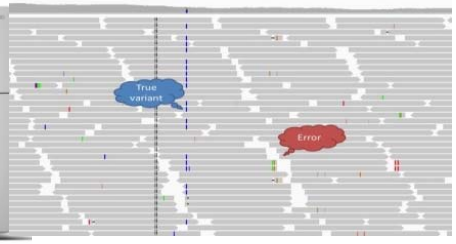


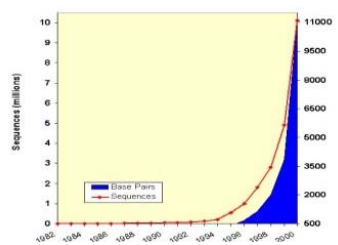
Unit 3: Data Mining: Differential Expression and Clustering

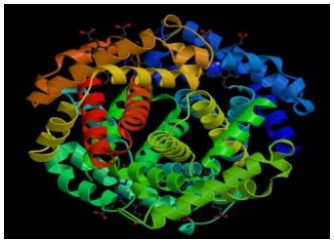
北京大学生物信息学中心 高歌

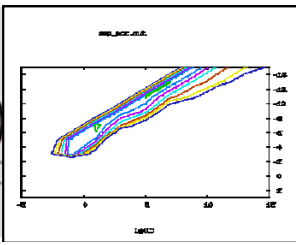
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University









Additional Information

How many non-coding transcripts?

What are the functional roles of those ncRNAs?

microRNA (miRNA)

- single-stranded RNAs of 21-23 (or some say 20-25) bp RNAs with regulatory fu
- In or bl im

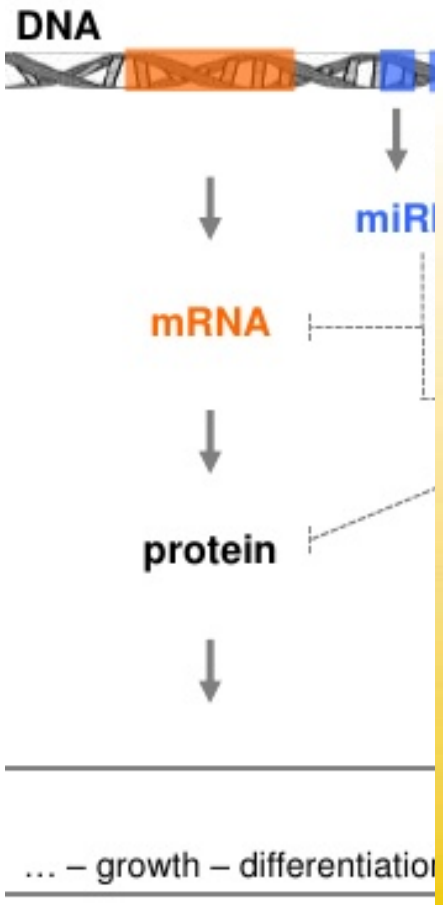
Table 4.2 Computational algorithms for microRNA target prediction

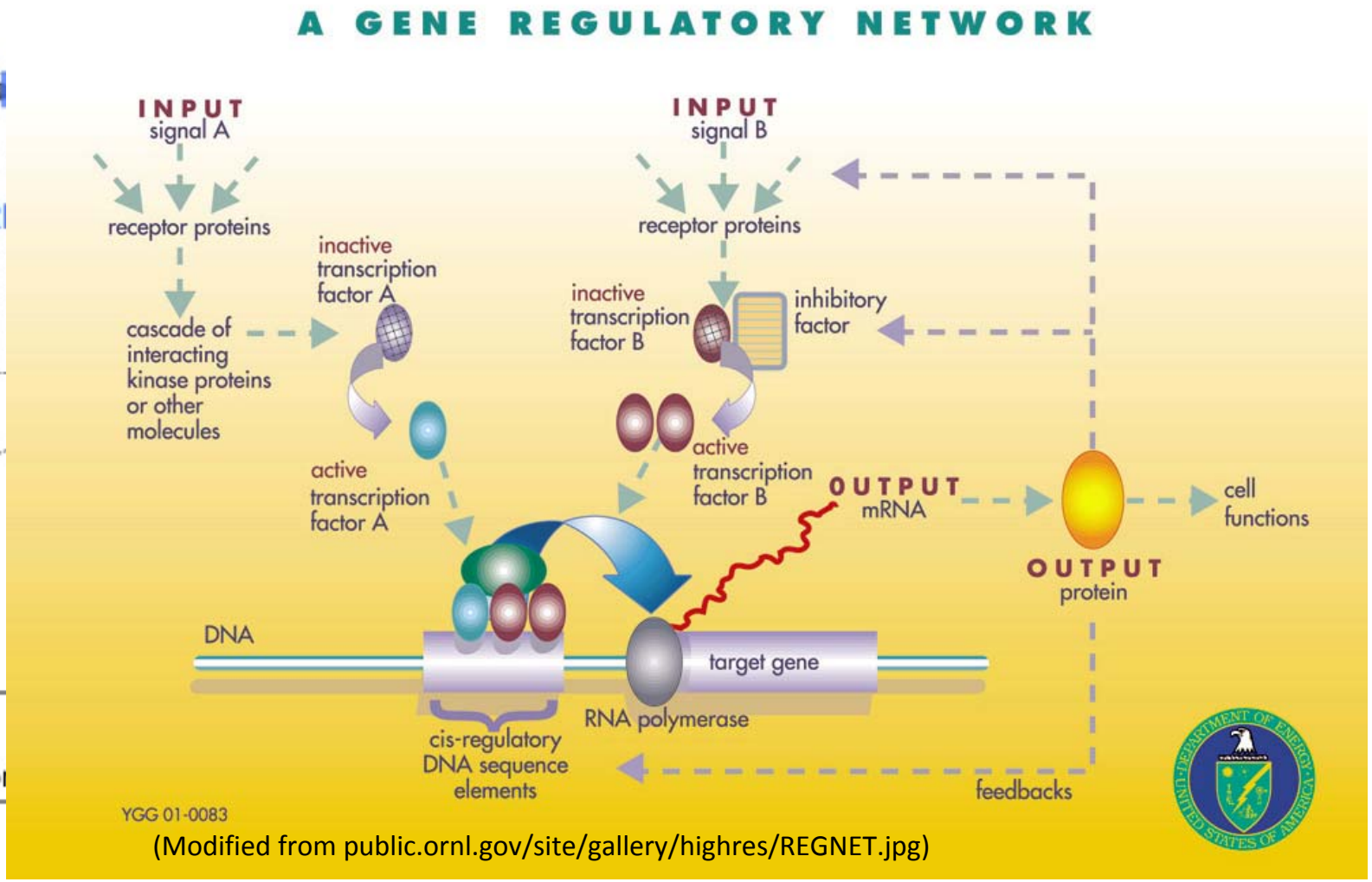
Name of the software	URL or availability	Supported organism(s)	Reference(s)
TargetScan, TargetScanS	http://genes.mit.edu/targetscan/	Vertebrates	Lewis *et al.*, 2003, 2005
miRanda	http://www.microrna.org/	Flies, vertebrates	Enright *et al.*, 2003, John *et al.*, 2004
DIANA-microT	http://diana.pcbi.upenn.edu/DIANA-microT/	Vertebrates	Kiriakidou *et al.*, 2004
RNAhybrid	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	Flies	Rehmsmeier *et al.*, 2004
GUUGle	http://bibiserv.techfak.uni-bielefeld.de/guugle/	Flies	Gerlach *et al.*, 2006
PicTar	http://pictar.bio.nyu.edu/	Nematodes, flies, vertebrates	Grun *et al.*, 2005, Krek *et al.*, 2005, Lall *et al.*, 2006
MicroInspector	http://mirna.imbb.forth.gr/microinspector/	Any	Rusinov *et al.*, 2005
MovingTargets	Available by request on DVD	Flies	Burgler *et al.*, 2005
FastCompare	http://tavazoielab.princeton.edu/mirnas/	Nematodes, flies	Chan *et al.*, 2005
miRU	http://bioinfo3.noble.org/miRNA/miRU.htm	Plants	Zhang 2005
TargetBoost	https://demo1.interagon.com/demo/	Nematodes, flies	Saetrom *et al.*, 2006
rna22	http://cbcsrv.watson.ibm.com/rna22.html	Nematodes, flies, vertebrates	Miranda *et al.*, 2006
miTarget	http://cbit.snu.ac.kr/~miTarget/	Any	Kim *et al.*, 2006

