Delivering Reproducible Gene Expression Analysis

Limsoon Wong

23 September 2011

Plan

- **Some issues in gene expression analysis**
- **Law of large numbers**
- **Using biological background information**
- **Finding more consistent disease subnetworks**

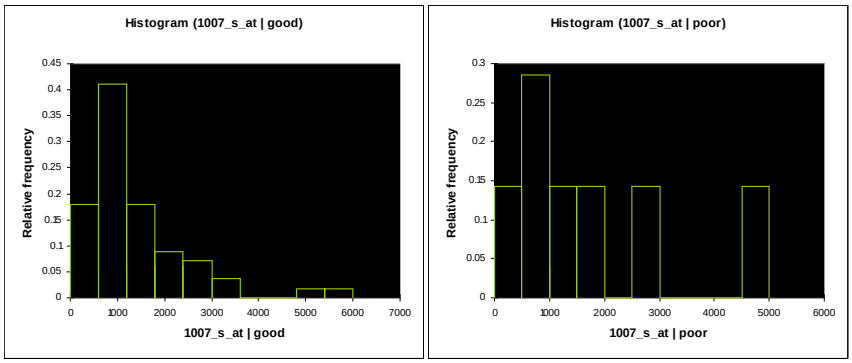
Some Issues in Gene Expression Analysis

Some Headaches

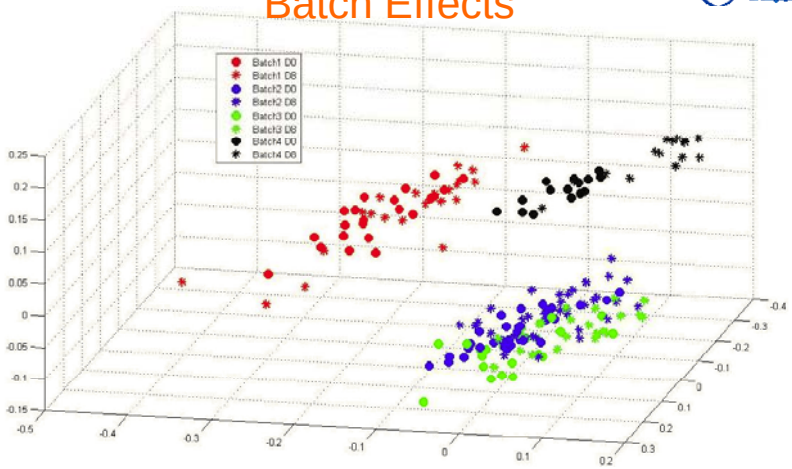
- **Natural fluctuations of gene expression in a person**
- **Noise in experimental protocols**
  - Numbers mean diff things in diff batches
  - Numbers mean diff things in data obtained from diff platforms

⇒ **Selected genes may not be meaningful**
  - Diff genes get selected in diff expts

Natural Fluctuations



Batch Effects



- **Samples from diff batches are grouped together, regardless of subtypes and treatment response**

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

Law of Large Numbers

Law of Large Numbers

- **Suppose you are in a room with 365 other people**
- **Q: What is prob that a specific person in the room has the same birthday as you?**
- **A: 1/365 = 0.3%**
- **Q: What is prob that there is a person in the room having same birthday as you?**
- **A: $1 - (364/365)^{365} = 63\%$**
- **Q: What is prob that there are two persons in the room having same birthday?**
- **A: 100%**

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

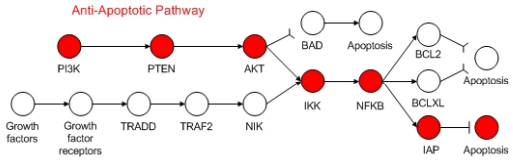
⇒ **Even more false positives?**

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

Using Biological Background Information

13

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**

14

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, << $1/2^6$
- **# 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) = $1000 * (1/2^6)^5 = 9.3*10^{-7}$

Microarray Workshop for Gene Expression Profiling, NUH, 23/9/2011
Copyright 2011 © Limsoon Wong

Towards More Meaningful Genes

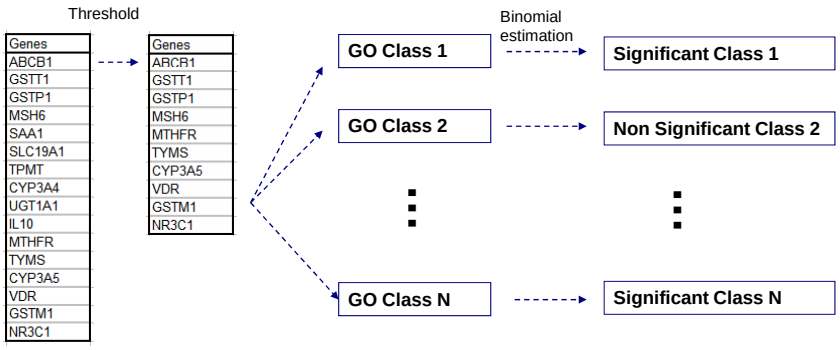
- **ORA**
  - Khatri et al
  - *Genomics*, 2002
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - *PNAS*, 2005
- **SNet**
  - Soh et al
  - *BMC Genomics*, 2011

