Enabling More Reproducible Gene Expression Analysis

Limsoon Wong

Plan

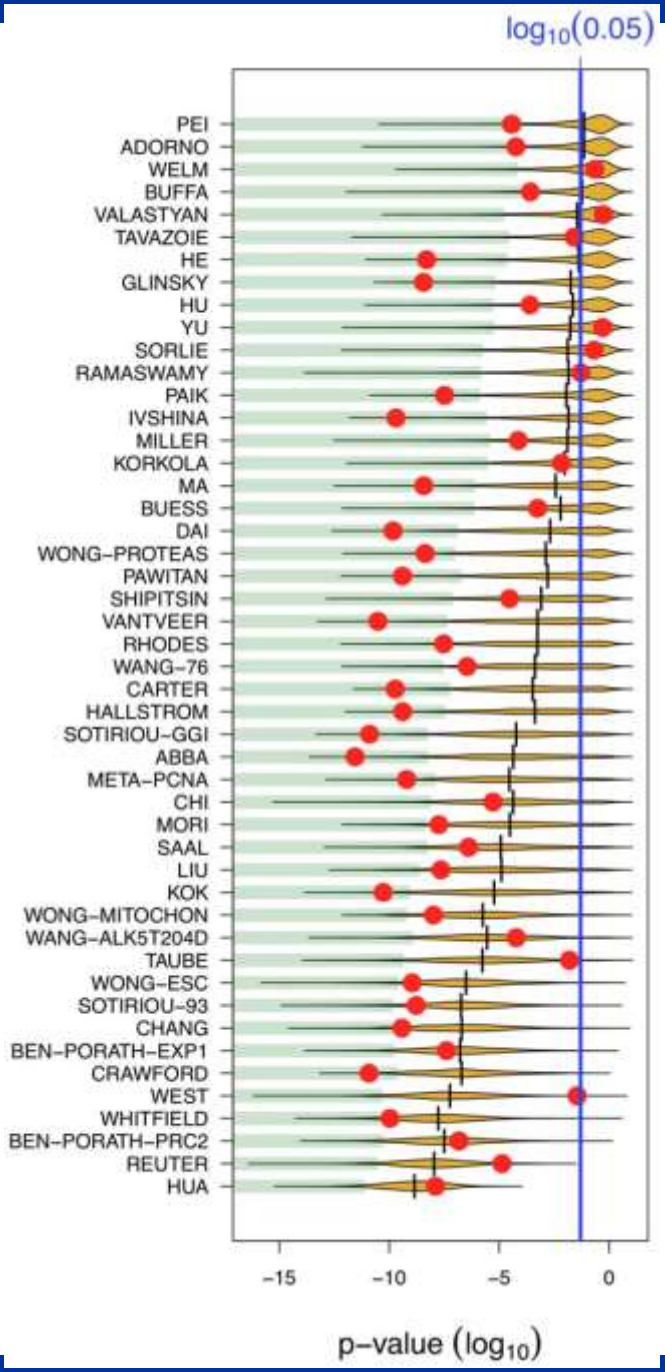
- **Motivation**
- **SNet: A network-based approach**
- **PFSNet: Two refinements to SNet**
- **Remarks**

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



"Most random gene expression signatures are significantly associated with breast cancer outcome"

Venet et al., *PLoS Comput Biol*, 7(10):e1002240, 2011.

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?**

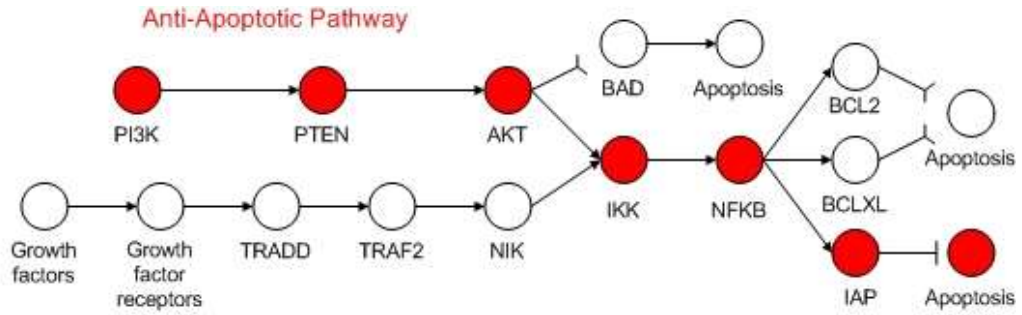
- **Prob(a gene is correlated) = $1/2^6$**
- **# of genes on array = 100,000**

$\Rightarrow$ **E(# of correlated genes) = 1,562**

$\Rightarrow$ **Many false positives**

- **These cannot be eliminated based on pure statistics!**

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**

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
- **PFSNet**

Overlap Analysis

Direct-Group Analysis

Network-Based Analysis

GSEA: Key Points

- **"Enrichment score"**
  - The degree that the genes in gene set C are enriched in the extremes of ranked list of all genes
  - Measured by Komogorov-Smirnov statistic

