

# Bioinformatics and Biomarker Discovery

## *Part 3: Examples*

**Limsoon Wong**  
**3 September 2014**



# Outline

- **ALL**
  - Gene expression profile classification
  - Beyond diagnosis and prognosis
  
- **WEKA**
  - Breast cancer
  - Dermatology
  - Pima Indians
  - Echocardiogram
  - Mammography

# Gene Expression Profile Classification

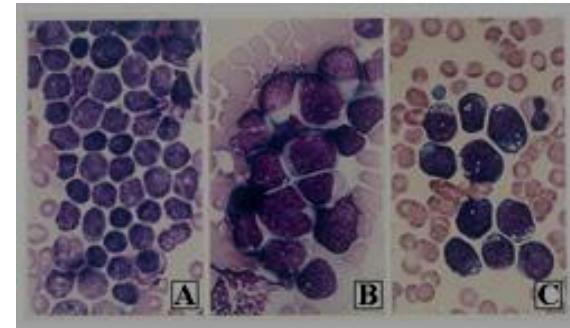
**Diagnosis of Childhood Acute  
Lymphoblastic Leukemia and Optimization  
of Risk-Benefit Ratio of Therapy**



# Childhood ALL

- Major subtypes: T-ALL, E2A-PBX, TEL-AML, BCR-ABL, MLL genome rearrangements, Hyperdiploid >50
- Diff subtypes respond differently to same Tx
- Over-intensive Tx
  - Development of secondary cancers
  - Reduction of IQ
- Under-intensive Tx
  - Relapse

- The subtypes look similar

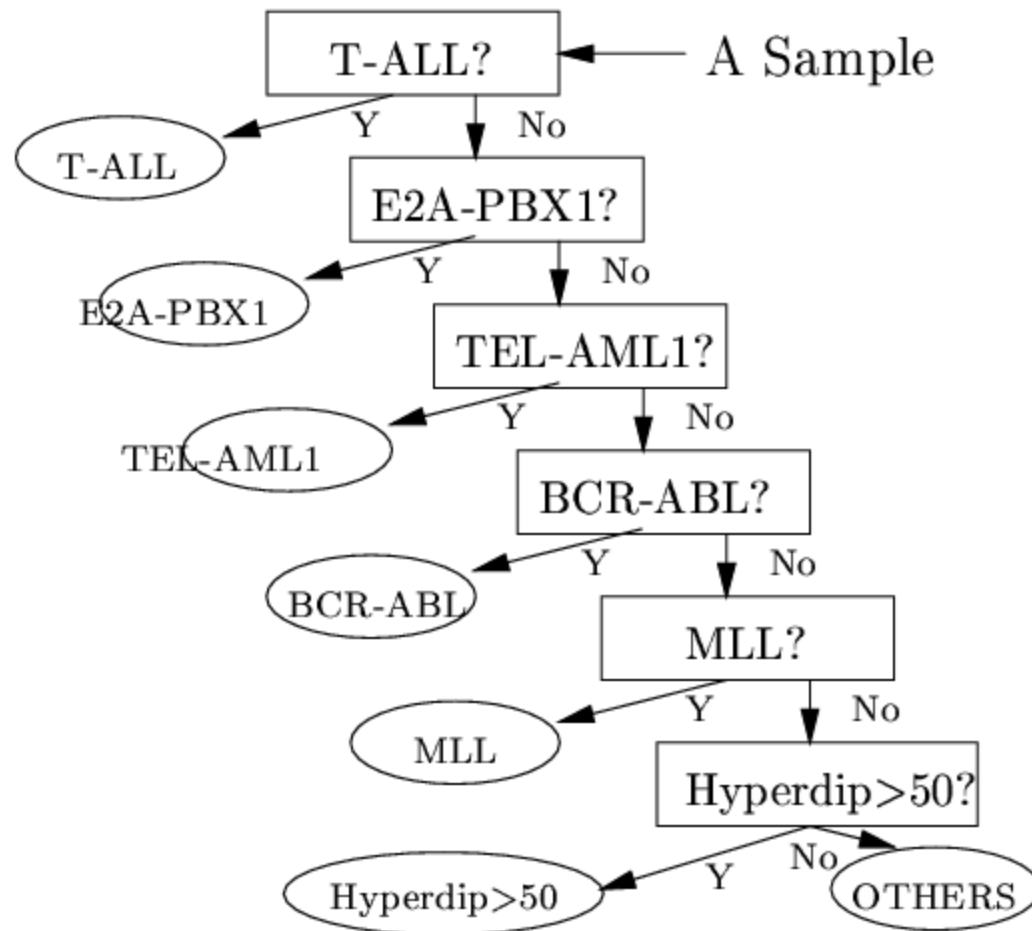


- Conventional diagnosis
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- Unavailable in most ASEAN countries

# Subtype Diagnosis by Machine Learning

- **Gene expression data collection**
- **Gene selection by e.g.  $\chi^2$**
- **Classifier training by e.g. emerging pattern**
- ~~**Classifier tuning (optional for some machine learning methods)**~~
- **Apply classifier for diagnosis of future cases by e.g. PCL**