(Source: *Methods Enzymol.* 427:65)

(Source: *Cell* 116:281)

- the transcriptome





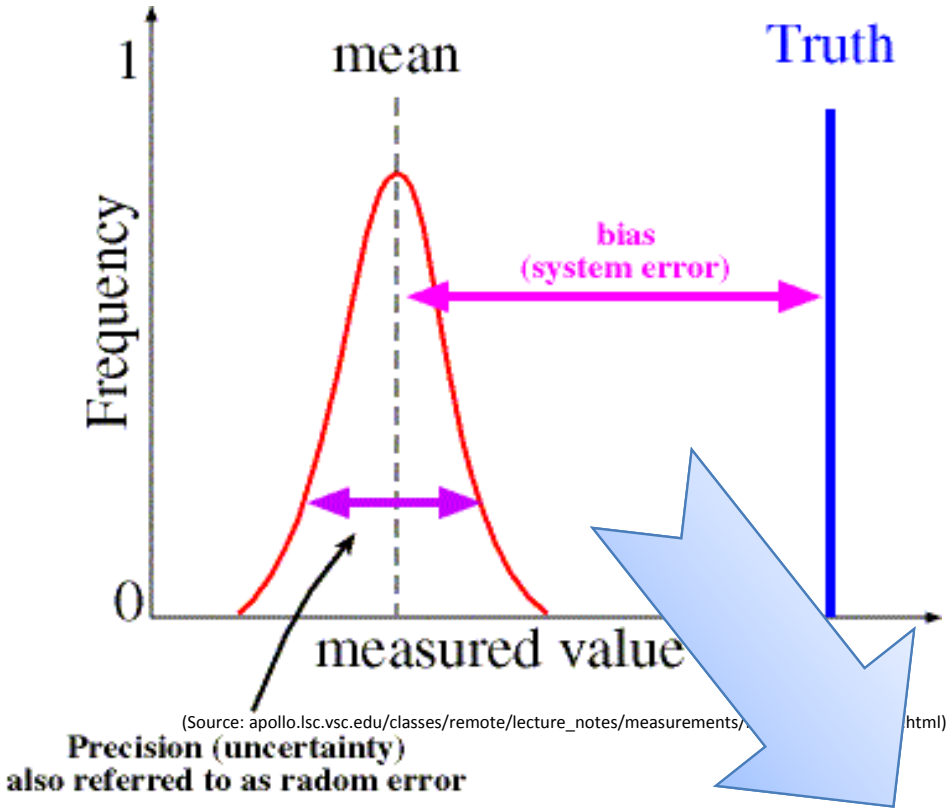
(Modified from public.ornl.gov/site/gallery/highres/REGNET.jpg)

- Differentially expressed genes
- Co-expressed genes

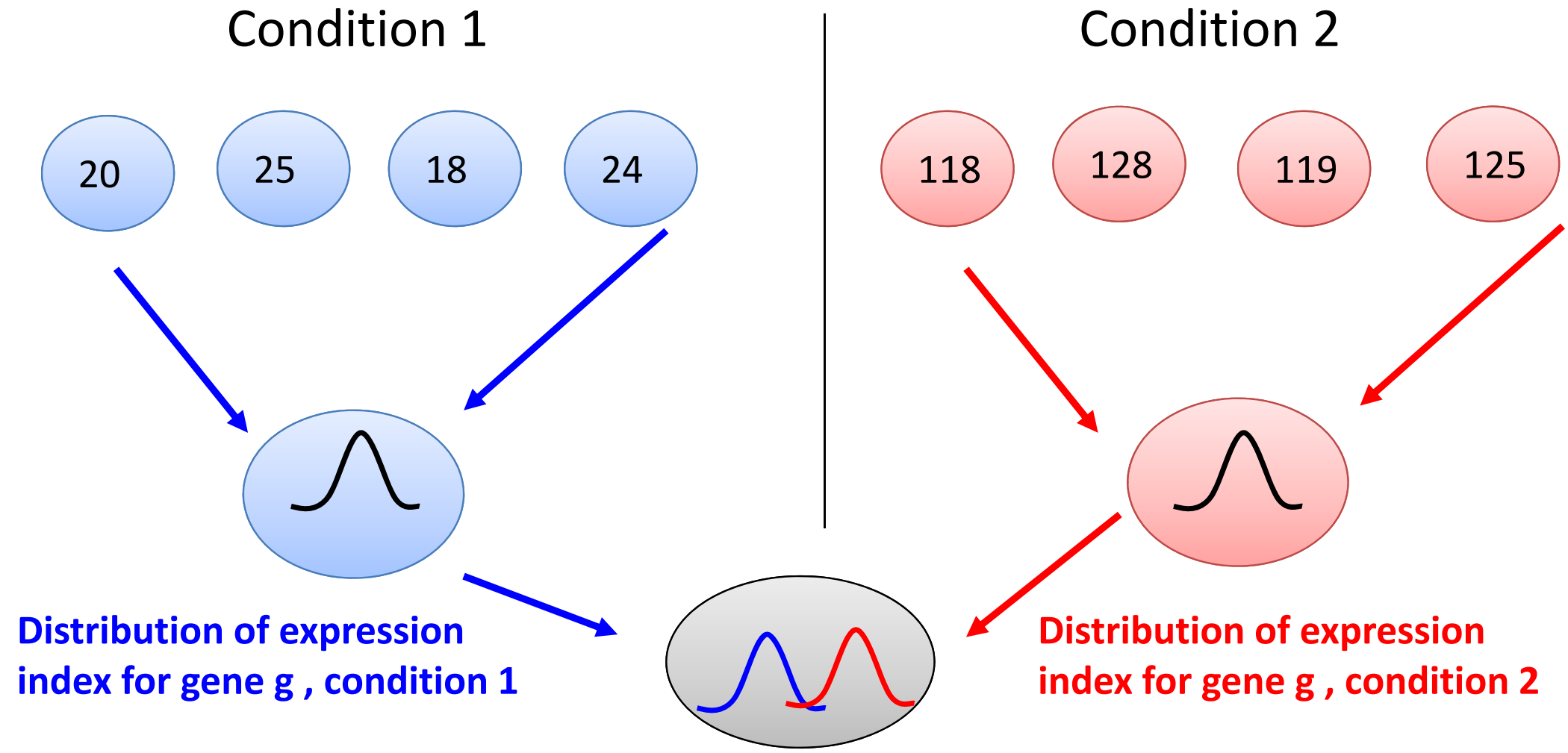
Data Mining: Differentially Expression Calling