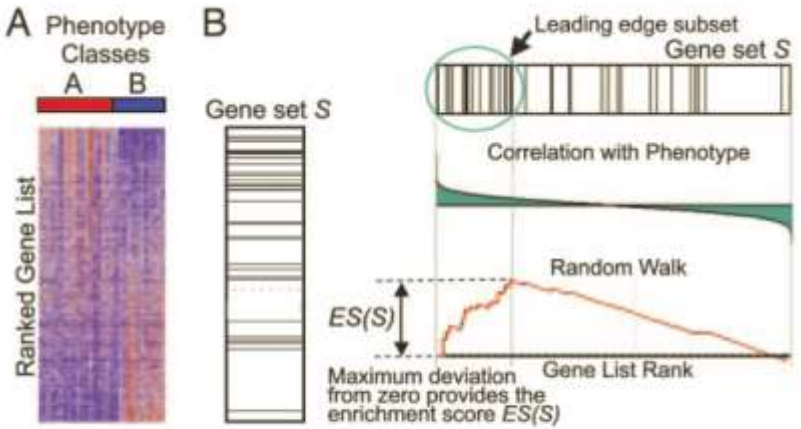


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

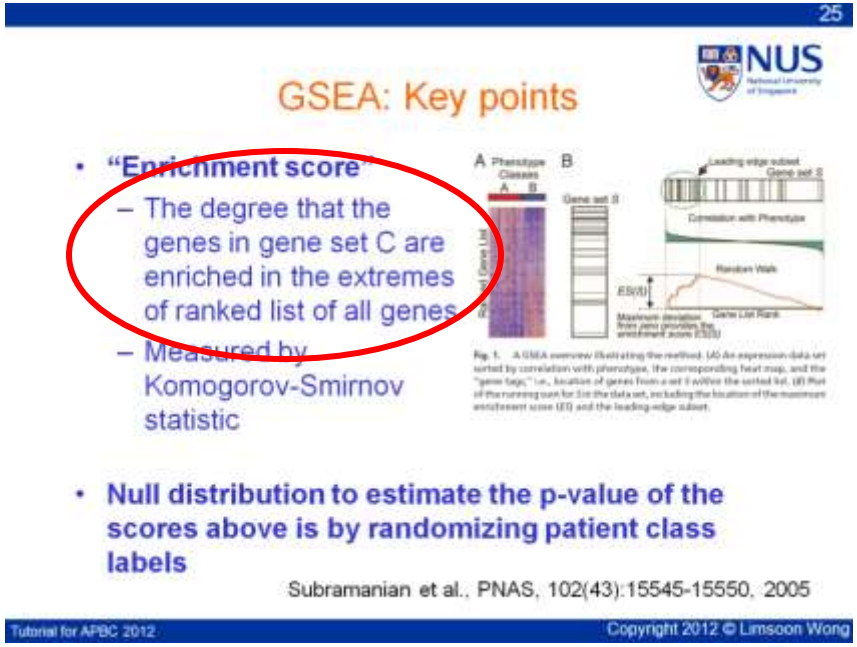
Subramanian et al., *PNAS*, 102(43):15545-15550, 2005

- **Null distribution to estimate the p-value of the scores above is by randomizing patient class labels**

Wong. “Using Biological Networks in Protein Function Prediction and Gene Expression Analysis”. *Internet Mathematics*, 7(4):274--298, 2011.

NUS National University of Singapore

A problem w/ GSEA



- **Its enrichment score considers all genes in C**
- **But …**
  - Not all branches of a large pathway have to “go wrong”

  ⇒ Cannot detect if only a small part of a pathway malfunctions
- **Solution: Break pathways into subnetworks**

Plan

- Motivation
- **SNet: A subnetwork-based approach**
- **PFSNet: Two refinements to SNet**
- **Remarks**

Soh et al. "Finding Consistent Disease Subnetworks Across Microarray Datasets". *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

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-statistic on the two sets of scores**
- **Determine significant subnetworks by permutations**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