# Childhood ALL Subtype Diagnosis Workflow



A tree-structured diagnostic workflow was recommended by our doctor collaborator

# Training and Testing Sets

Paired datasets	Ingredients	Training	Testing
T-ALL vs OTHERS1	OTHERS1 = {E2A-PBX1, TEL-AML1, BCR-ABL, Hyperdip>50, MLL, OTHERS}	28 vs 187	15 vs 97
E2A-PBX1 vs OTHERS2	OTHERS2 = {TEL-AML1, BCR-ABL Hyperdip>50, MLL, OTHERS}	18 vs 169	9 vs 88
TEL-AML1 vs OTHERS3	OTHERS3 = {BCR-ABL Hyperdip>50, MLL, OTHERS}	52 vs 117	27 vs 61
BCR-ABL vs OTHERS4	OTHERS4 = {Hyperdip>50, MLL, OTHERS}	9 vs 108	6 vs 55
MLL vs OTHERS5	OTHERS5 = {Hyperdip>50, OTHERS}	14 vs 94	6 vs 49
Hyperdip>50 vs OTHERS	OTHERS = {Hyperdip47-50, Pseudodip, Hypodip, Normo}	42 vs 52	22 vs 27

## Signal Selection by $\chi^2$

The  $\chi^2$  value of a signal is defined as:

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{(A_{ij} - E_{ij})^2}{E_{ij}},$$

where  $m$  is the number of intervals,  $k$  the number of classes,  $A_{ij}$  the number of samples in the  $i$ th interval,  $j$ th class,  $R_i$  the number of samples in the  $i$ th interval,  $C_j$  the number of samples in the  $j$ th class,  $N$  the total number of samples, and  $E_{ij}$  the expected frequency of  $A_{ij}$  ( $E_{ij} = R_i * C_j / N$ ).



# Accuracy of Various Classifiers

Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS <sup>1</sup>	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS <sup>2</sup>	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS <sup>3</sup>	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS <sup>4</sup>	2:0	3:0	1:4	2:0
MLL vs OTHERS <sup>5</sup>	0:1	0:0	0:0	0:0
Hyperdiploid>50 vs OTHERS	2:6	0:2	0:2	0:1
Total Errors	14	6	8	4

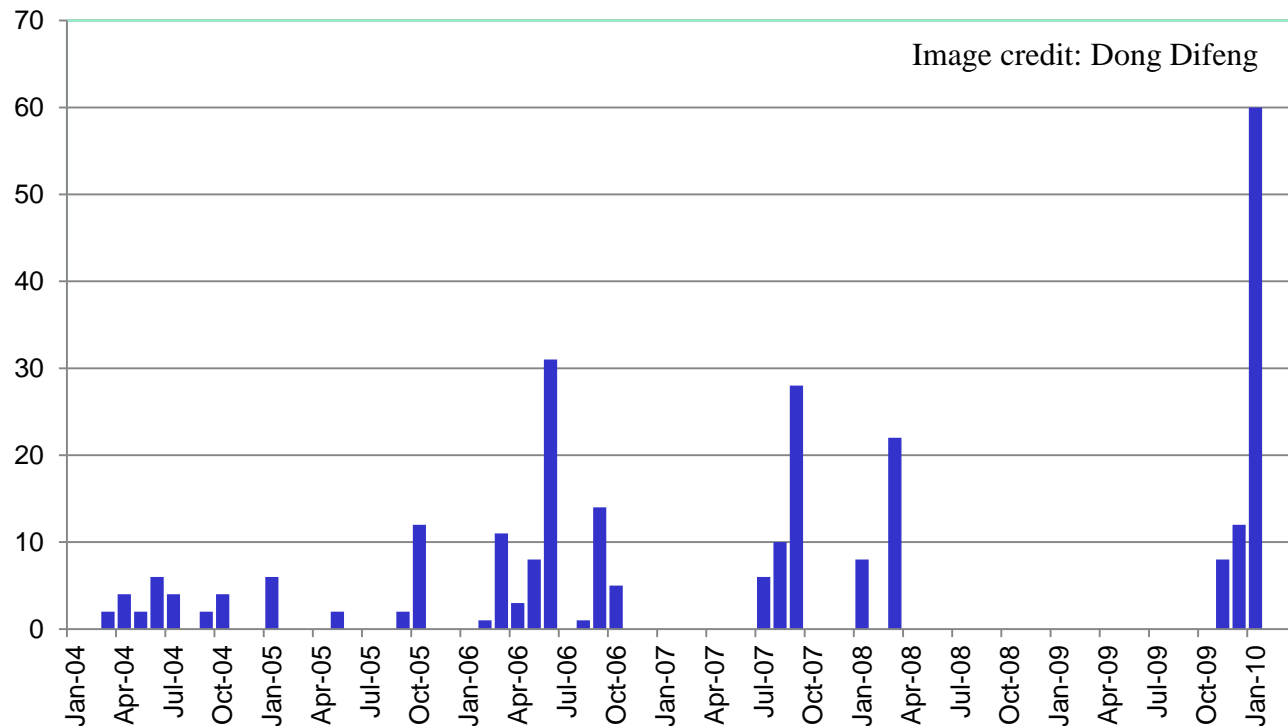
The classifiers are all applied to the 20 genes selected by  $\chi^2$  at each level of the tree

# Normalization



Sometimes, a gene expression study may involve batches of data collected over a long period of time...

### Time Span of Gene Expression Profiles



In such a case, batch effect may be severe... to the extent that you can predict the batch that each sample comes!

