Mining testable hypotheses from bio big data: Use of context in gene expression analysis

Wong Limsoon

NUS
National University of Singapore

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?**

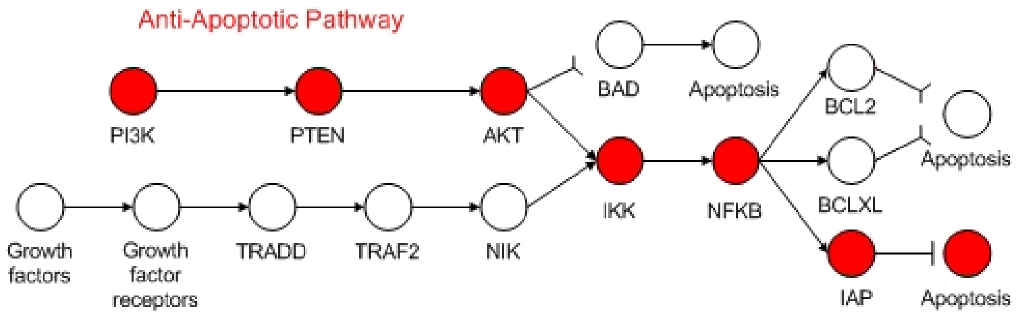
- **Prob(a gene is correlated) = $1/2^6$**
- **# of genes on array = 100,000**