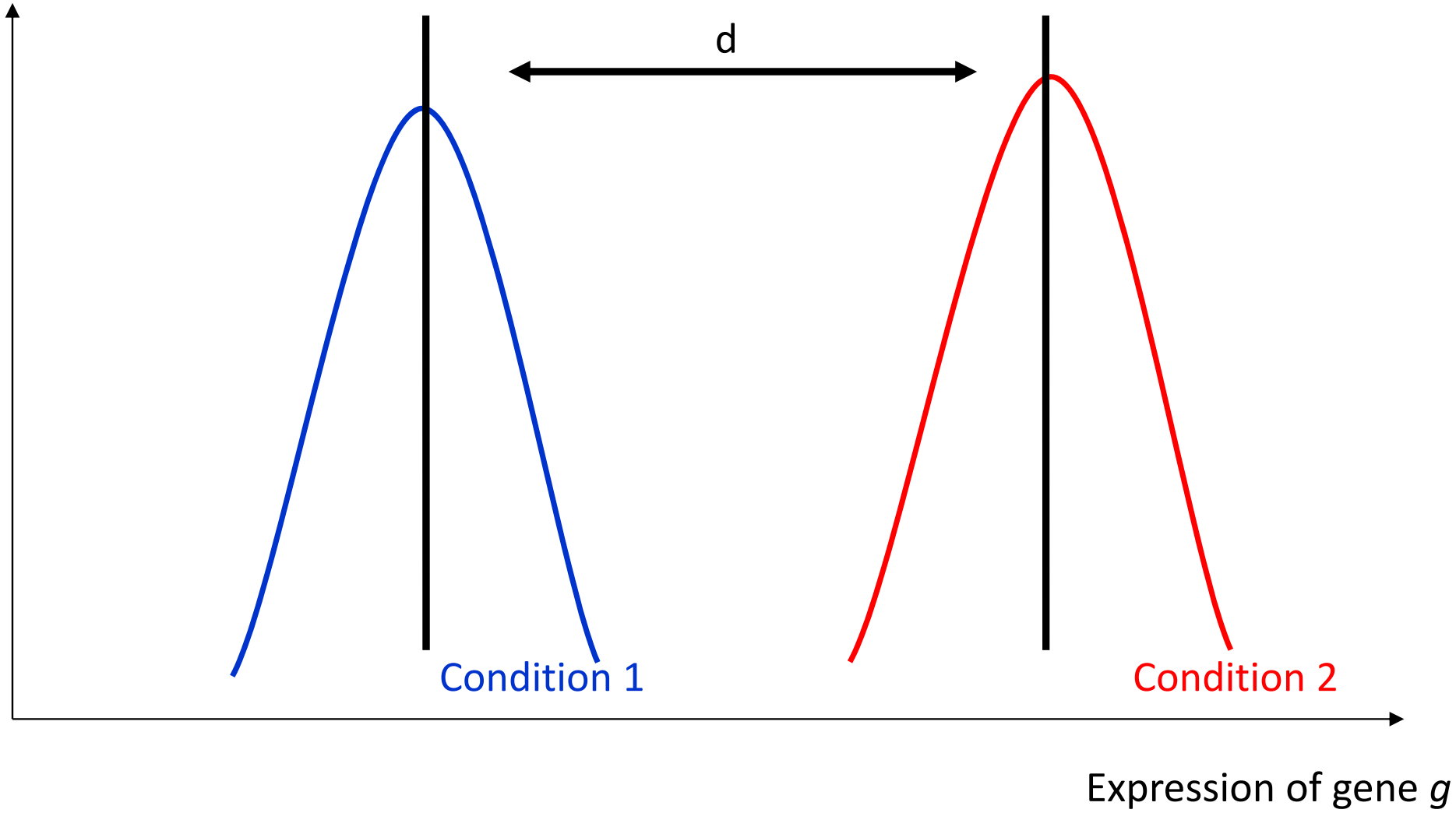
- Identify the genes with biological-significant difference in expression levels across samples

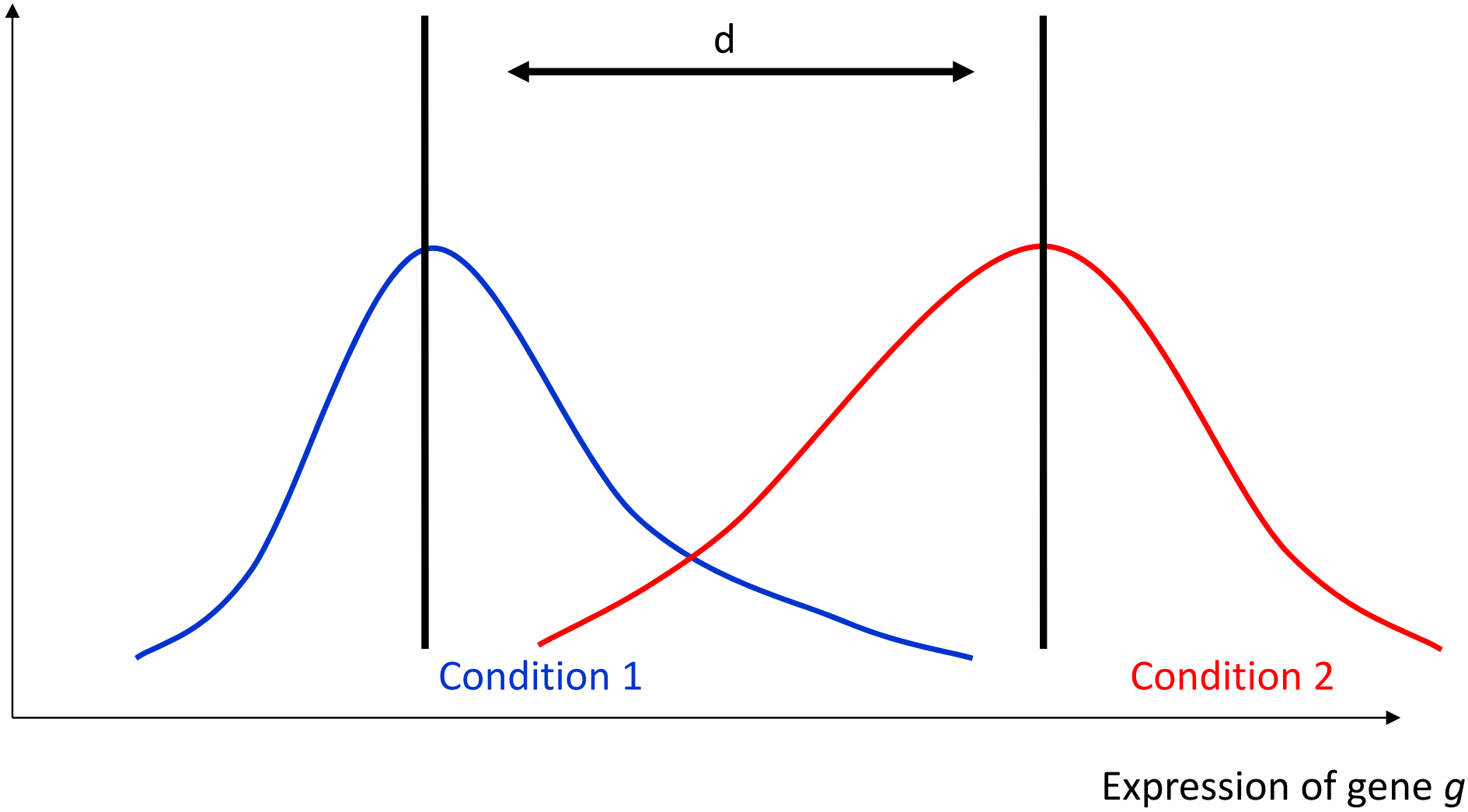
- Differences in expression values can result from many non-biological sources (e.g. experiment error/bias)
  - The 'real' differences are the differences that can NOT be explained by the various errors introduced during the experimental phase



- Random errors arise from random fluctuations in the measurements
- It could be reduced by repeating experiment many times (and get a mean value)
- Random errors could be modeled statistically by variance.







