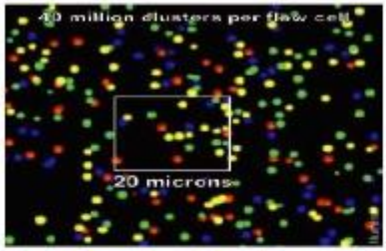
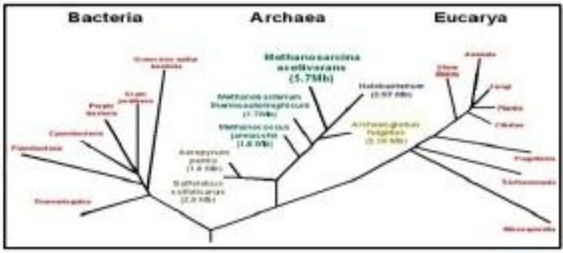
Bioinformatics: Introduction and Methods

Le Zhang

Computer Science Department, Southwest University

TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



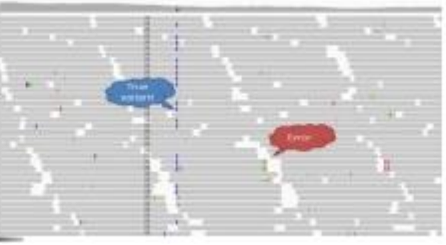


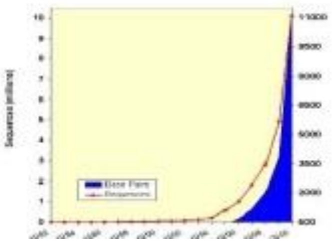
Transcriptome Analysis with noncoding RNAs

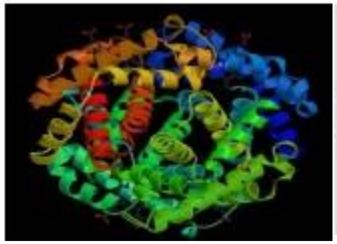
Le Zhang, Ph. D.