⇒ E(# of correlated genes) = 1,562

⇒ Many false positives

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

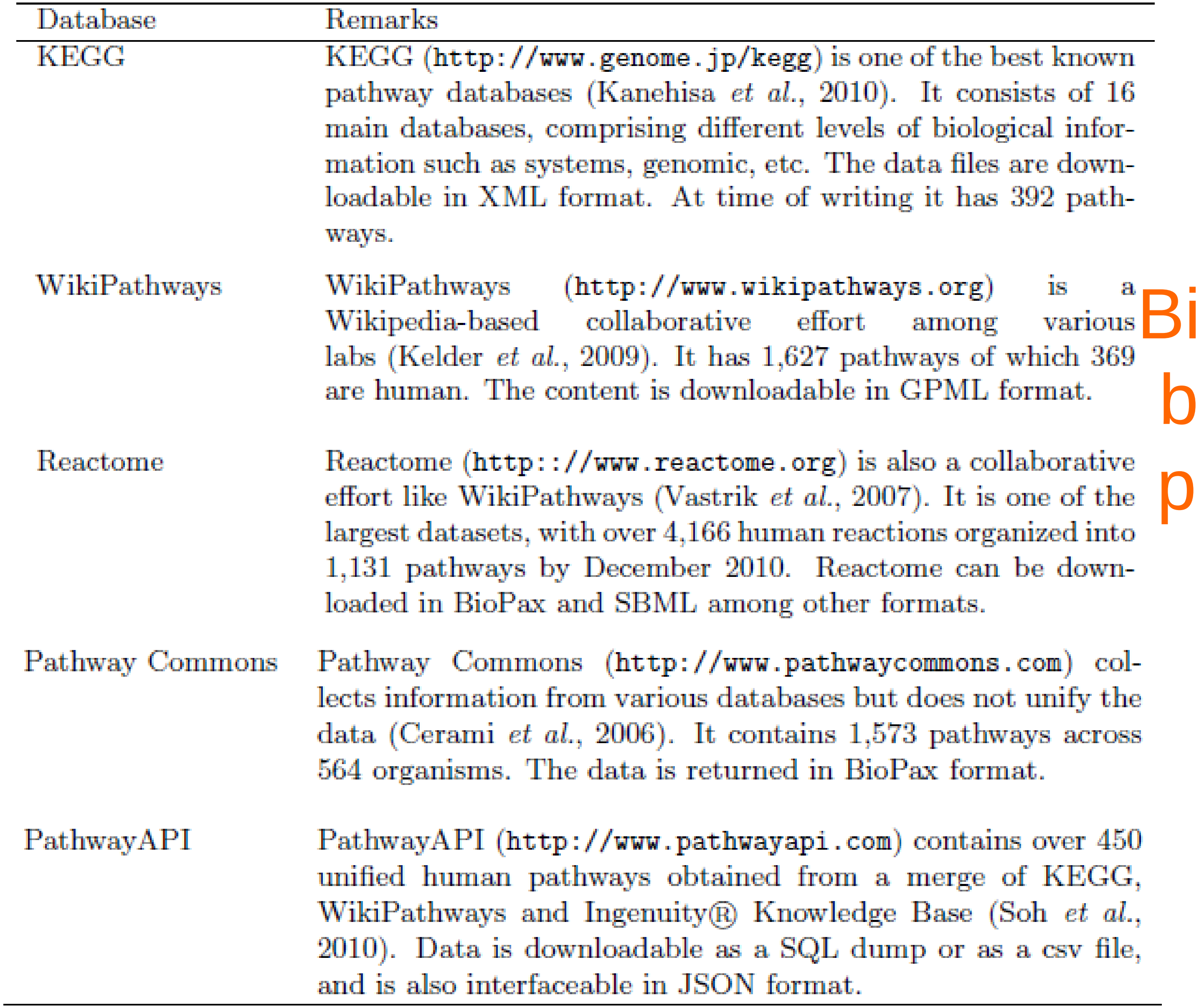
Gene Regulatory Circuits



- **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**

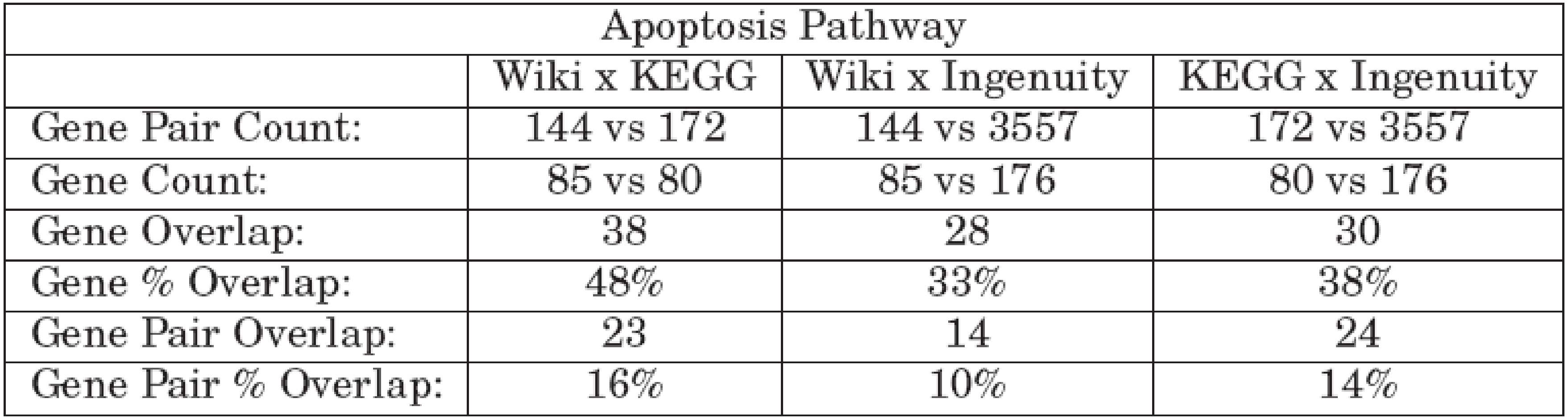
Big data of biological pathways



Database	Remarks
KEGG	KEGG (http://www.genome.jp/kegg) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.
WikiPathways	WikiPathways (http://www.wikipathways.org) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.
Reactome	Reactome (http::/www.reactome.org) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.
Pathway Commons	Pathway Commons (http://www.pathwaycommons.com) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.
PathwayAPI	PathwayAPI (http://www.pathwayapi.com) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh *et al.*, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.

Source: Goh et al. “How advancement in biological network analysis methods empowers proteomics”. *Proteomics*, accepted.

