Bioinformatics and Biomarker Discovery
Part 3: Examples

Limsoon Wong

5 September 2012

Outline

- **ALL**
  - Gene expression profile classification
  - Beyond diagnosis and prognosis

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

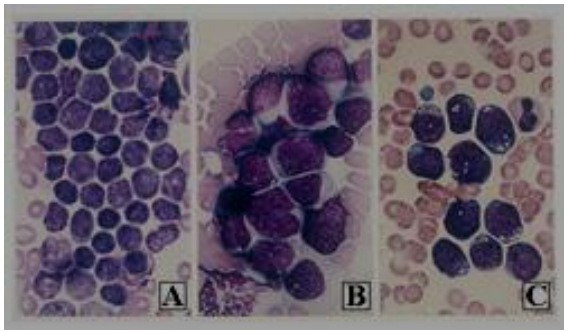
Gene Expression Profile Classification

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-intensiveTx**
  - Relapse

- **The subtypes look similar**



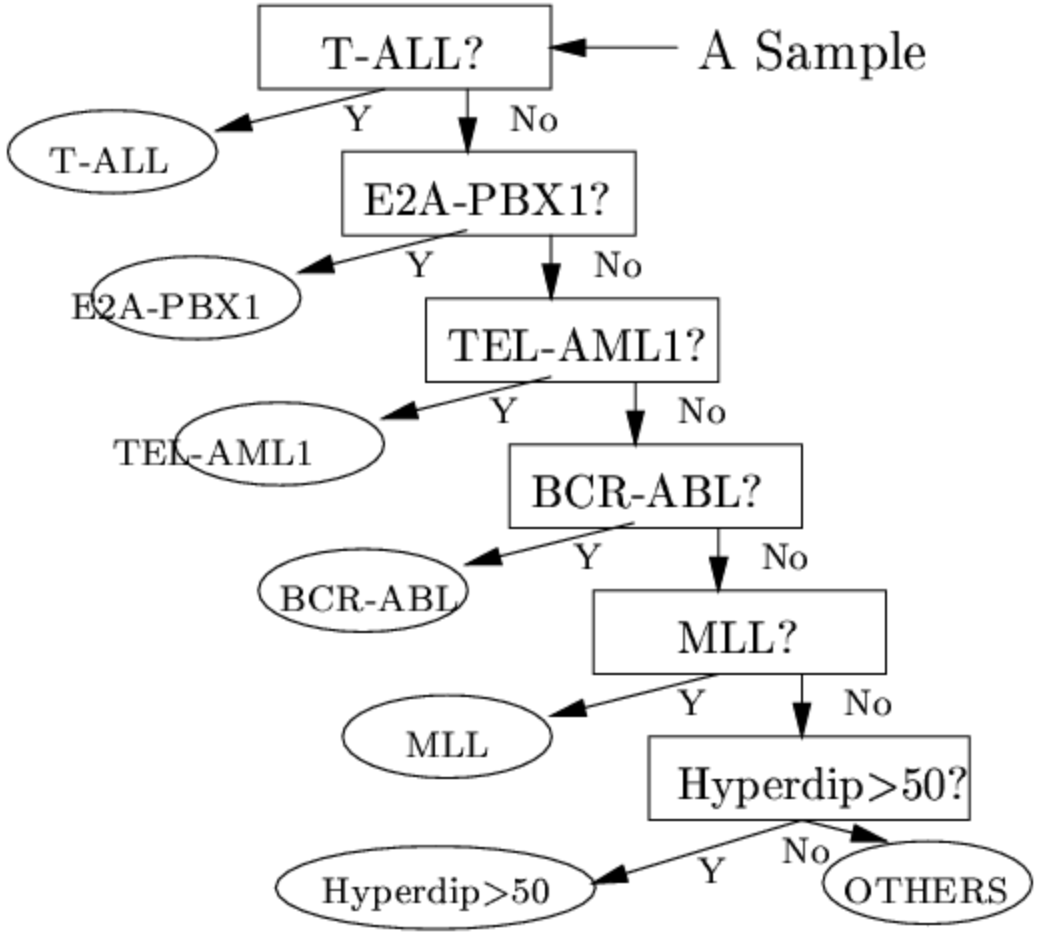
- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics
- **Unavailable in most ASEAN countries**