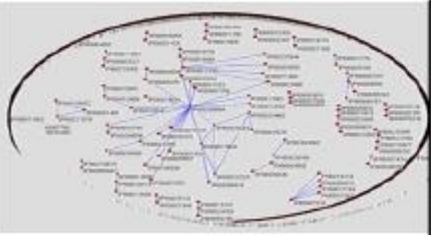
Computer Science Department

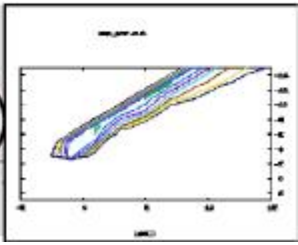
Southwest University



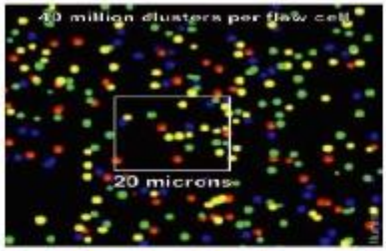


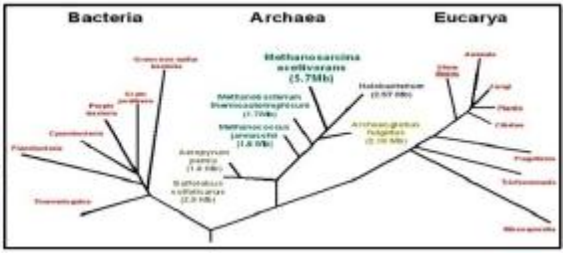






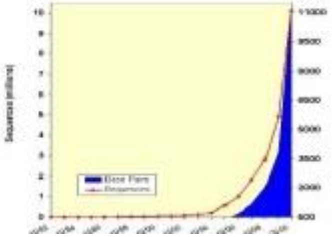
```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```

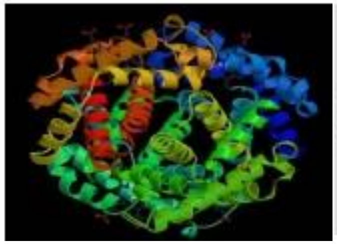


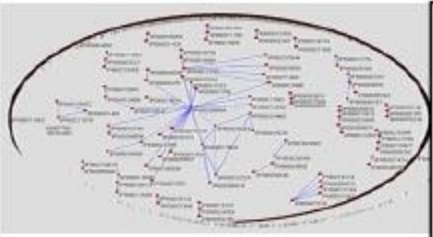


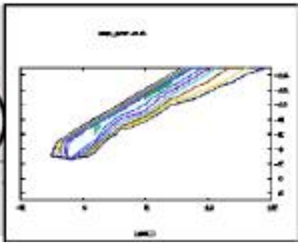
Unit 1:
From Information to Knowledge

Le Zhang, Ph. D.
Computer Science Department
Southwest University

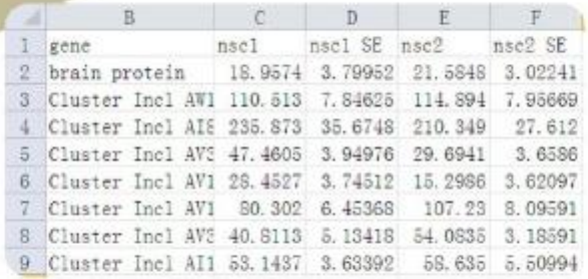






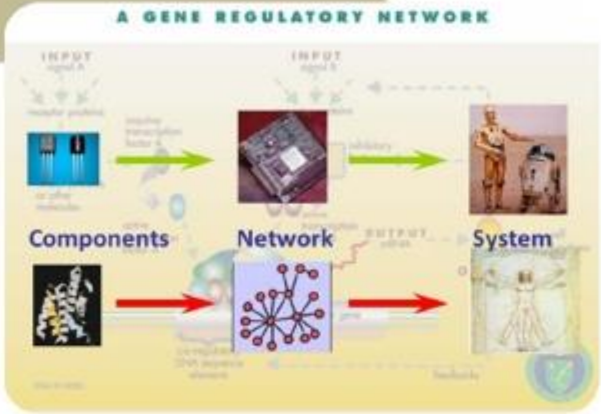


(Sequencing) Data

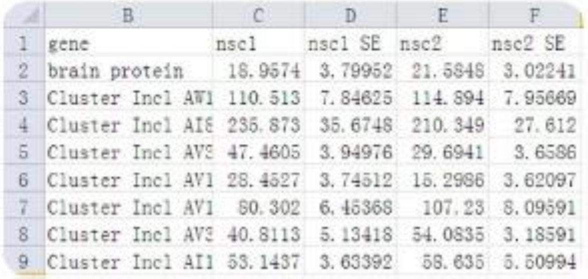


	B	C	D	E	F
1	gene	nsc1	nsc1 SE	nsc2	nsc2 SE
2	brain protein	18.9574	3.79952	21.5848	3.02241
3	Cluster Incl AW1	110.513	7.84625	114.894	7.95669
4	Cluster Incl AI8	235.873	35.6748	210.349	27.612
5	Cluster Incl AV3	47.4605	3.94976	29.6941	3.6586
6	Cluster Incl AV1	28.4527	3.74512	15.2986	3.62097
7	Cluster Incl AV1	80.302	6.45368	107.23	8.09591
8	Cluster Incl AV3	40.8113	5.13418	54.0835	3.18591
9	Cluster Incl AI1	53.1437	3.63392	58.635	5.50994

(Expression) Information



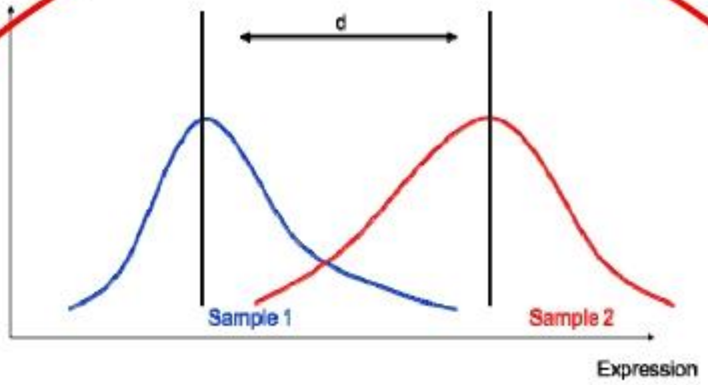
(Biological) Knowledge



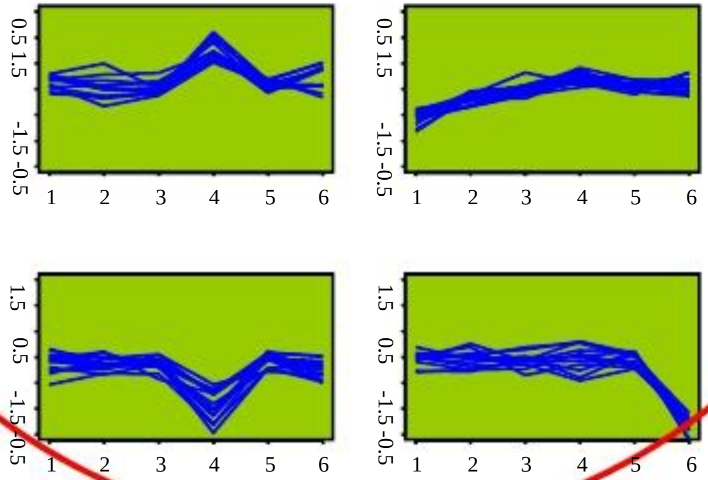
	B	C	D	E	F
1	gene	nsc1	nsc1 SE	nsc2	nsc2 SE
2	brain protein	18.9574	3.79952	21.5848	3.02241
3	Cluster Incl AW1	110.513	7.84625	114.894	7.95669
4	Cluster Incl AI8	235.873	35.6748	210.349	27.612
5	Cluster Incl AV3	47.4605	3.94976	29.6941	3.6586
6	Cluster Incl AV1	28.4527	3.74512	15.2986	3.62097
7	Cluster Incl AV1	80.302	6.45368	107.23	8.09591
8	Cluster Incl AV3	40.8113	5.13418	54.0835	3.18591
9	Cluster Incl AI1	53.1437	3.63392	58.635	5.50994

(Expression) Information

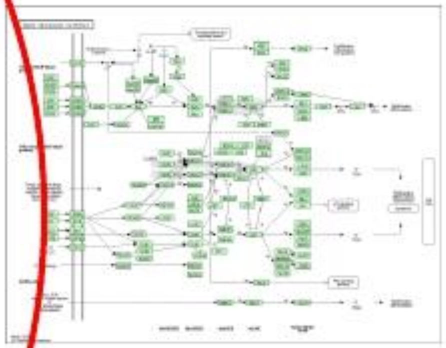
Differentially Expression Calling



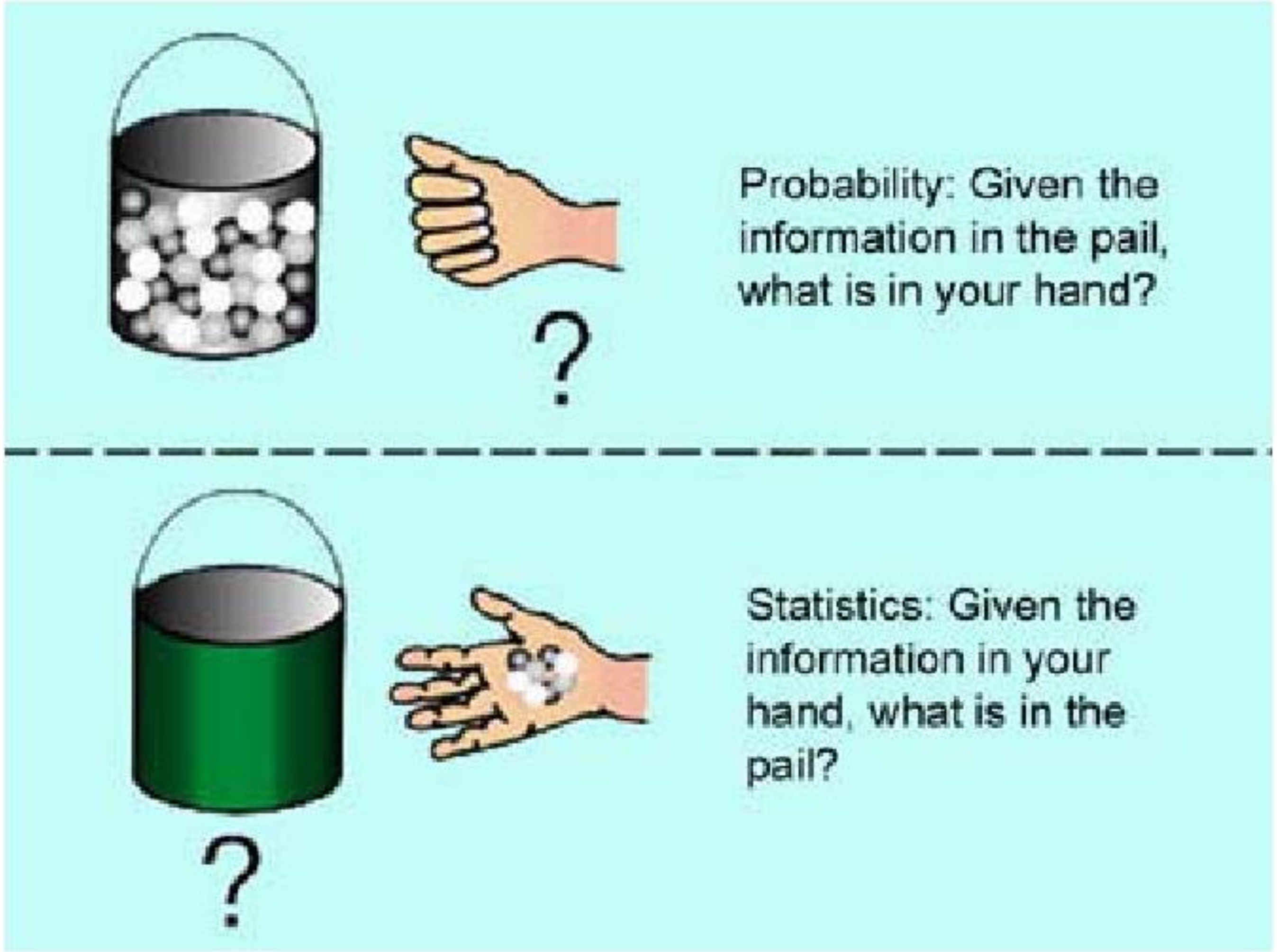
Clustering/Classifying



Pathway



- Functional assignment
- Pathway Enrichment
- Biological Processes

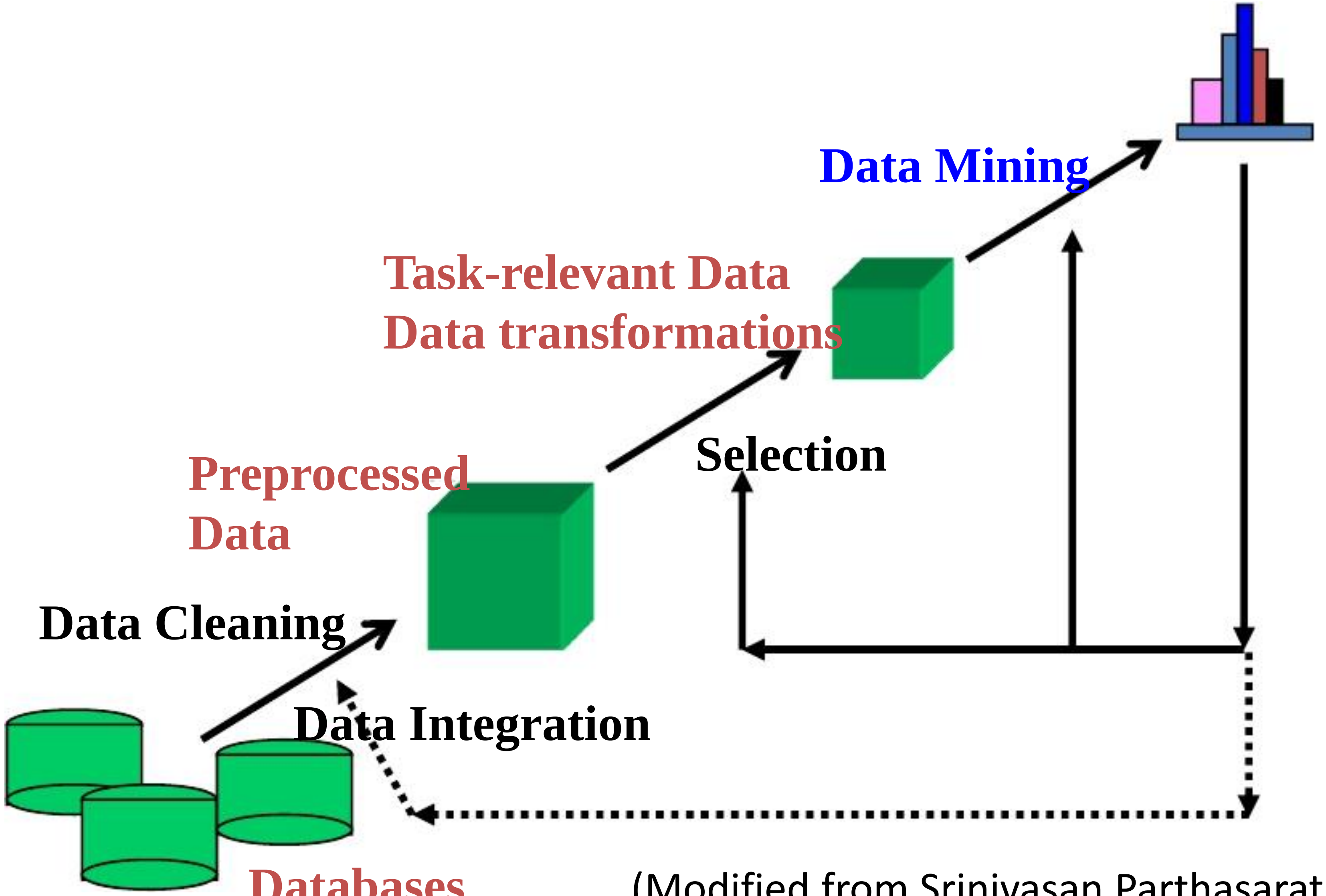


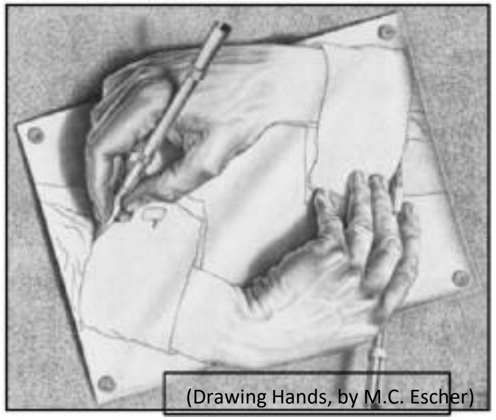
Probability: Given the information in the pail, what is in your hand?

Statistics: Given the information in your hand, what is in the pail?

(Figure Source: http://ocw.mit.edu/OcwWeb/Economics/14-30Spring-2006/CourseHome/index.htm)

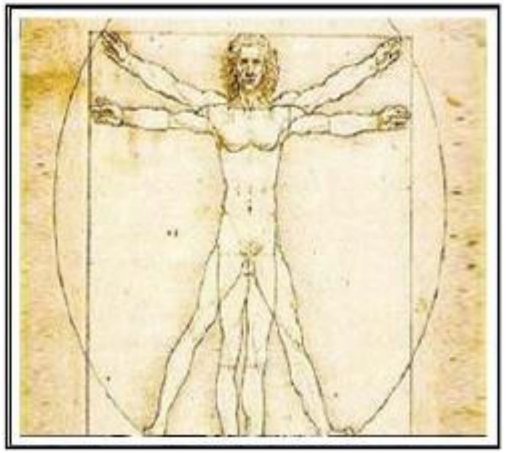
Statistical Learning-guided Mining





(Drawing Hands, by M.C. Escher)

(Modified from Srinivasan Parthasarathy, Ohio State Univ)



(Prior) biological knowledge

(Domain Knowledge)

Data

Model/Algorithm

Set Parameters

Inventory Parameters

Demand Rate 100

Replenishment Rate:

Lot Size 400

Fill Rate (%) 100

Setup Cost 100.0

Product Revenue 6

Product Cost 6

Holding Cost Rate (%) 1

Other Holding Cost 0.1

Backorder Cost 100.00

Lead Time 1

Optimization Parameters

Size 5

Rep. Time Weight 1

Setup Time 0.1

Stochastic Parameters

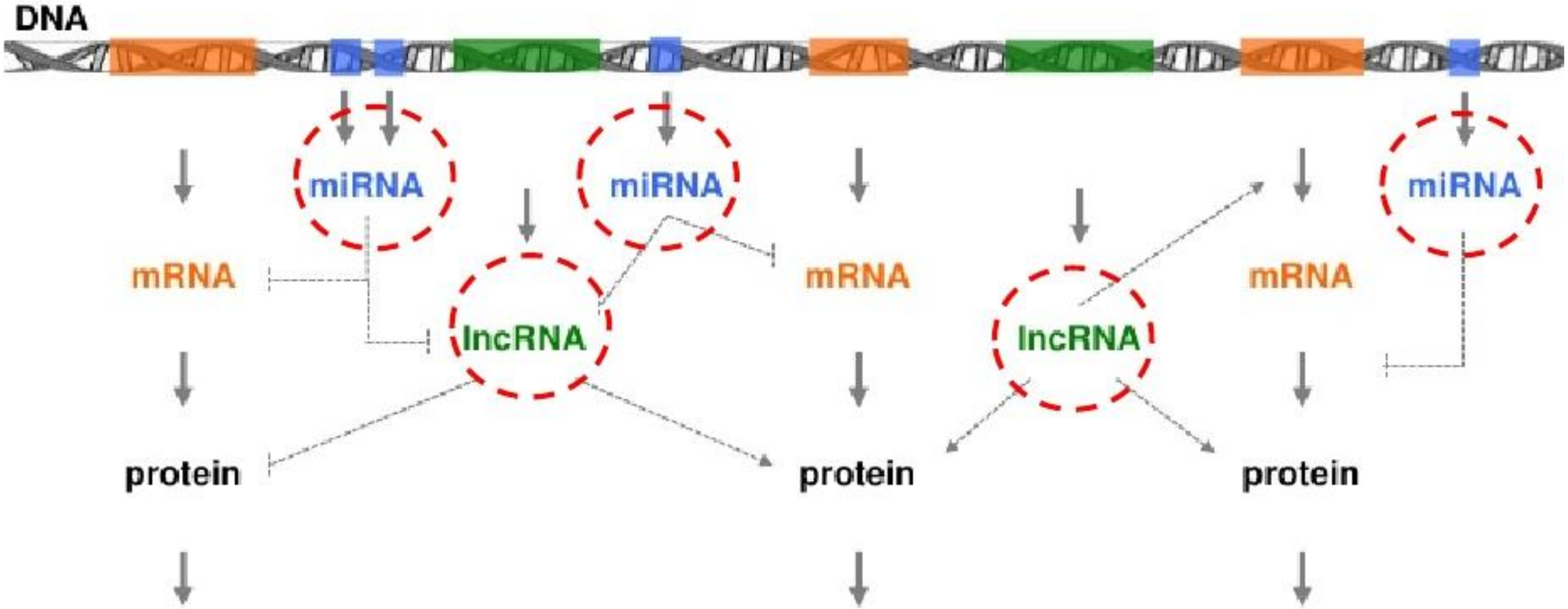
Dem. Mean 10

Dem. SD 10

Max. Prob. Short 0.5

Parameters

- the transcriptome



cellular functions and processes

... – growth – differentiation – apoptosis – migration – cell cycle regulation – signal transduction – transcription - ...

(Modified from http://www.slideshare.net/mateongenaert/05-mestdagh)

A **non-coding RNA (ncRNA)** is any RNA molecule that could function without being translated into a protein.

The DNA sequence from which a non-coding RNA is transcribed as the end product is often called an RNA gene or **non-coding RNA gene**.

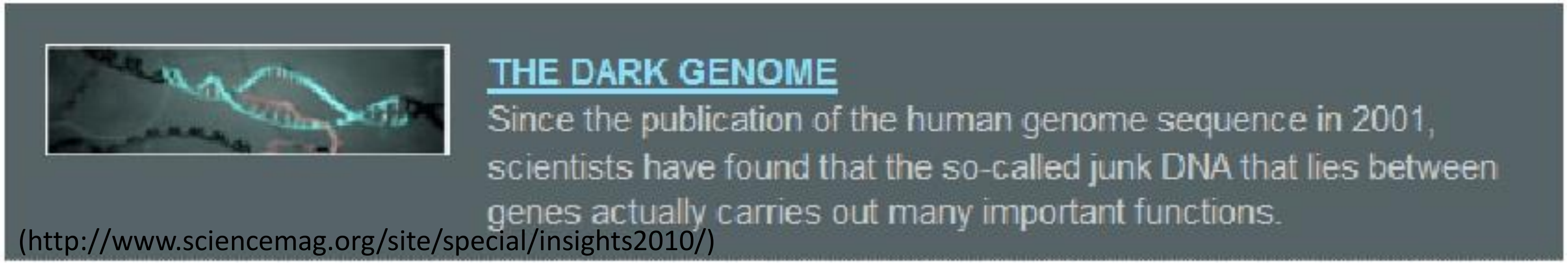
Early discovered ncRNAs are mostly housekeeping

- “Assist” in translation in a necessary, but passive roles
- Constitutively expressed
- Include
  - rRNA
  - tRNA
  - snRNA
  - snoRNA
  - tmRNA
  - telomerase RNA
  - …

Recently discovered regulatory ncRNAs since 2000

- actively regulate gene transcription and translation
- are involved in various gene regulations through multiple mechanisms
- Many have specific expression patterns
- are widely encoded in the genome
  - The ENCODE (ENCyclopedia Of DNA Elements) pilot project suggested that over 90% of the human genome may be represented in primary transcripts.
  - Over 95% of all transcripts are noncoding. Some estimate the number of ncRNAs to be ~30,000.

INSIGHTS OF THE DECADE

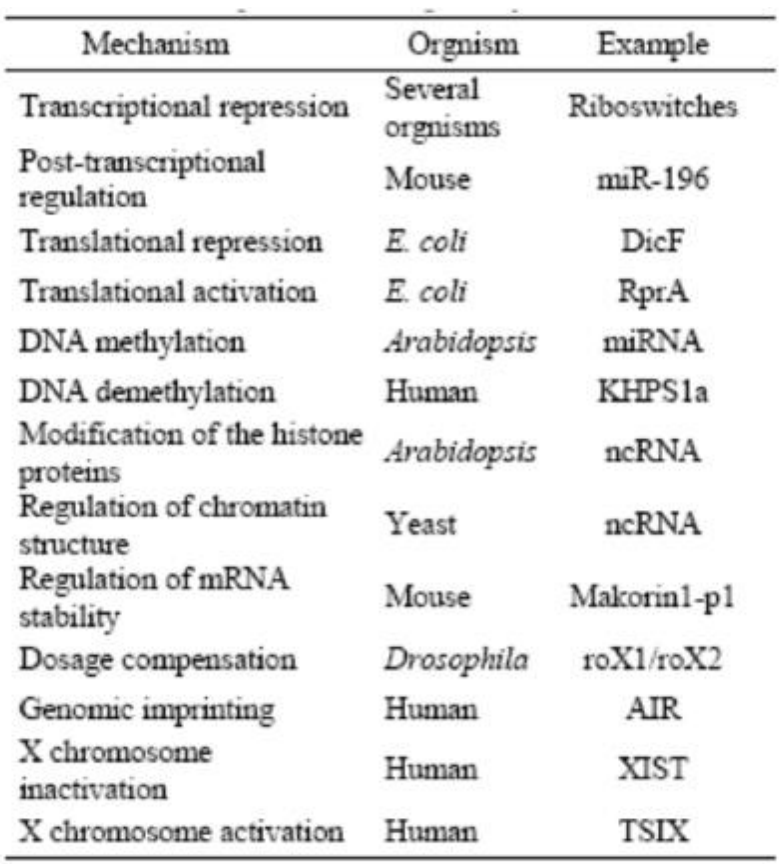


THE DARK GENOME

Since the publication of the human genome sequence in 2001, scientists have found that the so-called junk DNA that lies between genes actually carries out many important functions.

(http://www.sciencemag.org/site/special/insights2010/)

Representative Regulatory Mechanisms of ncRNAs



Mechanism	Orgnism	Example
Transcriptional repression	Several orgnisms	Riboswitches
Post-transcriptional regulation	Mouse	miR-196
Translational repression	*E. coli*	DicF
Translational activation	*E. coli*	RprA
DNA methylation	*Arabidopsis*	miRNA
DNA demethylation	Human	KHPS1a
Modification of the histone proteins	*Arabidopsis*	ncRNA
Regulation of chromatin structure	Yeast	ncRNA
Regulation of mRNA stability	Mouse	Makorin1-p1
Dosage compensation	*Drosophila*	roX1/roX2
Genomic imprinting	Human	AIR
X chromosome inactivation	Human	XIST
X chromosome activation	Human	TSIX

Qi, Sci China '06

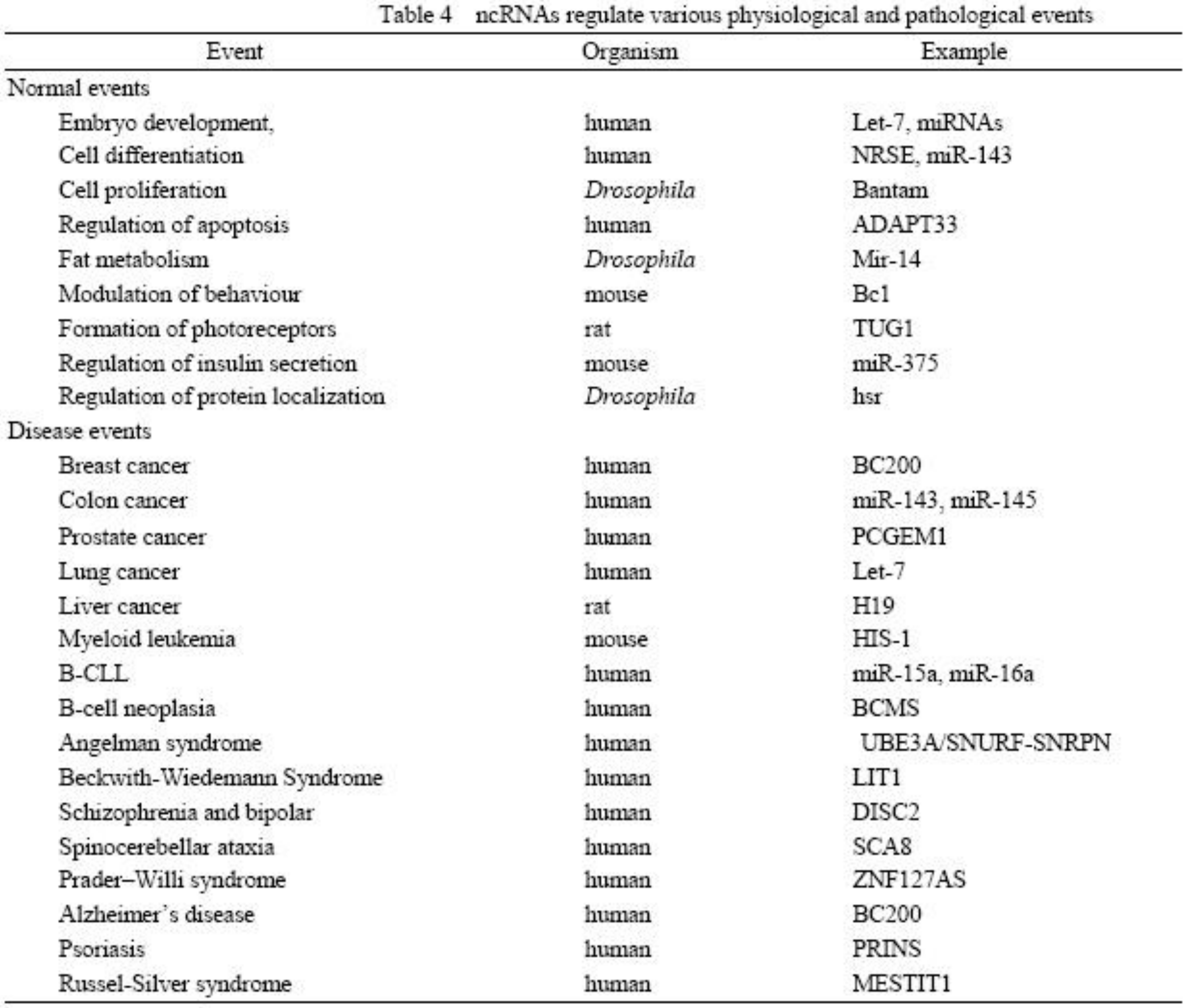
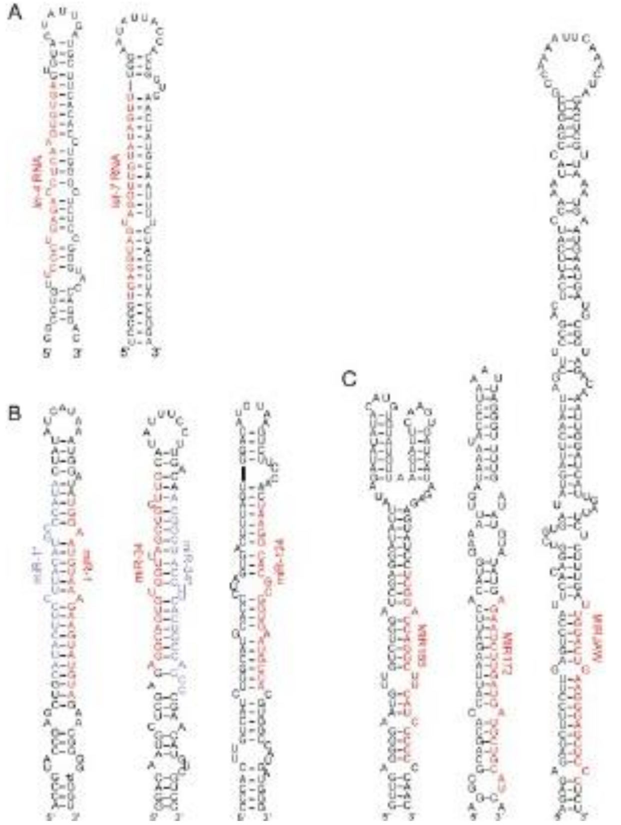


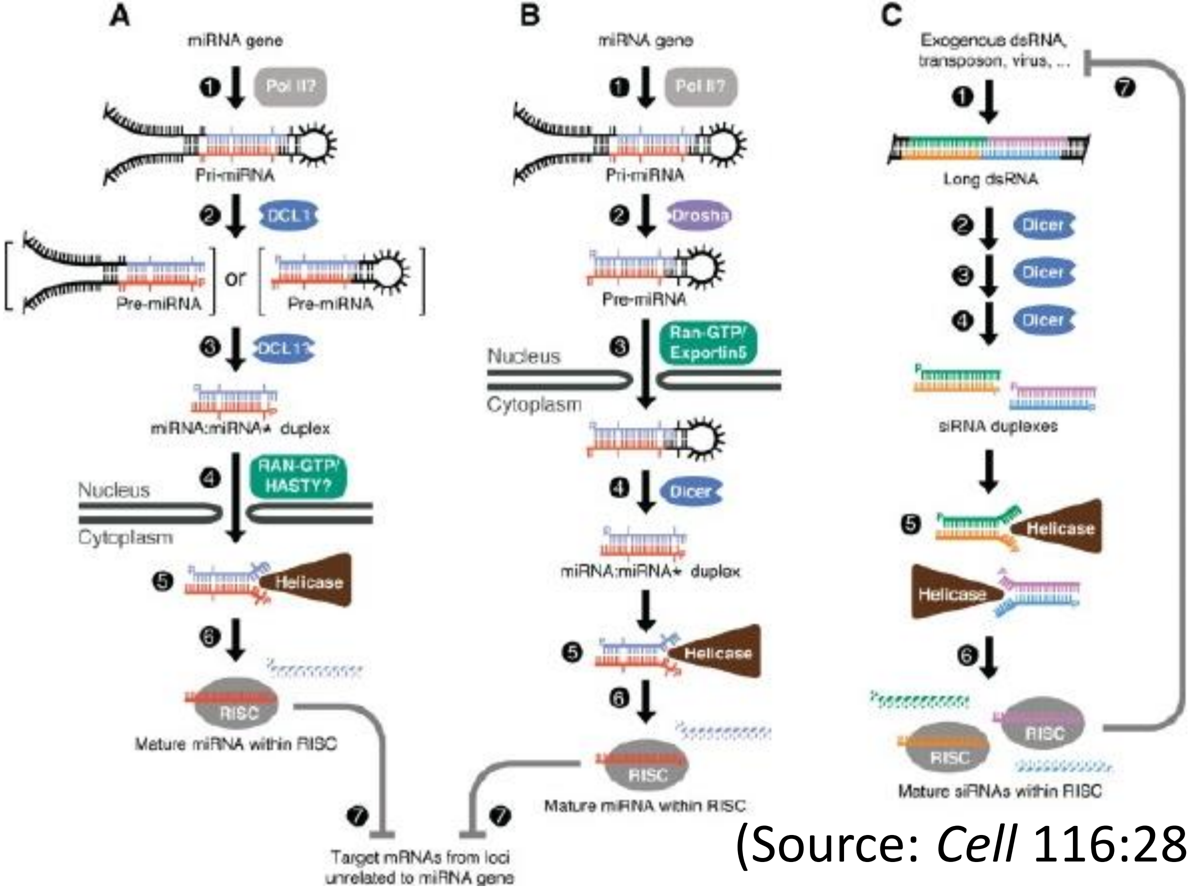
Table 4 ncRNAs regulate various physiological and pathological events

Event	Organism	Example
Normal events		
Embryo development,	human	Let-7, miRNAs
Cell differentiation	human	NRSE, miR-143
Cell proliferation	*Drosophila*	Bantam
Regulation of apoptosis	human	ADAPT33
Fat metabolism	*Drosophila*	Mir-14
Modulation of behaviour	mouse	Bc1
Formation of photoreceptors	rat	TUG1
Regulation of insulin secretion	mouse	miR-375
Regulation of protein localization	*Drosophila*	hsr
Disease events		
Breast cancer	human	BC200
Colon cancer	human	miR-143, miR-145
Prostate cancer	human	PCGEM1
Lung cancer	human	Let-7
Liver cancer	rat	H19
Myeloid leukemia	mouse	HIS-1
B-CLL	human	miR-15a, miR-16a
B-cell neoplasia	human	BCMS
Angelman syndrome	human	UBE3A/SNURF-SNRPN
Beckwith-Wiedemann Syndrome	human	LIT1
Schizophrenia and bipolar	human	DISC2
Spinocerebellar ataxia	human	SCA8
Prader–Willi syndrome	human	ZNF127AS
Alzheimer's disease	human	BC200
Psoriasis	human	PRINS
Russel-Silver syndrome	human	MESTIT1

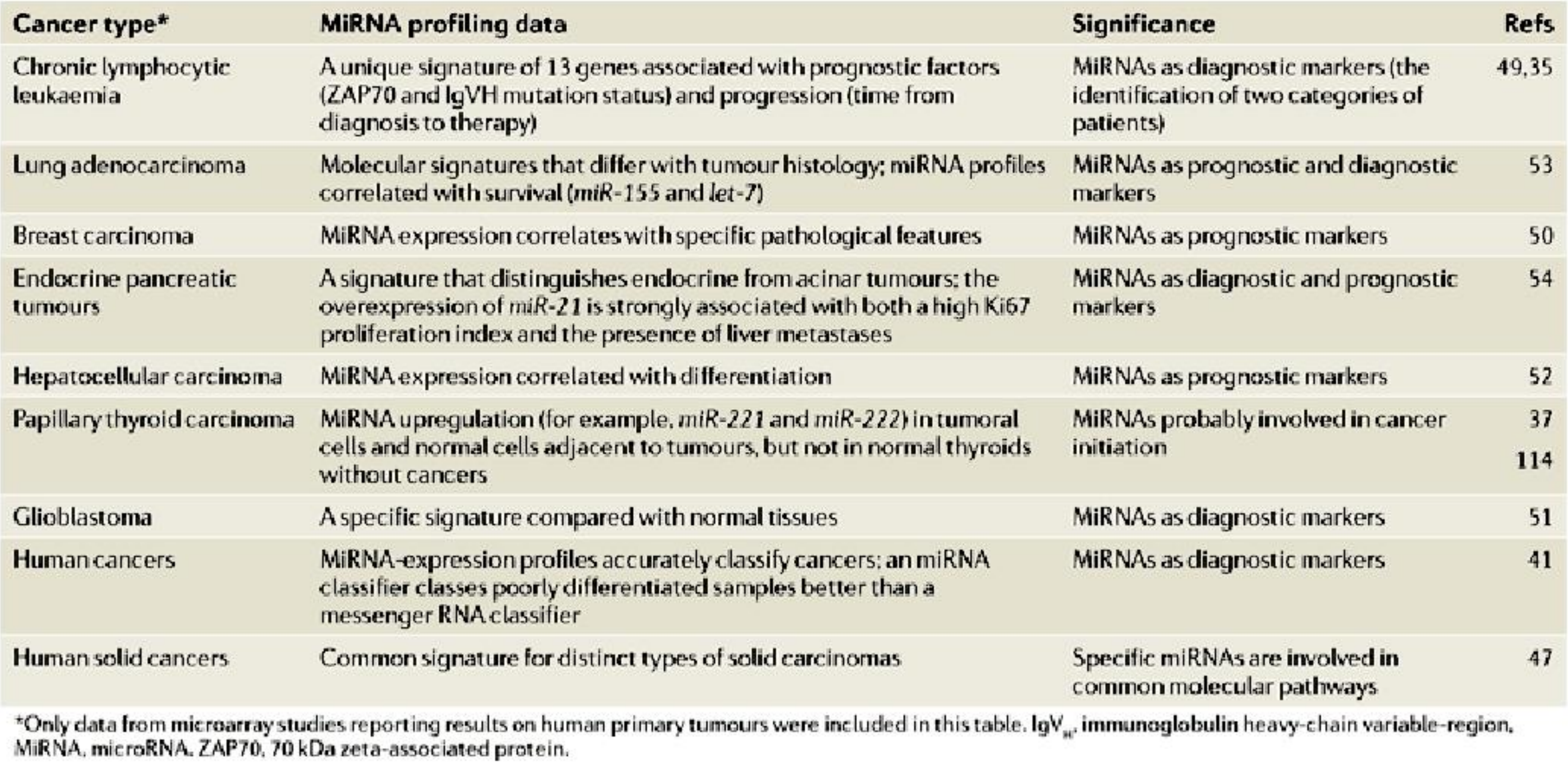
microRNA (miRNA)

- single-stranded RNAs of 21-23 (or some say 20-25) nt RNAs with regulatory functions when associated with a protein complex.
- In plants miRNAs can silence gene activity via destruction of homologous mRNA or blocking its translation. In animals, miRNAs inhibit translation by binding with imperfect homology to the 3’ untranslated region of mRNA.





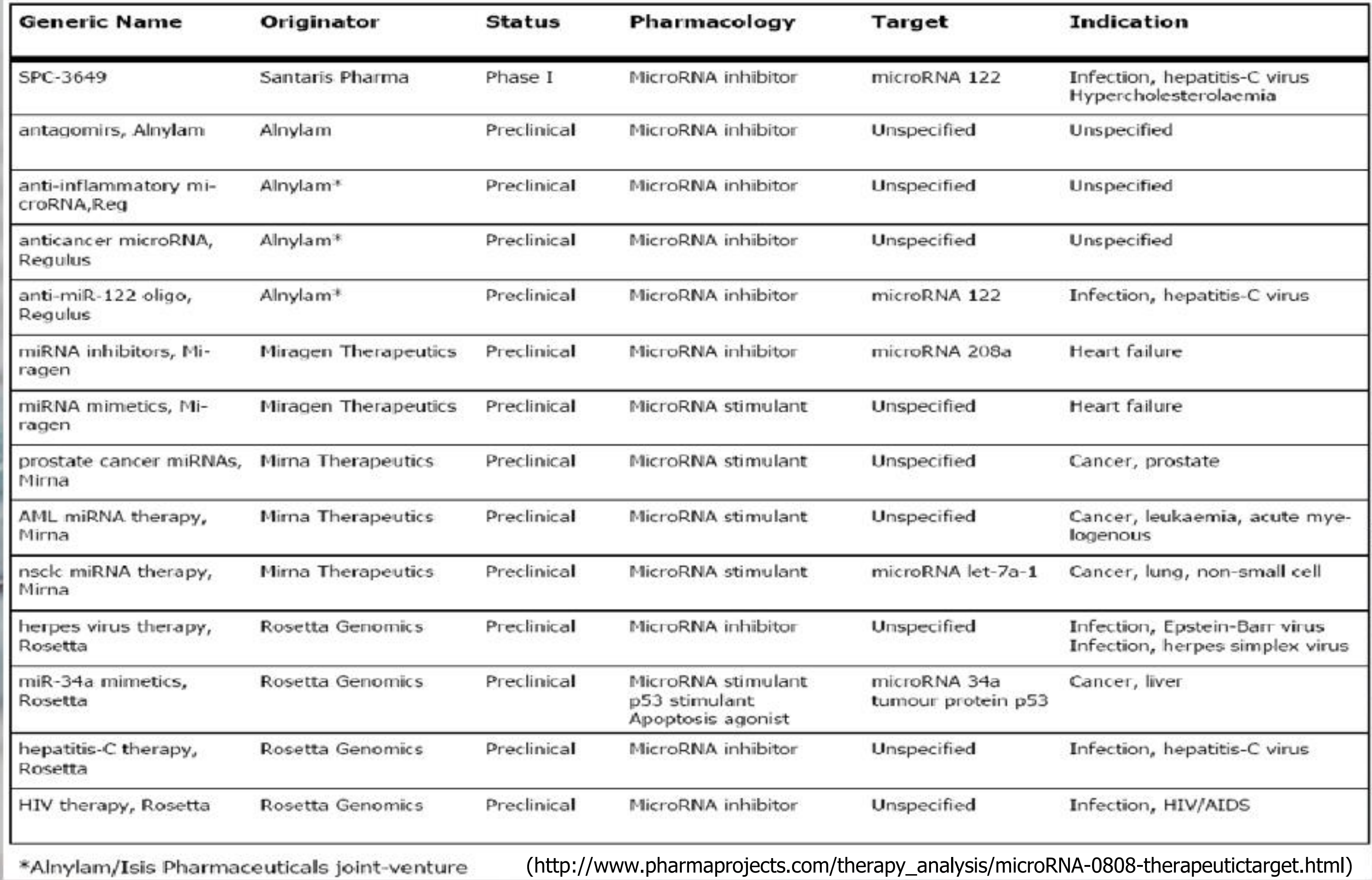
(Source: *Cell* 116:281)



Cancer type*	MiRNA profiling data	Significance	Refs
Chronic lymphocytic leukaemia	A unique signature of 13 genes associated with prognostic factors (ZAP70 and IgVH mutation status) and progression (time from diagnosis to therapy)	MiRNAs as diagnostic markers (the identification of two categories of patients)	49,35
Lung adenocarcinoma	Molecular signatures that differ with tumour histology; miRNA profiles correlated with survival (*miR-155* and *let-7*)	MiRNAs as prognostic and diagnostic markers	53
Breast carcinoma	MiRNA expression correlates with specific pathological features	MiRNAs as prognostic markers	50
Endocrine pancreatic tumours	A signature that distinguishes endocrine from acinar tumours; the overexpression of *miR-21* is strongly associated with both a high Ki67 proliferation index and the presence of liver metastases	MiRNAs as diagnostic and prognostic markers	54
Hepatocellular carcinoma	MiRNA expression correlated with differentiation	MiRNAs as prognostic markers	52
Papillary thyroid carcinoma	MiRNA upregulation (for example, *miR-221* and *miR-222*) in tumoral cells and normal cells adjacent to tumours, but not in normal thyroids without cancers	MiRNAs probably involved in cancer initiation	37 114
Glioblastoma	A specific signature compared with normal tissues	MiRNAs as diagnostic markers	51
Human cancers	MiRNA-expression profiles accurately classify cancers; an miRNA classifier classes poorly differentiated samples better than a messenger RNA classifier	MiRNAs as diagnostic markers	41
Human solid cancers	Common signature for distinct types of solid carcinomas	Specific miRNAs are involved in common molecular pathways	47

*Only data from microarray studies reporting results on human primary tumours were included in this table. IgV$_H$, immunoglobulin heavy-chain variable-region. MiRNA, microRNA. ZAP70, 70 kDa zeta-associated protein.

(Source: *Nat Rev Cancer* **6**, 857)