Human Apoptosis Pathway



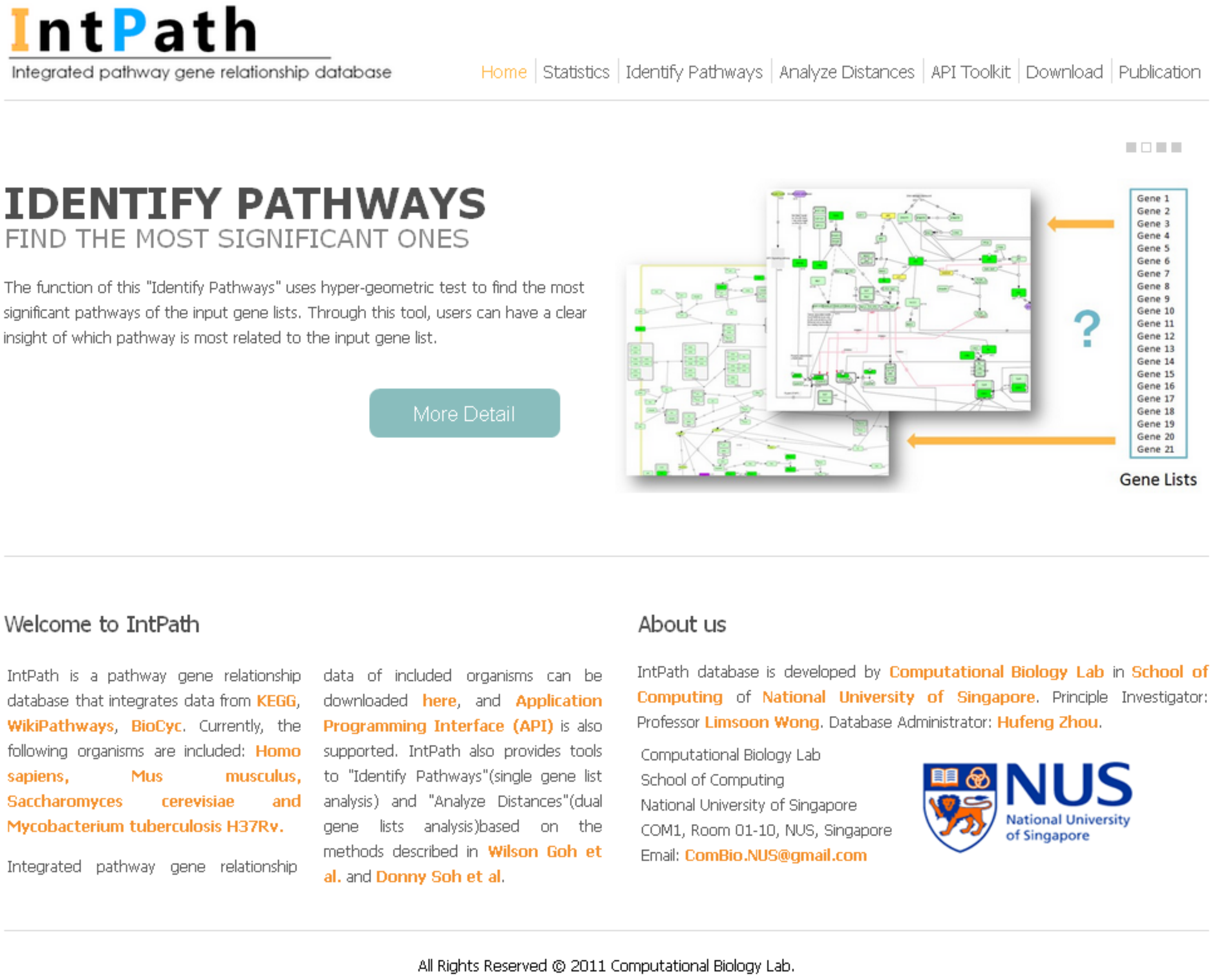
Apoptosis Pathway			
	Wiki x KEGG	Wiki x Ingenuity	KEGG x Ingenuity
Gene Pair Count:	144 vs 172	144 vs 3557	172 vs 3557
Gene Count:	85 vs 80	85 vs 176	80 vs 176
Gene Overlap:	38	28	30
Gene % Overlap:	48%	33%	38%
Gene Pair Overlap:	23	14	24
Gene Pair % Overlap:	16%	10%	14%

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

- **The various data sources have low overlap**

⇒ Good to unify them to get more complete pathways, right?