Additional Information

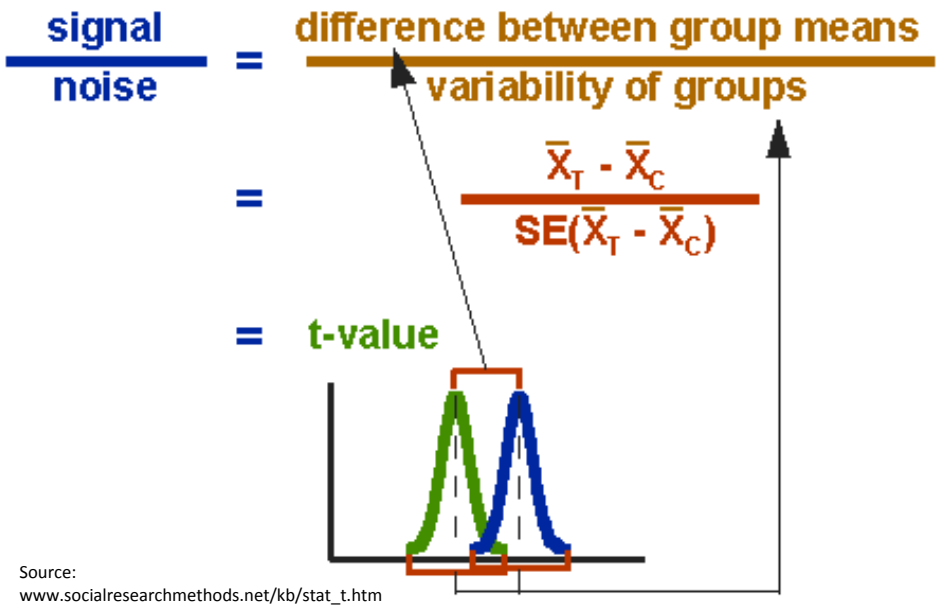
Statistical calling

1. Select a statistic which takes the variance into account, and will rank the genes in order of supporting strength for "differential expression".

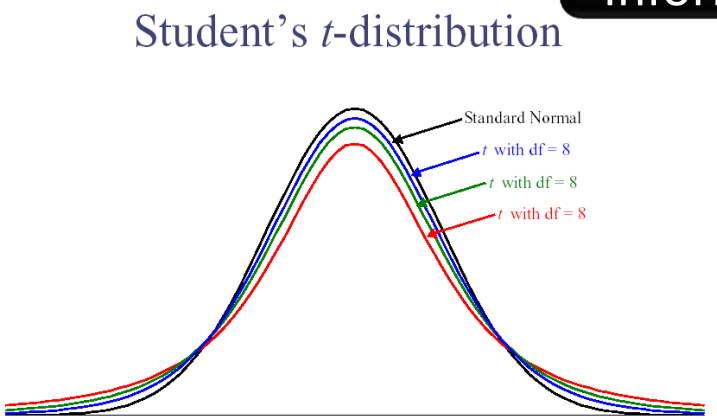
2. Derive the p-value for each gene, based on the NULL distribution of the statistic.

3. Choose a critical-value for the gene with p-value less than which being called as "being statistically significant".

Additional Information



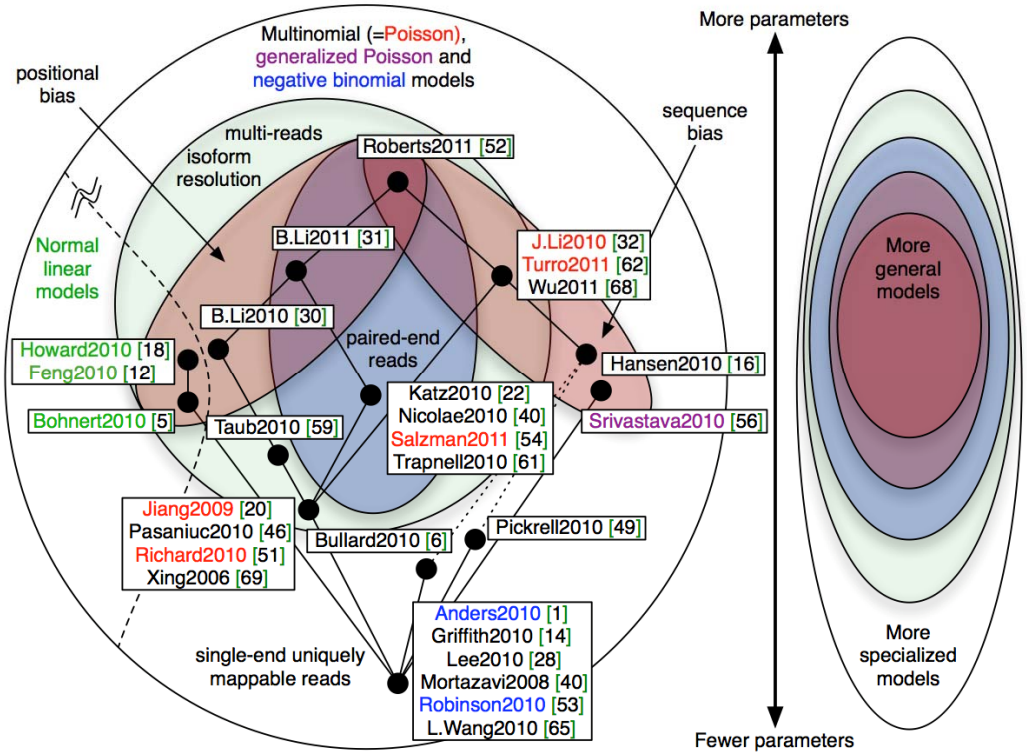
Source:
www.socialresearchmethods.net/kb/stat_t.htm