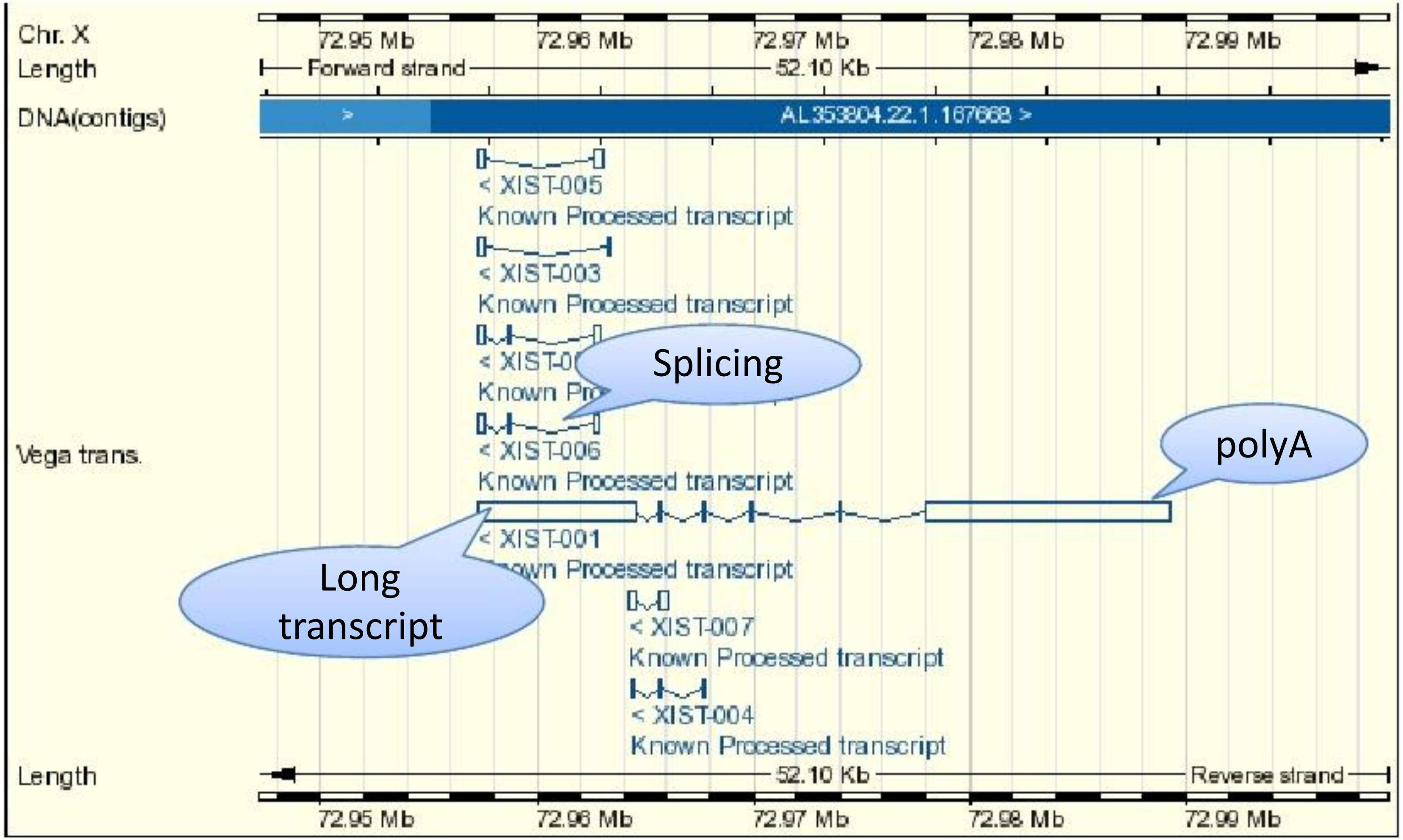


Generic Name	Originator	Status	Pharmacology	Target	Indication
SPC-3649	Santaris Pharma	Phase I	MicroRNA inhibitor	microRNA 122	Infection, hepatitis-C virus Hypercholesterolaemia
antagomirs, Alnylam	Alnylam	Preclinical	MicroRNA inhibitor	Unspecified	Unspecified
anti-inflammatory microRNA,Reg	Alnylam*	Preclinical	MicroRNA inhibitor	Unspecified	Unspecified
anticancer microRNA, Regulus	Alnylam*	Preclinical	MicroRNA inhibitor	Unspecified	Unspecified
anti-miR-122 oligo, Regulus	Alnylam*	Preclinical	MicroRNA inhibitor	microRNA 122	Infection, hepatitis-C virus
miRNA inhibitors, Miragen	Miragen Therapeutics	Preclinical	MicroRNA inhibitor	microRNA 208a	Heart failure
miRNA mimetics, Miragen	Miragen Therapeutics	Preclinical	MicroRNA stimulant	Unspecified	Heart failure
prostate cancer miRNAs, Mirna	Mirna Therapeutics	Preclinical	MicroRNA stimulant	Unspecified	Cancer, prostate
AML miRNA therapy, Mirna	Mirna Therapeutics	Preclinical	MicroRNA stimulant	Unspecified	Cancer, leukaemia, acute myelogenous
nsclc miRNA therapy, Mirna	Mirna Therapeutics	Preclinical	MicroRNA stimulant	microRNA let-7a-1	Cancer, lung, non-small cell
herpes virus therapy, Rosetta	Rosetta Genomics	Preclinical	MicroRNA inhibitor	Unspecified	Infection, Epstein-Barr virus Infection, herpes simplex virus
miR-34a mimetics, Rosetta	Rosetta Genomics	Preclinical	MicroRNA stimulant p53 stimulant Apoptosis agonist	microRNA 34a tumour protein p53	Cancer, liver
hepatitis-C therapy, Rosetta	Rosetta Genomics	Preclinical	MicroRNA inhibitor	Unspecified	Infection, hepatitis-C virus
HIV therapy, Rosetta	Rosetta Genomics	Preclinical	MicroRNA inhibitor	Unspecified	Infection, HIV/AIDS

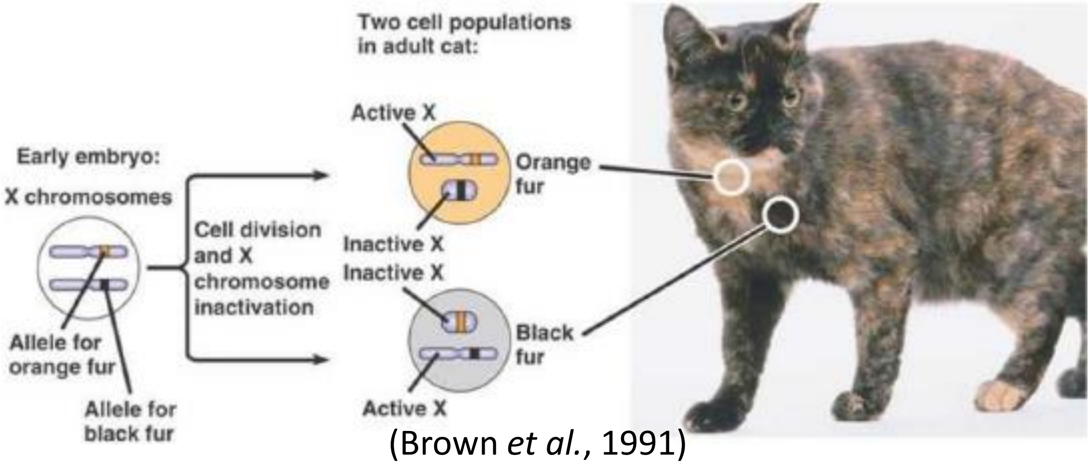
*Alnylam/Isis Pharmaceuticals joint-venture

(http://www.pharmaprojects.com/therapy_analysis/microRNA-0808-therapeutictarget.html)

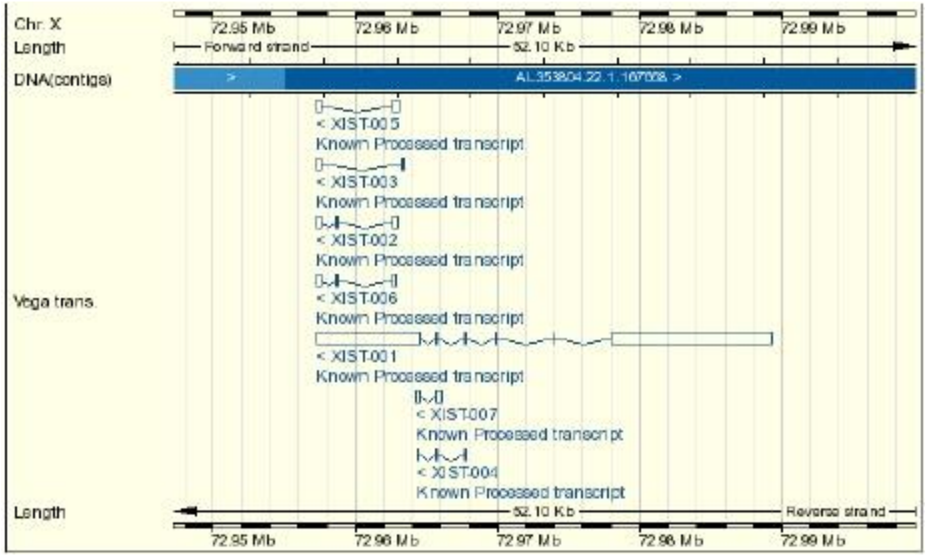
Xist : Beyond "small" ncRNA

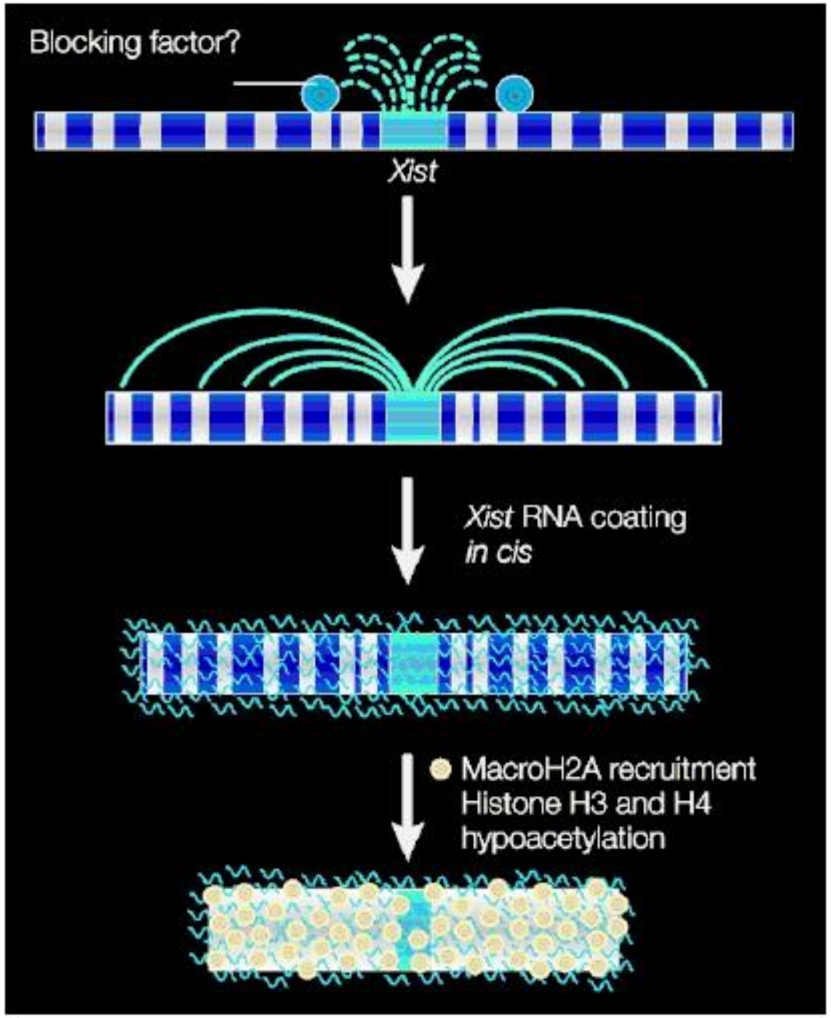


Xist – X inactive-specific transcript



(Brown *et al.*, 1991)

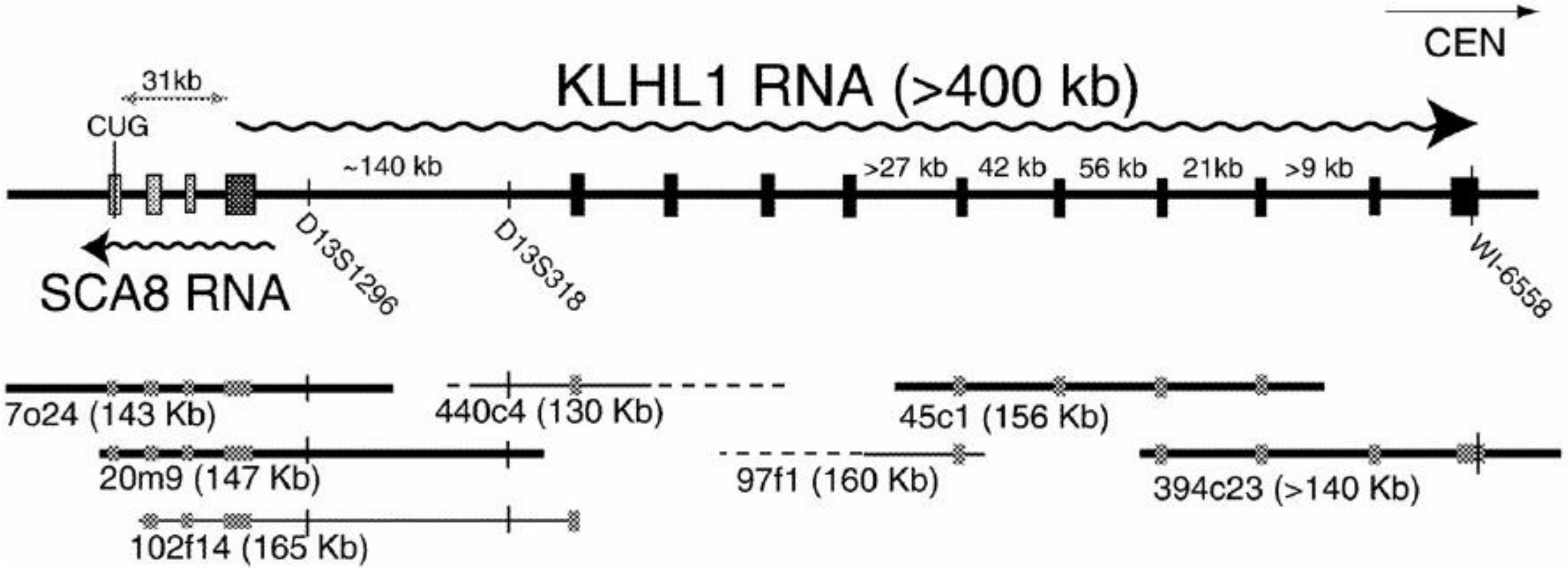




(Avner *et al.*, 2001)

SCA8:
Long ncRNA in Human Disease

- SCA8 is mutated in one form of spinal cerebella ataxia



(Nemes, J. P. *et al*. 2000)

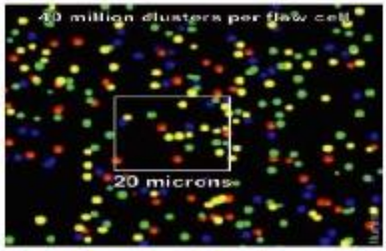
Long ncRNAs

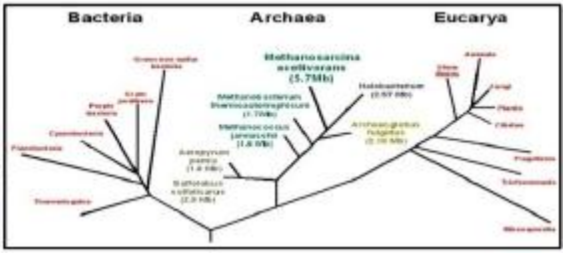
- Estimated ~2000+ in human.
- Some, but not all, are mRNA-like, with Poly(A) tails.
- Most have unknown function. Many may function via *cis* or *trans* antisense pairing.
  - Dosage compensation (e.g. XIST)
  - Neuron development (e.g. SCA8)
  - Genetic imprinting (e.g. IGF/H19)
  - Post-transcriptional regulation
    - mRNA degradation or stabilization
  - Translational regulation
  - Modulate protein function by directly binding to the protein

How many non-coding transcripts?

What are the functional roles of those ncRNAs?

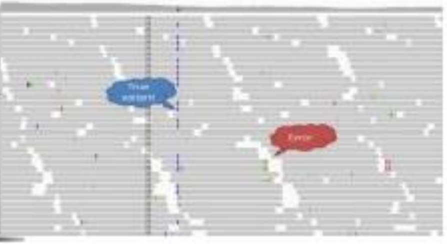
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA

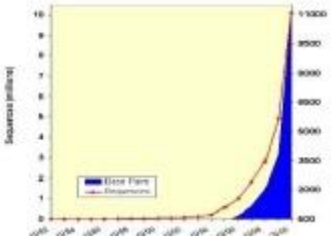


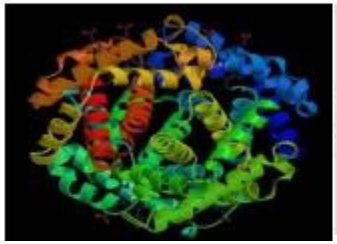


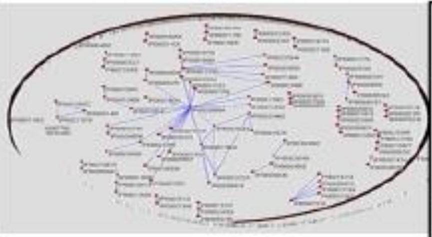
Unit 2:
Data Mining: Identify long ncRNAs

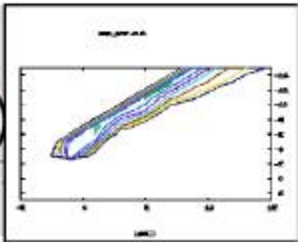
Le Zhang, Ph. D.
Computer Science Department
Southwest University





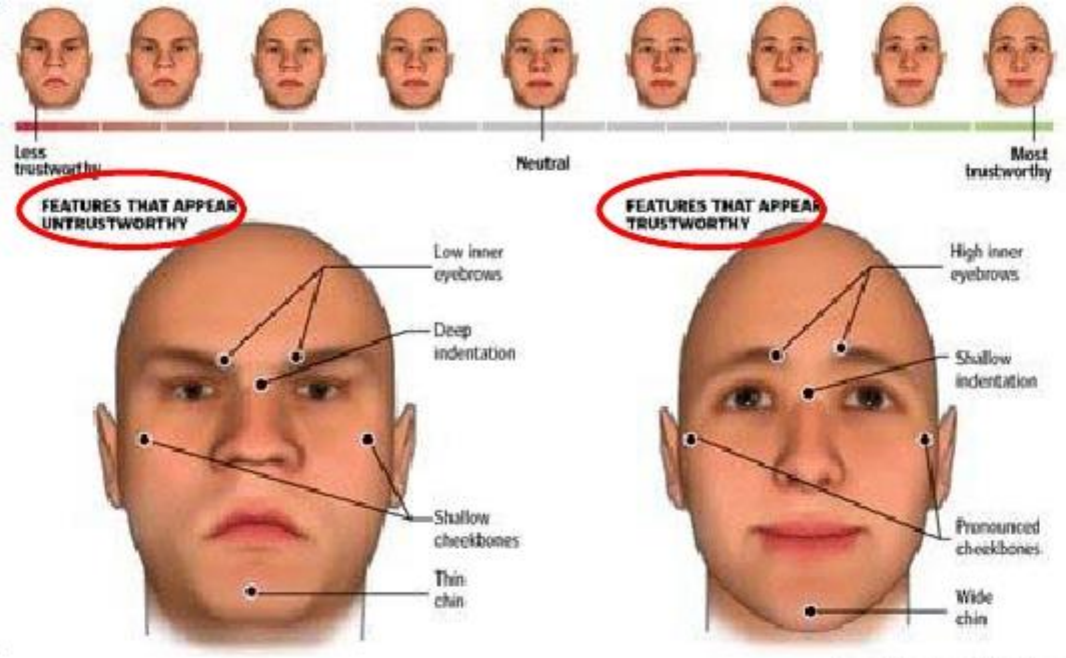






Identification

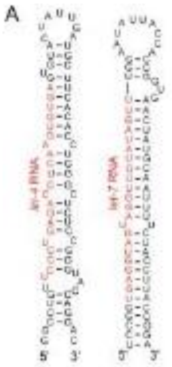
(Source: www.lkalop.com)



(Source: www.lemondrop.com/2009/01/22/certain-facial-features-found-to-create-a-feeling-of-trust/) The Boston Globe

Features ~ property of an entity

Structural features



(*Cell* 116:281)

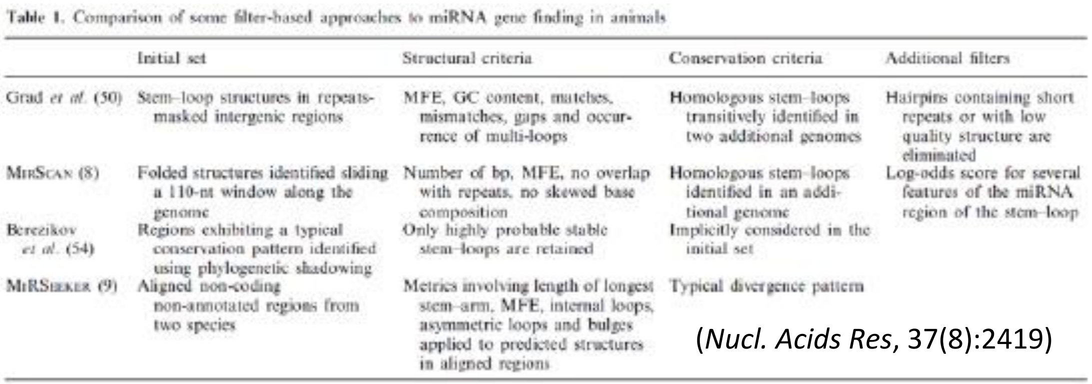
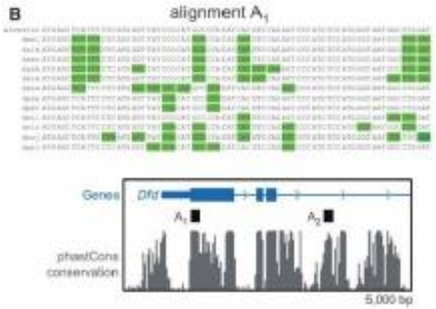


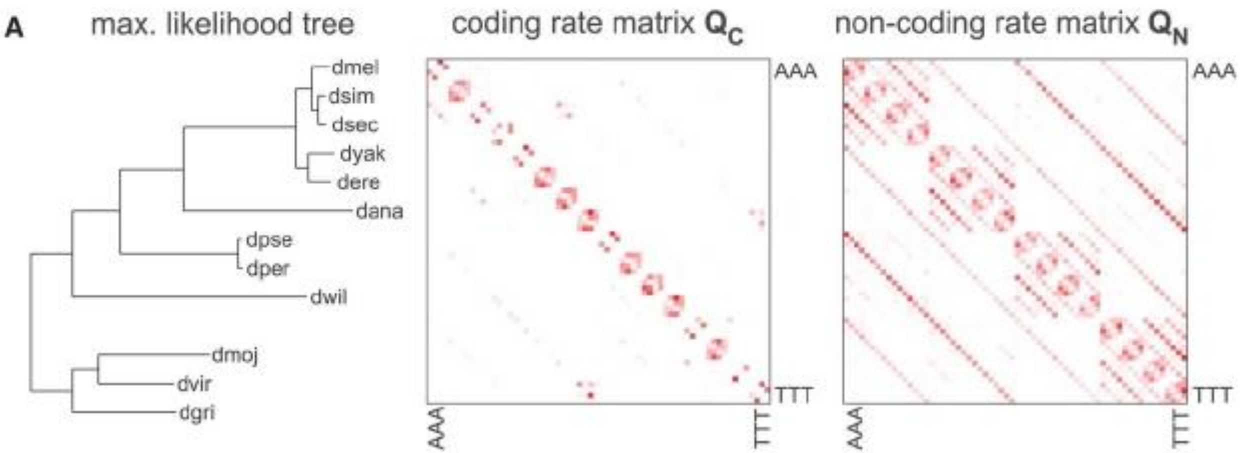
Table 1. Comparison of some filter-based approaches to miRNA gene finding in animals

	Initial set	Structural criteria	Conservation criteria	Additional filters
Grad *et al.* (50)	Stem–loop structures in repeats-masked intergenic regions	MFE, GC content, matches, mismatches, gaps and occurrence of multi-loops	Homologous stem–loops transitively identified in two additional genomes	Hairpins containing short repeats or with low quality structure are eliminated
MirScan (8)	Folded structures identified sliding a 110-nt window along the genome	Number of bp, MFE, no overlap with repeats, no skewed base composition	Homologous stem–loops identified in an additional genome	Log-odds score for several features of the miRNA region of the stem–loop
Berezikov *et al.* (54)	Regions exhibiting a typical conservation pattern identified using phylogenetic shadowing	Only highly probable stable stem–loops are retained	Implicitly considered in the initial set	
MirSeeker (9)	Aligned non-coding non-annotated regions from two species	Metrics involving length of longest stem–arm, MFE, internal loops, asymmetric loops and bulges applied to predicted structures in aligned regions	Typical divergence pattern	

(*Nucl. Acids Res*, 37(8):2419)

Evolutionary features





(*Bioinformatics* 27:i275)

Sequence features only

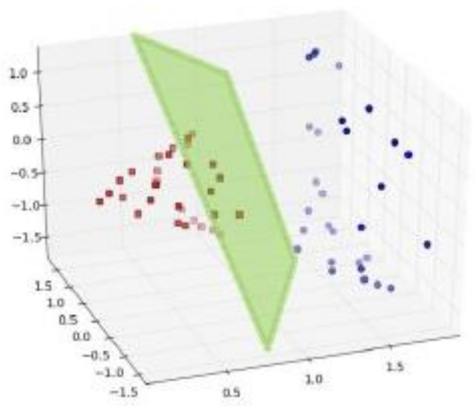
Mechanism neutral: works for both long and small ncRNAs

Accurate and Fast

SVM classifier

- **SVM – support vector machine**

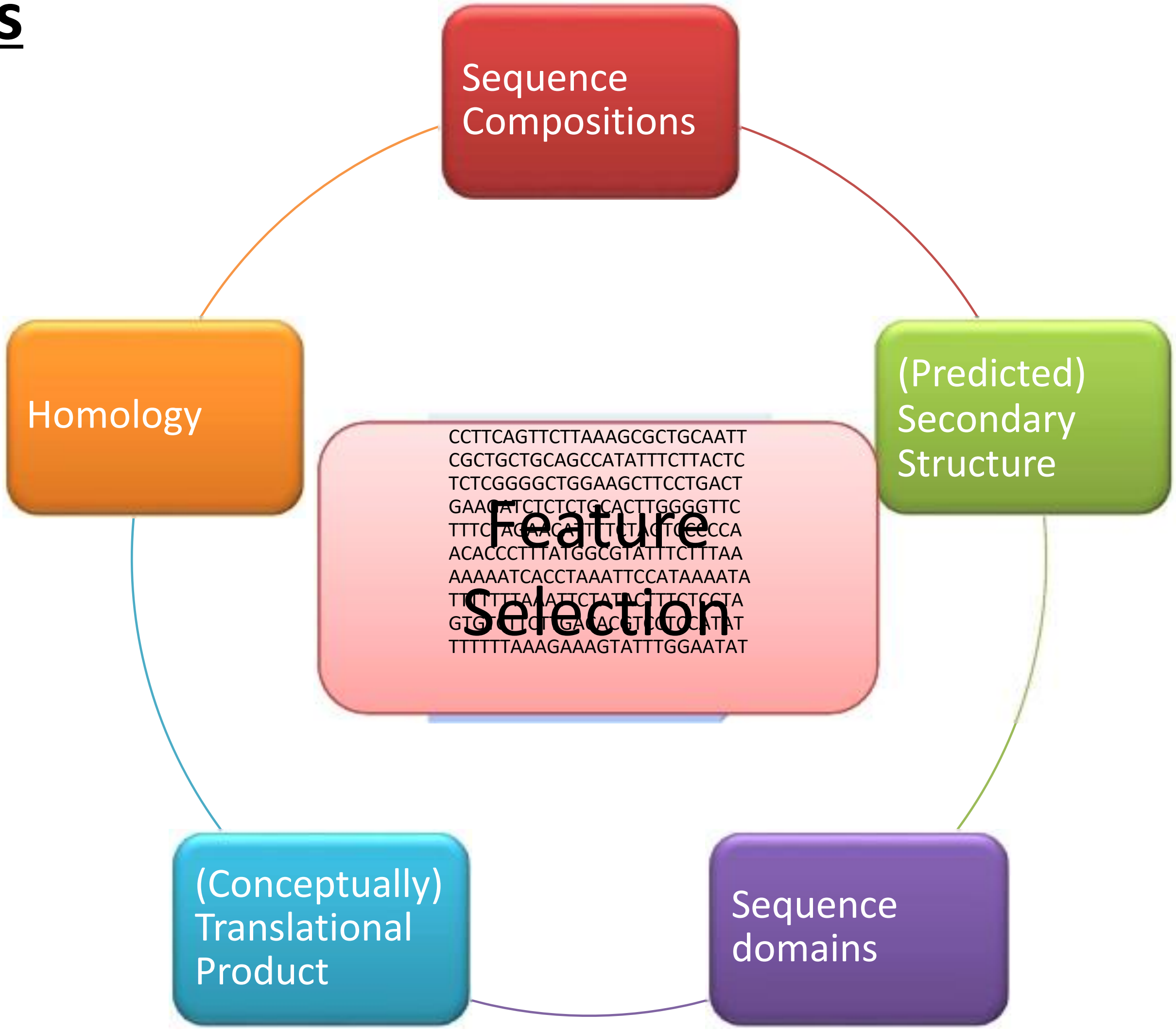
  Separate transformed data with a hyper plane in a high-dimensional space



- **Kernel function – Radial Basis Function(RBF)**

- **Grid-search to select proper values of parameter**

Sequence features



Sequence Compositions

(Predicted) Secondary Structure

Sequence domains

(Conceptually) Translational Product

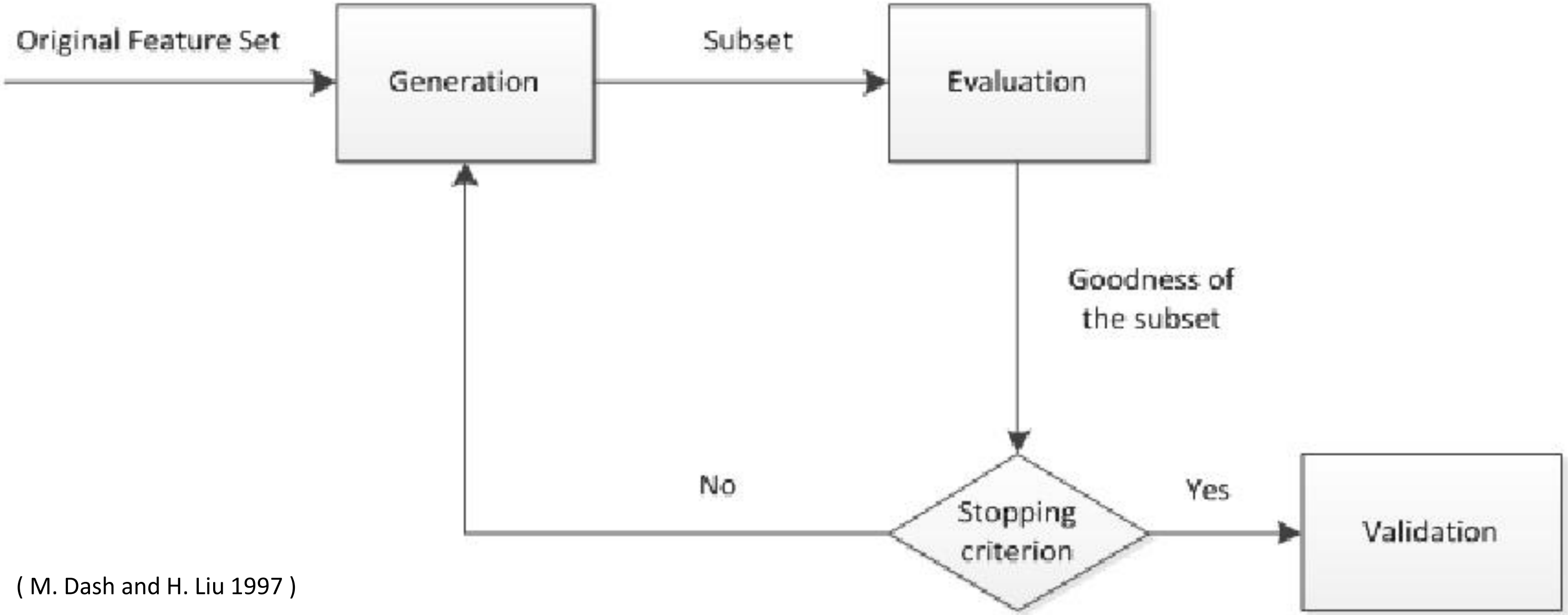
Homology

Feature Selection

CCTTCAGTTCTTAAAGCGCTGCAATT
CGCTGCTGCAGCCATATTTCTTACTC
TCTCGGGGCTGGAAGCTTCCTGACT
GAACATCTCTCTGCACTTGGGGTTC
TTTCTAGAACATTTTCTACTCCCCA
ACACCCTTTATGGCGTATTTCTTTAA
AAAAATCACCTAAATTCCATAAAATA
TTCTTTTAAATTCTATACTTTCTCCTA
GTGTCTTCTTGACACGTCCTCCATAT
TTTTTTAAAGAAAGTATTTGGAATAT

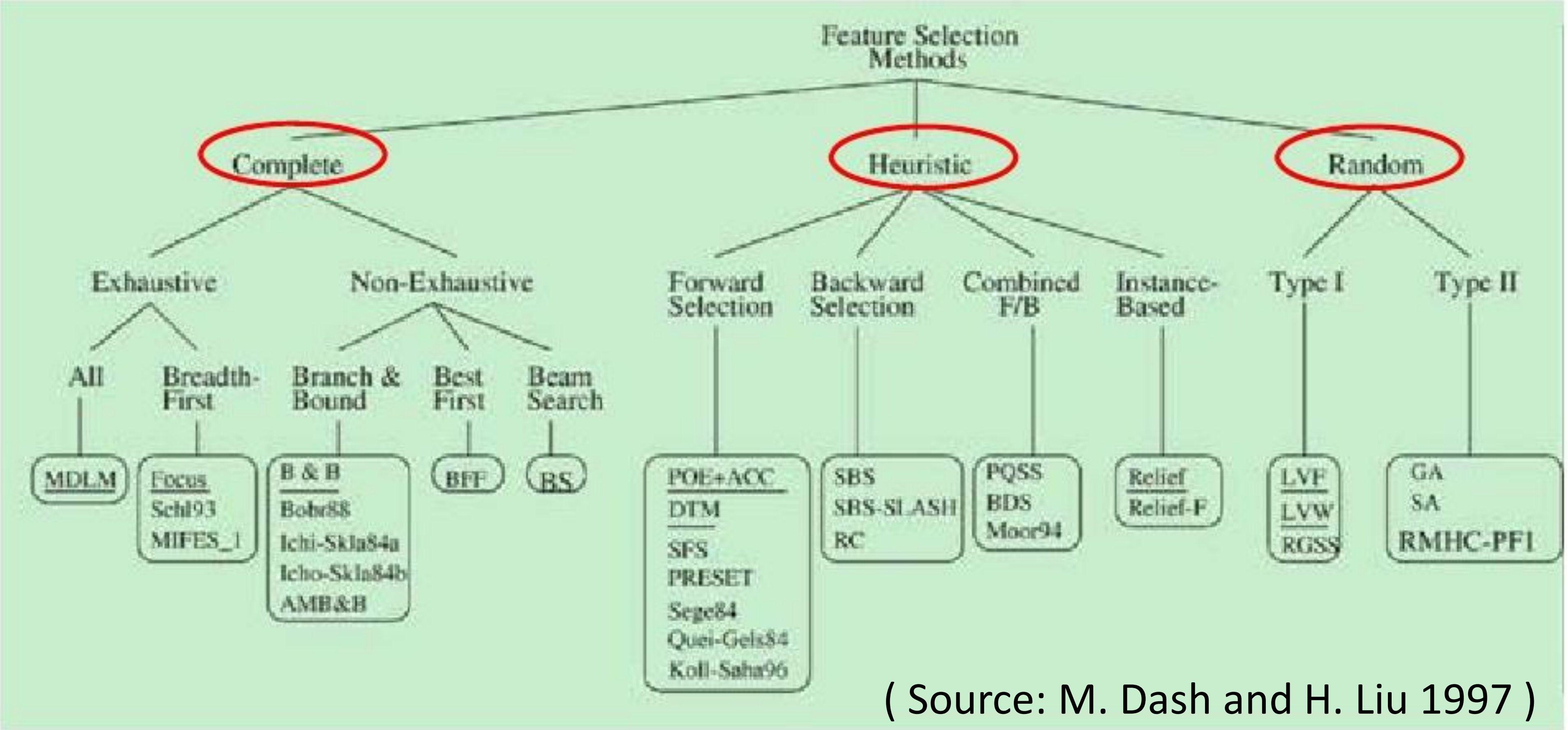
Feature Selection

Purpose： Choose the best feature set in term of accuracy, speed, and computing space

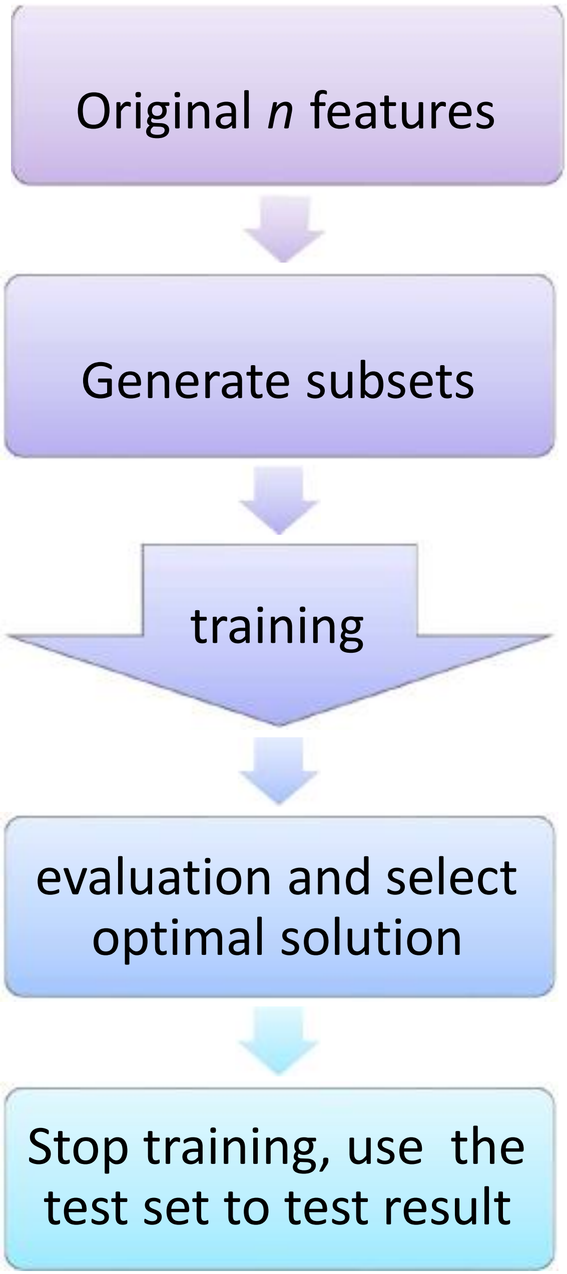


(M. Dash and H. Liu 1997)

Find The Optimal Subset



(Source: M. Dash and H. Liu 1997)



Original *n* features

Generate subsets

training

evaluation and select optimal solution

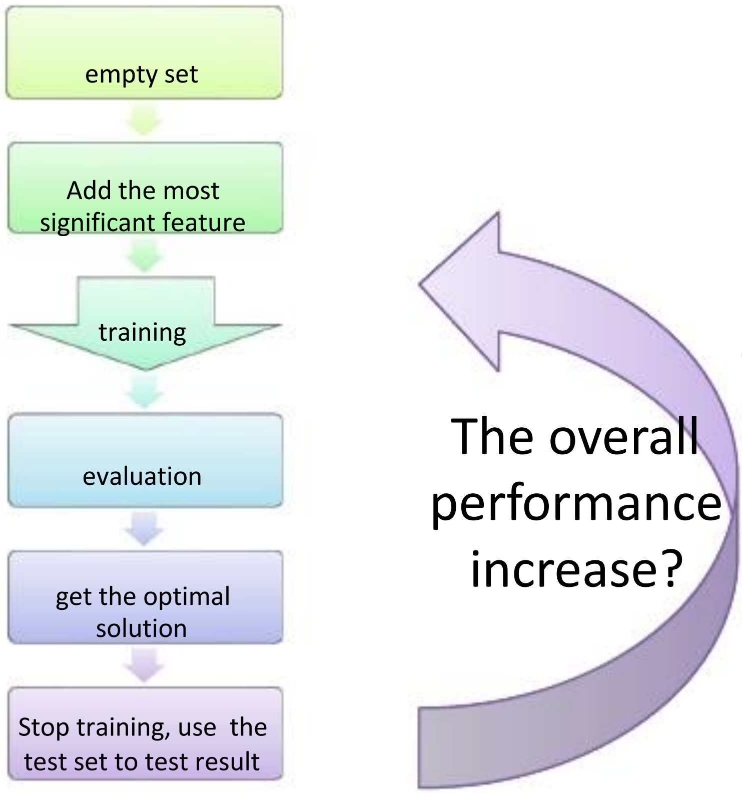
Stop training, use the test set to test result

Complete Search: Breadth First

The breadth-first traversal of all variables

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{n-1} + \binom{n}{n}$$



empty set

Add the most significant feature

training

evaluation

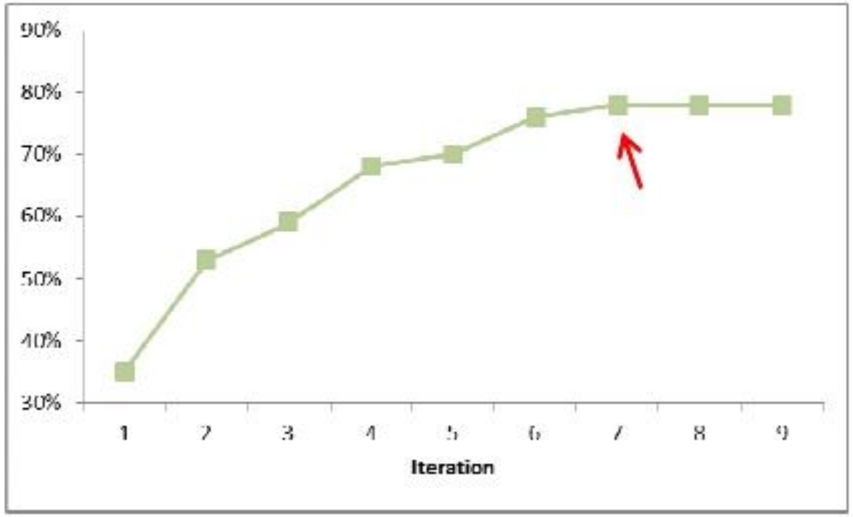
get the optimal solution

Stop training, use the test set to test result

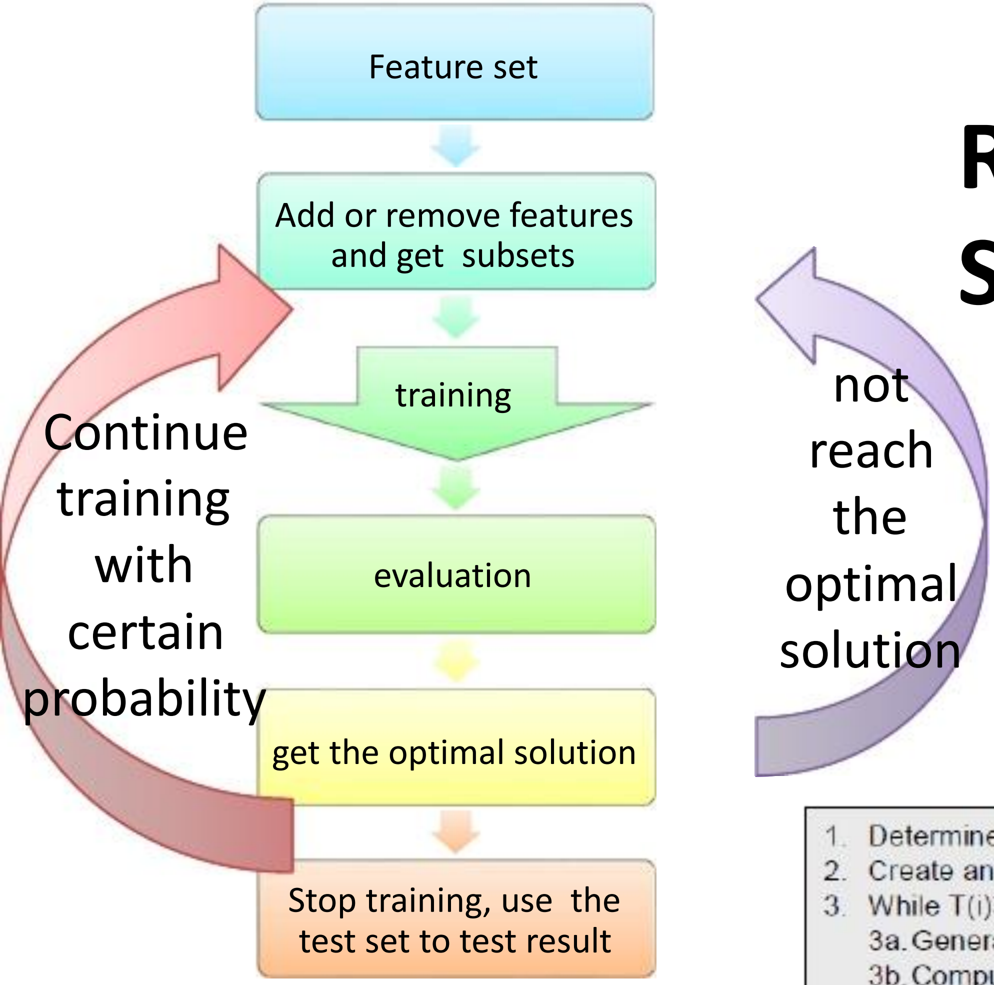
The overall performance increase?

Heuristic Search: Sequential Forward Selection

Features added greedily until the addition of further features does not increase the overall performance.



Random Search: Simulated Annealing



Feature set

Add or remove features and get subsets

training

evaluation

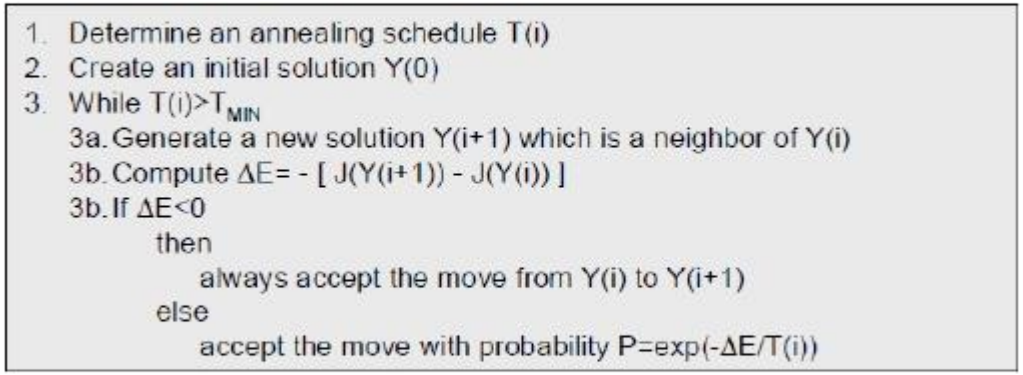
get the optimal solution

Stop training, use the test set to test result

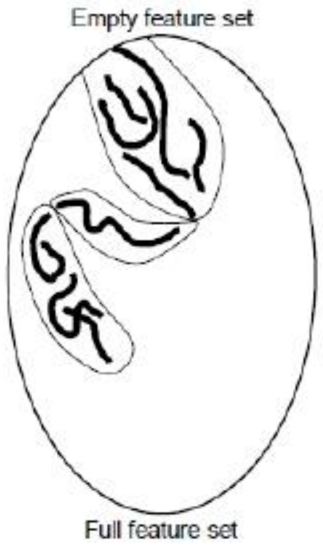
Continue training with certain probability

not reach the optimal solution

adding or removing features based on an "annealing-like" probability

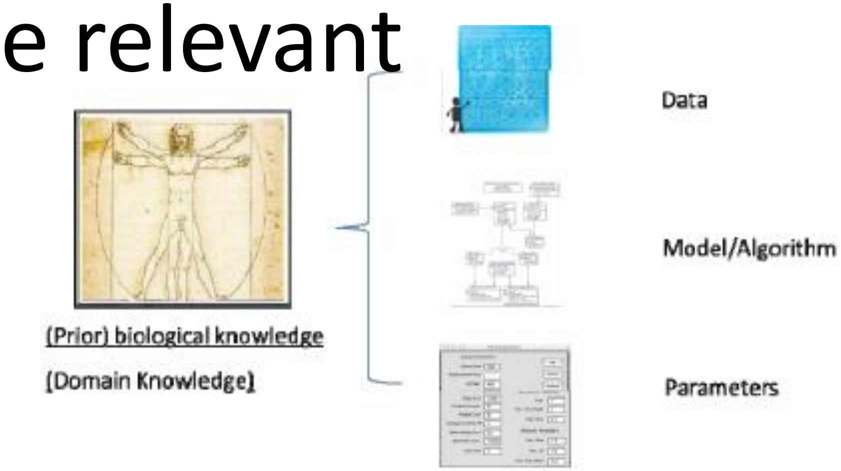


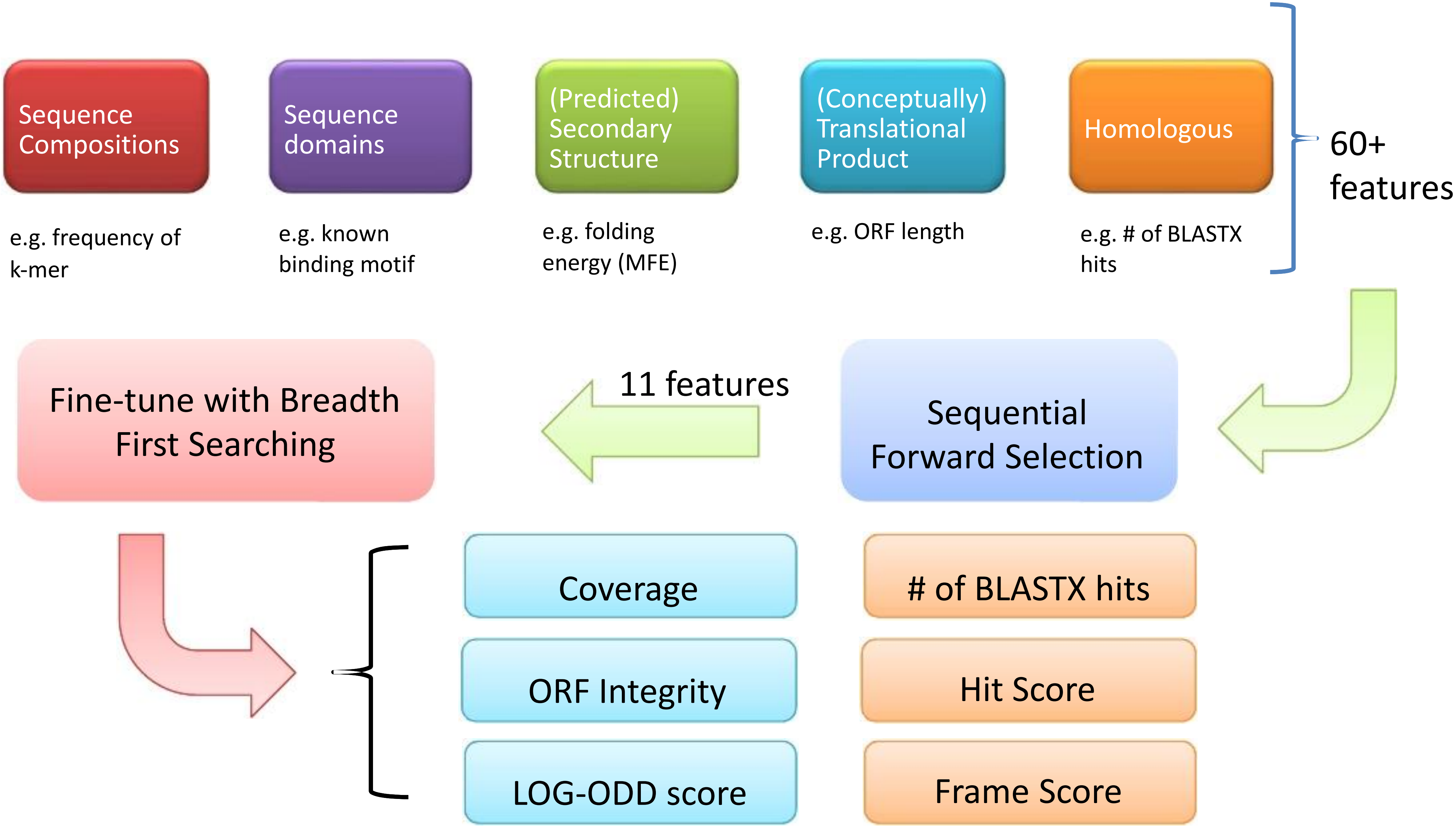
```
1. Determine an annealing schedule T(i)
2. Create an initial solution Y(0)
3. While $T(i)>T_{MIN}$
   3a. Generate a new solution Y(i+1) which is a neighbor of Y(i)
   3b. Compute $\Delta E = -[J(Y(i+1)) - J(Y(i))]$
   3b. If $\Delta E<0$
         then
            always accept the move from Y(i) to Y(i+1)
         else
            accept the move with probability $P=\exp(-\Delta E/T(i))$
```



Initialized feature set

- Properties of entity
- Speculate based on existed knowledge
- Certain statistic established by predecessors
- The data that is thought to be relevant





Sequence Compositions

e.g. frequency of k-mer

Sequence domains

e.g. known binding motif

(Predicted) Secondary Structure

e.g. folding energy (MFE)

(Conceptually) Translational Product

e.g. ORF length

Homologous

e.g. # of BLASTX hits

60+ features

Sequential Forward Selection

11 features

Fine-tune with Breadth First Searching

Coverage

ORF Integrity

LOG-ODD score

# of BLASTX hits

Hit Score

Frame Score

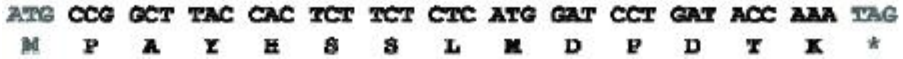
(Conceptually) Translated Product

Coverage

$$Coverage(S) = \frac{L_{ORF} - (L_{mismatch} + 2 * L_{frameshift})}{\text{Total Length}}$$

ORF Integrity

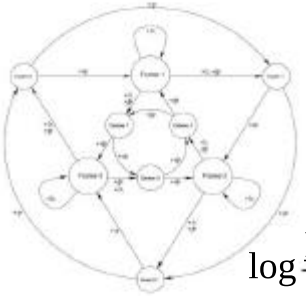
indicates whether the predicted ORF begins with a start codon and ends with an in-frame stop codon