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 $\mathcal{X}^2$ value of a signal is defined as:

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

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

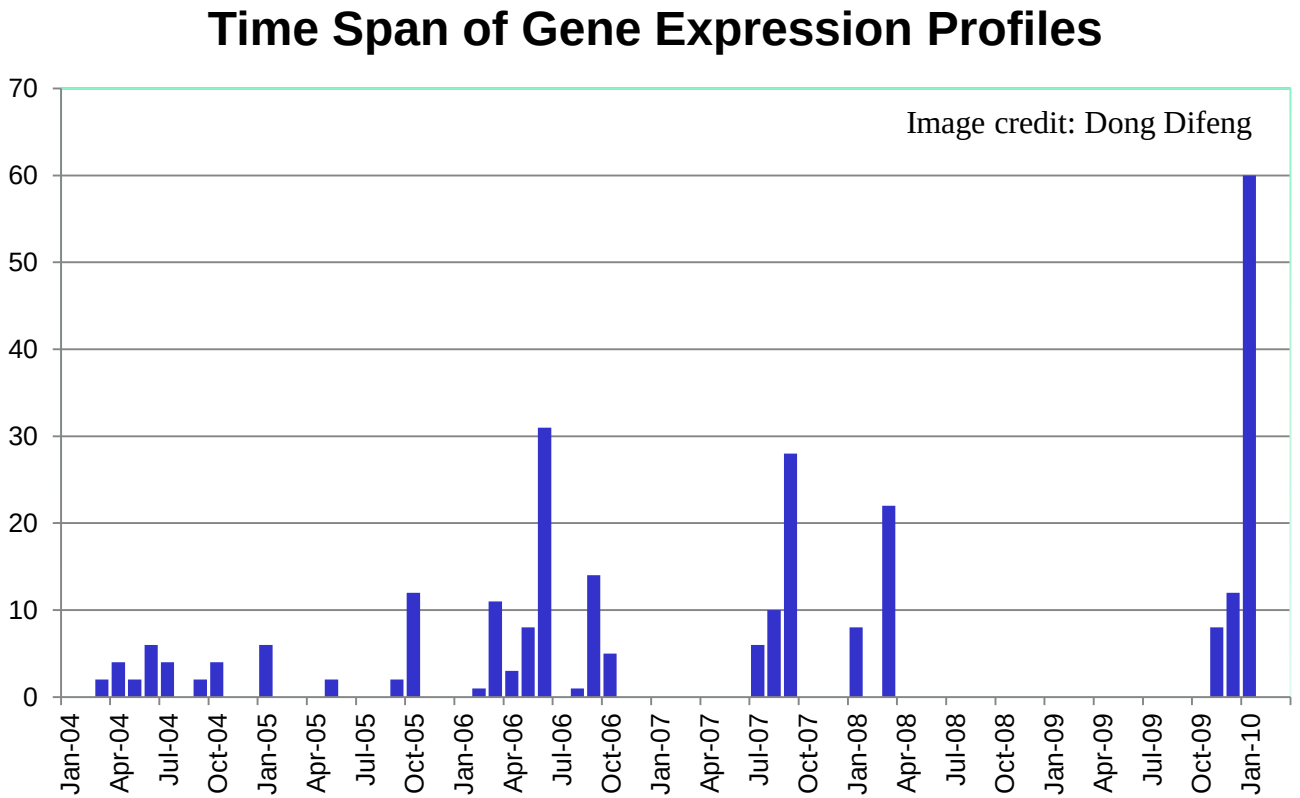
Accuracy of Various Classifiers

Testing Data	Error rate of different models			
	C4.5	SVM	NB	PCL
T-ALL vs OTHERS1	0:1	0:0	0:0	0:0
E2A-PBX1 vs OTHERS2	0:0	0:0	0:0	0:0
TEL-AML1 vs OTHERS3	1:1	0:1	0:1	1:0
BCR-ABL vs OTHERS4	2:0	3:0	1:4	2:0
MLL vs OTHERS5	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…