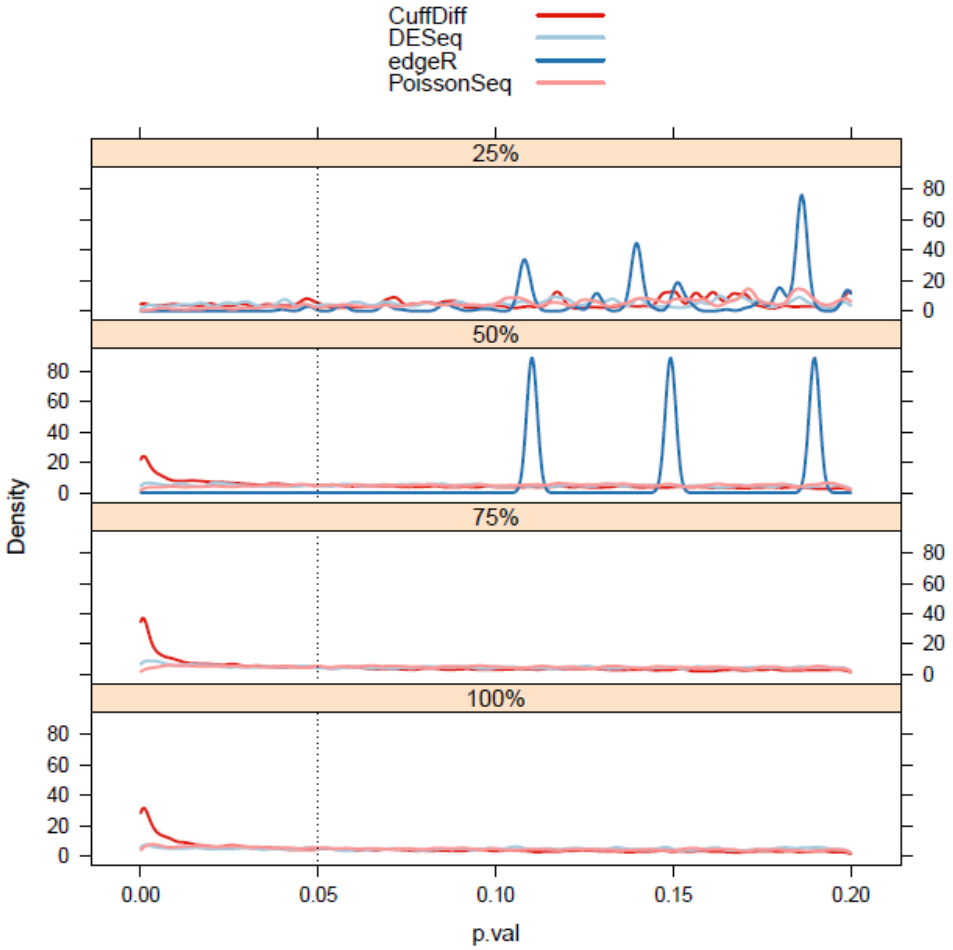
Source: projectile.sv.cmu.edu/research/public/talks/t-test.htm

- The t-test assesses whether the means of two groups are statistically different from each other
  - Take the variance into account through Standard Error (SE)
- Need to estimate the SE correctly
  - But the correct estimation depends on prior distribution (Normal) as well as the number of replicates (>10)

Model the data in RNA-Seq



Patcher 2011, arXiv:1104.3889 [q-bio.GN]



(*Genome Biology* 14:R95)

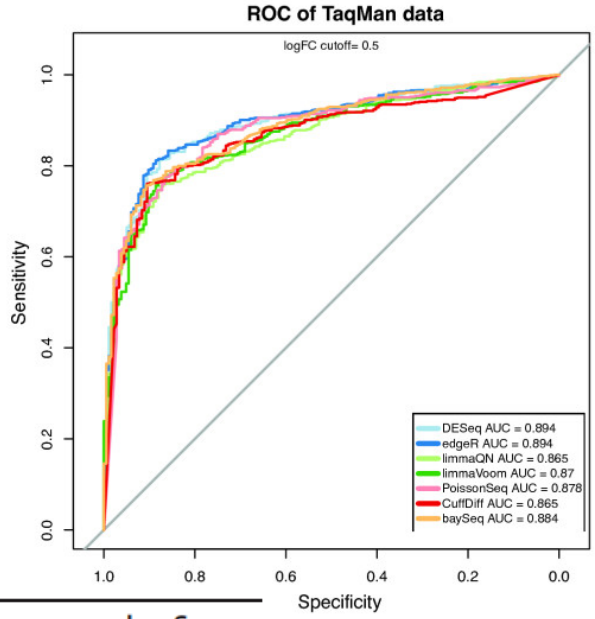
Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

Genome Biology

METHOD Open Access

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

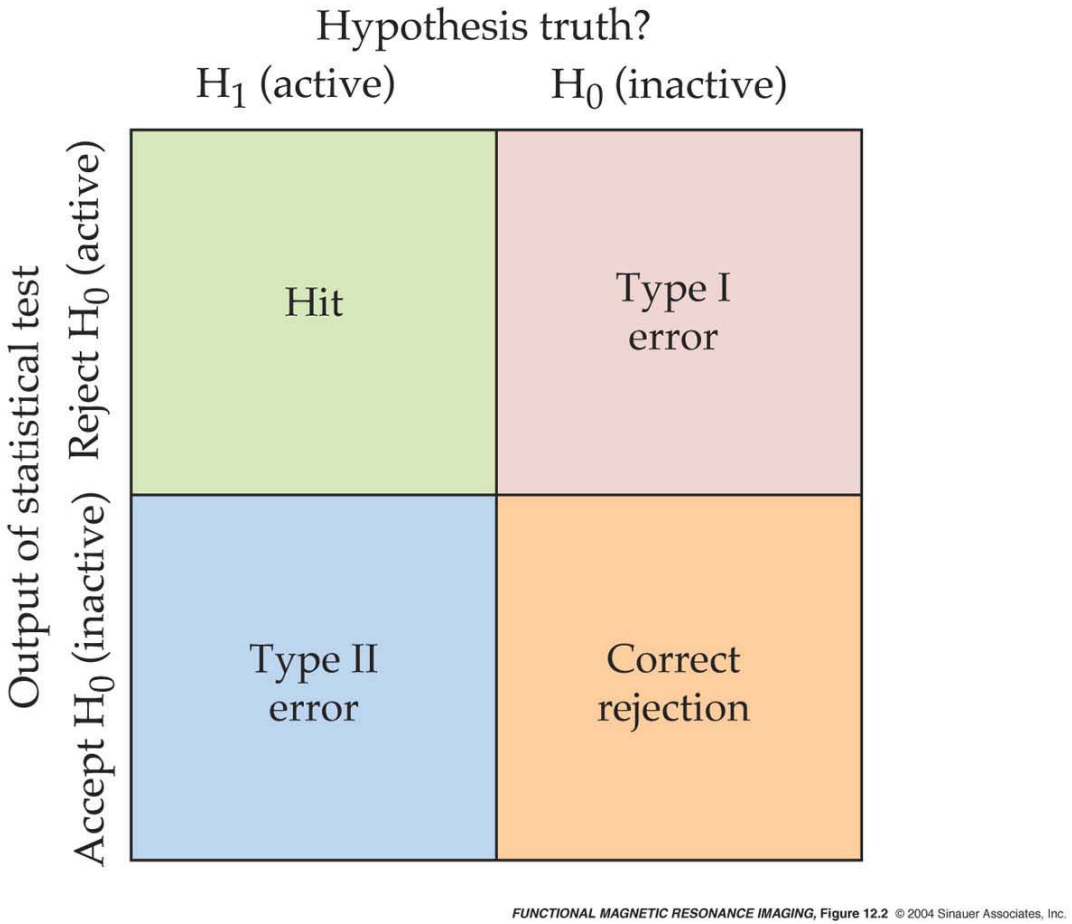
Franck Rapaport[1], Raya Khanin[1], Yupu Liang[1], Mono Pirun[1], Azra Krek[1], Paul Zumbo[2,3], Christopher E Mason[2,3], Nicholas D Socci[1] and Doron Betel[3,4*]