```
ATG CCG GCT TAC CAC TCT TCT CTC ATG GAT CCT GAT ACC AAA TAG
 M   P   A   Y   H   S   S   L   M   D   P   D   T   K   *
```

LOG-ODD score

indicator of the quality of a predicted ORF. The higher the score, the better the quality of the ORF



$$\log \frac{\Pr(D \mid M)}{\Pr(D \mid R)}$$

(*Nucleic Acids Res.* 35:W345)

Homologous

# of BLASTX hits

A true protein-coding transcript is likely to have more hits with known proteins than a non-coding transcript does

Hit Score

For a true protein-coding transcript, the hits are also likely to have higher quality

$$S_i = \operatorname*{mean}_{j}\{-\log_{10} E_{ij}\}, \quad i \in [0,1,2]$$

$$\text{HIT SCORE} = \operatorname*{mean}_{i\in\{0,1,2\}}\{S_i\} = \frac{\sum_{i=0}^{2} Si}{3},$$

Frame Score

For a true protein-coding transcript, most of the hits are likely to reside within one frame, whereas for a true non-coding transcript, even if it matches certain known protein sequence segments by chance, these chance hits are likely to scatter in any of the three frames

$$\text{FRAME SCORE} = \operatorname*{variance}_{i\in\{0,1,2\}}\{S_i\} = \frac{\sum_{i=0}^{2}(S_i - \bar{S})^2}{2}$$

(*Nucleic Acids Res.* 35:W345)

Coverage

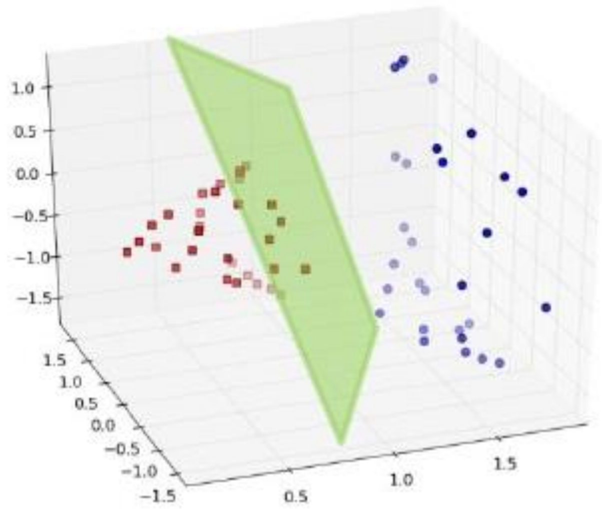
ORF Integrity

LOG-ODD score

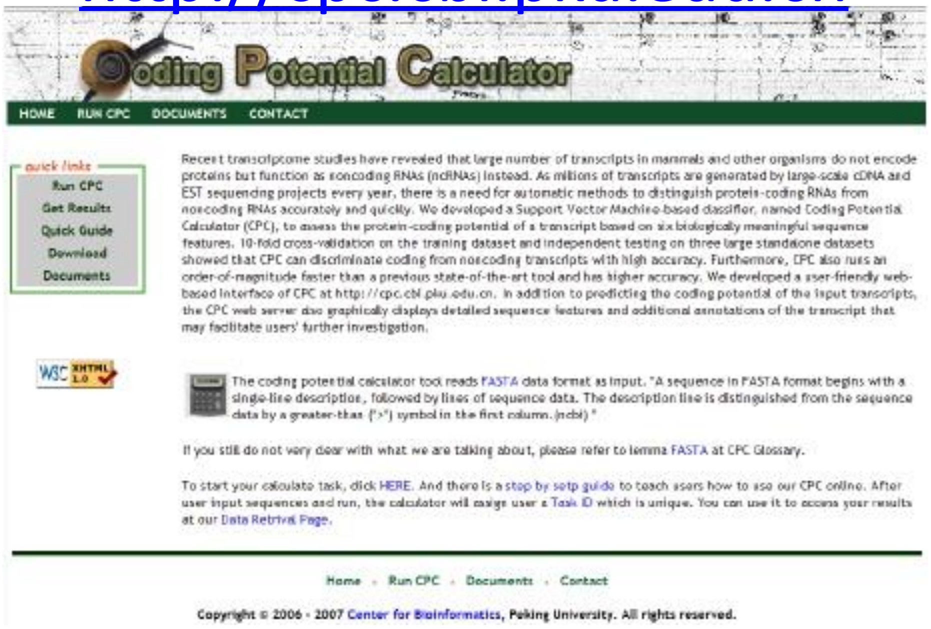
# of BLASTX hits

Hit Score

Frame Score



http://cpc.cbi.pku.edu.cn



Recent transcriptome studies have revealed that large number of transcripts in mammals and other organisms do not encode proteins but function as noncoding RNAs (ncRNAs) instead. As millions of transcripts are generated by large-scale cDNA and EST sequencing projects every year, there is a need for automatic methods to distinguish protein-coding RNAs from noncoding RNAs accurately and quickly. We developed a Support Vector Machine-based classifier, named Coding Potential Calculator (CPC), to assess the protein-coding potential of a transcript based on six biologically meaningful sequence features. 10-fold cross-validation on the training dataset and independent testing on three large standalone datasets showed that CPC can discriminate coding from noncoding transcripts with high accuracy. Furthermore, CPC also runs an order-of-magnitude faster than a previous state-of-the-art tool and has higher accuracy. We developed a user-friendly web-based interface of CPC at http://cpc.cbi.pku.edu.cn. In addition to predicting the coding potential of the input transcripts, the CPC web server also graphically displays detailed sequence features and additional annotations of the transcript that may facilitate users' further investigation.

The coding potential calculator tool reads FASTA data format as input. "A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (">") symbol in the first column. (ncbi) "

If you still do not very clear with what we are talking about, please refer to lemma FASTA at CPC Glossary.

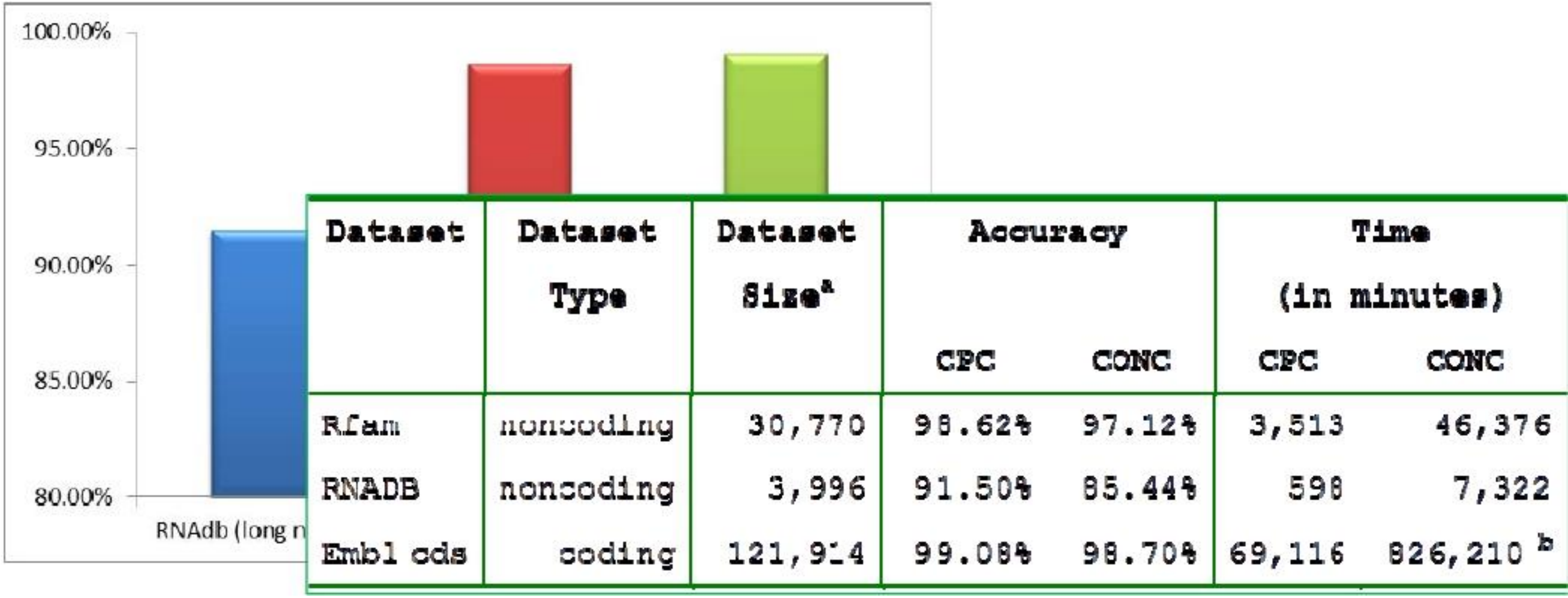
To start your calculate task, click HERE. And there is a step by setp guide to teach users how to use our CPC online. After user input sequences and run, the calculator will asign user a Task ID which is unique. You can use it to access your results at our Data Retrival Page.

Home · Run CPC · Documents · Contact

Copyright © 2006 - 2007 Center for Bioinformatics, Peking University. All rights reserved.

Coding Potential Calculator

http://cpc.cbi.pku.edu.cn



Dataset	Dataset Type	Dataset Size[a]	Accuracy		Time (in minutes)	
			CPC	CONC	CPC	CONC
Rfam	noncoding	30,770	98.62%	97.12%	3,513	46,376
RNADB	noncoding	3,996	91.50%	85.44%	598	7,322
Embl cds	coding	121,914	99.08%	98.70%	69,116	826,210 [b]

(*Nucleic Acids Res.* 35:W345)

Coding Potential Calculator

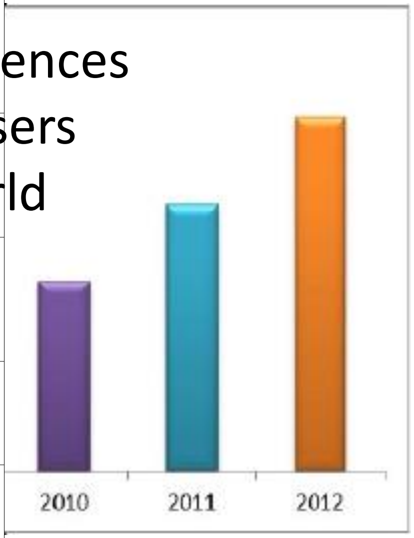
http://cpc.cbi.pku.edu.cn

Gene Regulation	FunctionofncRNA	HVanBakel*etal*.**PLoSBiology**,2010
	LongncRNA	HJia*etal*.,**RNA**,2010 TGBelard*etal*.,**Neuron**,2011 IUlitsky*etal*.**Cell**,2011 RSYoung*etal*.**GenomeBiolEvol**,2012
	ShortPeptide	XYang*etal*.,**GenomeRes**,2011
StemCell	Self-Renewal	JSMohamed*etal*.,**RNA**,2010
	Neurondevelopment	SYNg*etal*.,**EMBOJournal**,2011
Disease	Heartdiseases	JHLee*etal*.,**CircRes**,2011
	CancerMarker	BPMello*etal*.,**NucleicAcidRes**,2009
	Tumormechanism	ACTahira*etal*.,**MolecularCancer**,2011 RJFlockhart*etal*.,**GenomeRes**,2012
Evolution	Newgenes	DRose*etal*.,**JBioinformComptBio.**,2008 JFSousa*etal*.,**PLoSOne**,2010
	Functiondivergence ofduplicatedgenes	JTWang*etal*.,**BMCGenomics**,2012

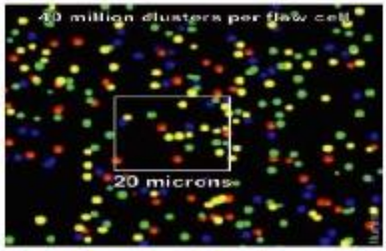
32 million sequences

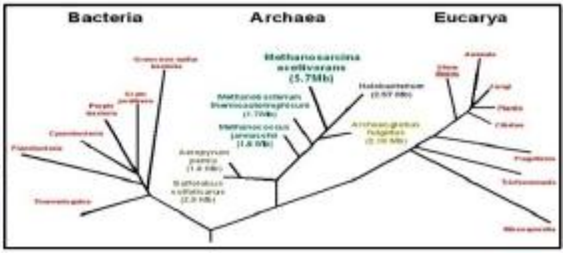
from 50000+ users

around the world