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

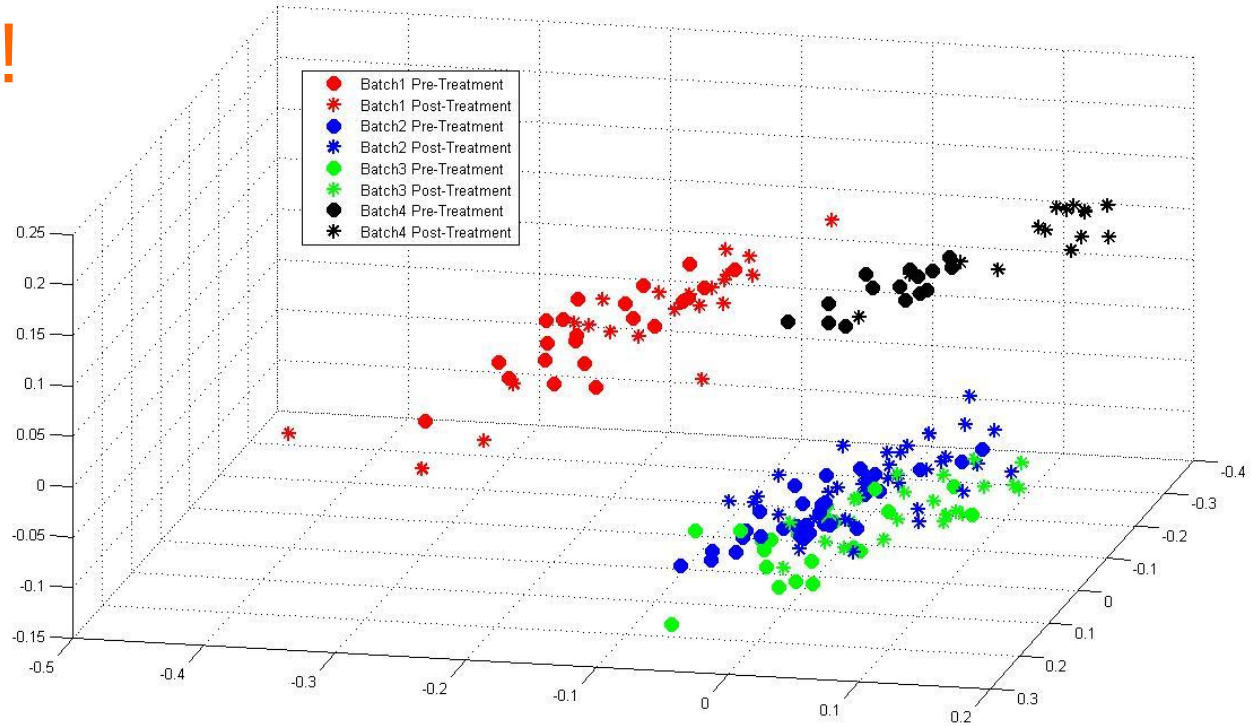


Image credit: Dong Difeng

⇒ Need normalization to correct for batch effect

Approaches to Normalization

- **Aim of normalization:**

  Reduce variance w/o increasing bias

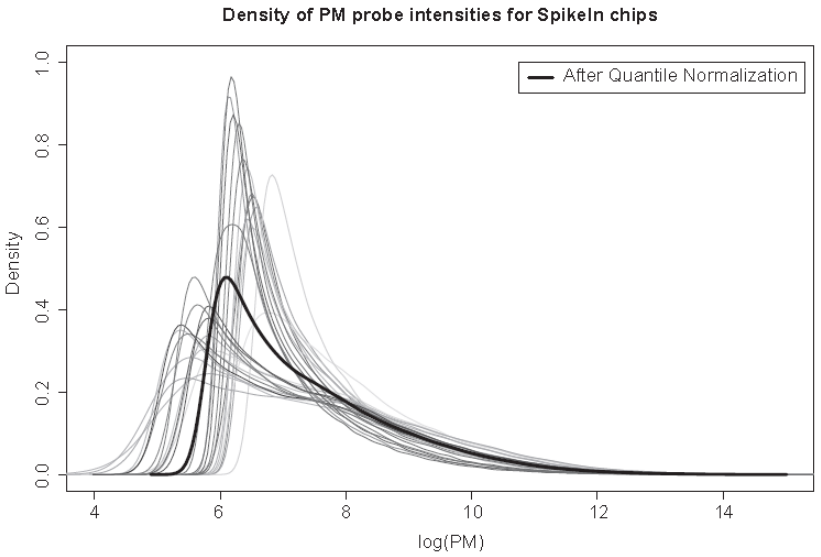
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