A unified database of biological pathways



IntPath

Integrated pathway gene relationship database

Home | Statistics | Identify Pathways | Analyze Distances | API Toolkit | Download | Publication

IDENTIFY PATHWAYS

FIND THE MOST SIGNIFICANT ONES

The function of this "Identify Pathways" uses hyper-geometric test to find the most significant pathways of the input gene lists. Through this tool, users can have a clear insight of which pathway is most related to the input gene list.

Welcome to IntPath

IntPath is a pathway gene relationship database that integrates data from KEGG, WikiPathways, BioCyc. Currently, the following organisms are included: Homo sapiens, Mus musculus, Saccharomyces cerevisiae and Mycobacterium tuberculosis H37Rv.

Integrated pathway gene relationship data of included organisms can be downloaded here, and Application Programming Interface (API) is also supported. IntPath also provides tools to "Identify Pathways"(single gene list analysis) and "Analyze Distances"(dual gene lists analysis)based on the methods described in Wilson Goh et al. and Donny Soh et al.

About us

IntPath database is developed by Computational Biology Lab in School of Computing of National University of Singapore. Principle Investigator: Professor Limsoon Wong. Database Administrator: Hufeng Zhou.

Computational Biology Lab
School of Computing
National University of Singapore
COM1, Room 01-10, NUS, Singapore
Email: ComBio.NUS@gmail.com

H. Zhou, et al. **IntPath---an integrated pathway gene relationship database for model organisms and important pathogens**. *BMC Systems Biology*, 6(Suppl 2):S2, 2012.

Using biology background: GSEA

- **"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

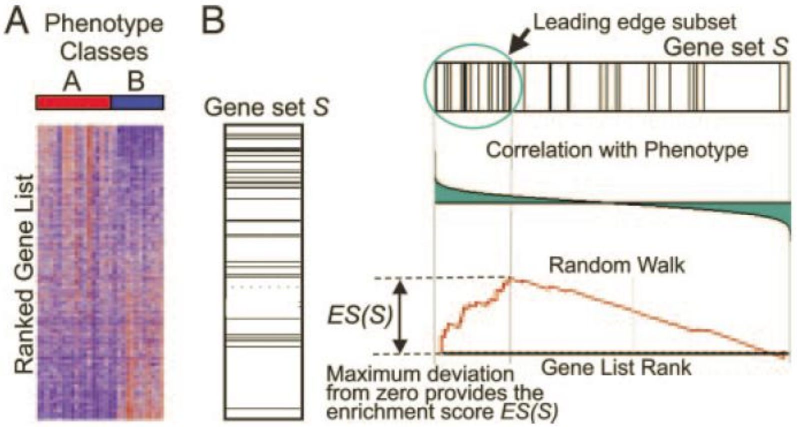


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location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