Evaluation	Cuffdiff	DESeq	edgeR	limmaVoom	PoissonSeq	baySeq
Normalization and clustering	All methods performed equally well					
DE detection accuracy measured by AUC at increasing qRT-PCR cutoff	Decreasing	Consistent	Consistent	Decreasing	Increases up to log expression change ≤ 2.0	Consistent
Null model type I error	High number of FPs	Low number of FPs	Low number of FPs	Low Number of FPs	Low number of FPs	Low number of FPs
Signal-to-noise vs *P* value correlation for genes detected in one condition	Poor	Poor	Poor	Good	Moderate	Good
Support for multi-factored experiments	No	Yes	Yes	Yes	No	No
Support DE detection without replicated samples	Yes	Yes	Yes	No	Yes	No
Detection of differential isoforms	Yes	No	No	No	No	No
Runtime for experiments with three to five replicates on a 12 dual-core 3.33 GHz, 100 G RAM server	Hours	Minutes	Minutes	Minutes	Seconds	Hours

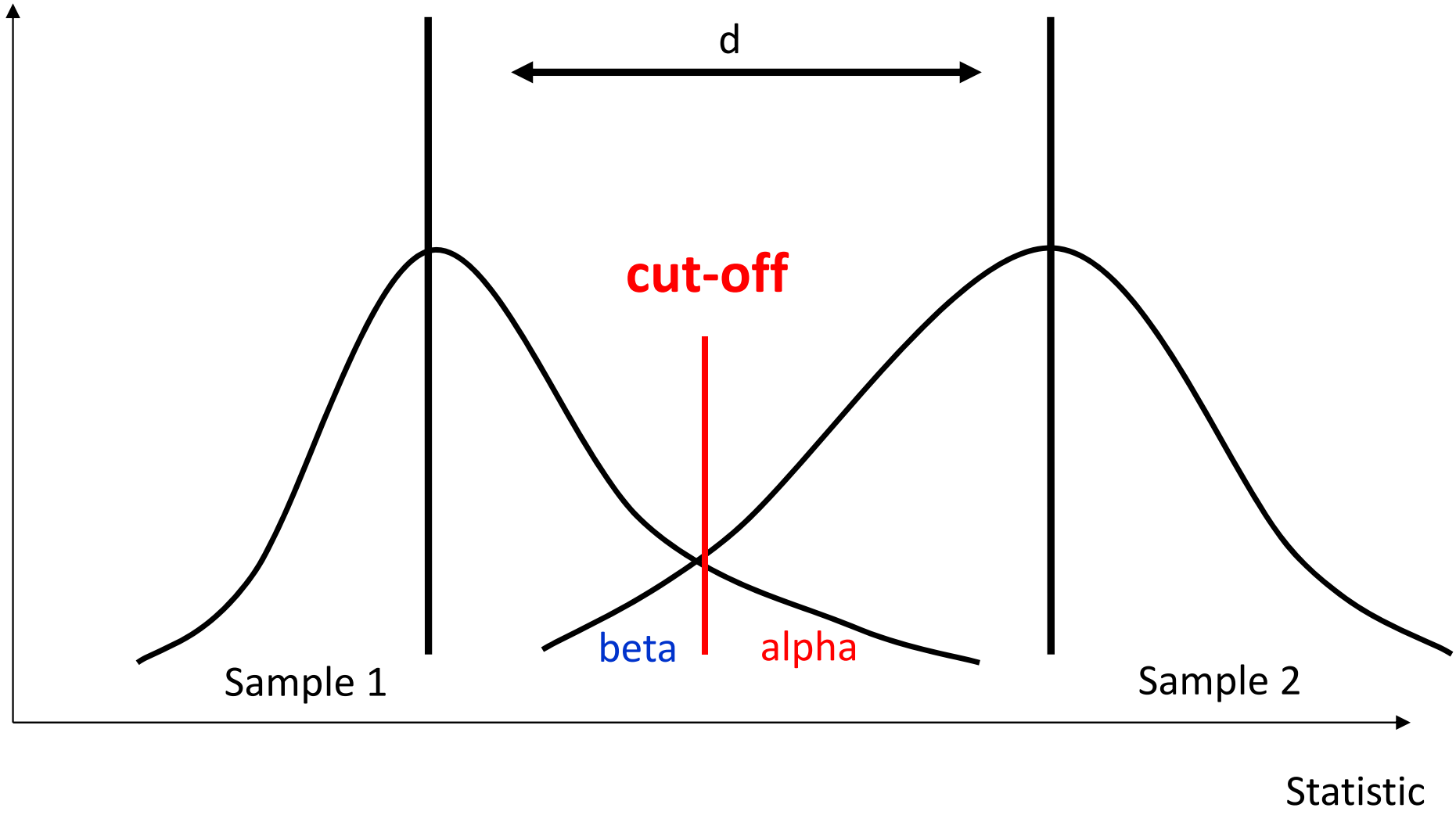
AUC, area under curve; DE, differential expression; FP, false positive.

(*Genome Biology* 14:R95)



FUNCTIONAL MAGNETIC RESONANCE IMAGING, Figure 12.2 © 2004 Sinauer Associates, Inc.

- Type I Error (False Positive): rejecting the null hypothesis when it is true
- Type II Error (False Negative): accepting the null hypothesis when it is false



Multiple Testing Issue

- If more than one test is made, then the collective FP value is greater than in the single-test
  - That is, overall Type I error increases

- E.g: you checked your RNA-Seq data and found 20 significantly different genes with a 0.05 threshold on each gene, then what is the chance that you making at least one error in overall?

- Pr(making a mistake) = 0.05
- Pr(not making a mistake) = 1 – 0.05 = 0.95
- Pr(not making any mistake) = 0.95^{20} = 0.358
- Pr(making at least one mistake) = 1 - 0.358 = 0.642

➔ There is a 64.2% chance of making at least one mistake

Multiple Testing Issue

Bonferroni Correction

- Most straightforward and plain
- For *n* hypothesis tests, only call p-values less than α/n as "being significant".
  - Or, adjust the raw p-value as min(n*p, 1)
- For example, if we want to have an experiment wide Type I error rate of 0.05 when we comparing 30000 genes, we'd need p-values less than $0.05/30000 = 1.67 \times 10^{-6}$ so that the gene(s) could be called as "being significant"

Additional Information

Type I (false positive) error rates

- Family-wise Error Rate

  FWER = p(V ≥ 1)

- Per-family Error Rate

  PFER = E(V)

- Per-comparison Error Rate

  PCER = E(V)/m

- False Discovery Rate

  FDR = E(V/R)

- False Positive Rate

Proportion of false positives among the genes that are flagged as differentially expressed.