NUS National University of Singapore

SNet: Score Subnetworks

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

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

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

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

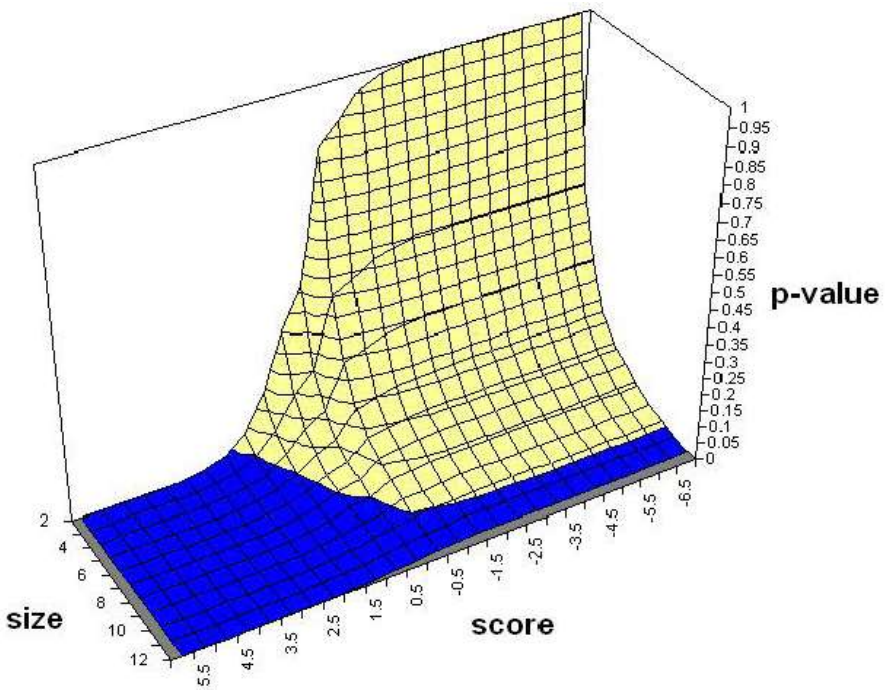
$G_{sn,j,d}^{score}$ refers to the score of the j^{th} gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G_{sn,j,d}^{score}$ is given by Equation 3) and is simply given by:

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

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, $SN_{sn,d}^{v\_score}$ and $SN_{sn,\neg d}^{v\_score}$ 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}.

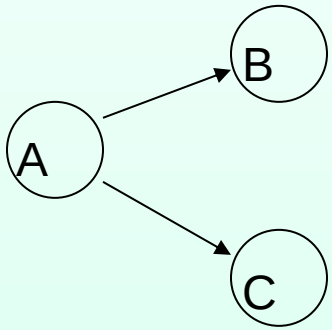
SNet: Significant Subnetworks

- **Randomize sample labels 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**



Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Key Insight # 1



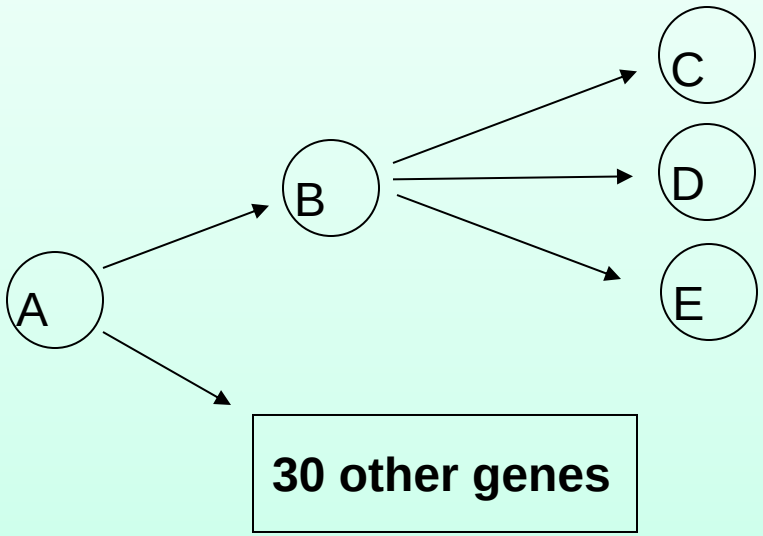
Genes A, B, C are high in phenotype *D*

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*

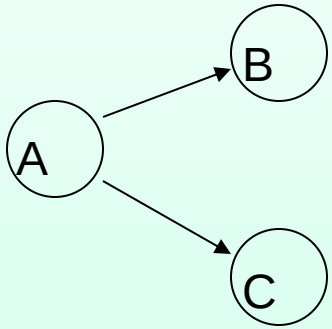
30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

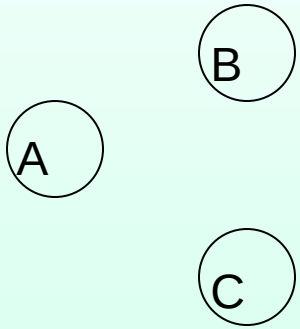
- **SNet: Able to capture the subnetwork branch within the pathway**

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**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

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.

Overlap = $|A \cap B| / \min(|A|,|B|)$

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**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

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.