```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```

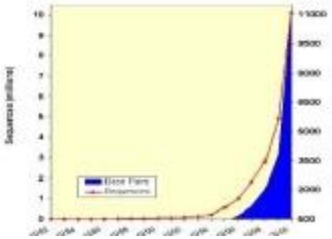


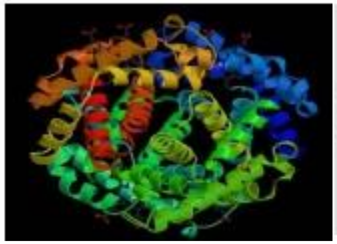


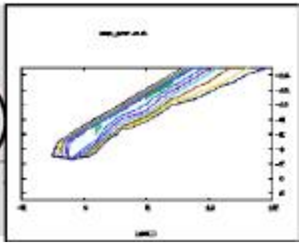
Unit 3:
Data Mining: Differential Expression and Clustering

Le Zhang, Ph. D.

Computer Science Department
Southwest 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 functions when associated with a protein complex.
- In plants miRNAs can silence gene activity via destruction of homologous mRNA or blocking its translation. In animals, miRNAs inhibit translation by binding with imperfect homology to the 3’ untranslated region of mRNA.

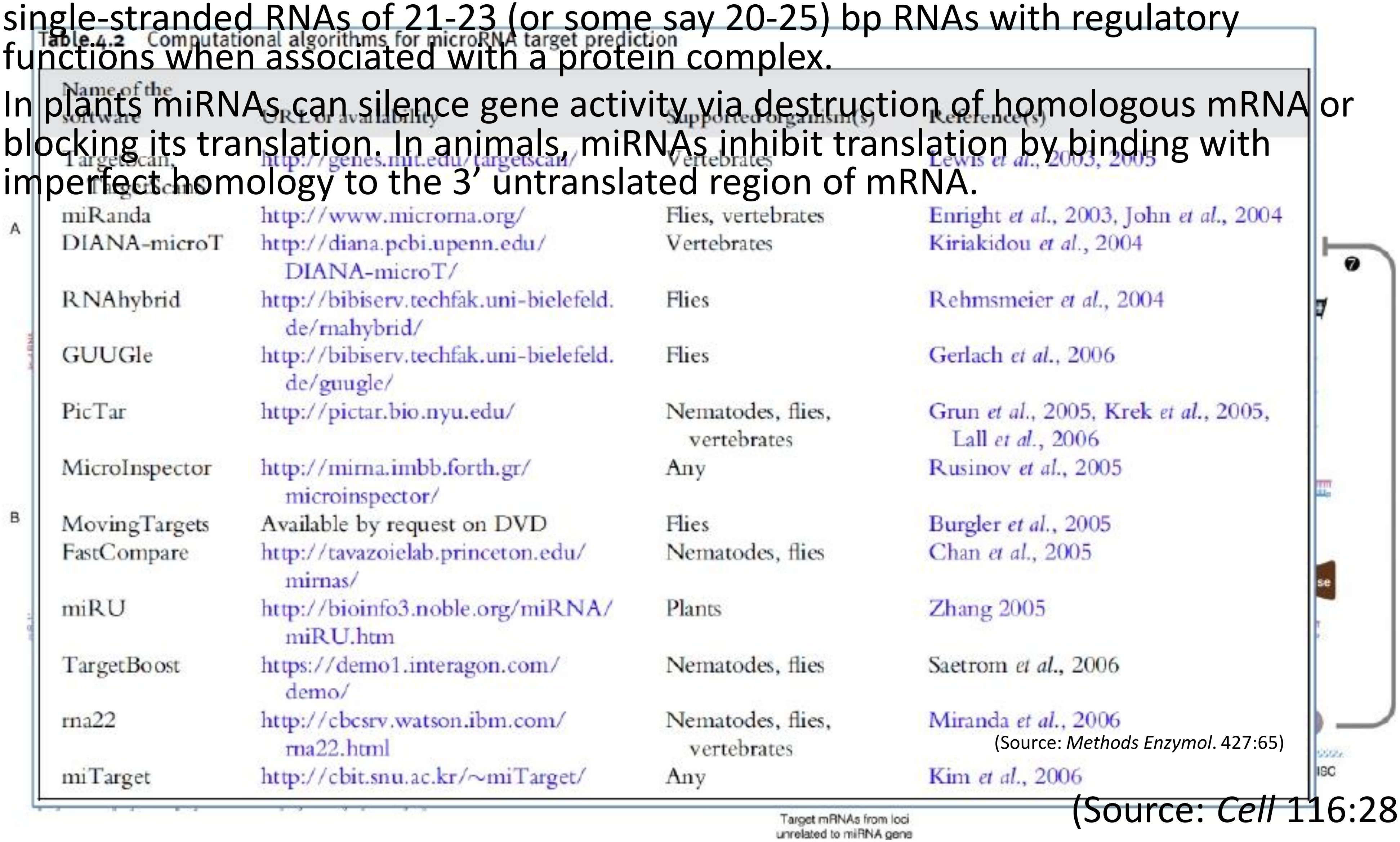


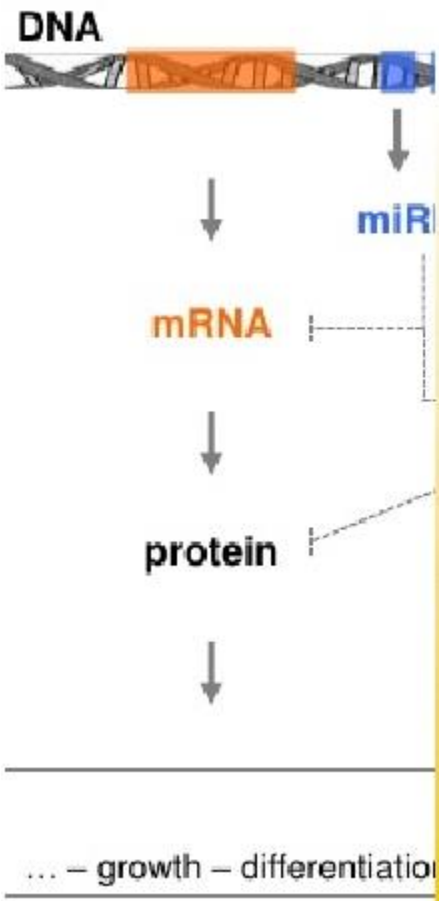
Table 4.2 Computational algorithms for microRNA target prediction

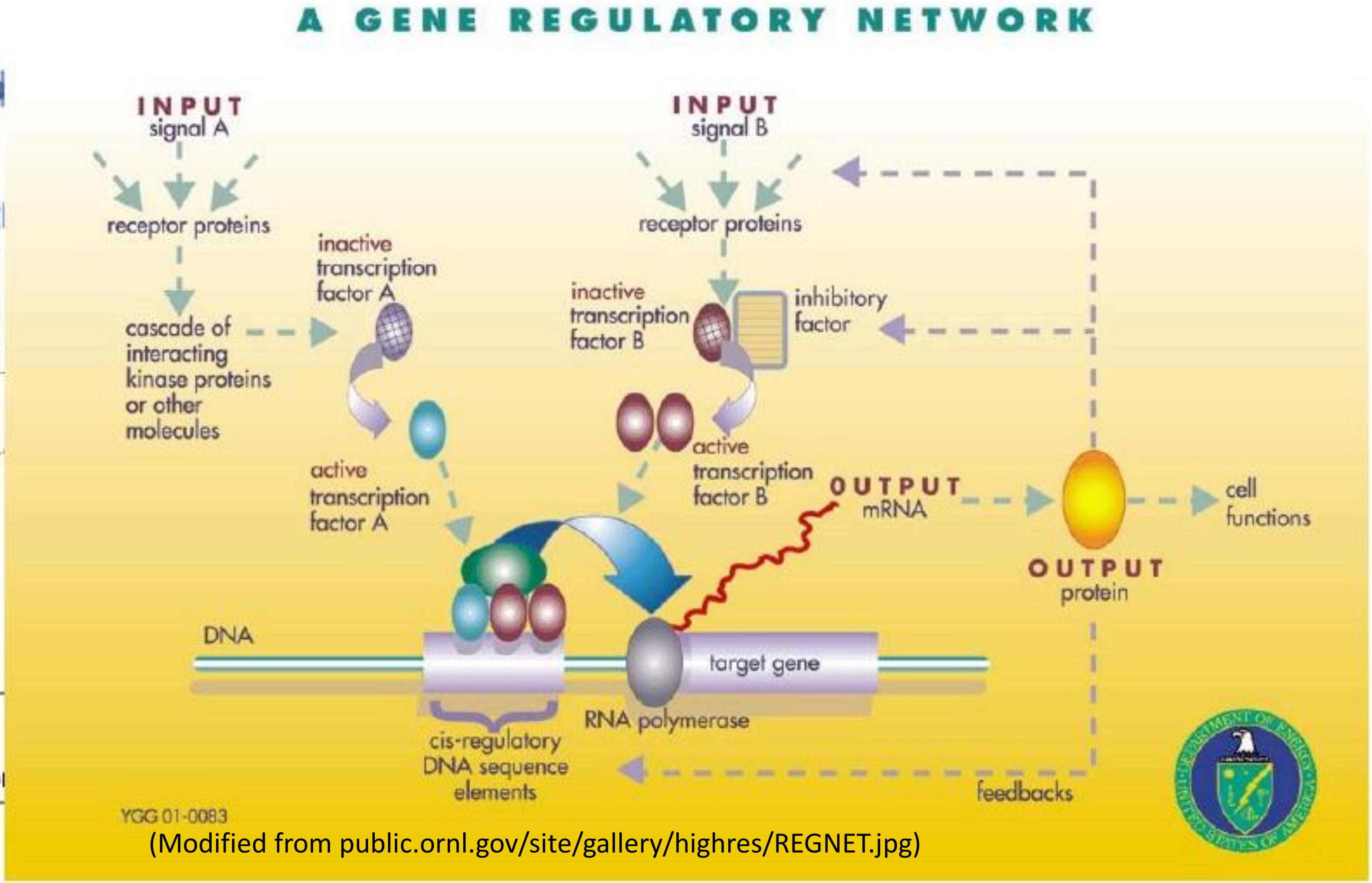
Name of the software	URL or availability	Supported organism(s)	Reference(s)
TargetScan, TargetScanS	http://genes.mit.edu/targetscan/	Vertebrates	Lewis *et al.*, 2003, 2005
miRanda	http://www.microrna.org/	Flies, vertebrates	Enright *et al.*, 2003, John *et al.*, 2004
DIANA-microT	http://diana.pcbi.upenn.edu/DIANA-microT/	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





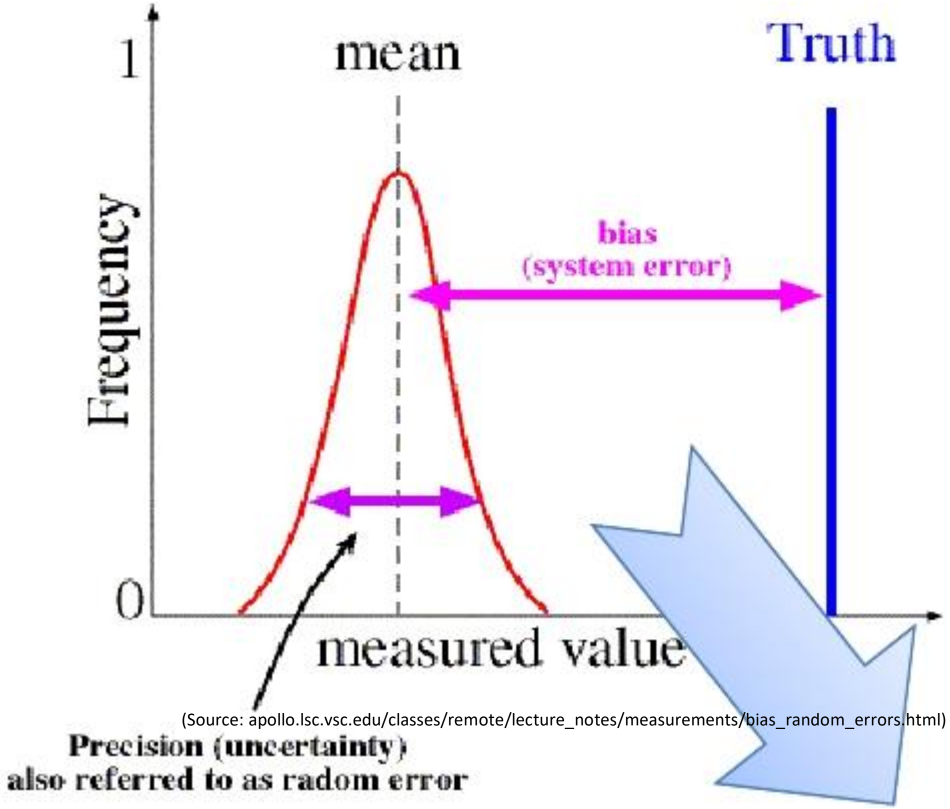
(Modified from public.ornl.gov/site/gallery/highres/REGNET.jpg)

- Differentially expressed genes
- Co-expressed genes

Data Mining: Differentially Expression Calling

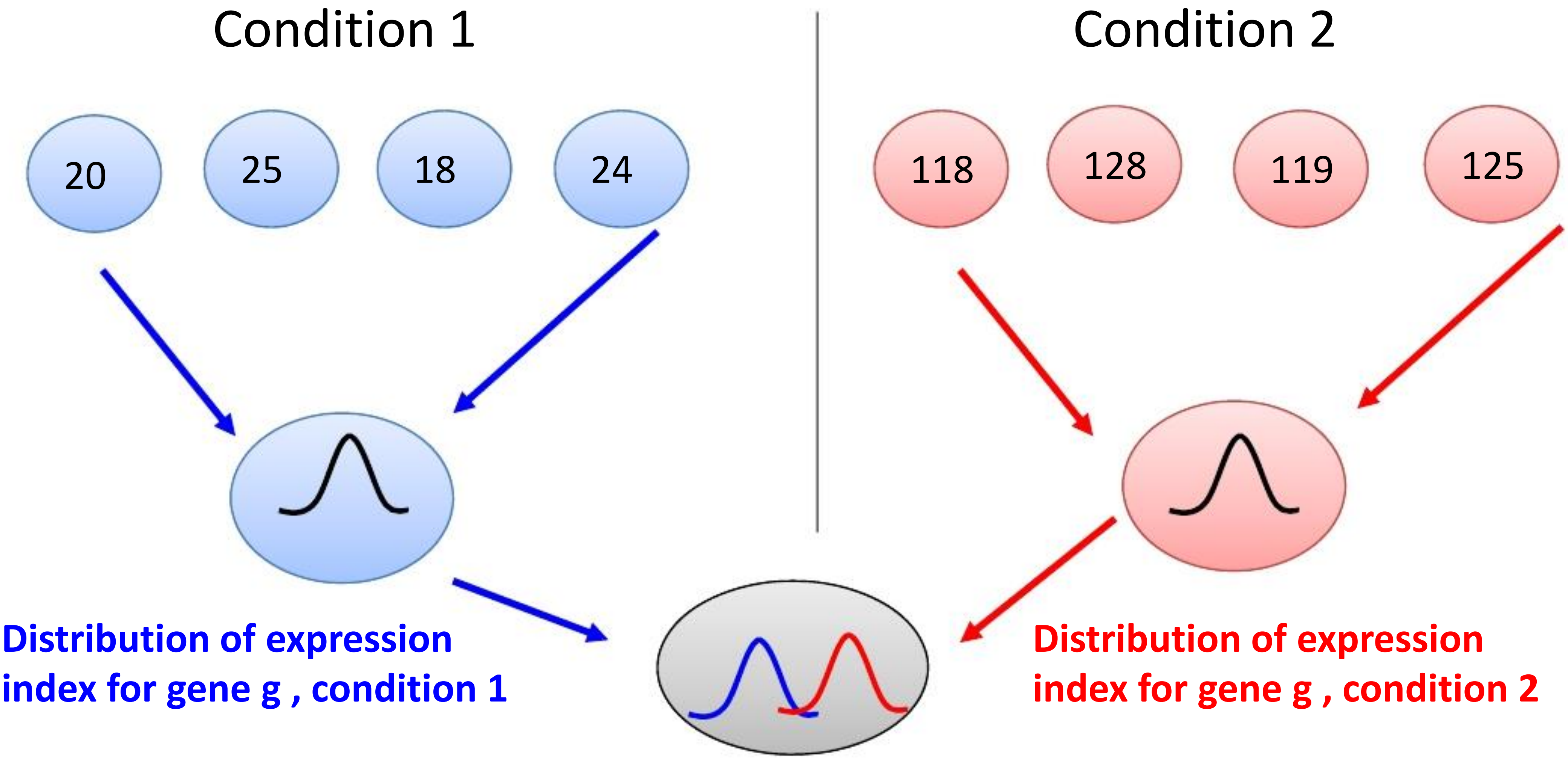
- Identify the genes with biological-significant difference in expression levels across samples

- 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



(Source: apollo.lsc.vsc.edu/classes/remote/lecture_notes/measurements/bias_random_errors.html)

- 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.



Condition 1

20 25 18 24

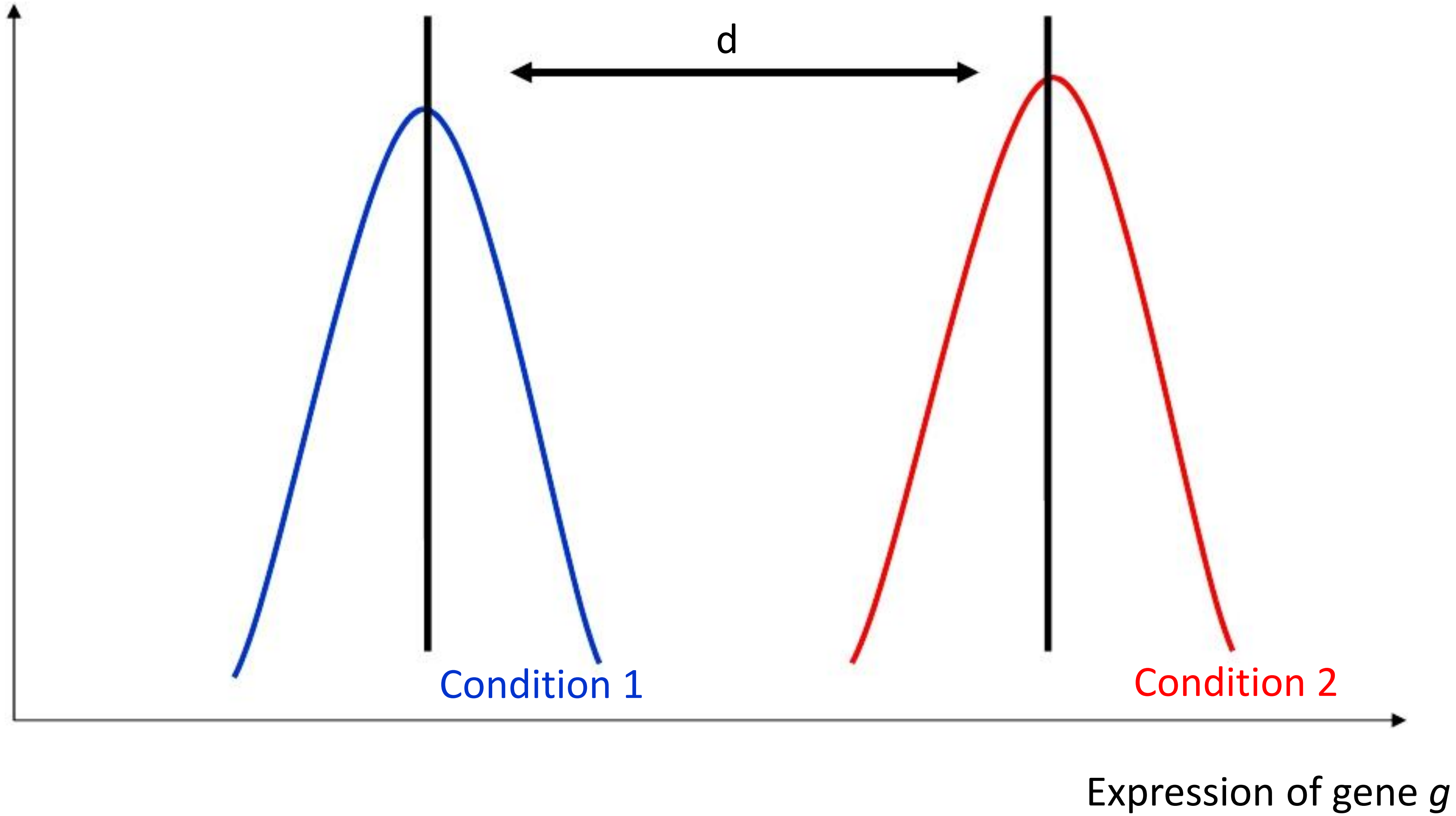
Distribution of expression index for gene g , condition 1

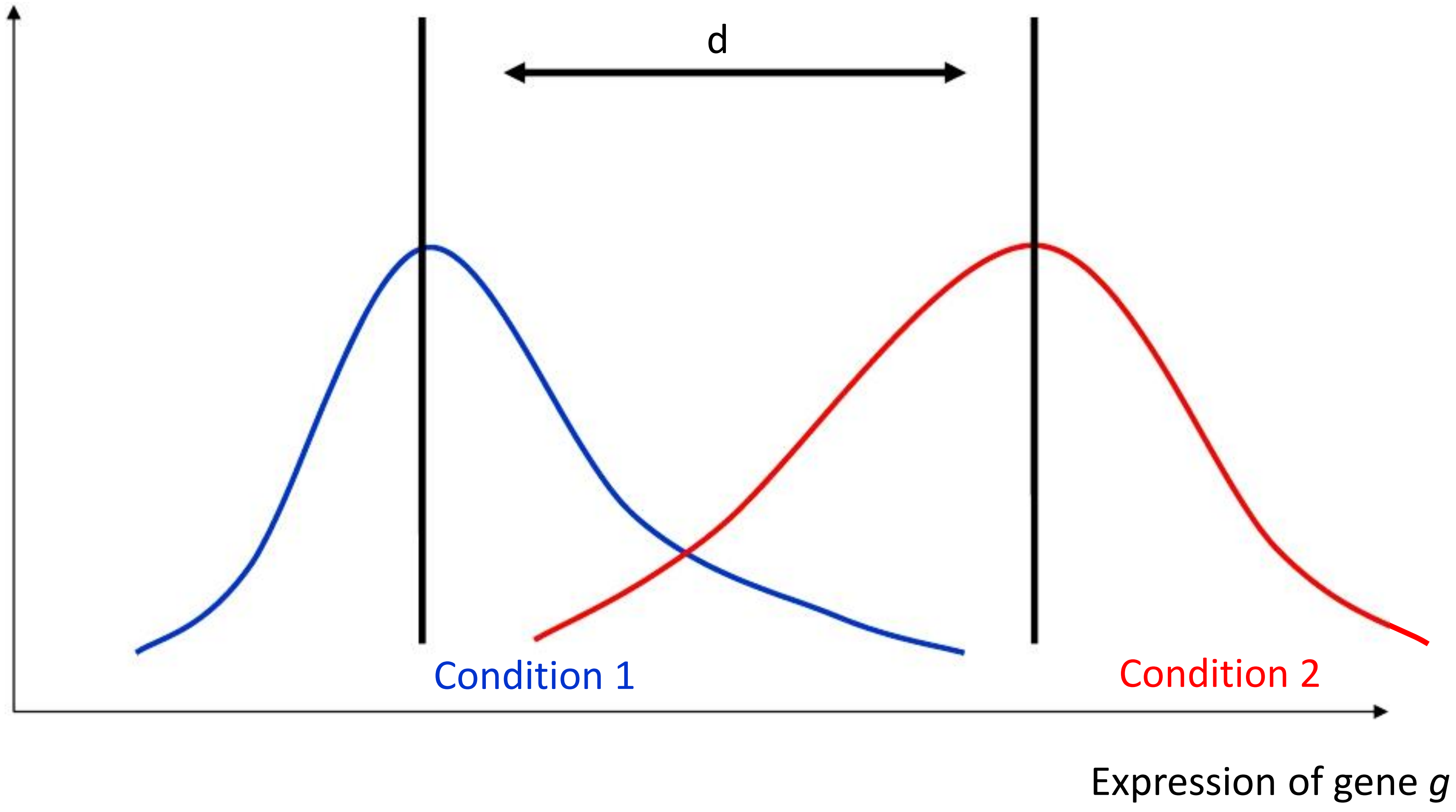
Condition 2

118 128 119 125

Distribution of expression index for gene g , condition 2

Distribution of differential expression statistic


d
Condition 1
Condition 2
Expression of gene *g*


d
Condition 1
Condition 2
Expression of gene *g*

Additional Information

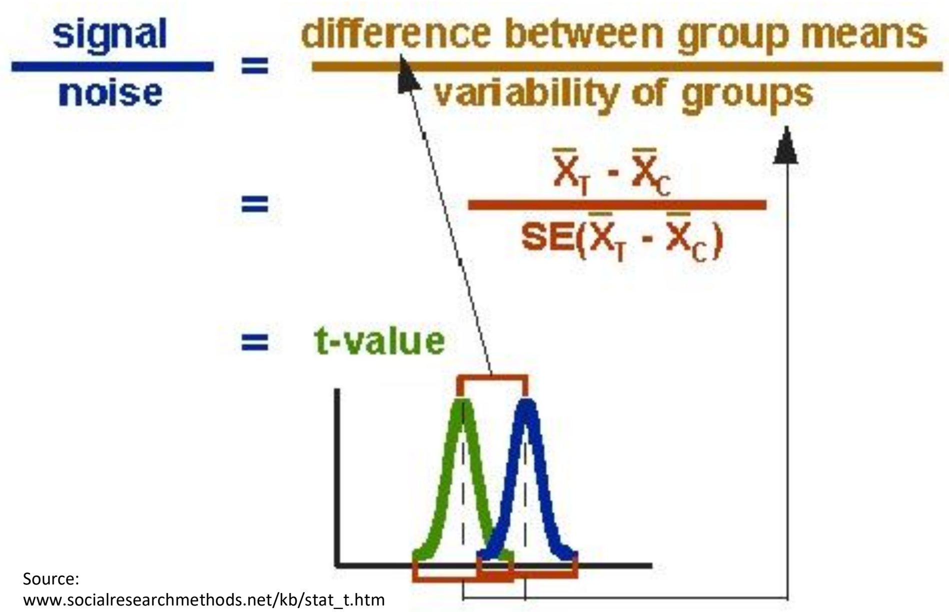
Statistical calling

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”.