	# not rejected	# rejected	totals
# true H	U	V (False Positive)	m_0
# non-true H	T (False Negative)	S	m_1
totals	m - R	R	m

Additional Information

q-value

- q-value is an measure of False Discovery Rate (FDR)
  - Proposed by Storey *et al.* in 2002 and tuned for microarray analysis

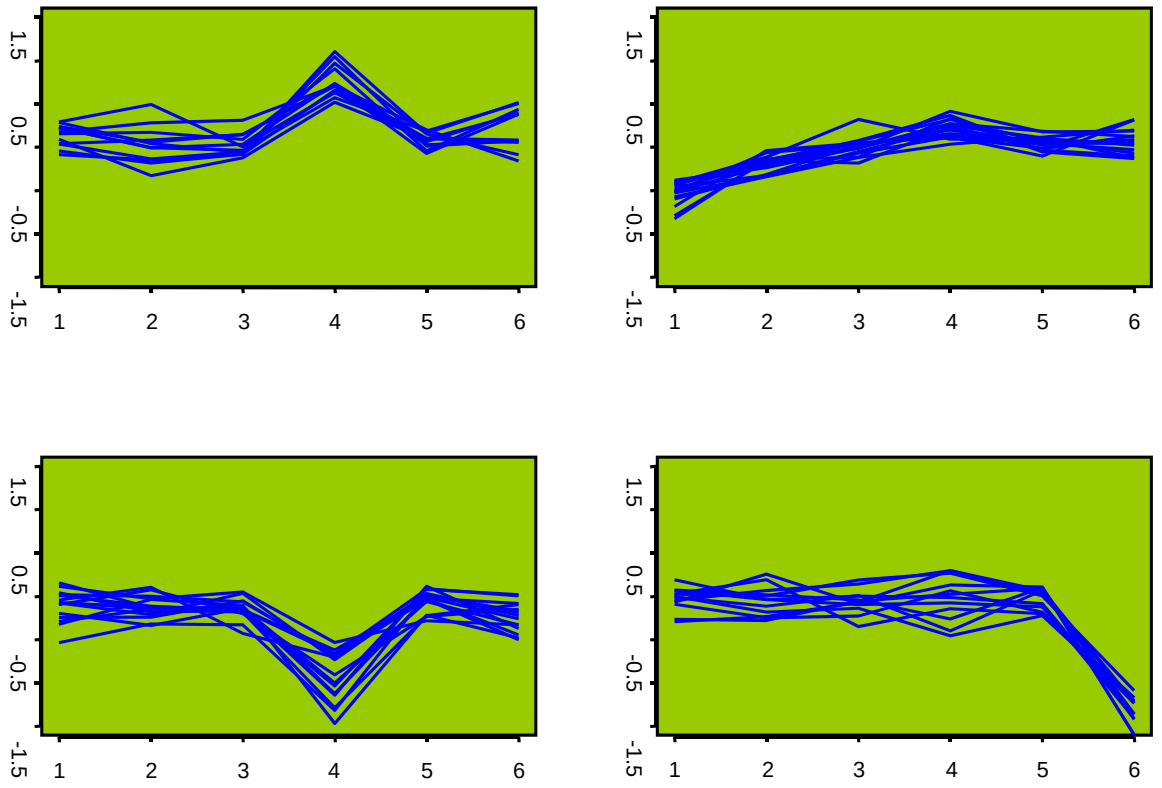
- The q-value for a particular gene *g* is the expected proportion of false positives incurred when calling that gene *g* "significant".

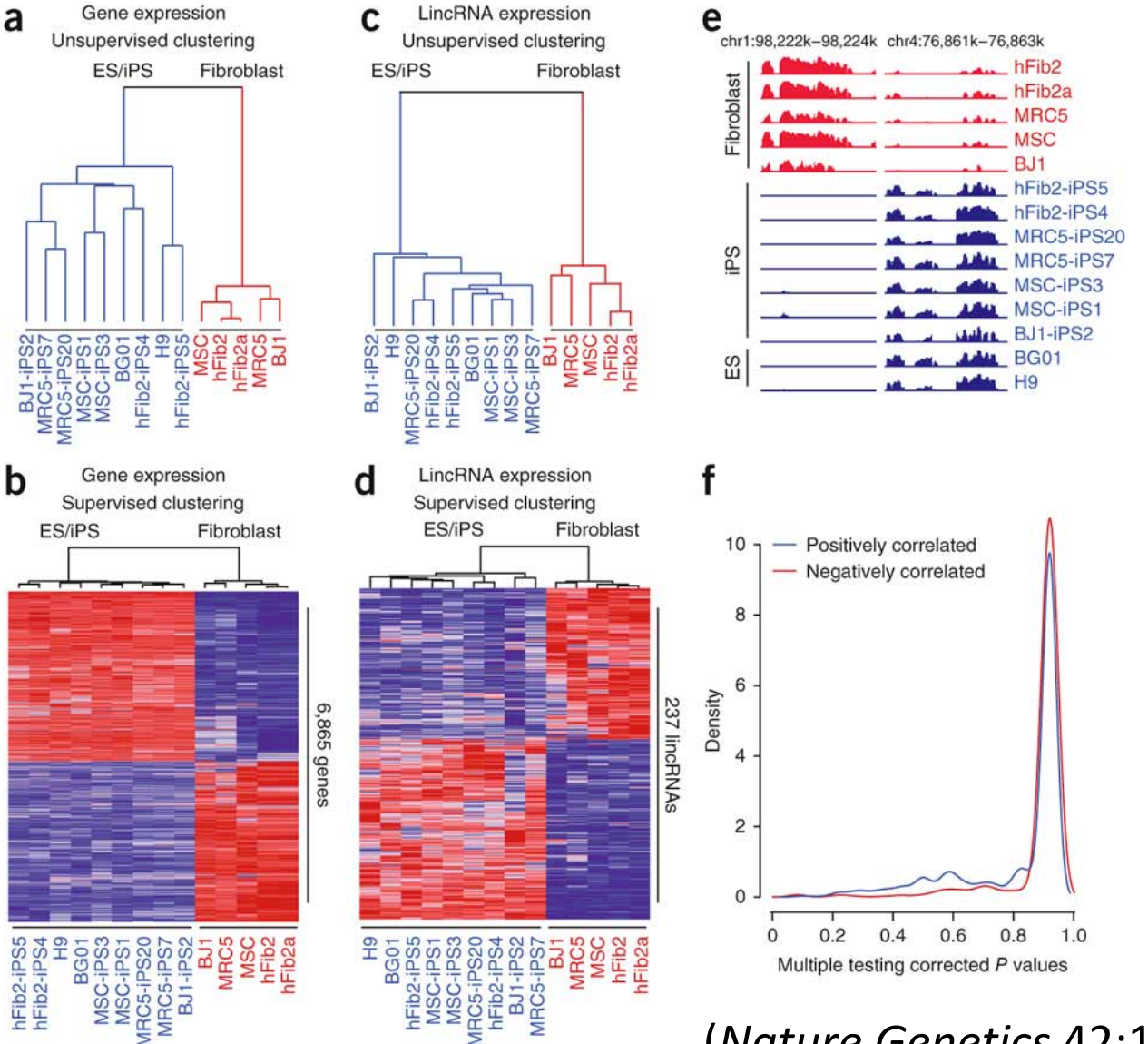
- In contrast, the p-value for a particular gene *g* is the probability that a randomly generated expression profile would be as or more extremely differentially expressed.