Overlap = $|A \cap B| / \min(|A|,|B|)$

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**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

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 $\geq$ 8 genes

Disease	γ	Num Genes (t-test)				Num Genes (SNet)			
		2	3	4	5	5	6	7	$\geq$ 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

Plan

- Motivation
- SNet: A subnetwork-based approach
- **PFSNet: Two refinements to SNet**
- **Remarks**

Issue #1 with SNet

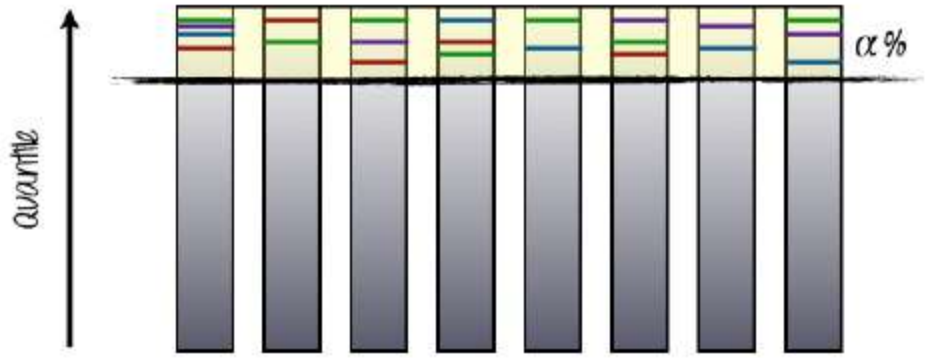


Fig. 2. In SNet, the top $\alpha\%$ of genes of each sample in phenotype D is highlighted in yellow. A subset of these genes that are thus highlighted in at least 50% of the samples are then taken to induce subnetworks.

- What if the real important genes are close to, but not in, the top $\alpha\%$ most highly expressed genes?
- Blindly increasing α does not help, as this will bring in lots of false-positive genes

Issue #2 with SNet

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

$G^{score}_{sn,j,d}$ 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:

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

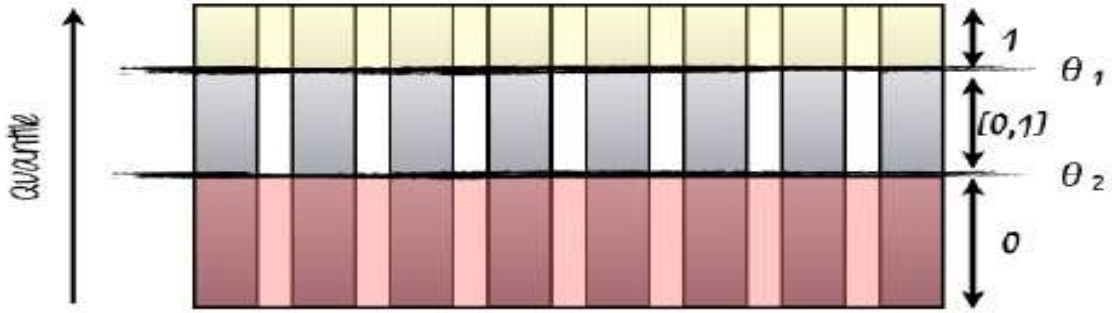
Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d.

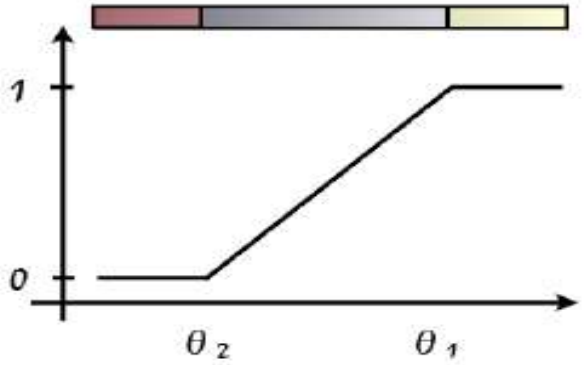
- **SNet weighs genes & scores subnetworks only on the basis of phenotype D**
- **Why not consider phenotype ~D as well?**

PFSNet

- **Deal with issue #1 of SNet using "fuzzification"**
- **Deal with issue #2 of SNet using paired t-test**

$\Rightarrow$ **PFSNet – Paired Fuzzy SNet**





Fuzzification

Our goal in this step is to compute a gene list, which segregates the pathways into smaller components. The voting criteria that determines whether the gene g_i is accepted into this gene list is given below:

$$\sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|} > \beta \qquad (1)$$

where D is the phenotype for which the subnetwork is generated, p_j ranges over the patients of phenotype D and fs is the fuzzy function which converts the gene expression value e_{g_i,p_j} to a value between 0 and 1.