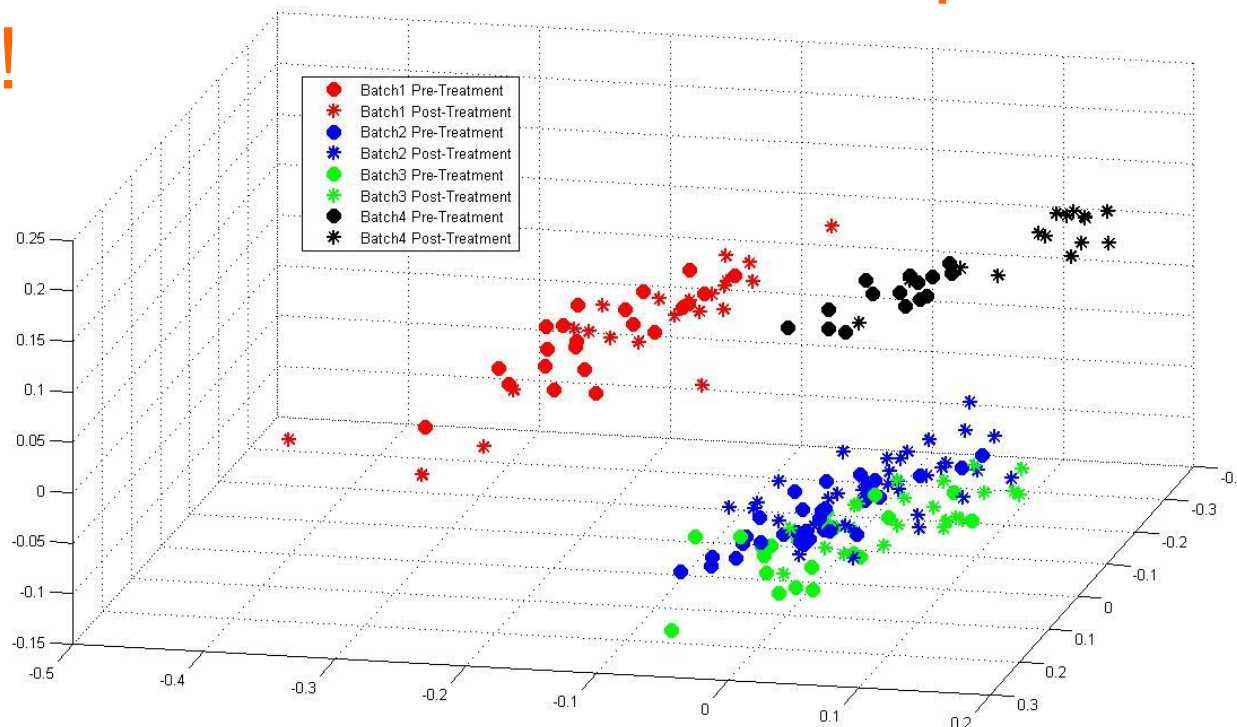


Image credit: Dong Difeng

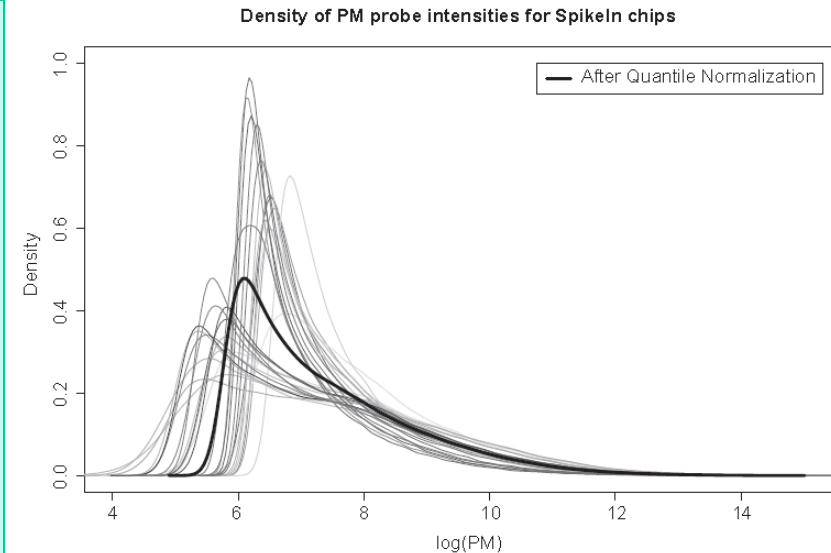
⇒ **Need normalization to correct for batch effect**

# Approaches to Normalization

- **Aim of normalization:**  
**Reduce variance w/o increasing bias**
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's
- **Xform data so that distribution of probe intensities is same on all arrays**
  - E.g.,  $(x - \mu) / \sigma$
- **Quantile normalization**

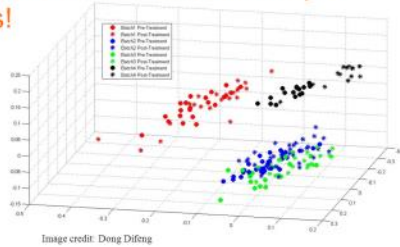
# Quantite Normalization

- Given  $n$  arrays of length  $p$ , form  $X$  of size  $p \times n$  where each array is a column
- Sort each column of  $X$  to give  $X_{\text{sort}}$
- Take means across rows of  $X_{\text{sort}}$  and assign this mean to each elem in the row to get  $X'_{\text{sort}}$
- Get  $X_{\text{normalized}}$  by arranging each column of  $X'_{\text{sort}}$  to have same ordering as  $X$



- Implemented in some microarray s/w, e.g., EXPANDER

In such a case, batch effect may be severe... to the extent that you can predict the batch that each sample comes!