Overlap Analysis

Direct-Group Analysis

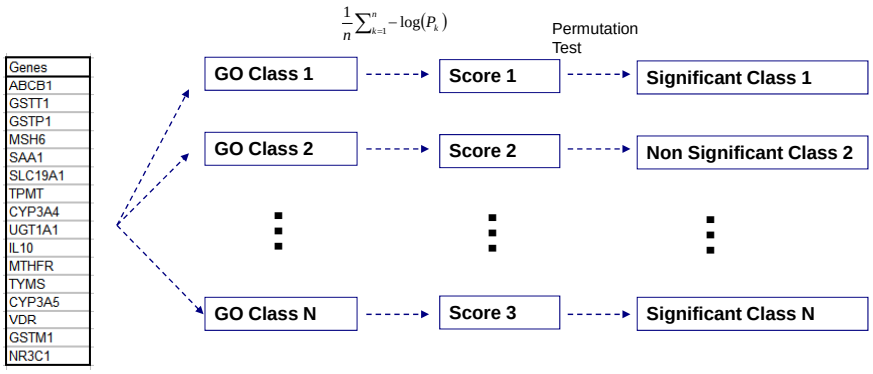
Network-Based Analysis

Overlap Analysis: ORA



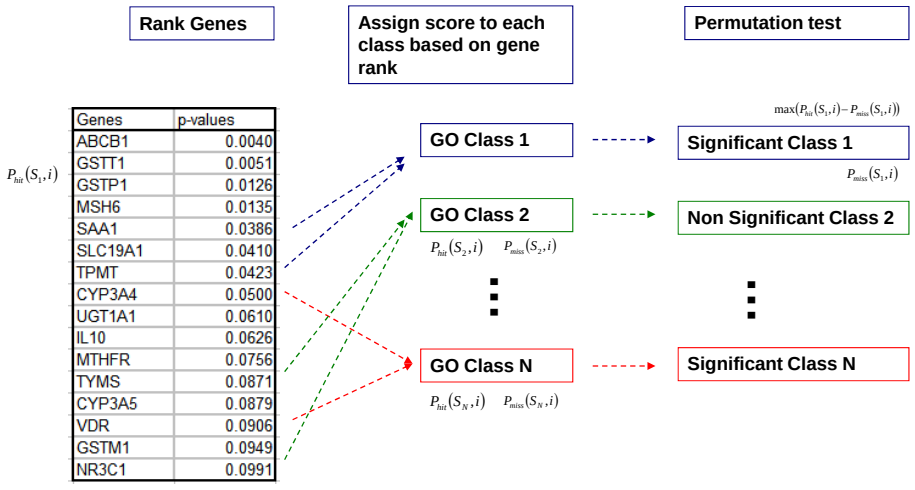
S Draghici et al. "Global functional profiling of gene expression". *Genomics*, 81(2):98-104, 2003.

Direct-Group Analysis: FCS



P Pavlidis et al. "Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex". *Neurochem Res*., 29(6):1213-1222, 2004.

Direct-Group Analysis: GSEA



A Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome wide expression profiles". *PNAS*, 102(43):15545-15550, 2005

More Consistent Disease Subnetworks

Network-Based Analysis: SNet

- **Group samples into type D and ¬D**
- **Extract & score subnetworks for type D**
  - Get list of genes highly expressed in most D samples
    - **These genes need not be differentially expressed!**
  - Put these genes into pathways
  - Locate connected components (ie., candidate subnetworks) from these pathway graphs
  - Score subnetworks on D samples and on ¬D samples
- **For each subnetwork, compute t-statistics on the two sets of scores**
- **Determine significant subnetworks by permutations**

SNet: Score Subnetworks

Step 2: Subnetwork Scoring We assign a score vector $SN^{u\_score}_{sn,d}$ with respect to phenotype d to each subnetwork sn within SN^{List} according to Equation 1.

$$SN^{u\_score}_{sn,d} = \langle SN^{i\_score}_{sn,1,d}, SN^{i\_score}_{sn,2,d}, ..., SN^{i\_score}_{sn,n,d} \rangle \quad (1)$$