In PFSNet, instead of computing the gene scores with respect to phenotype D, we also compute the gene scores with respect to phenotype $\neg D$. Hence, each node is given scores which we denote as $\beta_1^*(g_i)$ and $\beta_2^*(g_i)$, computed as follows:

$$\beta_1^*(g_i) = \sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|}, \quad \beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{fs(e_{g_i,p_j})}{|\neg D|} \qquad (4)$$

Accordingly, for every subnetwork S, each patient of phenotype D can be scored under β_1^* and β_2^*, as follows:

$$Score_1^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_1^*(g_i), \qquad (5)$$

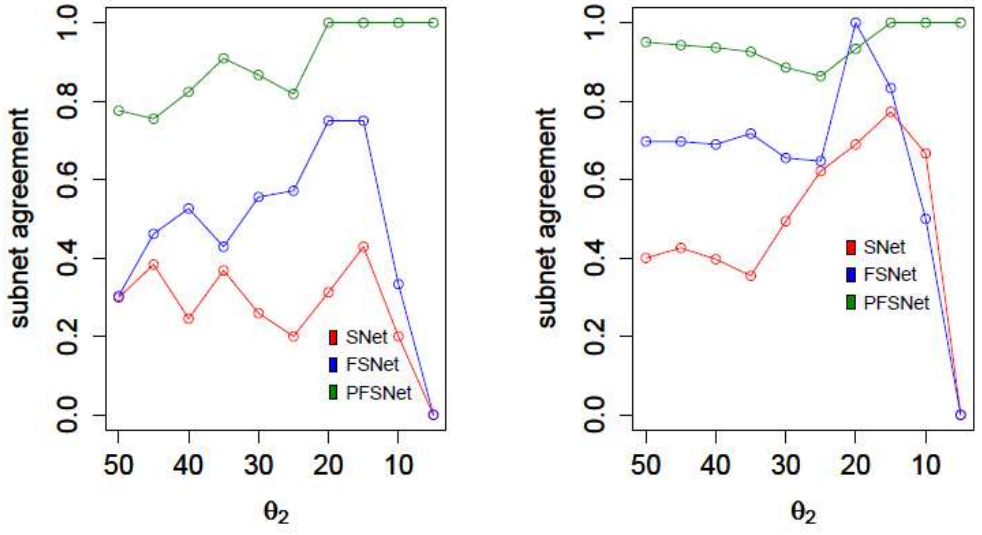
$$Score_2^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_2^*(g_i) \qquad (6)$$

Paired T-Test

- **$Score^{Pk}_1(S)$ and $Score^{Pk}_2(S)$ are computed for the same sample Pk and subnetwork S**

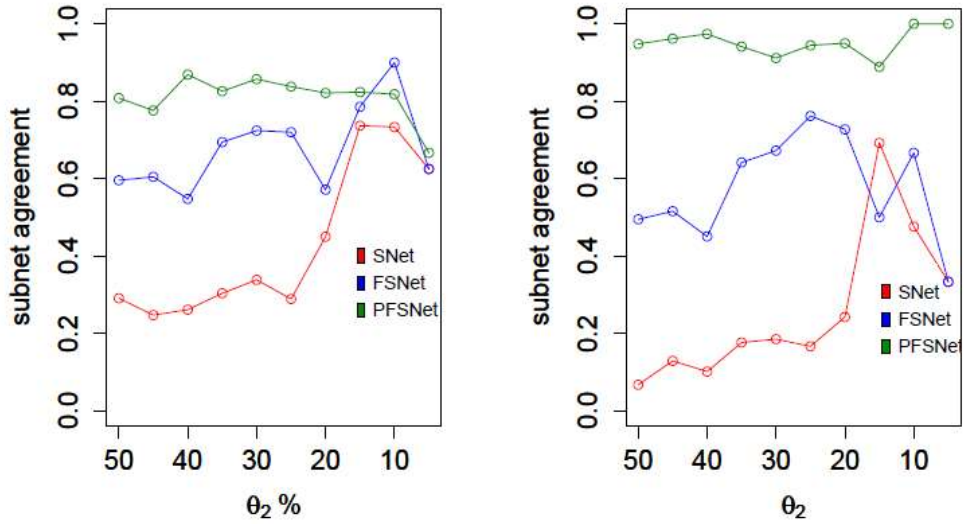
$\Rightarrow$ **Can do paired t-test**

  - Null hypothesis: If S is irrelevant to D vs ~D, we expect $Score^{Pk}_1(S) - Score^{Pk}_2(S)$ to be around 0