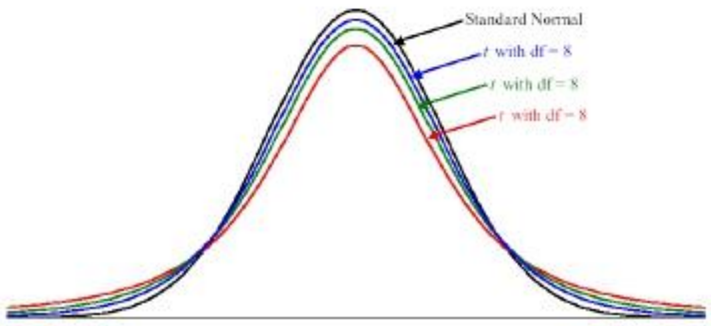
Additional Information



$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)} = \text{t-value}$$

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

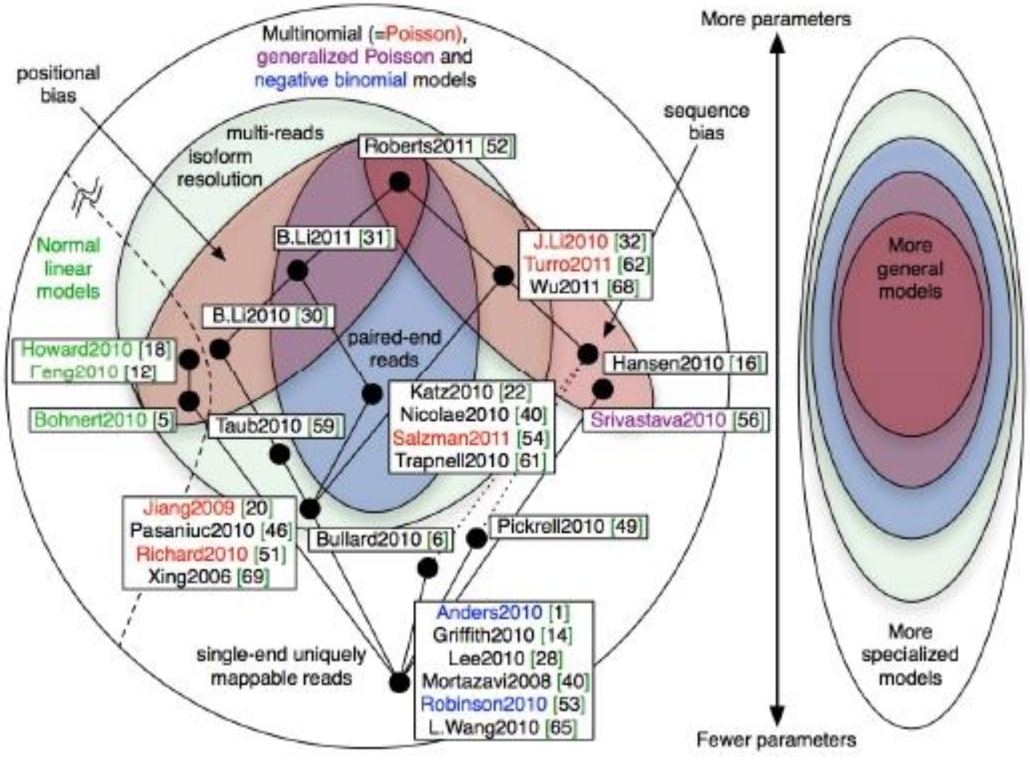
Student's *t*-distribution



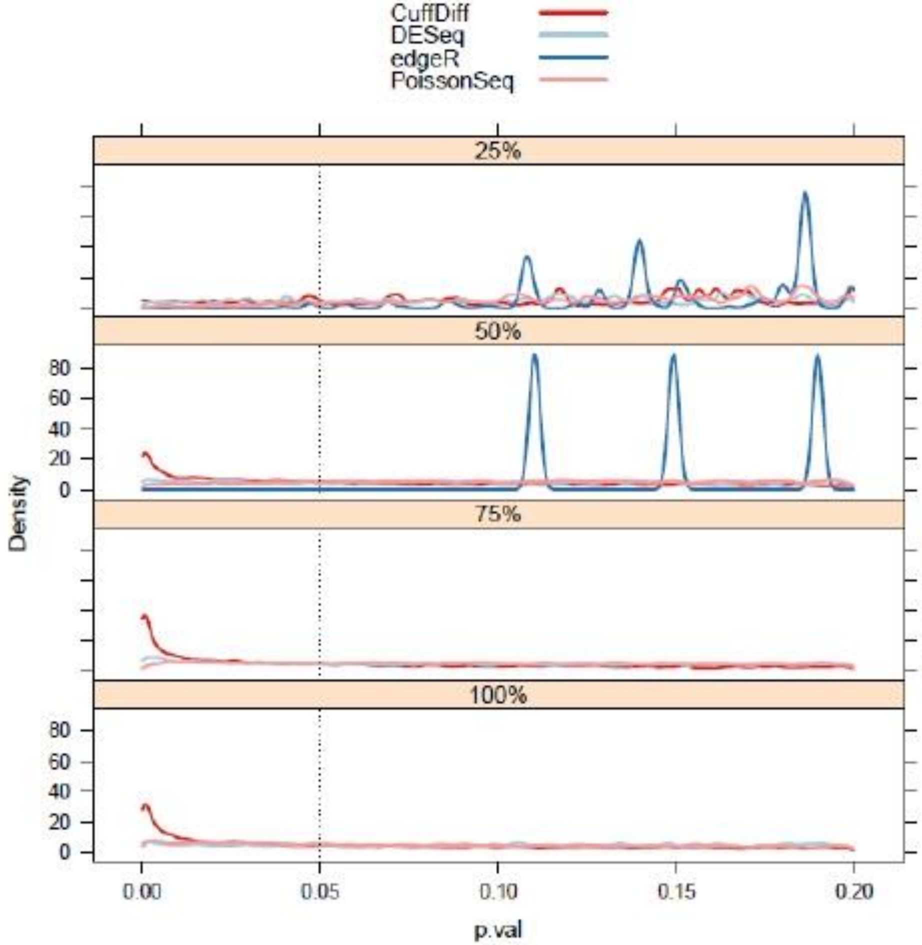
Source: projectile.sv.cmu.edu/research/public/talks/t-test.htm

- 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)

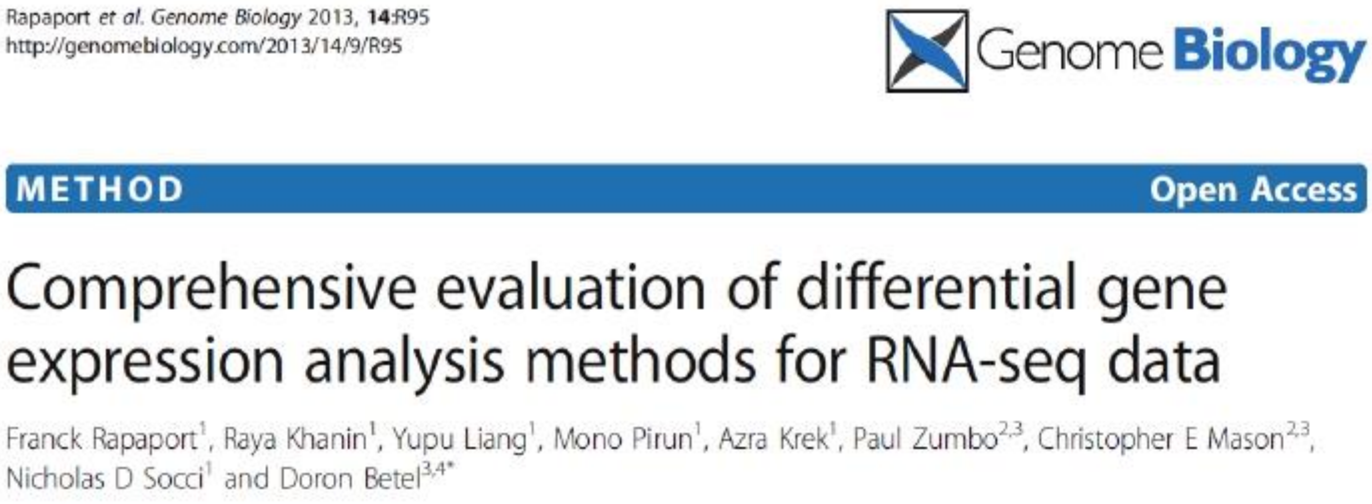
Model the data in RNA-Seq



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



(*Genome Biology* 14:R95)



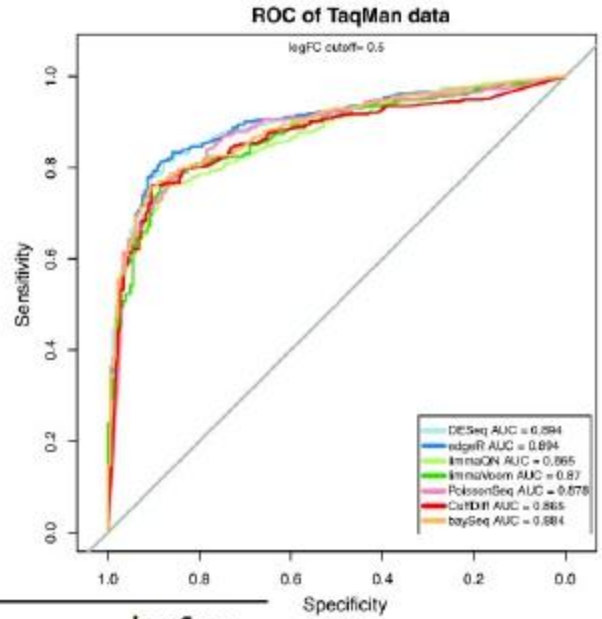
Rapaport *et al. Genome Biology* 2013, **14**:R95
http://genomebiology.com/2013/14/9/R95

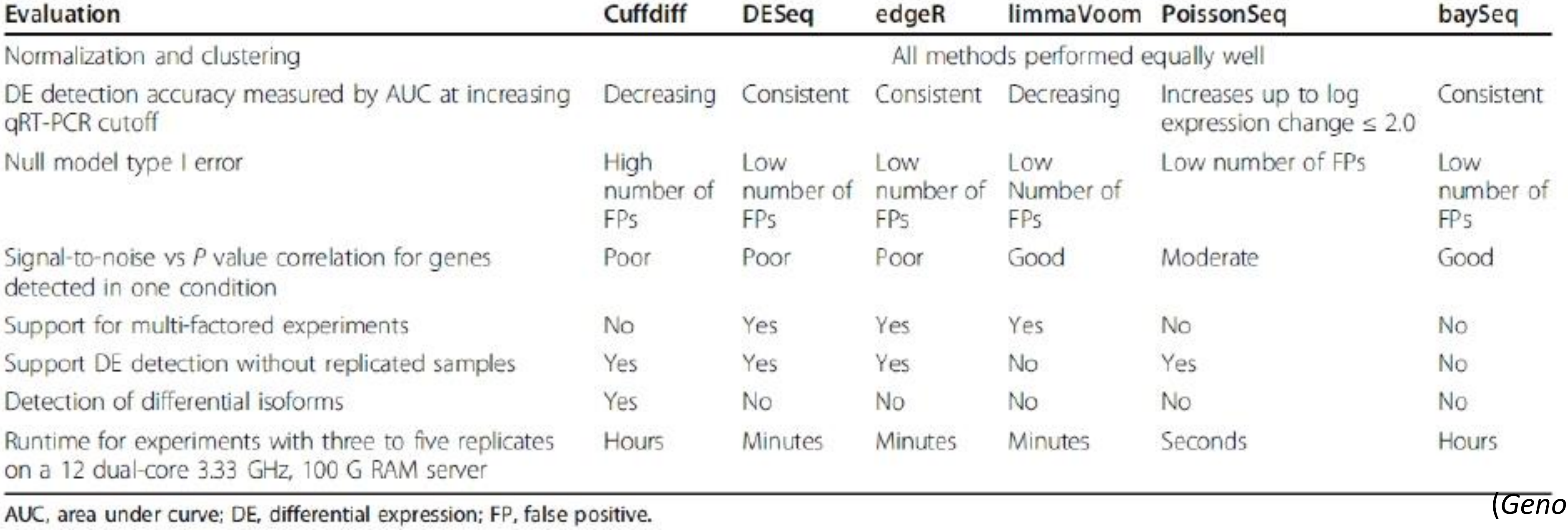
Genome **Biology**

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*]

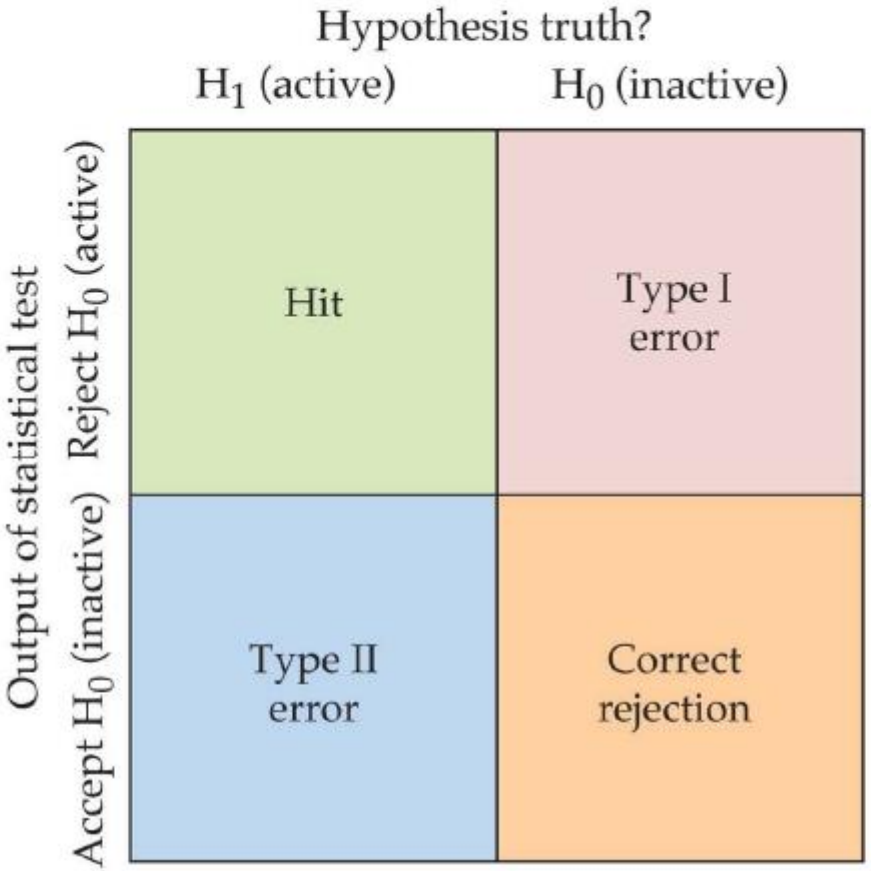

ROC of TaqMan data



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.

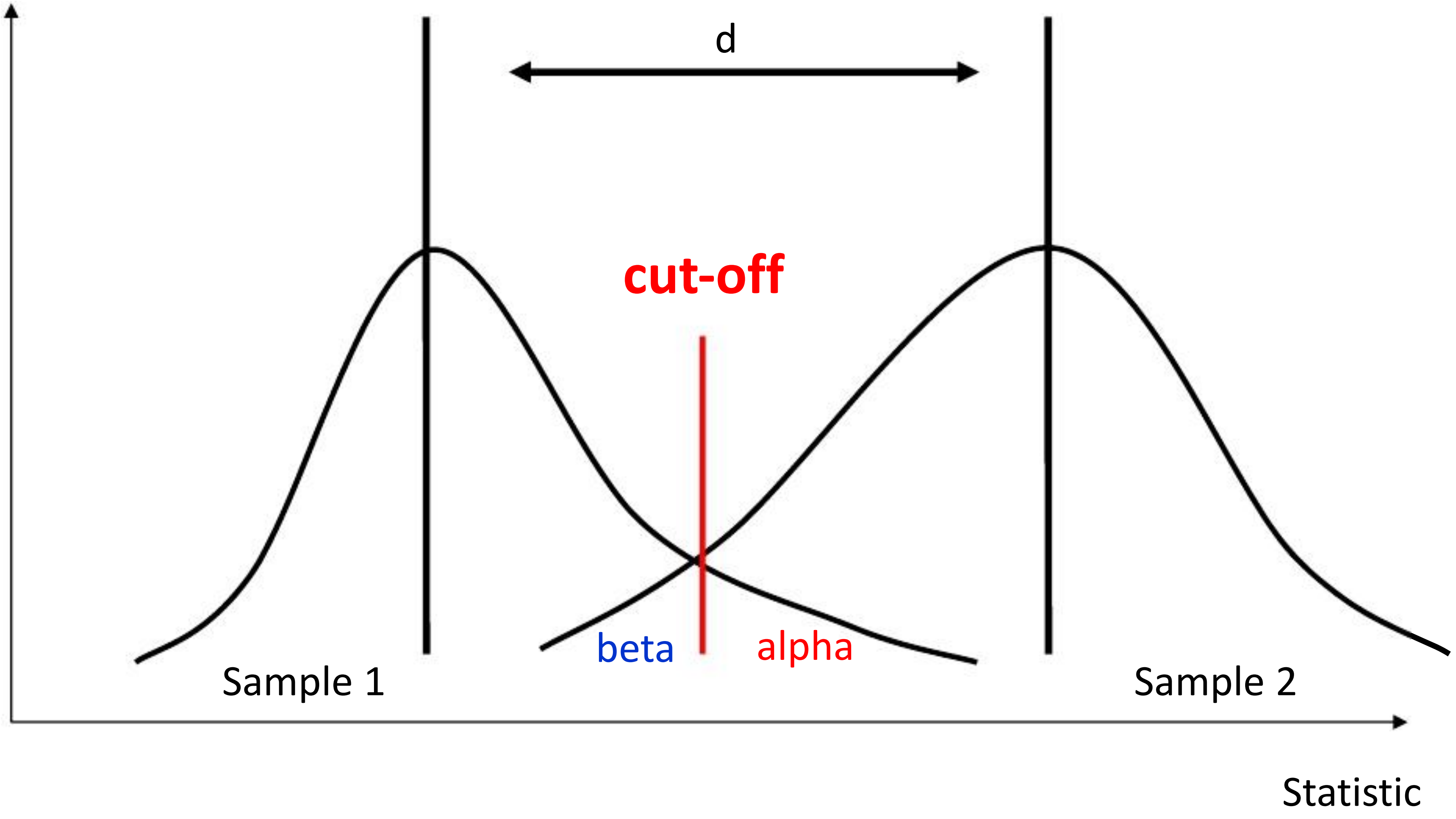
(*Genome Biology* 14:R95)



		Hypothesis truth?	
		H_1 (active)	H_0 (inactive)
Output of statistical test	Reject H_0 (active)	Hit	Type I error
	Accept H_0 (inactive)	Type II error	Correct rejection

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


d
cut-off
beta
alpha
Sample 1
Sample 2
Statistic

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

  FPR = E(V/m0)

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

[redacted]	#not rejected	#rejected	totals
#trueH	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”.

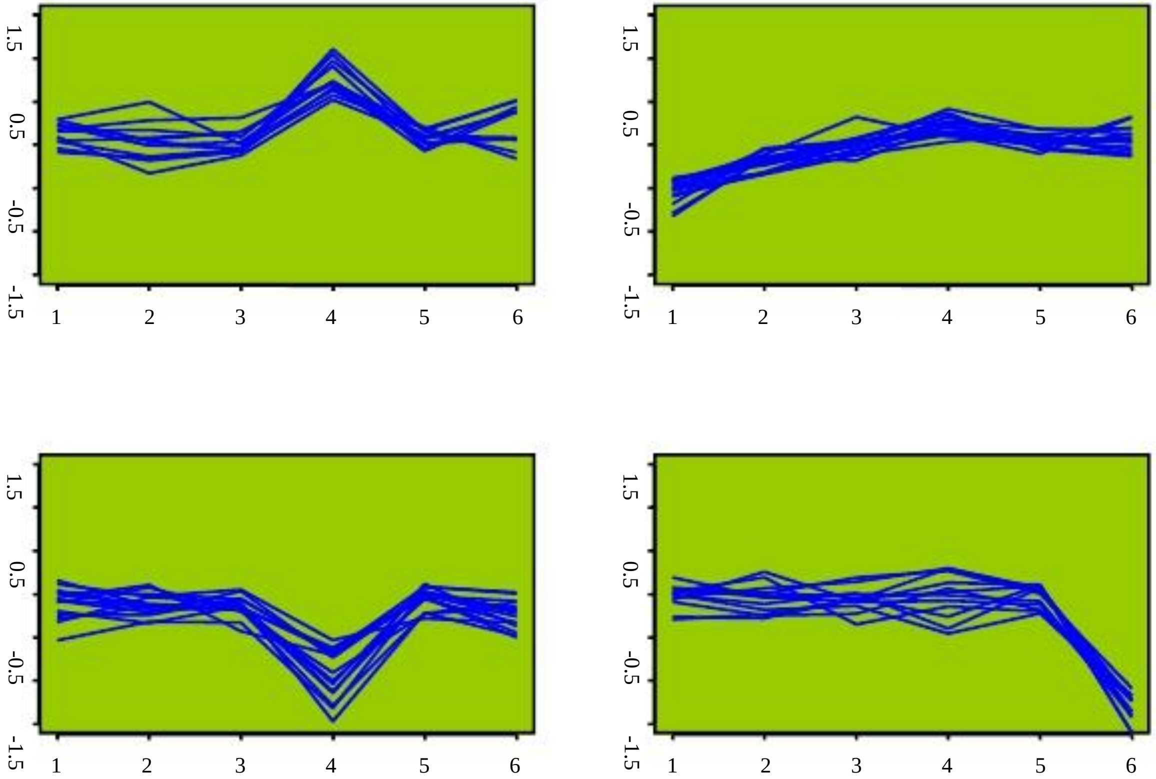
- In contrast, the p-value for a particular gene *g* is the probability that a randomly generated expression profile would be as or more extremely differentially expressed.

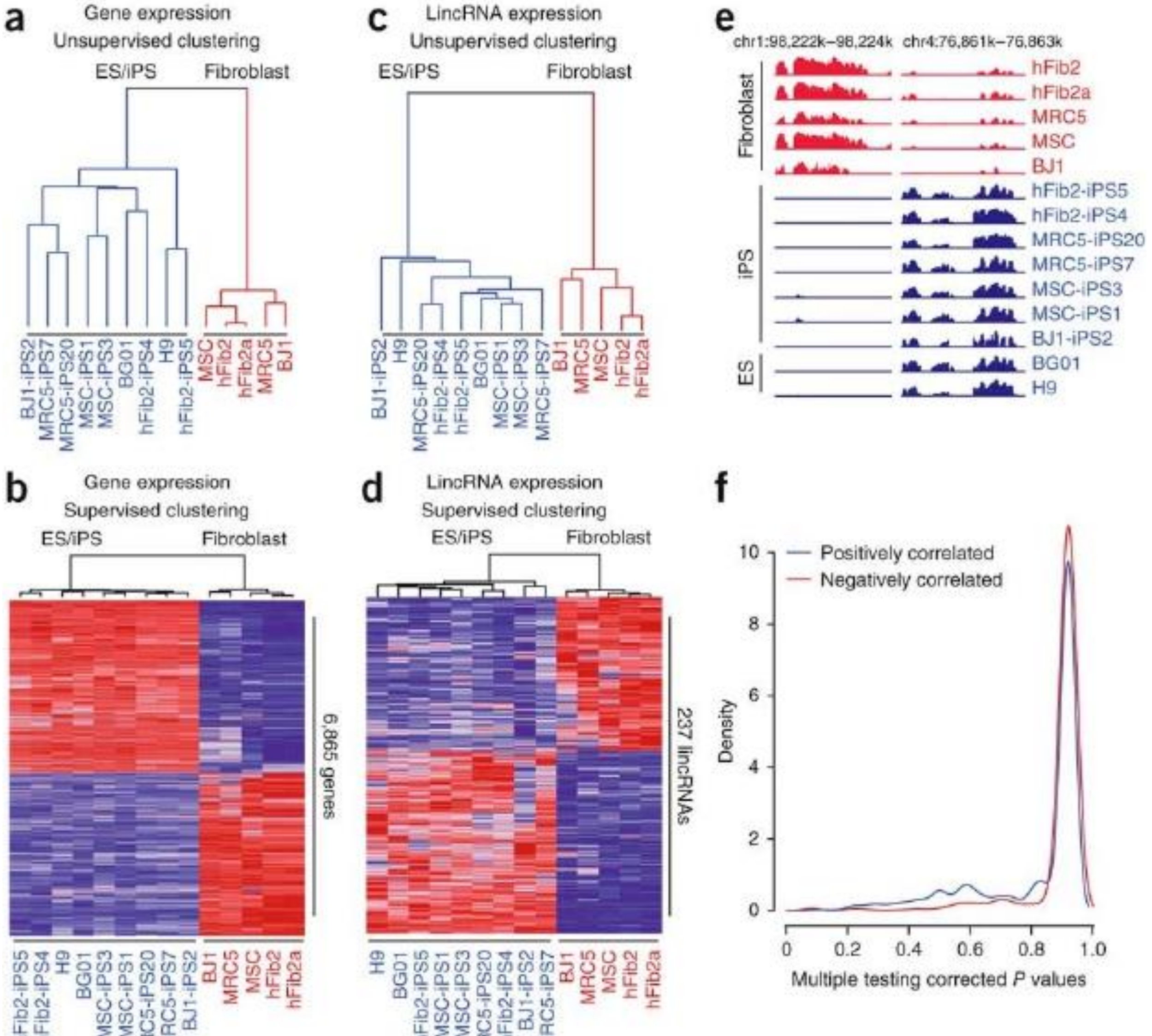
- Differentially expressed genes

- 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)

<u>Distance measurement</u>: how “similar” between two genes’ profile

Euclidean distance
(Absolution distance)

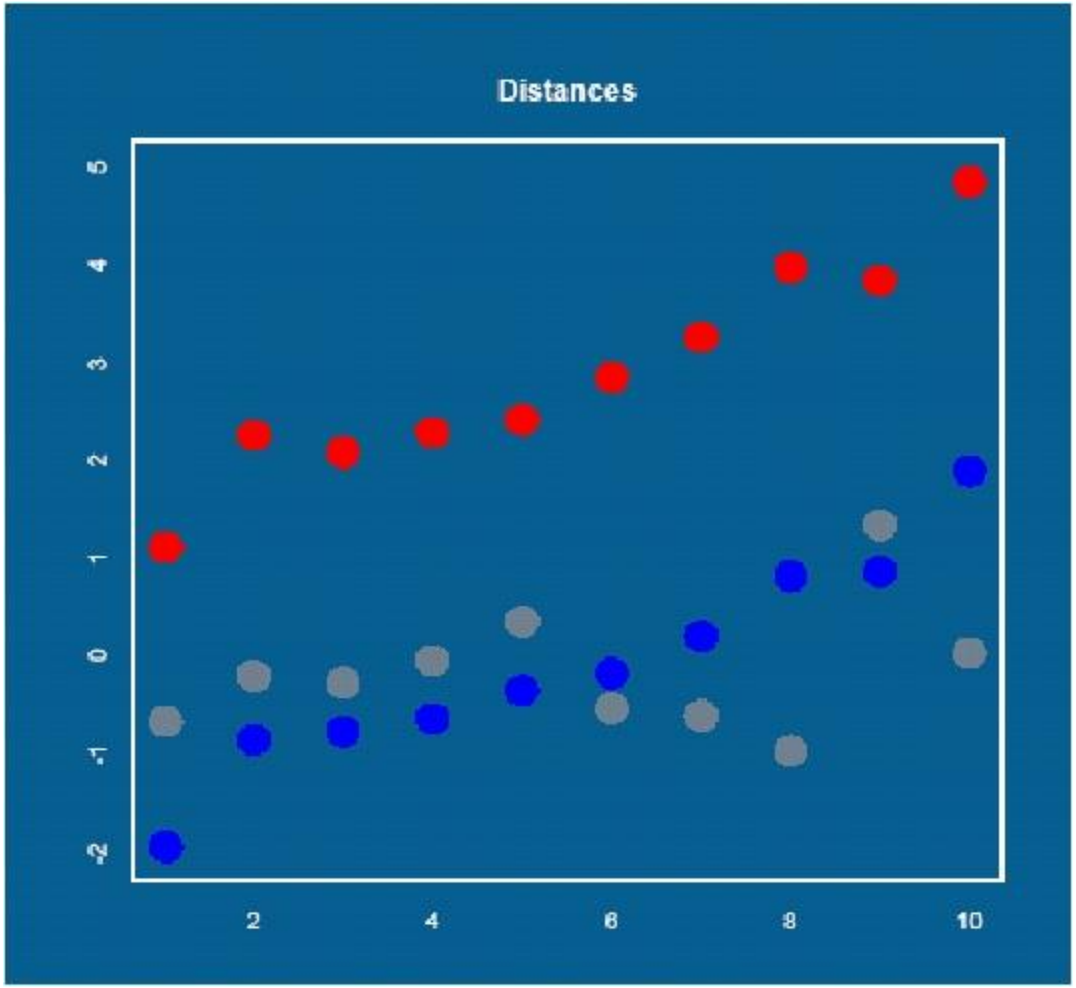
$$s(x_1,x_2) = \sqrt{\sum (x_{1k}^{2} - x_{2k})^2}$$

Pearson distance
(Correlation distance)