# After quantile normalization

⇒ Need normalization to correct for batch effect

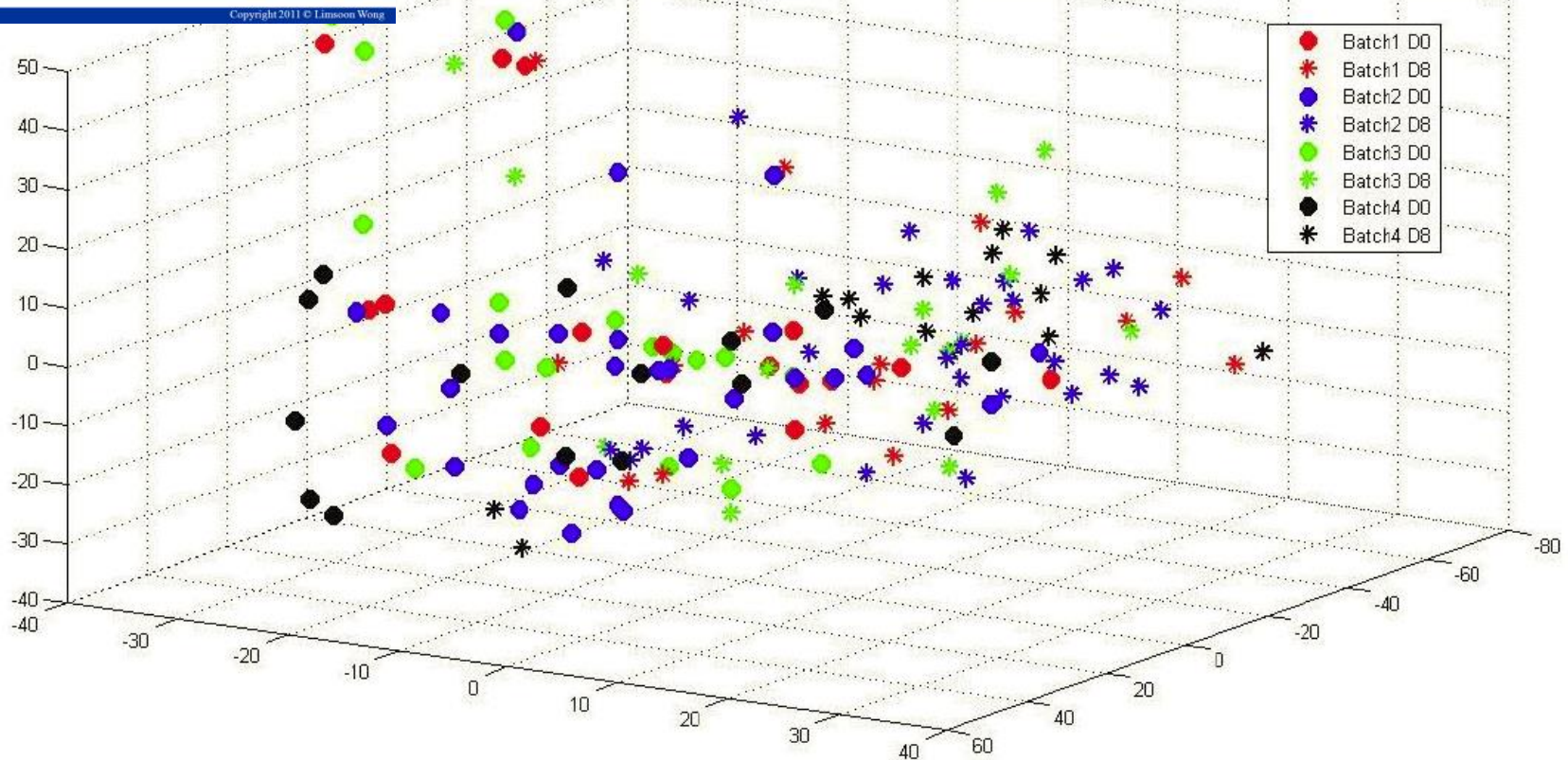


Figure 3.6: GEPs after the batch effects removing.

# Beyond Disease Diagnosis & Prognosis





# Percentage of Overlapping Genes

- **Low % of overlapping genes from diff expt in general**
  - Prostate cancer
    - Lapointe et al, 2004
    - Singh et al, 2002
  - Lung cancer
    - Garber et al, 2001
    - Bhattacharjee et al, 2001
  - DMD
    - Haslett et al, 2002
    - Pescatori et al, 2007

Datasets	DEG	POG
Prostate Cancer	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD	Top 10	0.20
	Top 50	0.42
	Top100	0.54