upregulated in ALL upregulated in AML

Fig. 4: Consistency of subnetworks in Leukemia dataset

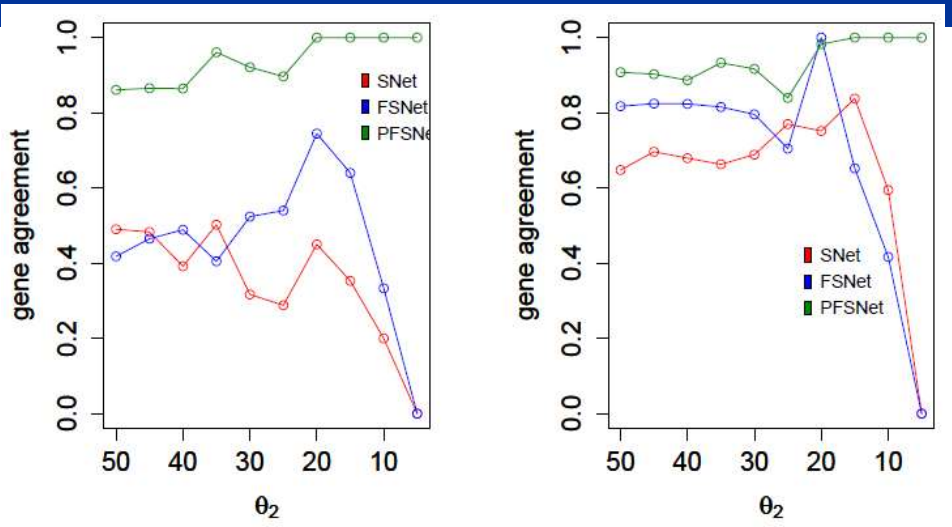


upregulated in DMD upregulated in NORM

Fig. 6: Consistency of subnetworks in DMD dataset

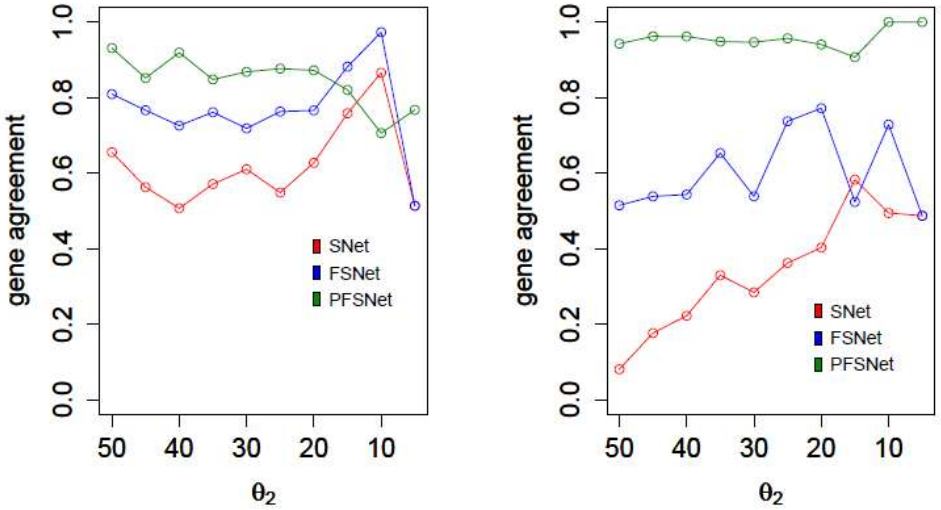
PSFNet vs SNet: Subnet Agreement

Overlap = |A∩B| / |A∪B|



upregulated in ALL upregulated in AML

Fig. 7: Consistency of genes in Leukemia dataset



upregulated in DMD upregulated in NORM

Fig. 9: Consistency of genes in DMD dataset

PSFNet vs SNet: Gene Agreement

Overlap = $|A \cap B| / |A \cup B|$

PFSNet vs GSEA & GGEA: Pathway Agreement

Dataset	PFSNet	FSNet	GSEA	GGEA
Leukemia	1.00	0.75	0.12	0.18
ALL (subtype)	0.56	0.38	0.34	0.37
DMD	0.82	0.79	0.57	0.51

For PFSNet and FSNet, threshold values of $\theta_1 = 0.95, \theta_2 = 0.85$ are used.

Overlap = $|A \cap B| / |A \cup B|$

PFSNet vs T-Test: Gene Agreement

Dataset	PFSNet		FSNet		SNet		t-test	
	D	¬D	D	¬D	D	¬D	D	¬D
Leukemia	1.00	0.81	0.64	0.42	0.35	0.58	0.21	0.20
ALL (subtype)	0.54	0.70	0.38	0.41	0.29	0.57	0.08	0.08
DMD	0.82	0.72	0.88	0.75	0.76	0.54	0.36	0.14

For PFSNet and FSNet, threshold values of $\theta_1 = 0.95, \theta_2 = 0.85$ are used. D represents subnetworks enriched in phenotype D and $\neg D$ represents subnetworks enriched in phenotype $\neg D$.