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

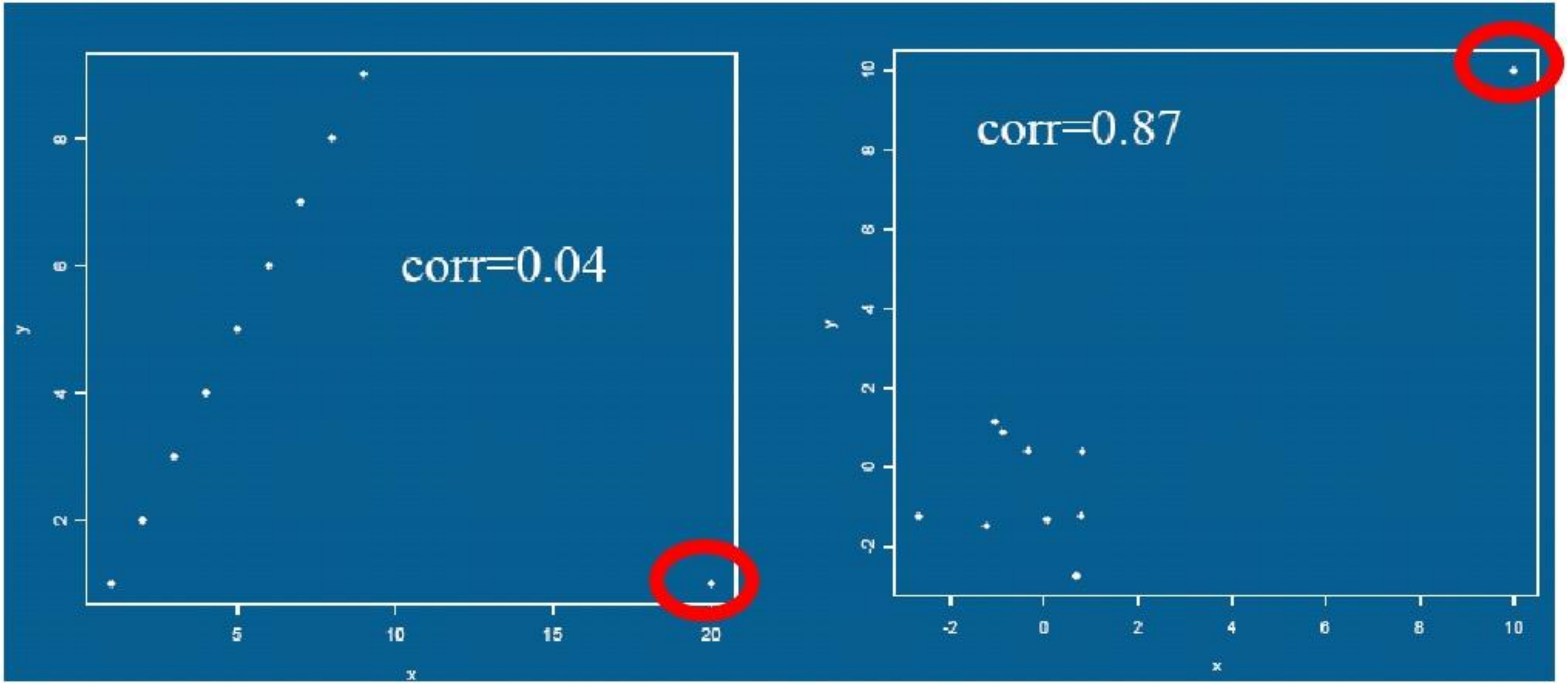
Pearson Distance:

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

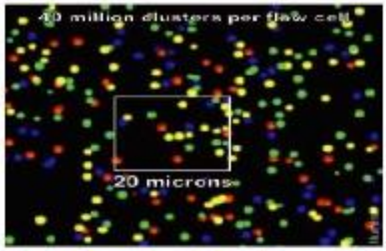


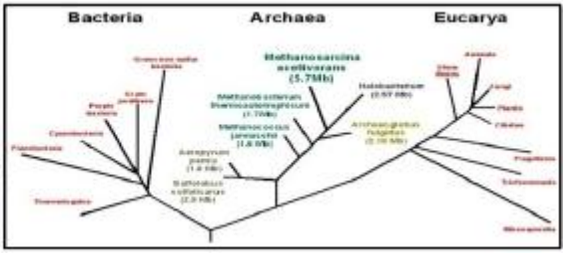
Eucl. Distance:

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


corr=0.04
y
2
4
6
8
x
5
10
15
20
corr=0.87
y
-2
0
2
4
6
8
10
x
-2
0
2
4
6
8
10

TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA





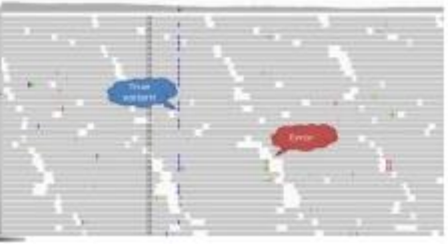
Unit 4:

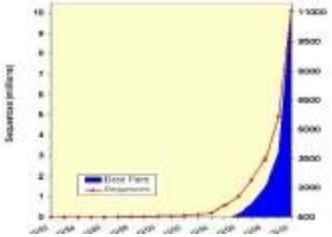
Computer Lab: Feature selection and Cluster analysis

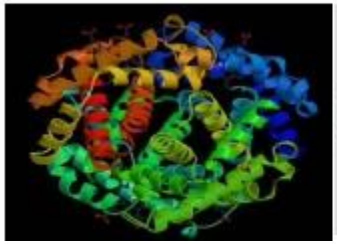
Le Zhang, Ph. D.

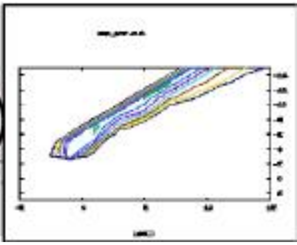
Computer Science Department

Southwest University

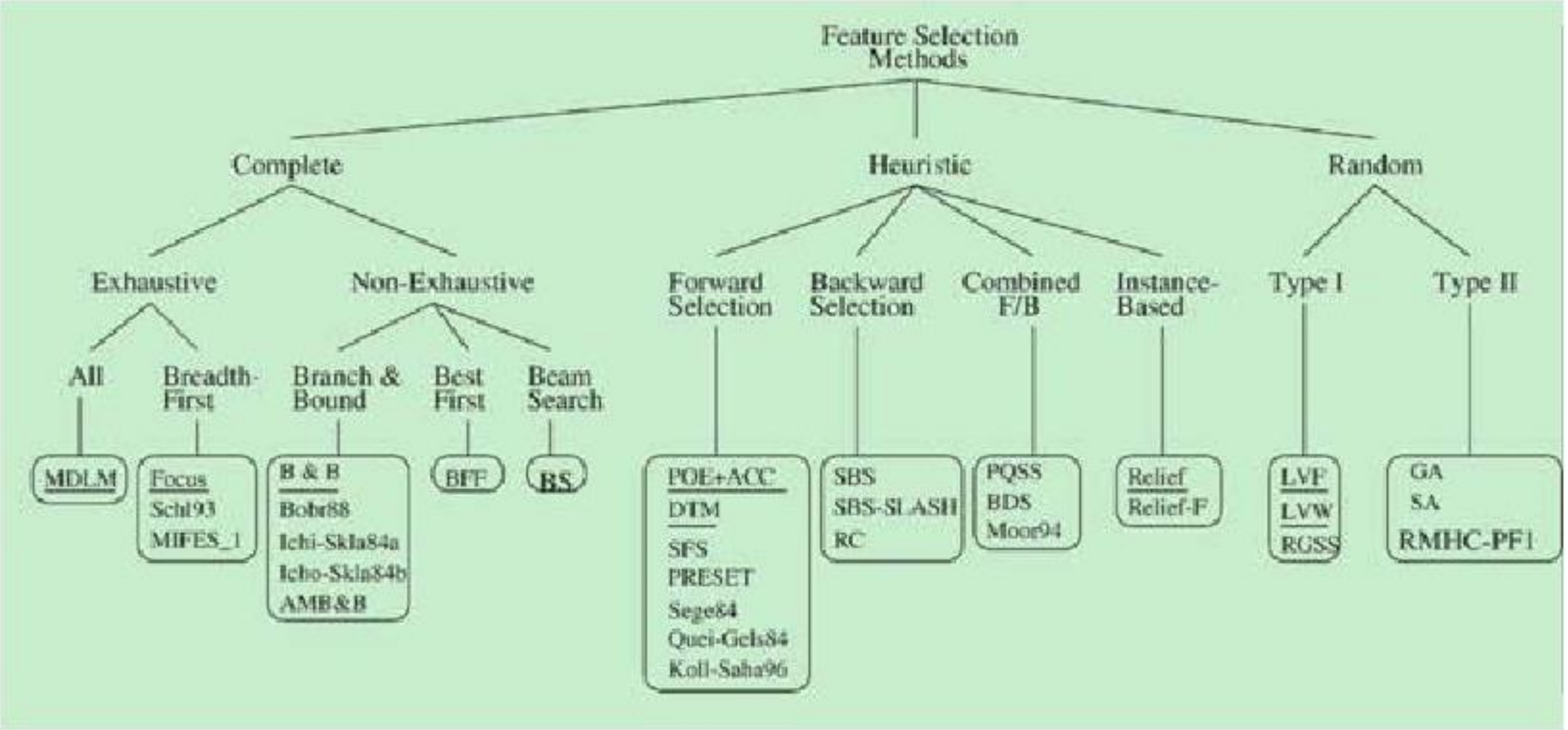








Find The Optimal Subset



Feature Selection Methods

Complete

Heuristic

Random

Exhaustive

Non-Exhaustive

Forward Selection

Backward Selection

Combined F/B

Instance-Based

Type I

Type II

All

Breadth-First

Branch & Bound

Best First

Beam Search

MDLM

Focus
Schl93
MIFES_1

B & B
Bobr88
Ichi-Skla84a
Icho-Skla84b
AMB&B

BFF

BS

POE+ACC
DTM
SFS
PRESET
Sege84
Quei-Gels84
Koll-Saha96

SBS
SBS-SLASH
RC

PQSS
BDS
Moor94

Relief
Relief-F

LVF
LVW
RGSS

GA
SA
RMHC-PF1

The way to find the optimal subset (M. Dash and H. Liu 1997)

Introduction Of Heuristic Search

- **SFS , Sequential Forward Selection**

  Set of variables starts from an empty set, each time we select a variable to join the subset and the optimal solution in the evaluation is selected. Each time select a optimal variable to join, a simple greedy algorithm.
- **SBS , Sequential Backward Selection**

  Set of variables starts from an set which has all variables ,each time we remove a variable from the subset and the optimal solution in the evaluation is selected.
- **BDS , Bidirectional Search**

  Using a sequence forward selection (SFS) starts from the empty set, while using the sequence backward selection (SBS) to start the search from the universal set, when the two are the same, stop the search.

Introduction Of Heuristic Search

- **LRS , Plus-L Minus-R Selection**

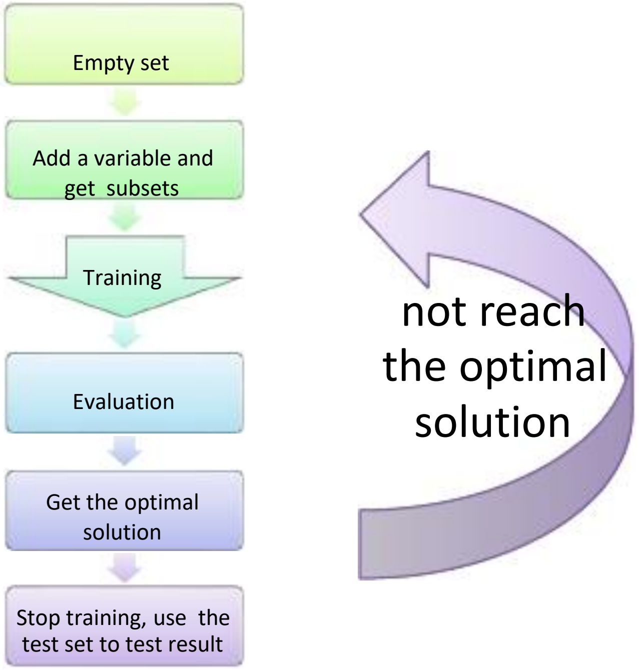
  Starts from the empty set, each time join L variables, and then remove R variables, the optimal solution in the evaluation is selected.(L > R)

  Starts from the universal set, each time remove R variables, and then join L variables, the optimal solution in the evaluation is selected. (L < R)

- **Sequential Floating Selection**

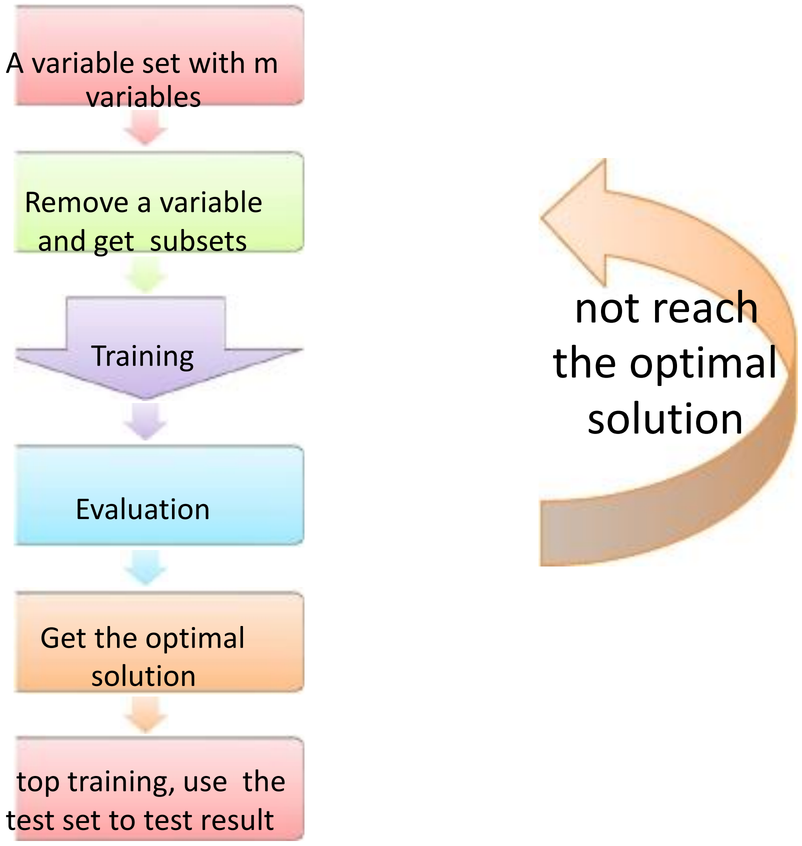
  Sequential Floating Selection is from the Plus-L Minus-R Selection , the differs is : the L and R is not fixed ,it will changing.

SFS


Empty set
Add a variable and get subsets
Training
Evaluation
Get the optimal solution
Stop training, use the test set to test result
not reach the optimal solution

Set of variables starts from an empty set, each time we select a variable to join the subset and the optimal solution in the evaluation is selected. Each time select a optimal variable to join, a simple greedy algorithm.

SBS



A variable set with m variables

Remove a variable and get subsets

Training

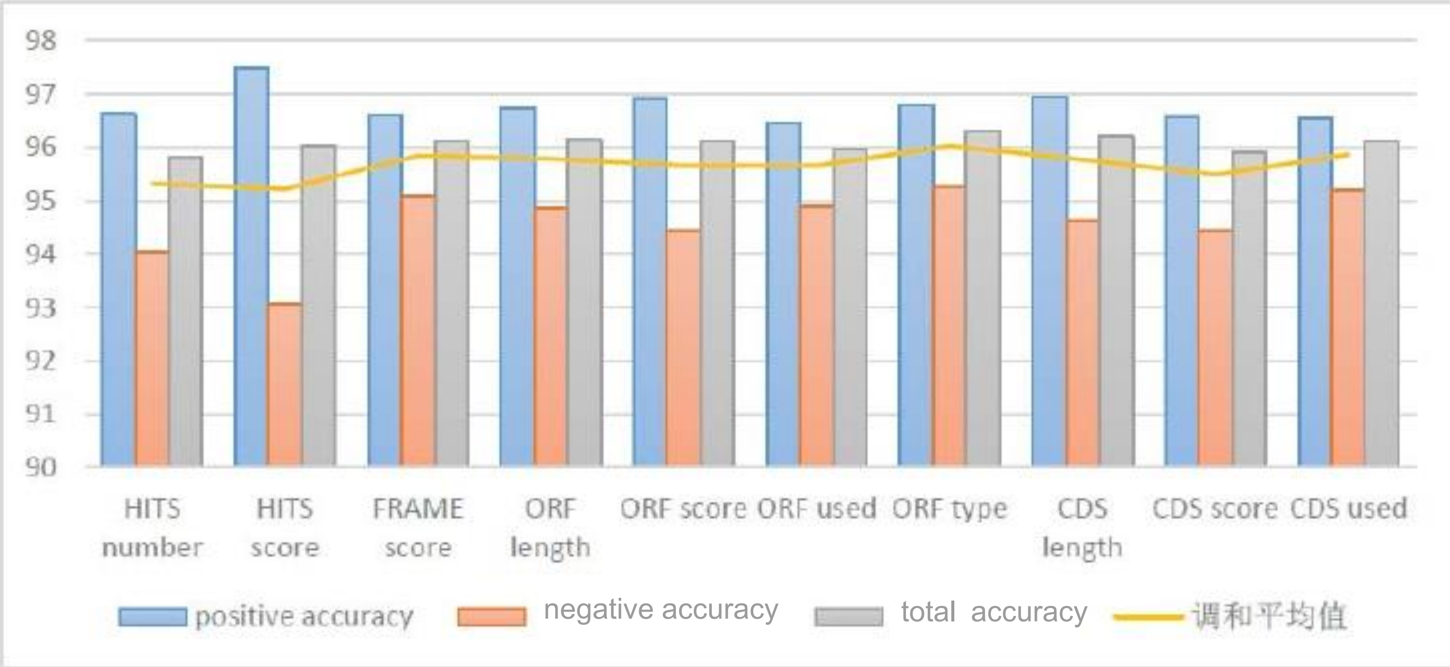
Evaluation

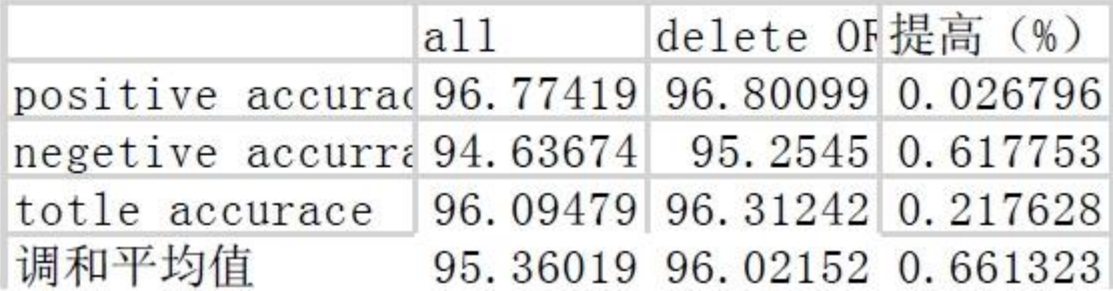
Get the optimal solution

top training, use the test set to test result

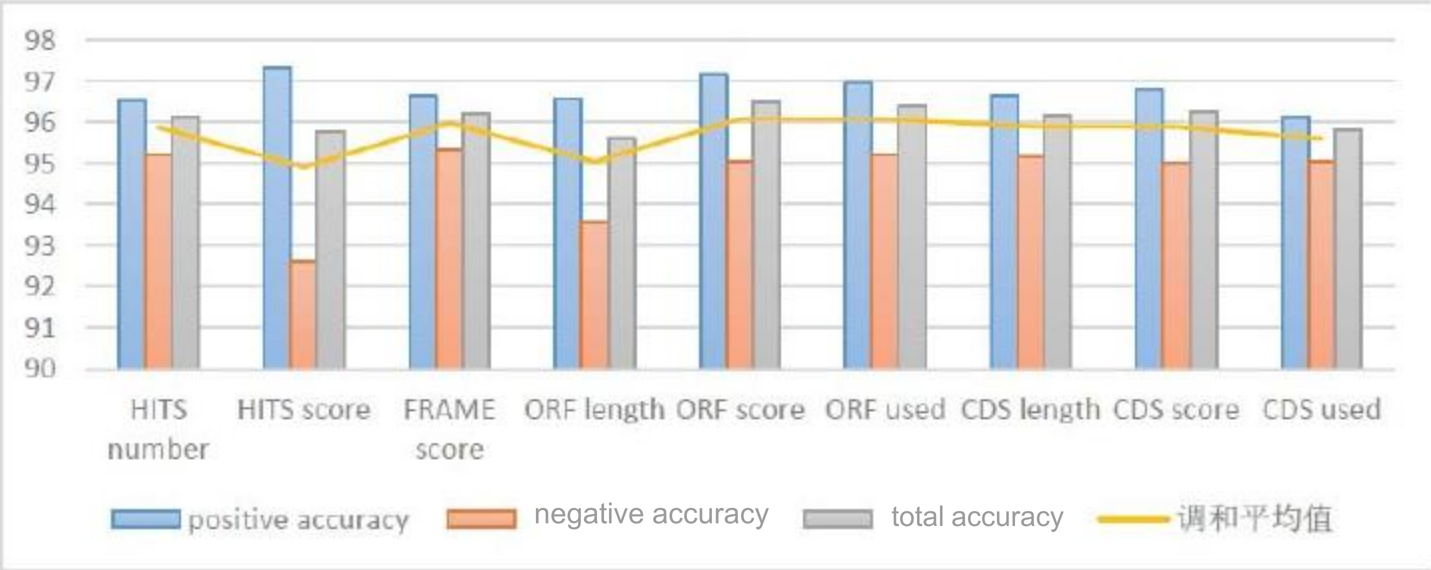
not reach the optimal solution

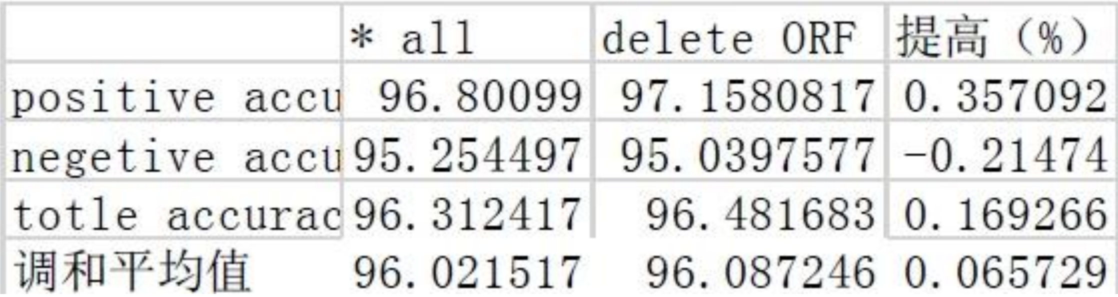
Set of variables starts from an set which has all variables ,each time we remove a variable from the subset and the optimal solution in the evaluation is selected.



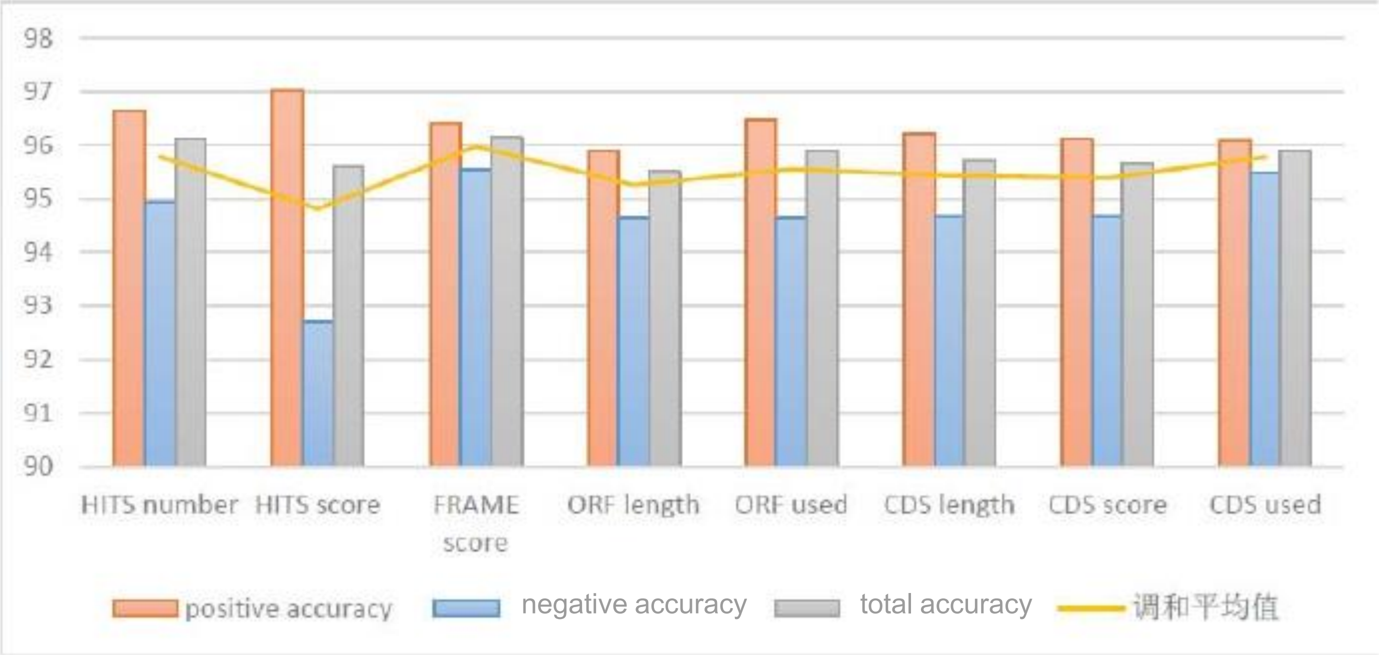


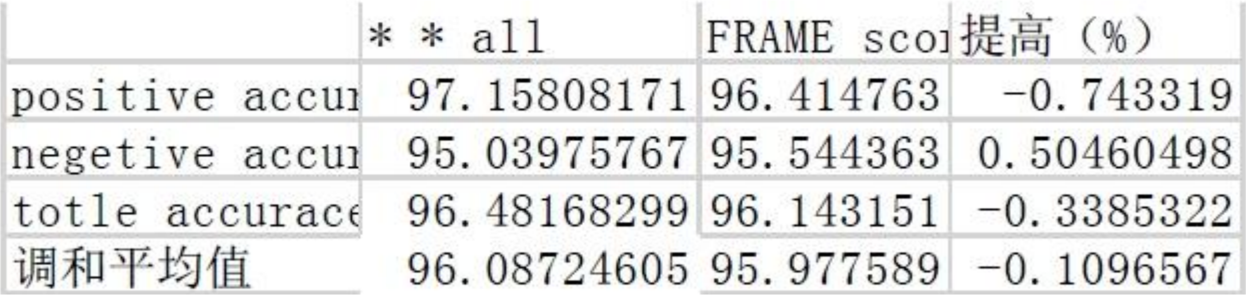
	all	delete OF	提高（%）
positive accurac	96.77419	96.80099	0.026796
negetive accurra	94.63674	95.2545	0.617753
totle accurace	96.09479	96.31242	0.217628
调和平均值	95.36019	96.02152	0.661323





	* all	delete ORF	提高（%）
positive accu	96.80099	97.1580817	0.357092
negetive accu	95.254497	95.0397577	-0.21474
totle accurac	96.312417	96.481683	0.169266
调和平均值	96.021517	96.087246	0.065729





	* * all	FRAME scor	提高（%）
positive accur	97.15808171	96.414763	-0.743319
negetive accur	95.03975767	95.544363	0.50460498
totle accurace	96.48168299	96.143151	-0.3385322
调和平均值	96.08724605	95.977589	-0.1096567

What is clustering

- 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 some sense or another) to each other than to those in other groups (clusters).

--from wikipedia

Distance

- Manhattan distance $(1) \quad \sum \quad | \qquad |$
- Euclidean distance $(2) \quad \sqrt{\sum \qquad\qquad}$
- Minkowski distance $(\ ) \quad \left[\sum \qquad\qquad \right]^{/}$
- Chebyshev distance $(\infty) \quad \max | \qquad |$
- Mahalanobis distance $(\ \ ) \quad \sqrt{\qquad (\ ) \qquad\qquad (\ )}$
- Lance and Williams distance $(\ ) \quad \sum \quad \frac{| \qquad |}{\qquad}$

Change to distance

- Using R
- **dist(x, method =“euclidean”, diag = FALSE, upper = FALSE, p=2)**
- **x** a numeric matrix, data frame or "dist" object.
- **method** the distance measure to be used. This must be one of "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski". Any unambiguous substring can be given.
- **diag** logical value indicating whether the diagonal of the distance matrix should be printed by print.dist.
- **upper** logical value indicating whether the upper triangle of the distance matrix should be printed by print.dist.
- **p** The power of the Minkowski distance.