Unfortunately, it doesn't always work

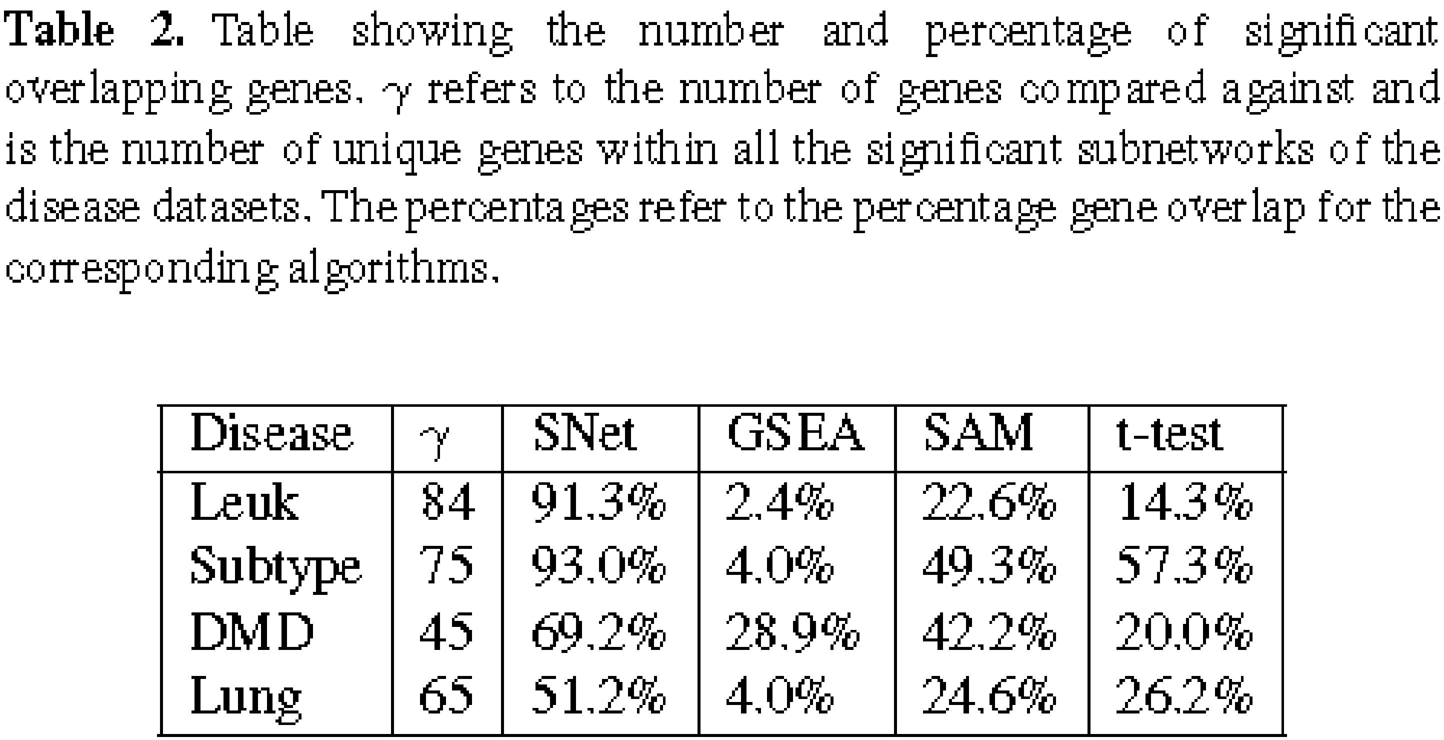


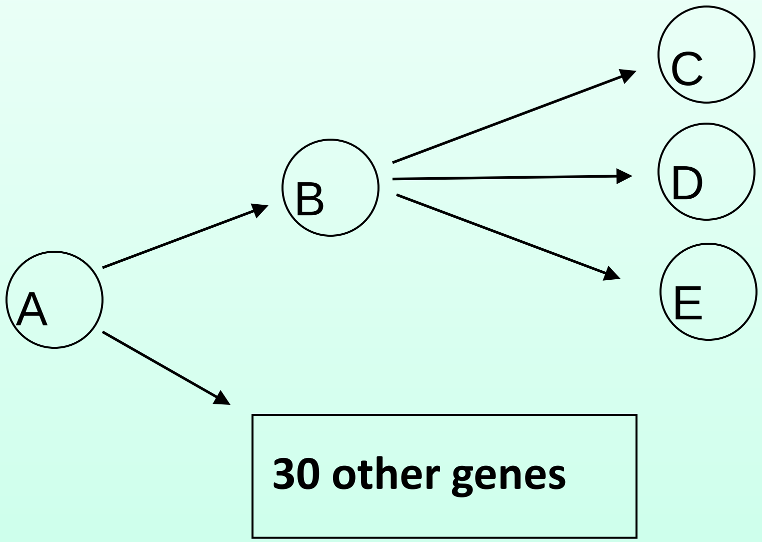
Table 2. Table showing the number and percentage of significant overlapping genes. γ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding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

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

- **Surprisingly, GSEA fails on large unified pathways!**

More is not always better, unless …



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

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

30 other genes not diff expressed

GSEA: Entire network is likely to be missed

- **Need to know how to capture the subnetwork branch within the pathway**