Overlap = $|A \cap B| / |A \cup B|$

PFSNet vs GSEA & GGEA: Pathway Agreement

Dataset	PFSNet	FSNet	GSEA	GGEA
Leukemia	1.00	0.75	0.12	0.18
ALL (subtype)	0.56	0.38	0.34	0.37
DMD	0.82	0.79	0.57	0.51

Testing subnets from PFSNet using GSEA & GGEA

	PFSNet	FSNet	SNet
Leukemia (GSEA)	0.50	0.00	0.00
Leukemia (GGEA)	0.67	0.50	0.50
ALL subtype (GSEA)	1.00	0.15	0.11
ALL subtype (GGEA)	1.00	0.47	0.35
DMD (GSEA)	0.90	0.57	0.50
DMD (GGEA)	0.54	0.71	0.45

Top 5 Subnets

Leukemia	ALL subtype	DMD
Proteasome Degradation	Wnt Signaling*	Striated Muscle Contraction*
IL-4 Signaling*	Antigen Processing	Integrin Signaling
Antigen Processing*	Jak-STAT Signaling*	VEGF Signaling*
B-Cell Receptor Signaling	T-Cell Receptor Signaling	Tight Junction
Wnt Signaling*	Adherens Junction*	Actin Cytoskeleton Signaling