Zhang et al, Bioinformatics, 2009

# Individual Genes

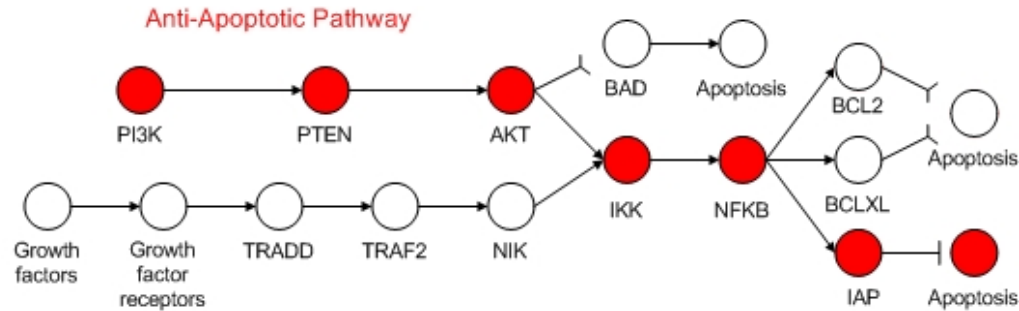
- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **How many genes on a microarray are expected to perfectly correlate to these samples?**
- **Prob(a gene is correlated) =  $1/2^6$**
- **# of genes on array = 30,000**
- ⇒ **E(# of correlated genes) = 468**
- ⇒ **Many false positives**
- **These cannot be eliminated based on pure statistics!**

# Group of Genes

- **Suppose**
    - Each gene has 50% chance to be high
    - You have 3 disease and 3 normal samples
  - **What is the chance of a group of 5 genes being perfectly correlated to these samples?**
  - **Prob(group of genes correlated) =  $(1/2^6)^5$** 
    - Good,  $\ll 1/2^6$
  - **# of groups =  $^{30000}C_5$**
  - ⇒ **E(# of groups of genes correlated) =  $^{30000}C_5 * (1/2^6)^5 = 2 * 10^{11}$**
- ⇒ **Even more false positives?**

  - **Perhaps no need to consider every group**

# Gene Regulatory Circuits



- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype

- **Uncertainty in selected genes can be reduced by considering biological processes of the genes**
- **The unifying biological theme is basis for inferring the underlying cause of disease subtype**

# Taming false positives by considering pathways instead of all possible groups



## Group of Genes



- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **What is the chance of a group of 5 genes being perfectly correlated to these samples?**

- **Prob(group of genes correlated) =  $(1/2^6)^5$** 
  - Good,  $\ll 1/2^6$

• ~~# of groups =  $30000 C_5$~~

⇒ ~~E(# of groups of genes correlated) =  $30000 C_5 * (1/2^6)^5$~~   
~~=  $2 * 10^{11}$~~

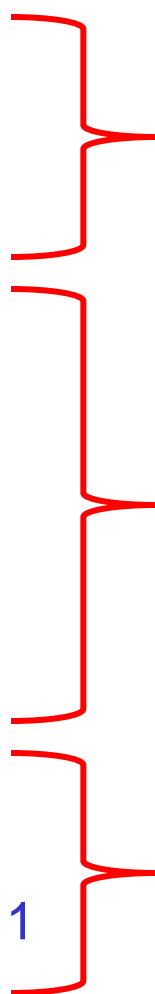
# of pathways = 1000

E(# of pathways correlated) =  $1000 * (1/2^6)^5 = 9.3 * 10^{-7}$

⇒ **Even more false positives?**

- **Perhaps no need to consider every group**

# Towards More Meaningful Genes

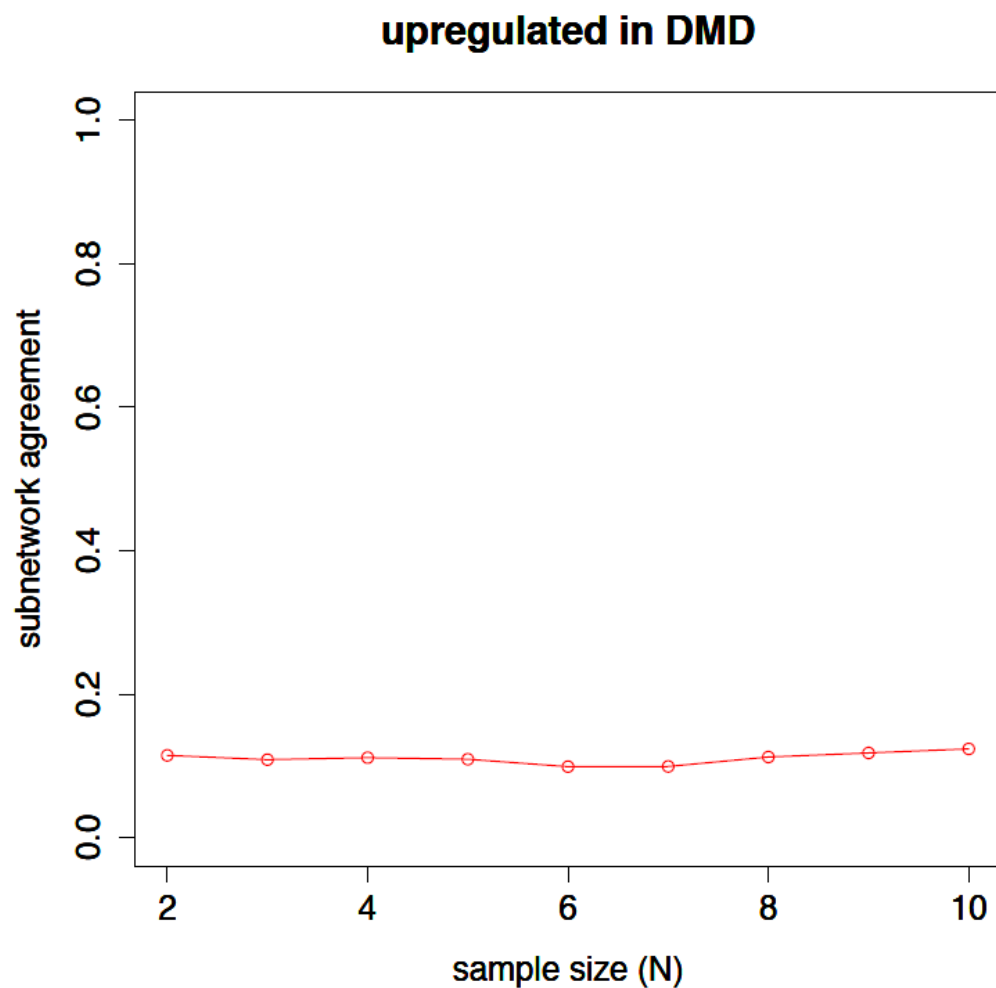
- **ORA**
    - Khatri et al
    - *Genomics*, 2002
  - **FCS**
    - Pavlidis & Noble
    - PSB 2002
  - **GSEA**
    - Subramanian et al
    - *PNAS*, 2005
  - **SNet**
    - Soh et al
    - *BMC Genomics*, 2011
- Overlap Analysis
- Direct-Group Analysis
- Network-Based Analysis
- 