Where n is the number of patients in phenotype d. The formula $SN^{i\_score}_{sn,i,d}$ for the i^{th} patient (also the i^{th} element of this vector) is given by:

$$SN^{i\_score}_{sn,i,d} = \sum_{j=1}^{g} G^{score}_{sn,j,d} \quad (2)$$

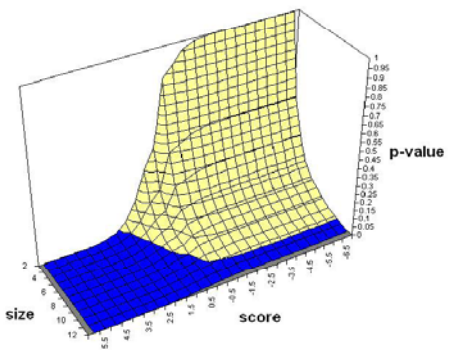
$G^{score}_{sn,j,d}$ refers to the score of the j^{th} gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G^{score}_{sn,j,d}$ is given by Equation 3) and is simply given by:

$$G^{score}_{sn,j,d} = k/n \quad (3)$$

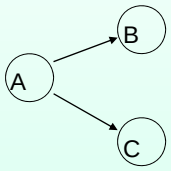
Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d. The entire Step 2 is repeated for the other disease phenotype $\neg d$, giving us the score vectors, $SN^{u\_score}_{sn,d}$ and $SN^{u\_score}_{sn,\neg d}$ for the same set of connected components. The t-test is finally calculated between these two vectors, creating a final t-score for each subnetwork sn within SN_{List}.

SNet: Significant Subnetworks

- **Randomize patient samples many times**
- **Get t-score for subnetworks from the randomizations**
- **Use these t-scores to establish null distribution**
- **Filter for significant subnetworks from real samples**



Key Insight # 1



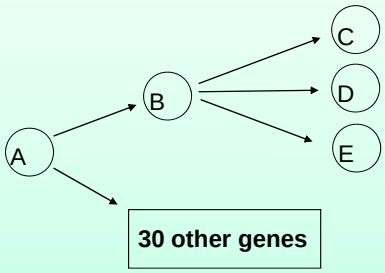
Genes A, B, C are high in phenotype *D*

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*

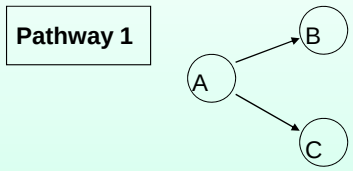
Genes C, D and E not high in phenotype *~D*

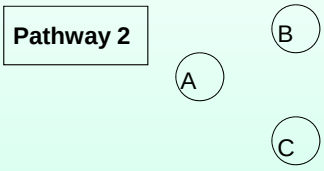
30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

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

Key Insight # 3





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**

Let's see whether SNet gives us subnetworks that are

(i) more consistent between datasets of the same types of disease samples

(ii) larger and more meaningful

Better Subnetwork Overlap

Table 1. Table showing the percentage overlap significant subnetworks between the datasets. Each row refers to a separate disease (as indicated in the first column). Each disease is tested against two datasets depicted in the second and third column. The overlap percentages refer to the pathway overlaps obtained from running SNet (column 4) and GSEA (column 5) The actual number of overlaps are parenthesized in the same columns.

Disease	Dataset 1	Dataset 2	SNet	GSEA
Leuk	Golub	Armstrong	83.3% (20)	0.0% (0)
Subtype	Ross	Yeoh	47.6% (10)	23.1% (6)
DMD	Haslett	Pescatori	58.3% (7)	55.6% (10)
Lung	Bhatt	Garber	90.9% (9)	0.0% (0)

- **For each disease, take significant subnetworks from one dataset and see if it is also significant in the other dataset**

Better Gene Overlaps

Table 2. Table showing the number and percentage of significant overlapping genes. γ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding algorithms.

Disease	γ	SNet	GSEA	SAM	t-test
Leuk	84	91.3%	2.4%	22.6%	14.3%
Subtype	75	93.0%	4.0%	49.3%	57.3%
DMD	45	69.2%	28.9%	42.2%	20.0%
Lung	65	51.2%	4.0%	24.6%	26.2%

- **For each disease, take significant subnetworks extracted independently from both datasets and see how much their genes overlap**

Larger Subnetworks

Table 3. Table comparing the size of the subnetworks obtained from the t-test and from SNet. The first column shows the disease and the second column shows the number of genes which comprised of the subnetworks. The third and fourth column depicts the number of genes present within each subnetwork for the t-test and SNet respectively. So for instance in the leukemia dataset, we have 8 subnetworks with size 2 genes, 1 subnetwork with size 3 genes for the t-test. For SNet, we have 2 subnetworks with size 5 genes, 3 subnetworks with size 6 genes, 2 subnetworks with size 7 genes and 1 subnetwork with a size of ≥ 8 genes

Disease	γ	Num Genes (t-test)				Num Genes (SNet)			
		2	3	4	5	5	6	7	≥ 8
Leuk	84	8	1	0	0	2	3	2	1
Subtype	75	5	1	1	1	1	0	1	6
DMD	45	3	1	0	0	1	0	0	5
Lung	65	3	2	1	0	5	3	0	1

Remarks

What have we learned?

- **Common headaches in gene expression analysis**
  - Natural fluctuation, protocol noise, batch effect
- **Use of biological background info to tame false positives**
- **Overlap analysis → direct-group analysis → network-based analysis**
- **SNet method yields more consistent and larger disease subnetworks**

Acknowledgements

Donny Soh

Difeng Dong

- **A*STAR AIP scholarship**
- **A*STAR SERC PSF grant**

Agency for Science, Technology and Research
SINGAPORE