- Differentially expressed genes
- Co-expressed genes

Clustering: Group cases (genes/samples) with similar expression pattern/levels (Unsupervised learning)

- Hierarchical Cluster, k-mean Cluster, Self-Organizing Maps (SOM), etc





(*Nature Genetics* 42:1113)

Distance measurement: how "similar" between two genes' profile

Euclidean distance
(Absolution distance)

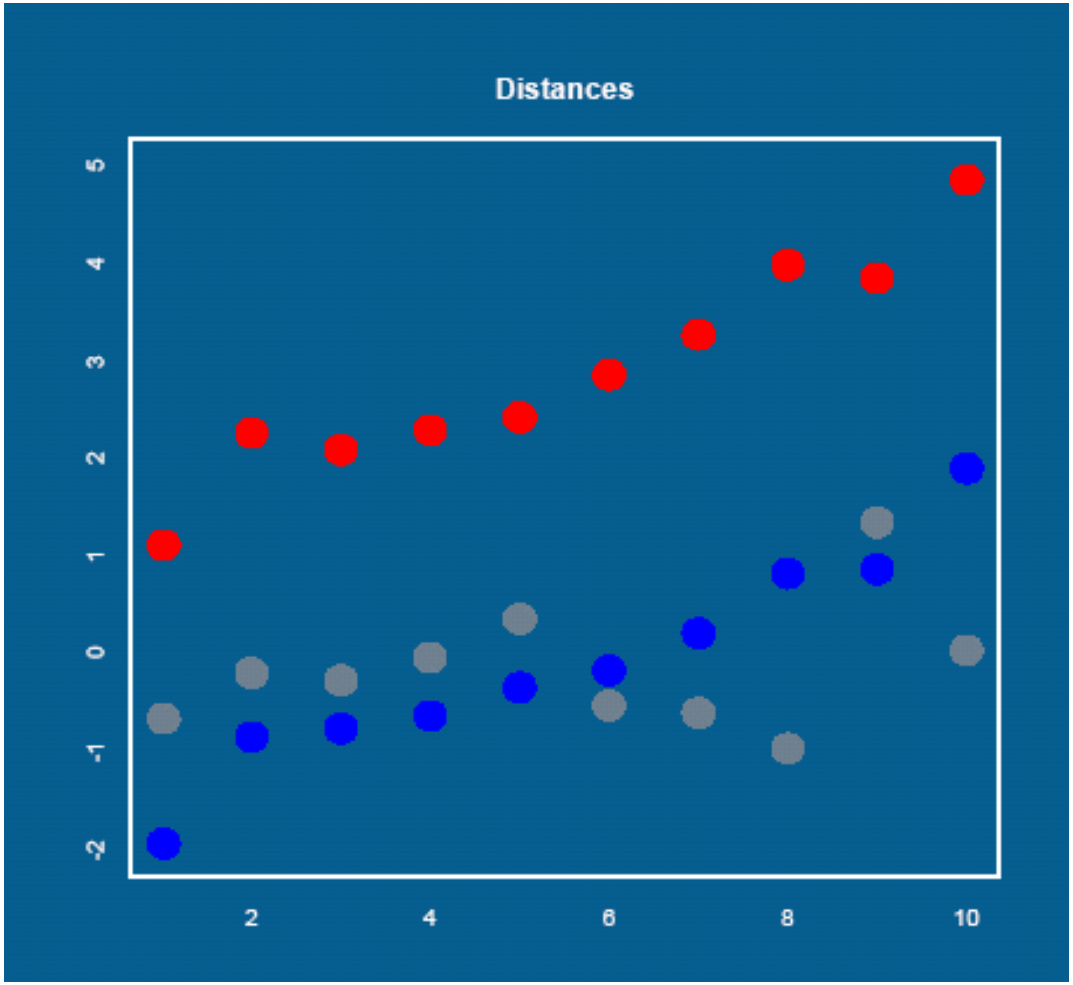
$$s(x_1, x_2) = \sqrt{\sum (x_{1k}^2 - x_{2k}^2)}$$

Pearson distance
(Correlation distance)

$$s(x_1, x_2) = \frac{\sum_{k=1}^{K} (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)}{\sqrt{\sum_{k=1}^{K} (x_{1k} - \bar{x}_1)^2 \sum_{k=1}^{K} (x_{2k} - \bar{x}_2)^2}}$$

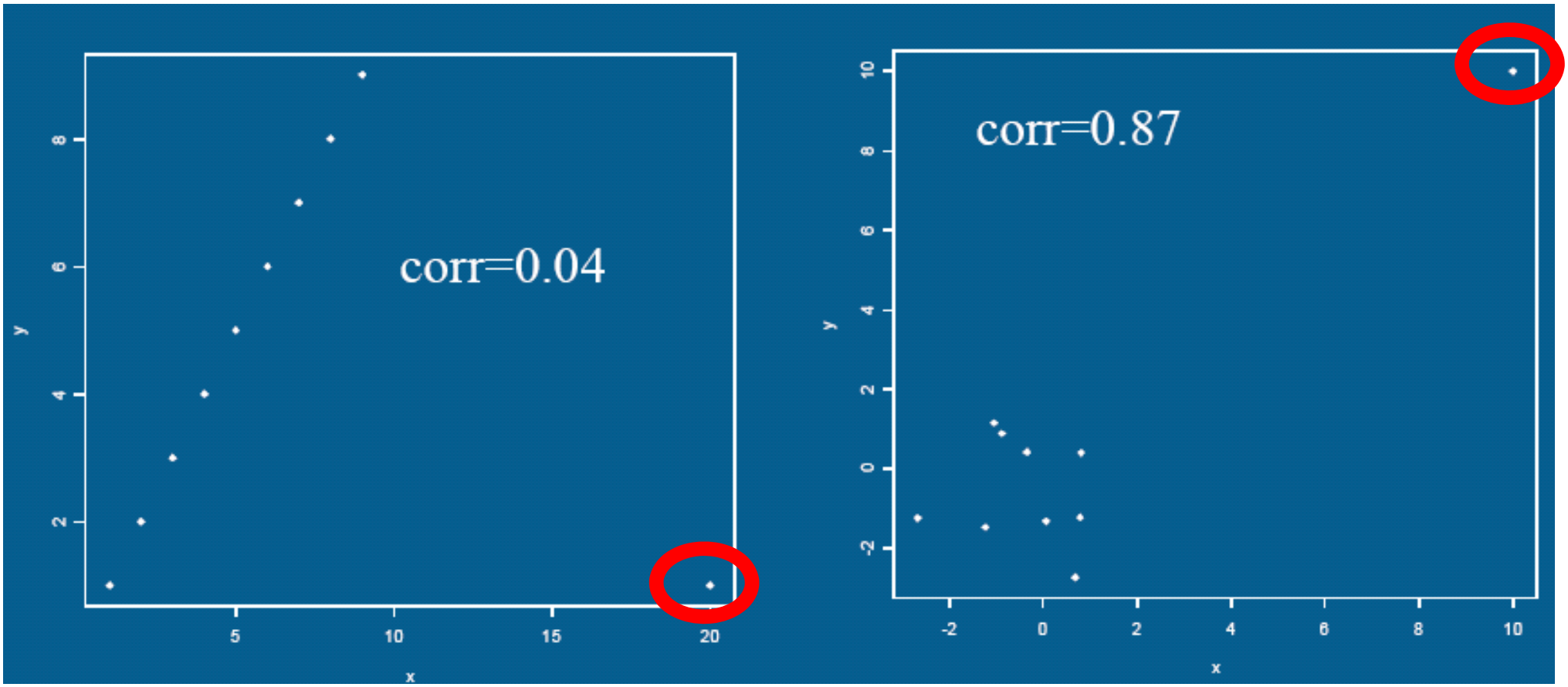
Pearson Distance:

- red-blue: .006
- red-gray: .768
- blue-gray: .7101



Eucl. Distance:

- red-blue: 9.45
- red-gray: 10.26
- blue-gray: 3.29



Summary Question

- Do you think the classical t-test could be used in differential expression calling? Explain.

- Spearman coefficient is a more robust correlation measurement than Pearson coefficient. Could you write the distance formula with Spearman coefficient?

生物信息学：导论与方法
Bioinformatics: Introduction and Methods