# Intersection Analysis (ORA)

- **Intersect the list of differentially expressed genes with a list of genes on a pathway**
- **If intersection is significant, the pathway is postulated as basis of disease subtype or treatment response**

**Exercise: What is a good test statistics to determine if the intersection is significant?**

# Disappointing Performance



DMD gene expression data

- Pescatori et al., 2007
- Haslett et al., 2002

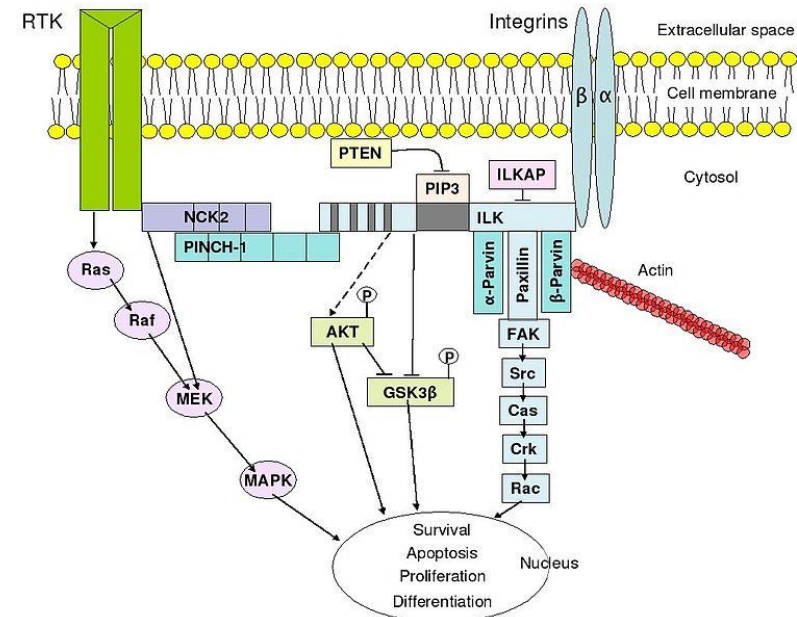
Pathway data

- PathwayAPI, Soh et al., 2010



# Issue #1 with ORA

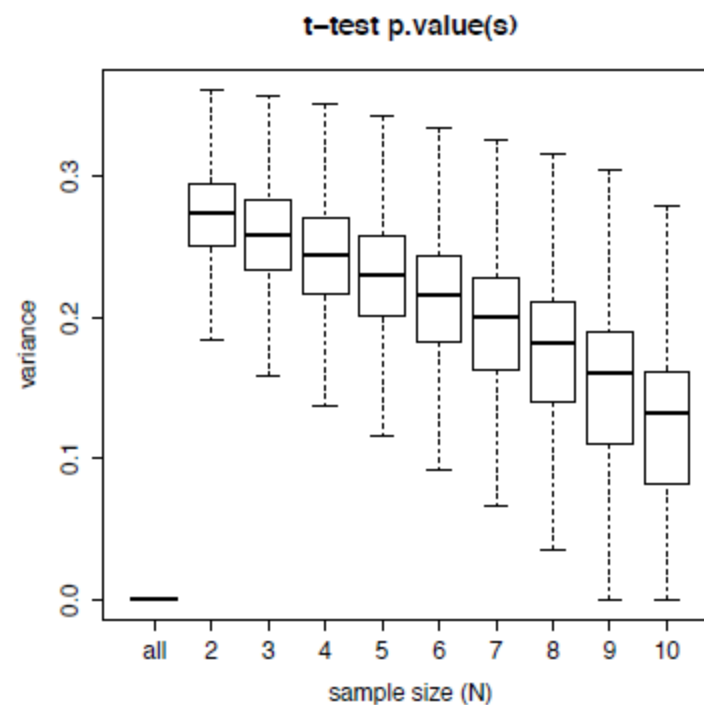
- Its null hypothesis basically says “Genes in the given pathway behaves **no differently** from randomly chosen gene sets of the same size”
- This null hypothesis is obviously false  
 ⇒ Lots of false positives



- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behaviour of genes in a pathway is more coordinated than random ones