The asterisk indicates subnetworks that were not found in SNet

DMD: Striated Muscle Contraction

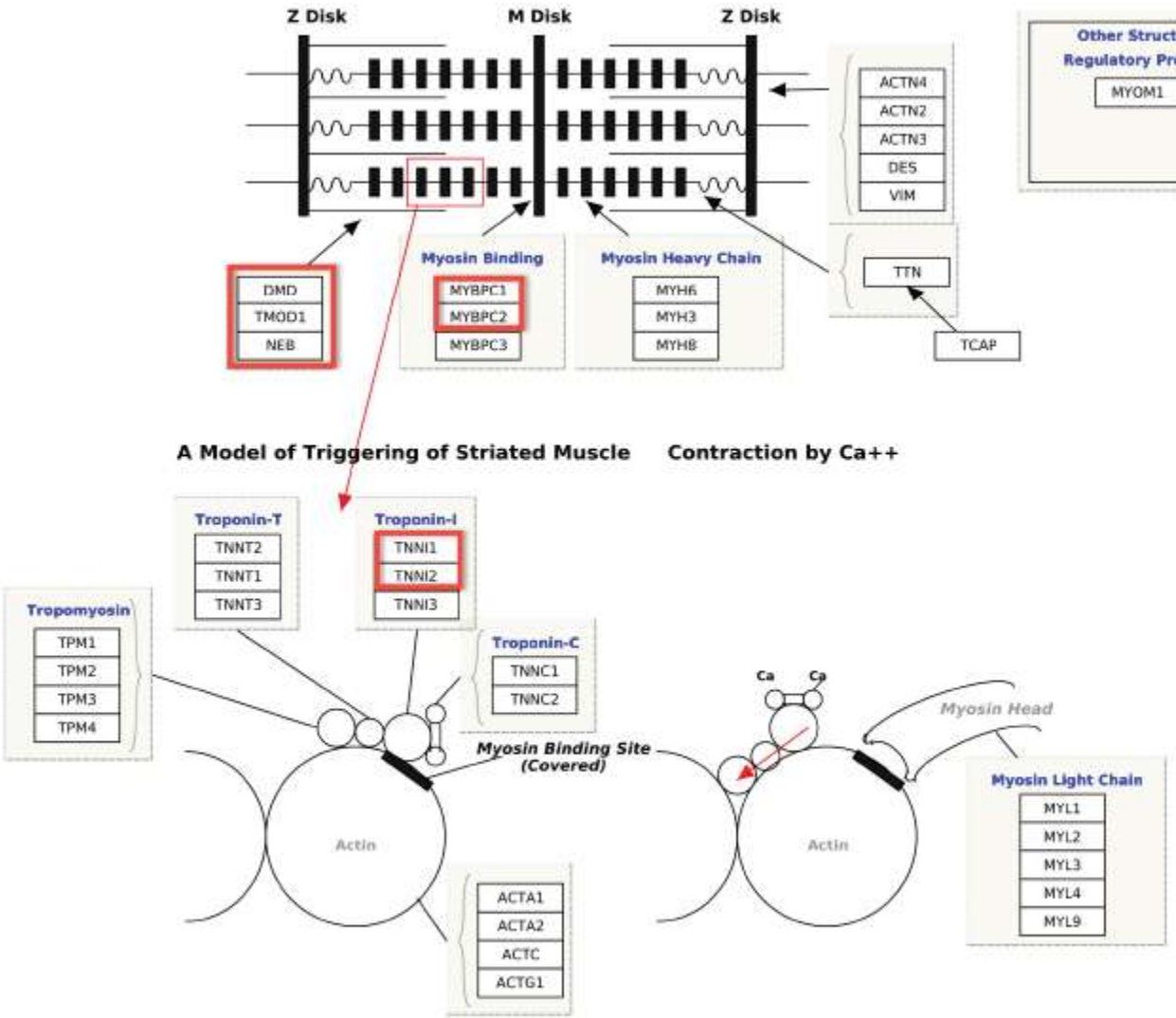
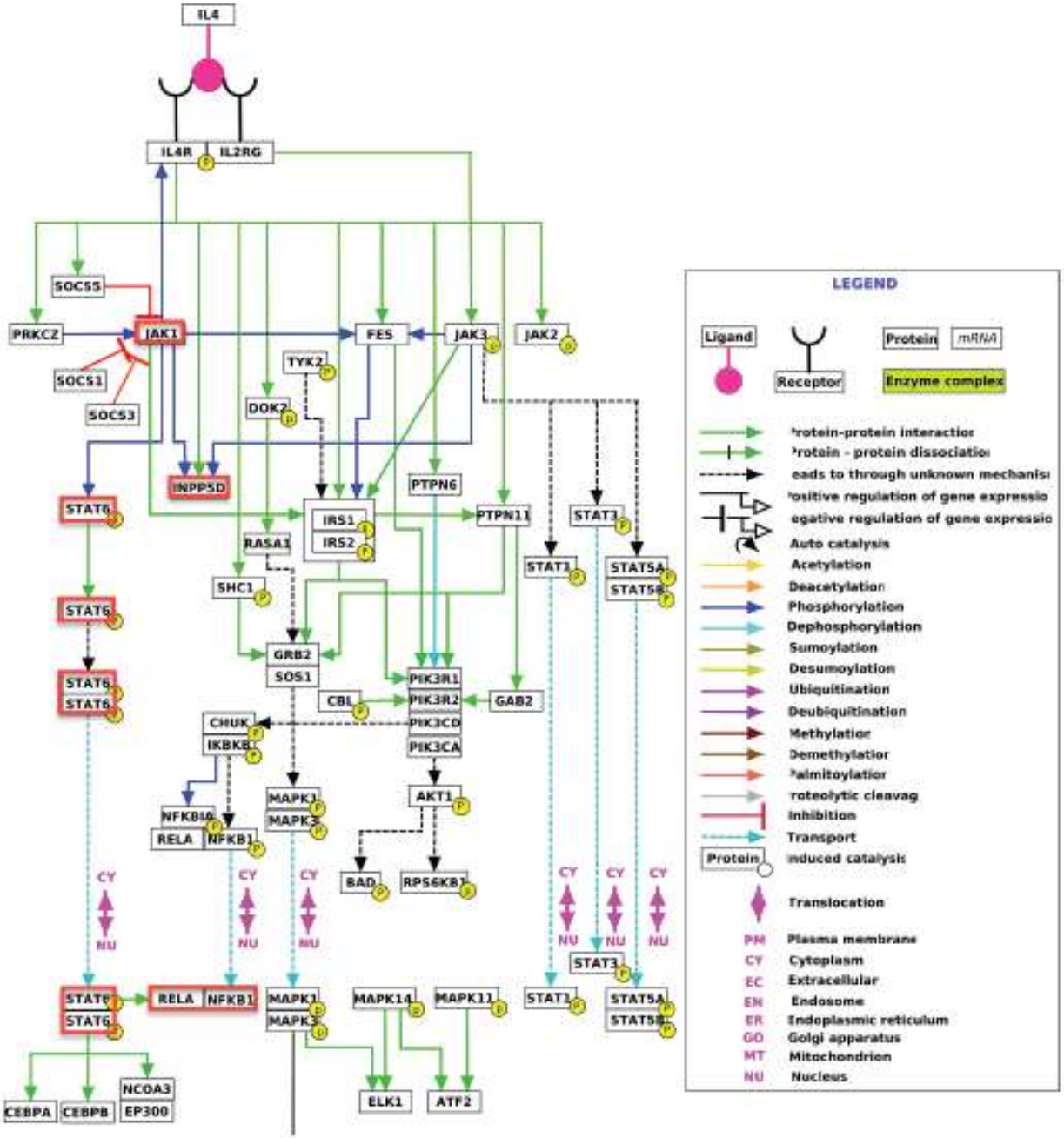


Fig. 5. An example of a biologically relevant pathway for DMD. The nodes from the induced subnetwork identified by PFSNet is highlighted with red boxes.

NUS National University of Singapore



Leukemias: IL-4 Signaling in ALL

What have we learned?

- **Use of biological background info to tame false positives**
- **Overlap analysis → direct-group analysis → network-based analysis**
- **Subnetwork-based methods yield more consistent and larger disease subnetworks**
- **Fuzzification and paired test help also**

Still a major challenge

- **Suppose there are very few samples, so few that you cannot estimate the p-value by permuting class labels**

- **What do you do?**

Acknowledgements

Donny Soh

Kevin Lim