- **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 × n where each array is a column*
- Sort each column of *X to give* X_{sort}
- Take means across rows of X_{sort} *and assign this* mean to each elem in the row to get X'_{sort}
- Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as *X*



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

48

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

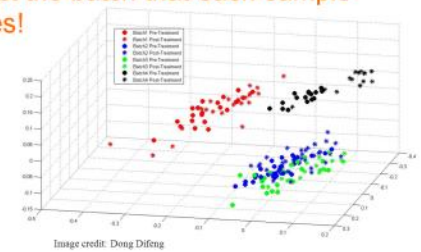


Image credit: Dong Difeng

⇒ Need normalization to correct for batch effect

After quantile normalization

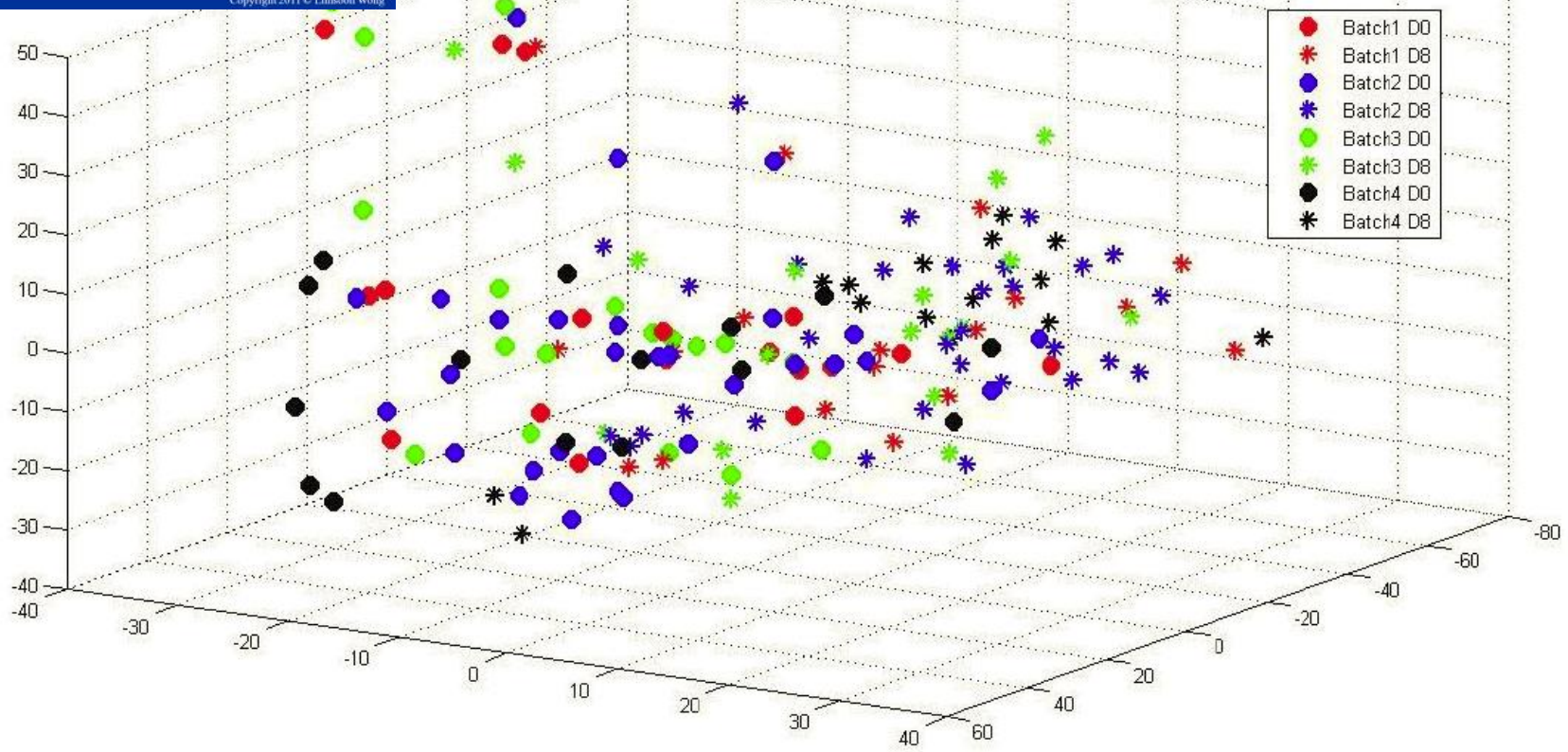


Figure 3.6: GEPs after the batch effects removing.