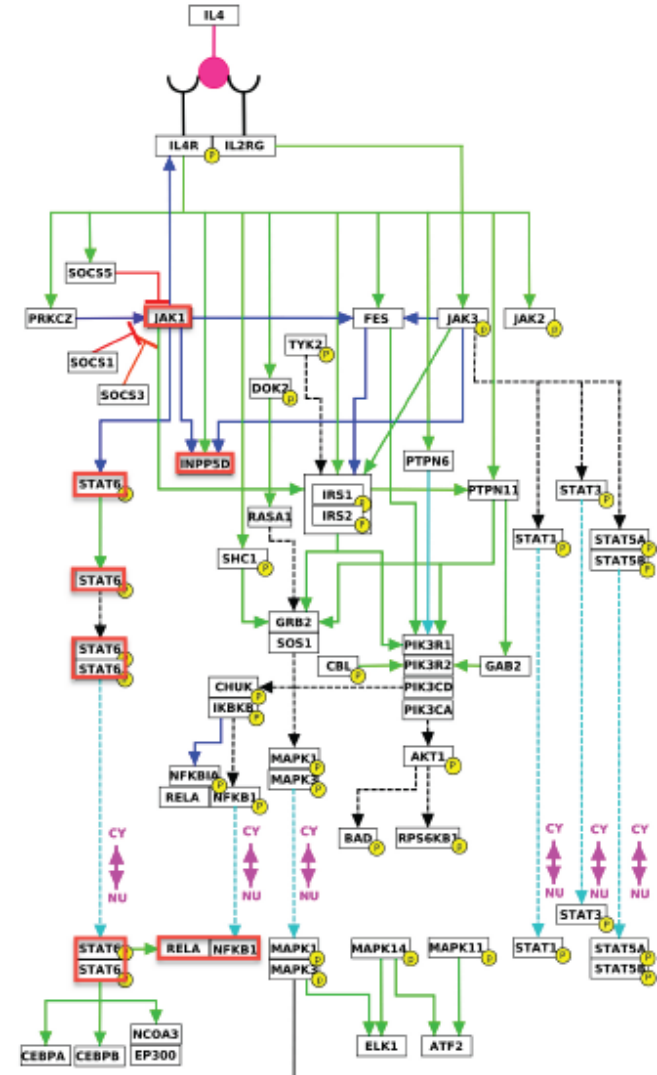
## Issue #2 with ORA

- It relies on a pre-determined list of DE genes
- This list is sensitive to the test statistic used and to the significance threshold used
- This list is unstable regardless of the threshold used when sample size is small



## Issue #3 with ORA

- It tests whether the entire pathway is significantly differentially expressed
- If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch



# ORA-Paired: Paired Test and New Null Hypothesis



- Let  $g_i$  be genes in a given pathway  $P$
- Let  $p_j$  be patients
- Let  $q_k$  be normals
- Let  $\Delta_{i,j,k} = \text{Expr}(g_i, p_j) - \text{Expr}(g_i, q_k)$
- Test whether  $\Delta_{i,j,k}$  is a distribution with mean 0

- **Issue #1 is solved**
  - The null hypothesis is now “If a pathway  $P$  is irrelevant to the difference between patients and normals, then the genes in  $P$  are expected to behave similarly in patients and normals”
- **Issue #2 is solved**
  - No longer need a pre-determined list of DE genes
- **Issue #3 is unsolved**
- **Is sample size now larger?**
  - $|\text{patients}| * |\text{normals}| * |\text{genes in } P|$

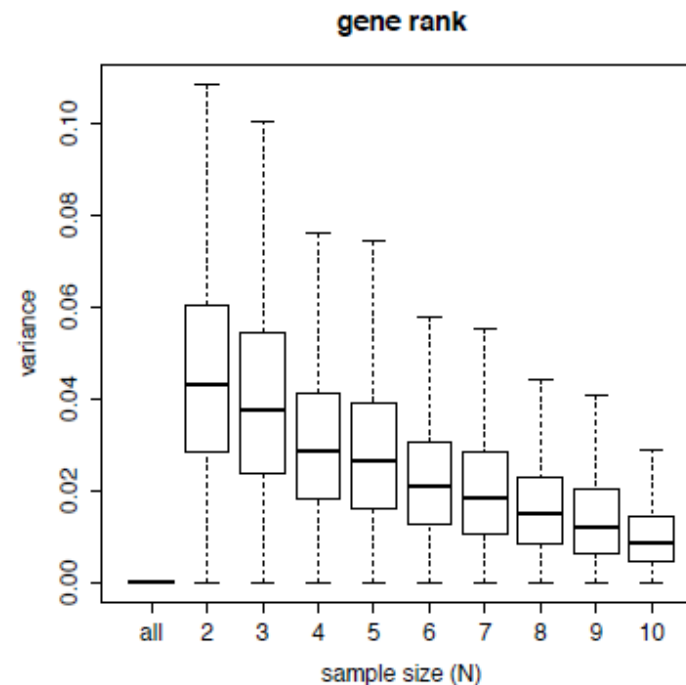
# NEA-Paired: Paired Test on Subnetworks

- **Given a pathway P**
- **Let each node and its immediate neighbourhood in P be a subnetwork**
- **Apply ORA-Paired on each subnetwork individually**

- **Issues #1 & #2 are solved as per ORA-Paired**
- **Issue #3 is partly solved**
  - Testing subnetworks instead of whole pathways
  - But subnetworks derived in fragmented way