Hierarchical clustering method

- Single linkage method $\min\{\quad,\quad\}$
- Complete linkage method $\mathrm{m}\quad\{\quad,\quad\}$
- Median method $\frac{}{}(\qquad)$
- Average linkage method $\frac{}{}\qquad\frac{}{}$
- Centroid method $\frac{}{}\qquad\frac{}{}\qquad\frac{}{}$
- Ward method $=\frac{}{}\qquad\frac{}{}\qquad\frac{}{}$

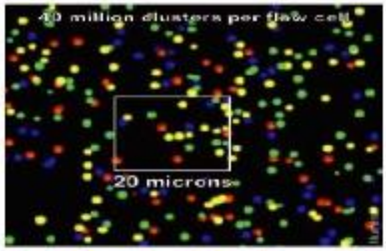
hclust

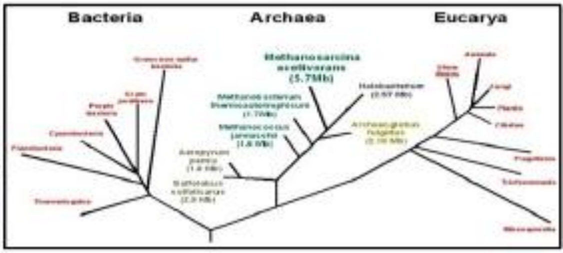
- hclust(d, method = "complete", members = NULL)
- d a dissimilarity structure as produced by dist.
- method the agglomeration method to be used. This should be (an unambiguous abbreviation of) one of "ward", "single", "complete", "average", "mcquitty", "median" or "centroid".

Reference

- 统计建模与R软件
- http://www.cnblogs.com/xiangshancuizhu/archive/2012/03/12/2392360.html
- http://en.wikipedia.org/wiki/Feature_selection
- http://en.wikipedia.org/wiki/Cluster_analysis
- http://www.biostars.org/p/14156/

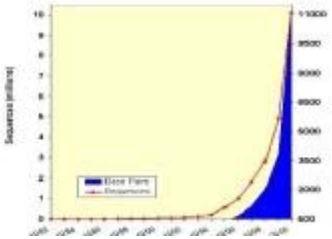
```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```

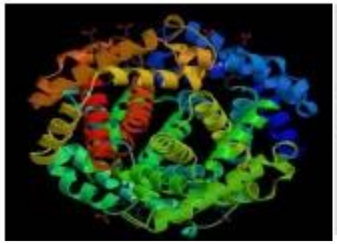


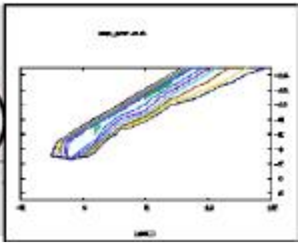


Unit 5:
Differential gene expression analysis

Le Zhang, Ph.D.
Computer Science Department
Southwest University







Background

- High-throughput sequencing technology is rapidly becoming the standard method for measuring RNA expression levels (aka RNA-seq).

- One of the main goals of these experiments is to identify the differentially expressed genes in two or more conditions.

Differential gene expression analysis

- 3 steps:
- 1. Normalization of counts
- 2. parameter estimation of the statistical model
- 3. Test for differential gene expression

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*]

Goal : Comparison of different analysis methods for RNA-seq data from different perspectives.

Such as, Cuffdiff, edgeR, DESeq, PoissonSeq, baySeq, and limma.

Datasets for Research

They used two benchmark datasets:

1 The first is the Sequencing Quality Control (SEQC) dataset, which includes replicated samples of the human whole body reference RNA and human brain reference RNA along with RNA spike-in controls.

2 The second dataset is RNA-seq data from biological replicates of three cell lines that were characterized as part of the ENCODE project.

The measures of their analysis

- The analysis in this paper focused on a number of measures that are most relevant for detection of differential gene expression from RNA-seq data
- i) normalization of count data;
- ii) sensitivity and specificity of DE detection;
- iii) performance on the subset of genes that are expressed in one condition but have no detectable expression in the other condition;
- iv) the effects of reduced sequencing depth and number of replicates on the detection of differential expression.

Normalized counts by log expression correlation

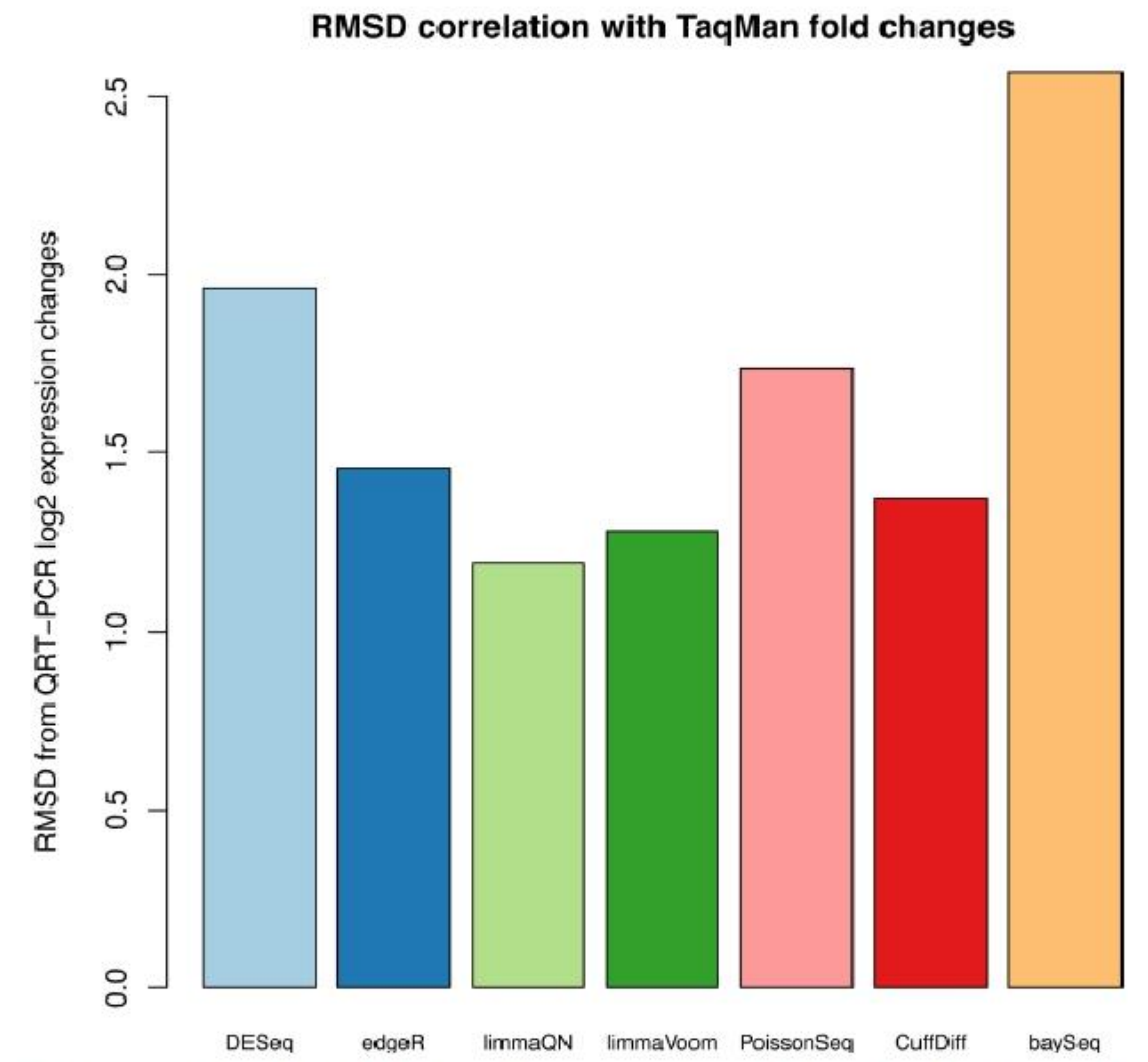


Figure 1 RMSD correlation between qRT-PCR and RNA-seq $\log_2$ expression changes computed by each method. Overall, there is good concordance between $\log_2$ values derived from the DE methods and the experimental values derived from qRT-PCR measures. Upper quartile normalization implemented in baySeq package is least correlated with qRT-PCR values. DE, differential expression; RMSD, root-mean-square deviation.

Differential expression analysis

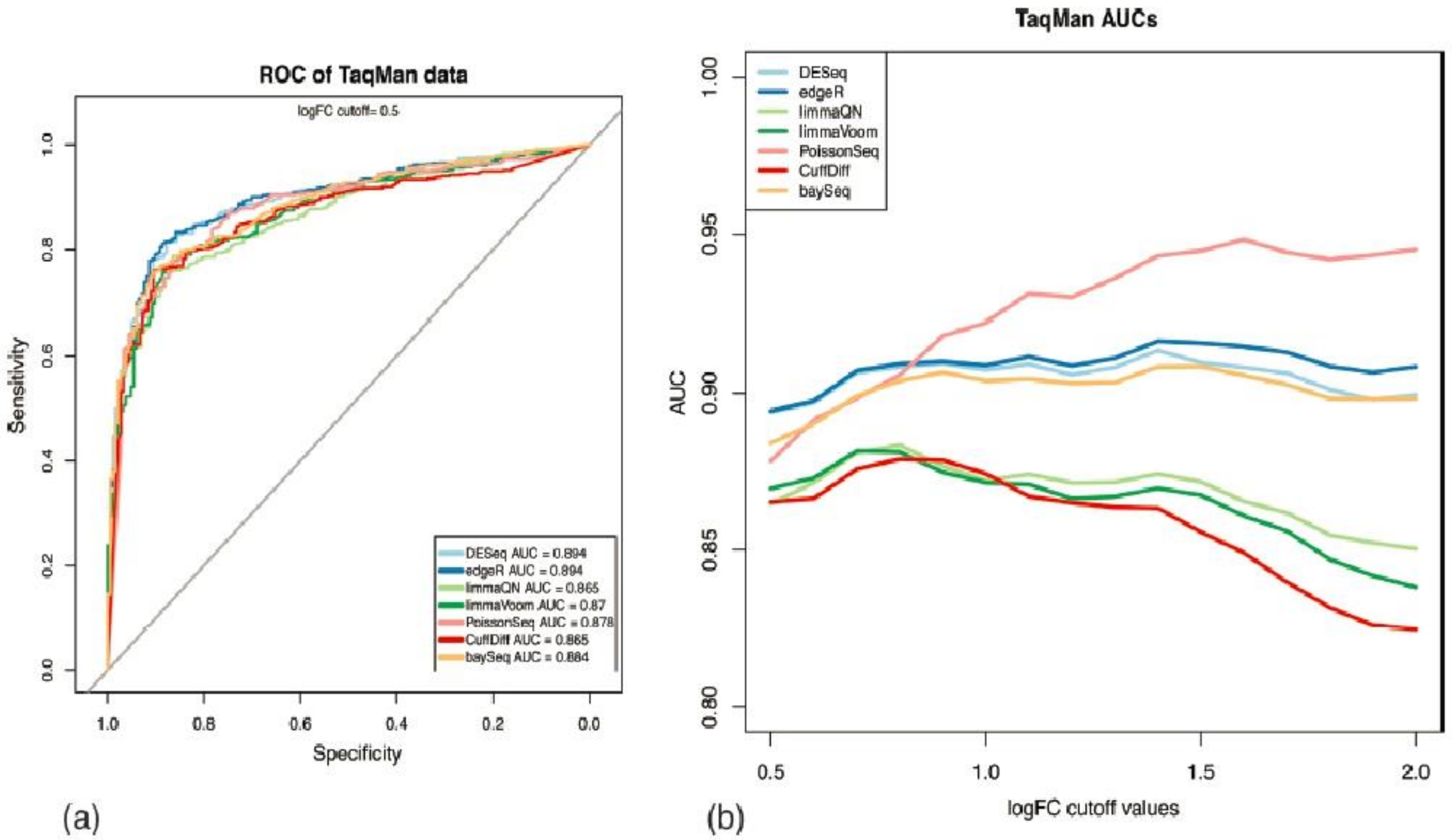
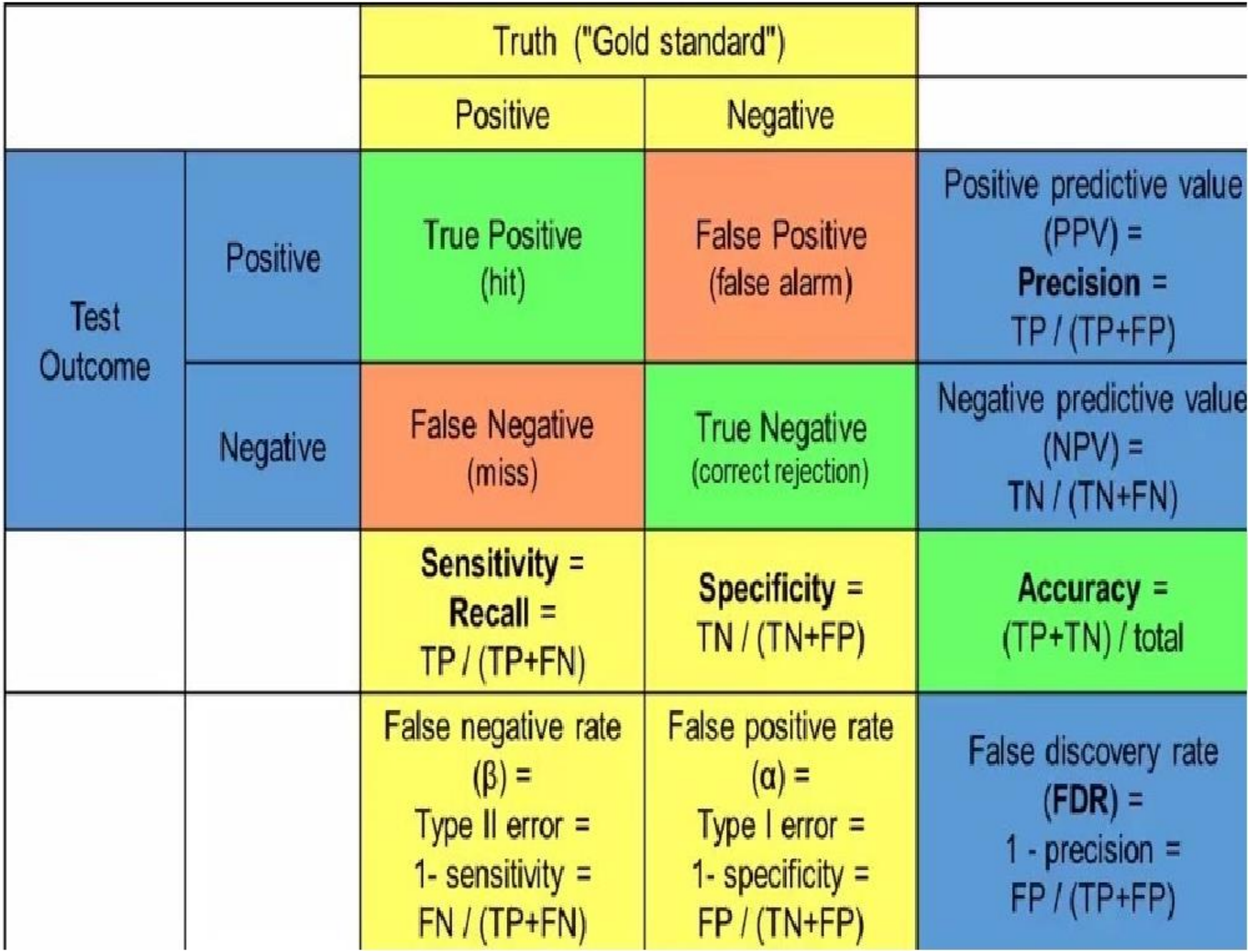
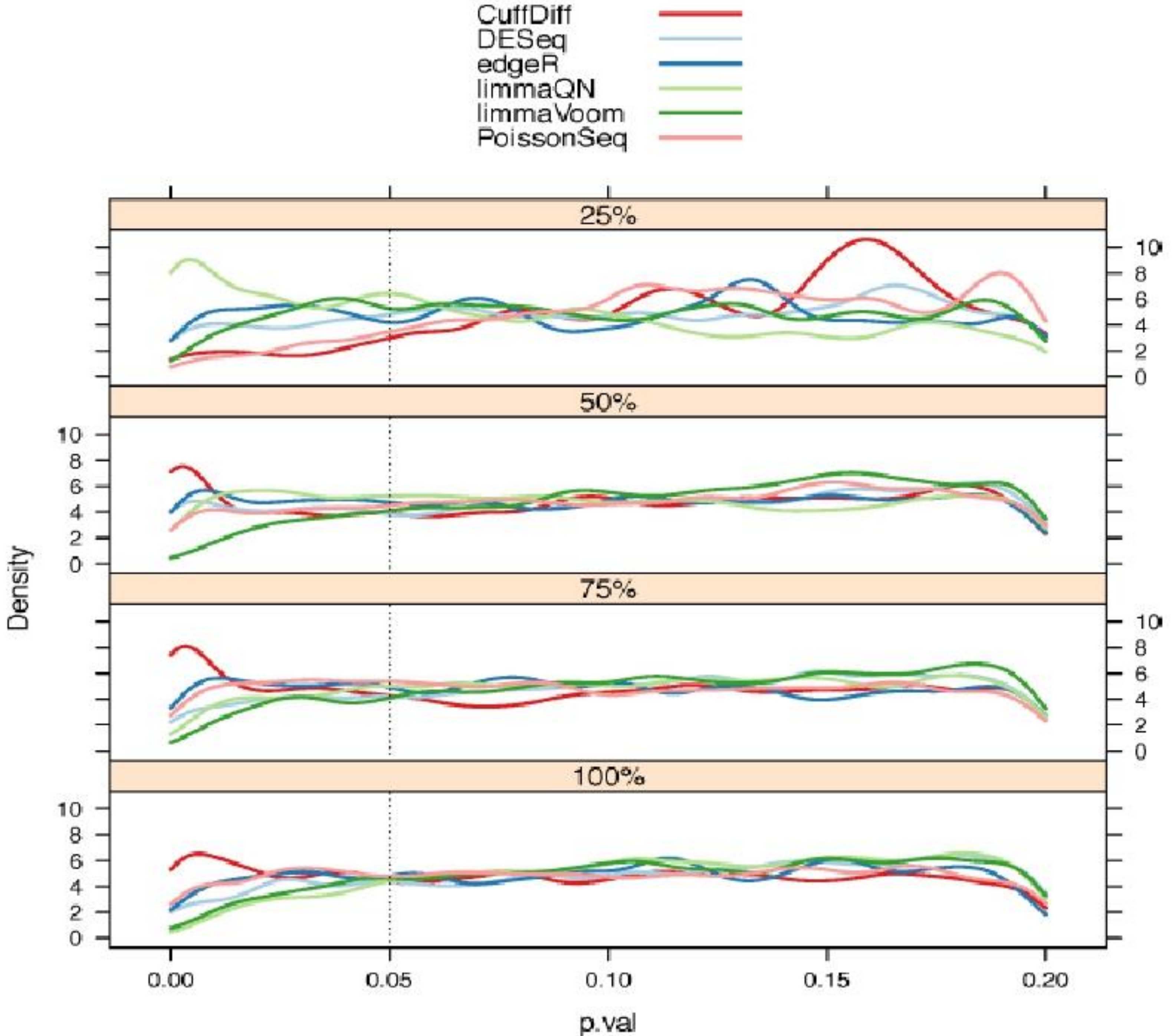


Figure 2 Differential expression analysis using qRT-PCR validated gene set. **(a)** ROC analysis was performed using a qRT-PCR $\log_2$ expression change threshold of 0.5. The results show a slight advantage for DESeq and edgeR in detection accuracy. **(b)** At increasing $\log_2$ expression ratios (incremented by 0.1), representing a more stringent cutoff for differential expression, the performances of the Cuffdiff and limma methods gradually reduce whereas PoissonSeq performance increases. AUC, area under the curve.

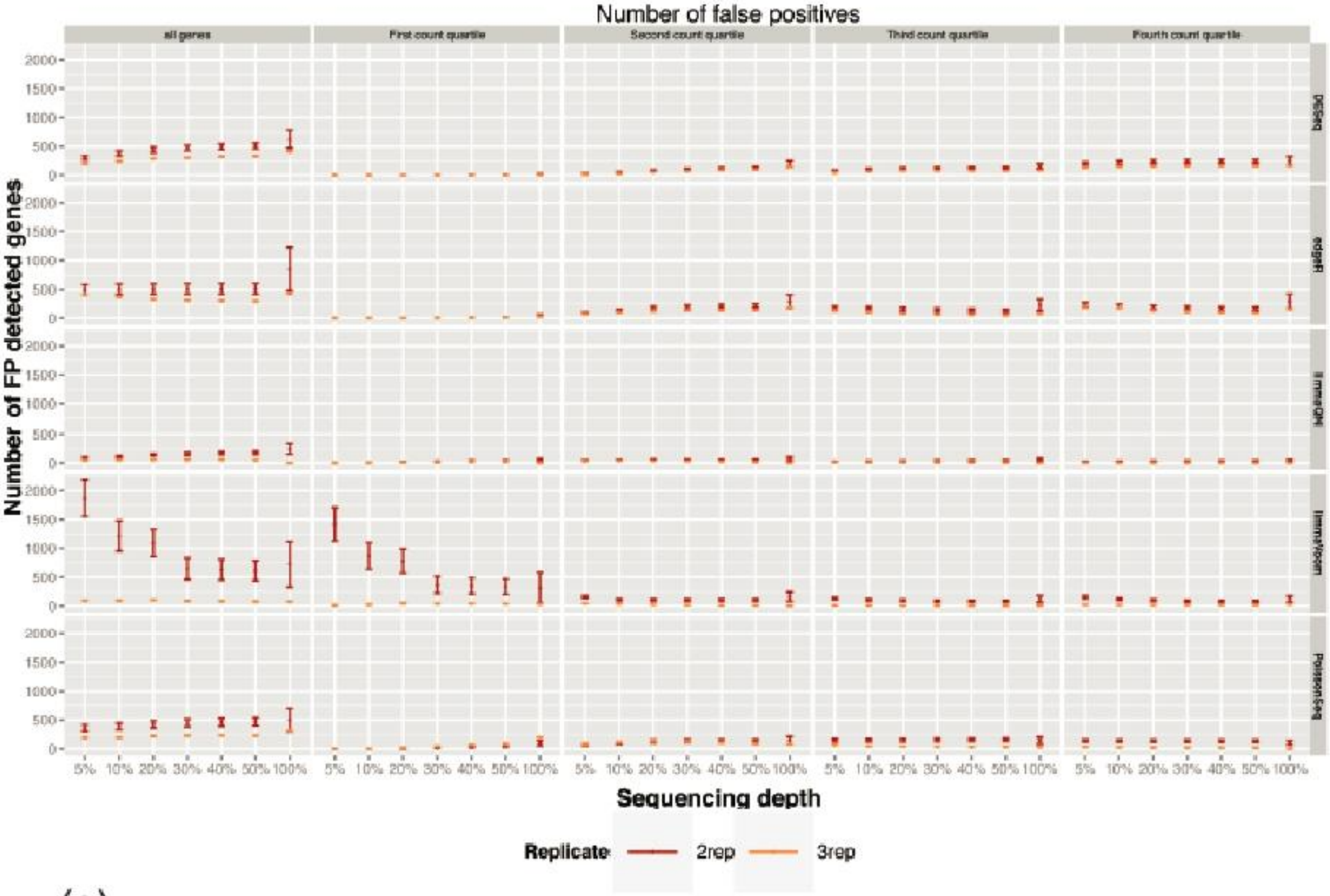


		Truth ("Gold standard")		
		Positive	Negative	
Test Outcome	Positive	True Positive (hit)	False Positive (false alarm)	Positive predictive value (PPV) = **Precision** = TP / (TP+FP)
	Negative	False Negative (miss)	True Negative (correct rejection)	Negative predictive value (NPV) = TN / (TN+FN)
		Sensitivity = **Recall** = TP / (TP+FN)	**Specificity** = TN / (TN+FP)	**Accuracy** = (TP+TN) / total
		False negative rate (β) = Type II error = 1- sensitivity = FN / (TP+FN)	False positive rate (α) = Type I error = 1- specificity = FP / (TN+FP)	False discovery rate (**FDR**) = 1 - precision = FP / (TP+FP)

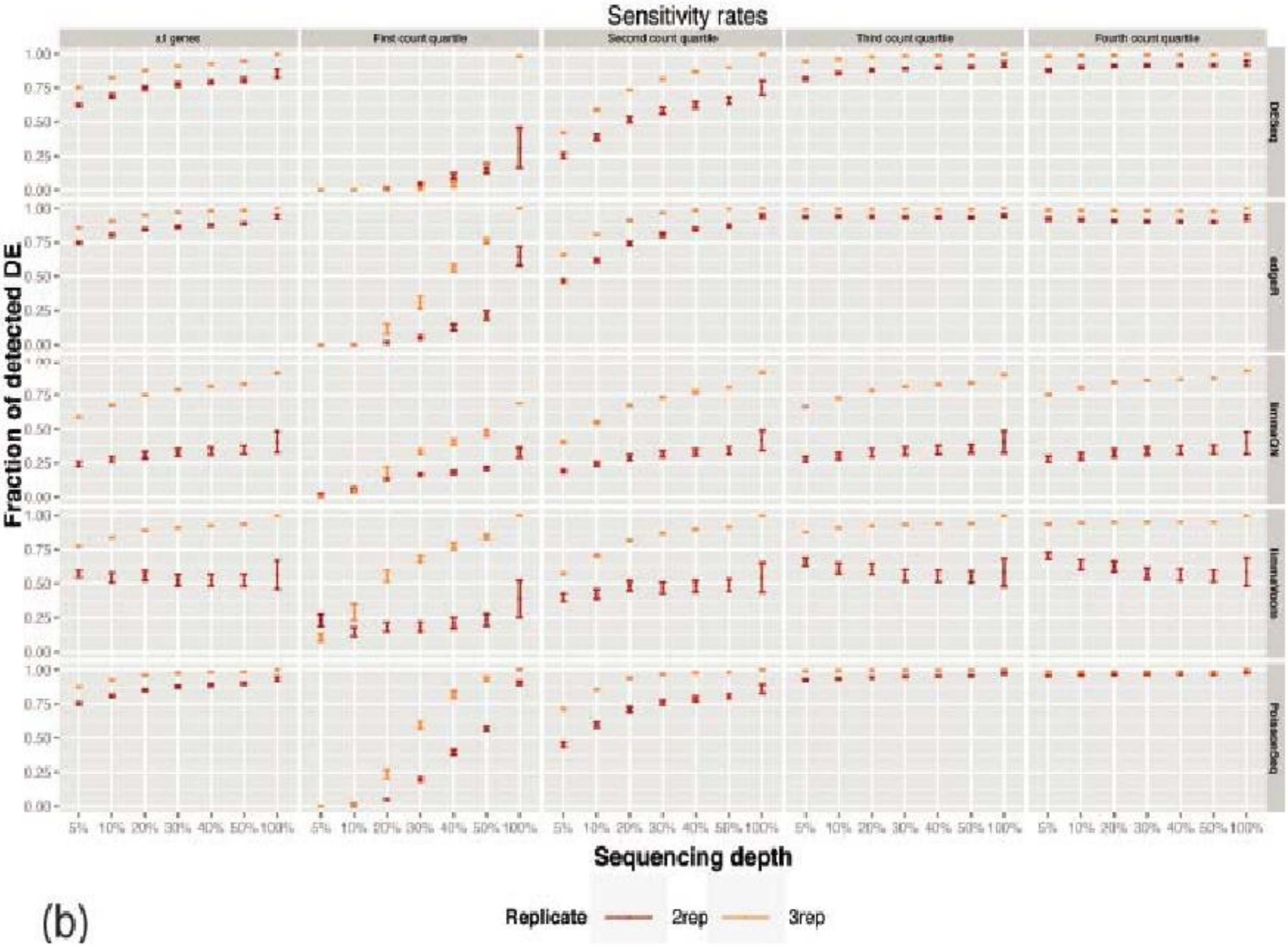
Null model evaluation of type I error



Impact of sequencing depth and number of replicate samples on DE analysis



(a)


Sensitivity rates
All genes
First count quartile
Second count quartile
Third count quartile
Fourth count quartile
DESeq
edgeR
limmaQN
limmaVoom
PoissonSeq
Fraction of detected DE
1.00
0.75
0.50
0.25
0.00
5%
10%
20%
30%
40%
50%
100%
Sequencing depth
Replicate
2rep
3rep

(b)

Conclusion

1 In most benchmarks Cuffdiff performed less favorably

✓with a higher number of false positives

✓without any increase in sensitivity.

2 Our results conclusively demonstrate that the addition of replicate samples provides substantially greater detection power of DE than increased sequence depth.

- Hence, including more replicate samples in RNA-seq experiments is always to be preferred over increasing the number of sequenced reads.

Bioinformatics: Introduction and Methods

Computer Science Department, Southwest University

Thank you