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 = 100,000**

⇒ **E(# of correlated genes) = 1,562**

⇒ **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?**

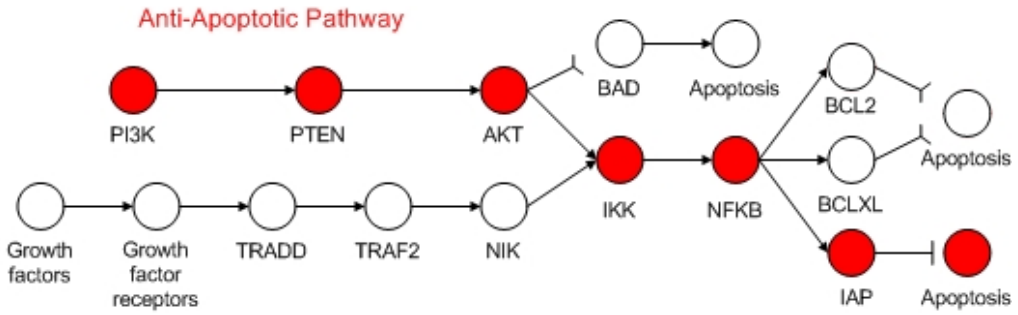
- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- **# of groups = $^{100000}C_5$**

$\Rightarrow$ **E(# of groups of genes correlated) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$**

$\Rightarrow$ **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

NUS National University of Singapore

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, << $1/2^6$
- ~~**# of groups = $^{100000}C_5$**~~
- ~~**E(# of groups of genes correlated) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$**~~

⇒ **Even more false positives?**

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

# of pathways = 1000

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

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**

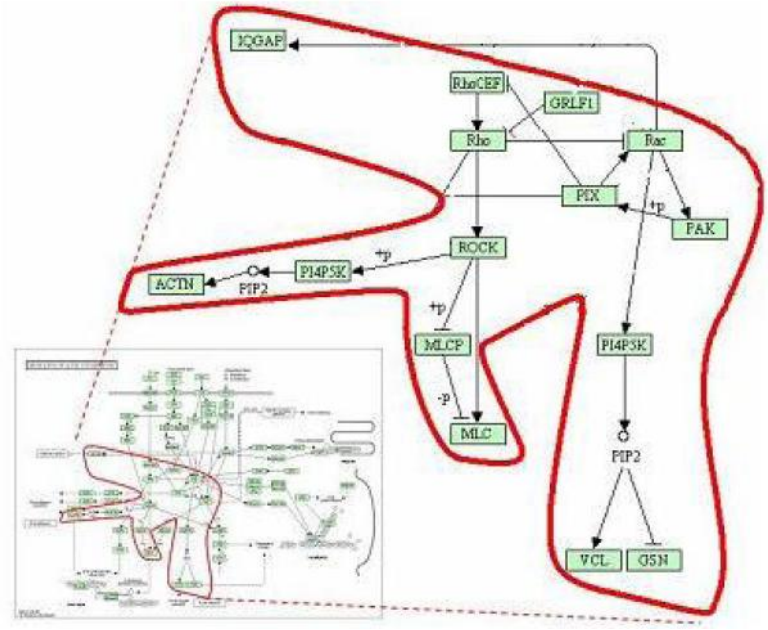
Exercise: What is a good test statistics to determine if the intersection is significant?