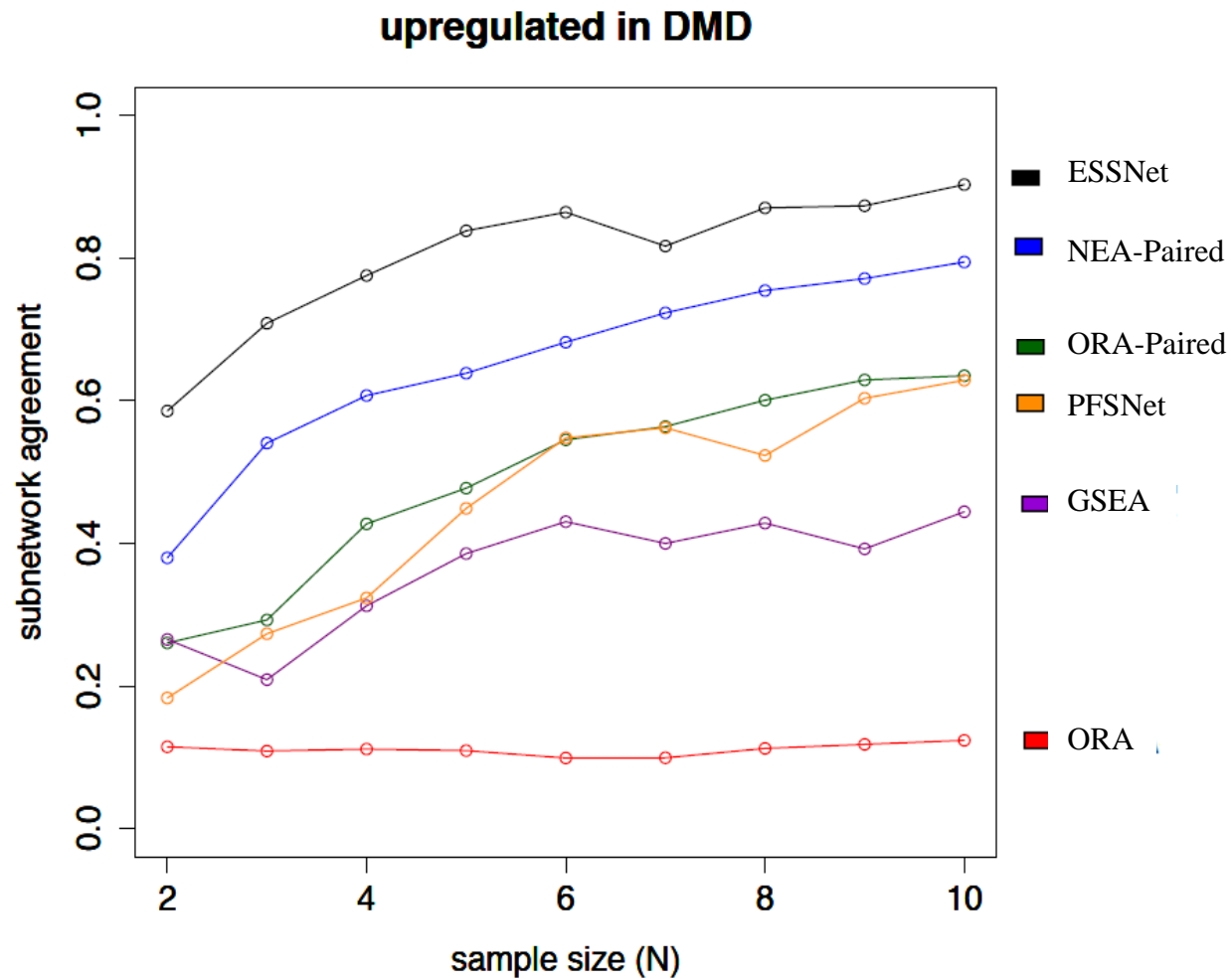
# ESSNet: Larger Subnetworks

- Compute the average rank of a gene based on its expression level in patients
- Use the top  $\alpha\%$  to extract large connected components in pathways
- Test each component using ORA-Paired



- Gene rank is very stable
- Issues #1 - #3 solved

# Fantastic Performance



## Concluding Remarks

- **Consistent successful gene expression profile analysis needs deep integration of background knowledge**
- **Most gene expression profile analysis methods fail to give reproducible results when sample size is small (and some even fail when sample size is quite large)**
- **Logical analysis to identify key issues and simple logical solution to the issues can give fantastic results**



# References

- E.-J. Yeoh et al., “Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling”, *Cancer Cell*, 1:133-143, 2002
- L.D. Miller et al., “Optimal gene expression analysis by microarrays”, *Cancer Cell* 2:353-361, 2002
- J. Li, L. Wong, “Techniques for Analysis of Gene Expression”, *The Practical Bioinformatician*, Chapter 14, pages 319-346, WSPC, 2004
- D. Soh, D. Dong, Y. Guo, L. Wong. “Finding Consistent Disease Subnetworks Across Microarray Datasets”. *BMC Bioinformatics*, 12(Suppl 13):S15, 2011
- K. Lim, L. Wong. “Finding consistent disease subnetworks using PFSNet”. *Bioinformatics*, 30(2):189--196, January 2014

# A Popular Software Package: WEKA





- <http://www.cs.waikato.ac.nz/ml/weka>
- **Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.**

**Exercise: Download a copy of WEKA. What are the names of classifiers in WEKA that correspond to C4.5 and SVM?**

## Let's try WEKA on ...

- **Breast cancer**
- **Dermatology**
- **Pima Indians**
- **Echocardiogram**
- **Mammography**