Caution:

- **Initial list of differentially expressed genes is defined using test statistics with arbitrary thresholds**
- **Diff test statistics and diff thresholds result in a diff list of differentially expressed genes**

⇒ **Outcome may be unstable**

Connected-Component Analysis (SNet)



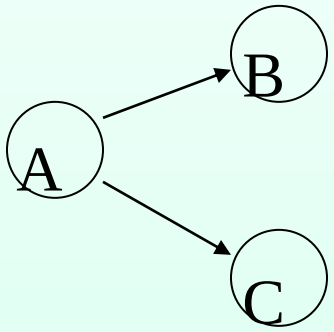
- **Select $C_{P,X}$ if $Scc_{P,X}$ is significant**

$$Scc_{P,X} = \sum_{j \in C_{P,X}} \frac{\#\ patients\ \_in\ \_X\ \_having\ \_high\ \_j}{\#\ patients\ \_in\ \_X}$$

Datasets	DEG	SAM POG	Our POG
Prostate Cancer			
	Top 10	0.30	
	Top 50	0.14	
	Top100	0.15	0.82
Lung Cancer			
	Top 10	0.00	
	Top 50	0.20	
	Top100	0.31	0.70
DMD			
	Top 10	0.20	
	Top 50	0.42	
	Top100	0.54	0.67

Zhang et al, Bioinformatics, 2009

Key Insight # 1



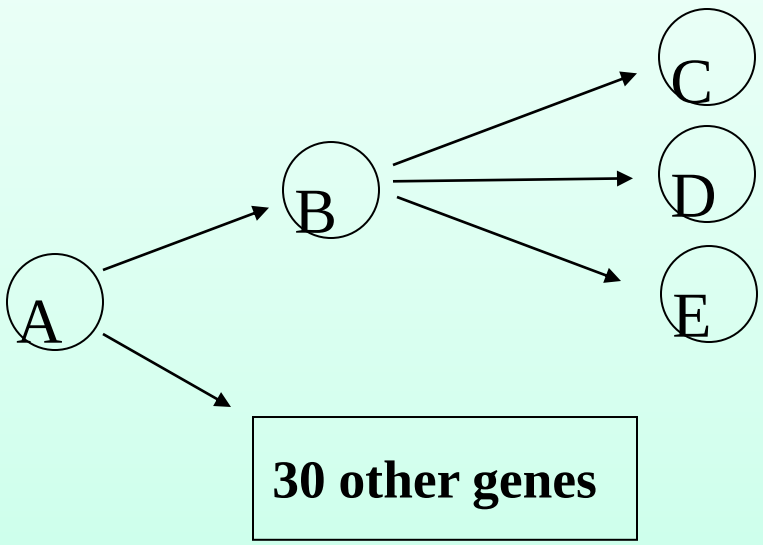
Genes A, B, C are high in phenotype *D*

A is high in phenotype ~*D* but B and C are not

Conventional techniques: Gene B and Gene C are selected. Possible incorrect postulation of mutations in gene B and C

- **SNet does not require all the genes in subnet to be diff expressed**
- **It only requires the subnet as a whole to be diff expressed**
- **Able to capture entire relationship, postulating a mutation in gene A**

Key Insight # 2



A branch within pathway consisting of genes A, B, C, D and E are high in phenotype *D*

Genes C, D and E not high in phenotype ~*D*

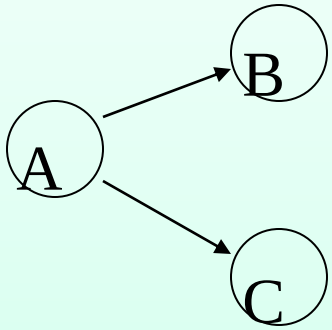
30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

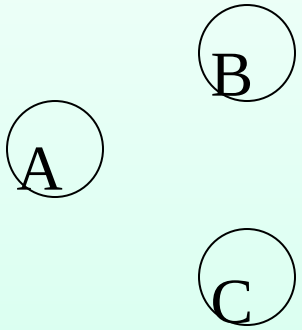
- **SNet: Able to capture the subnetwork branch within the pathway**

Key Insight # 3

Pathway 1



Pathway 2



Genes A, B and C are present in two separate pathways

A, B and C are high in phenotype *D*, but not high in phenotype ~*D*

Conventional techniques:

Both pathways are scored equally. So both got selected, resulting in pathway 2 being a false positive

- **SNet: Able to select only pathway 1, which has the relevant relationship**

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

A Popular Software Package: WEKA

NUS National University of Singapore

- 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**