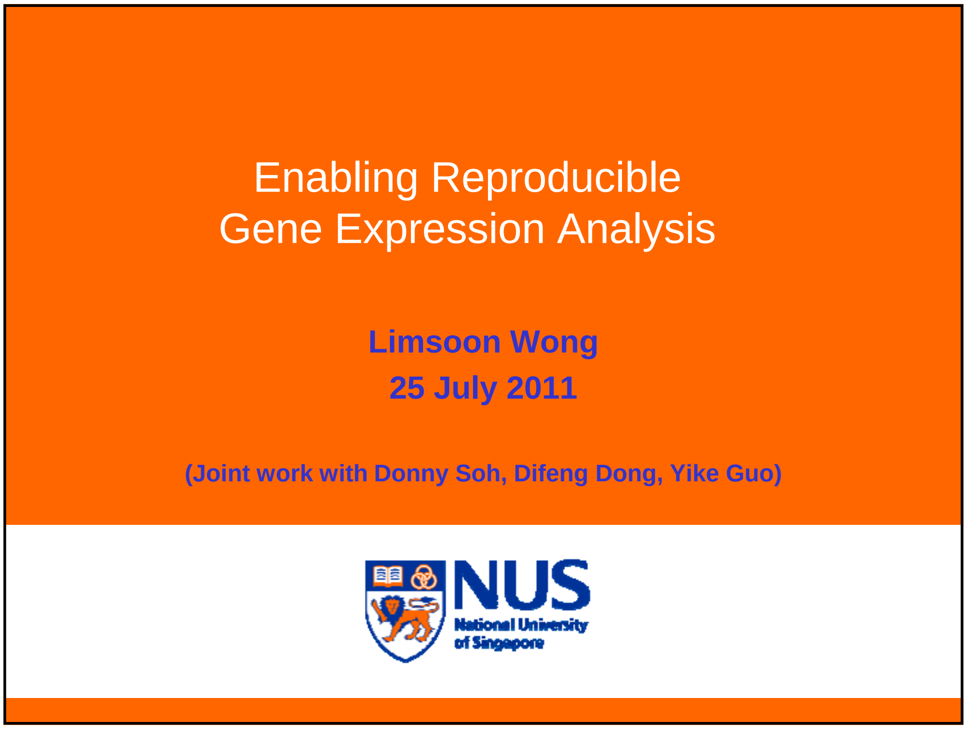
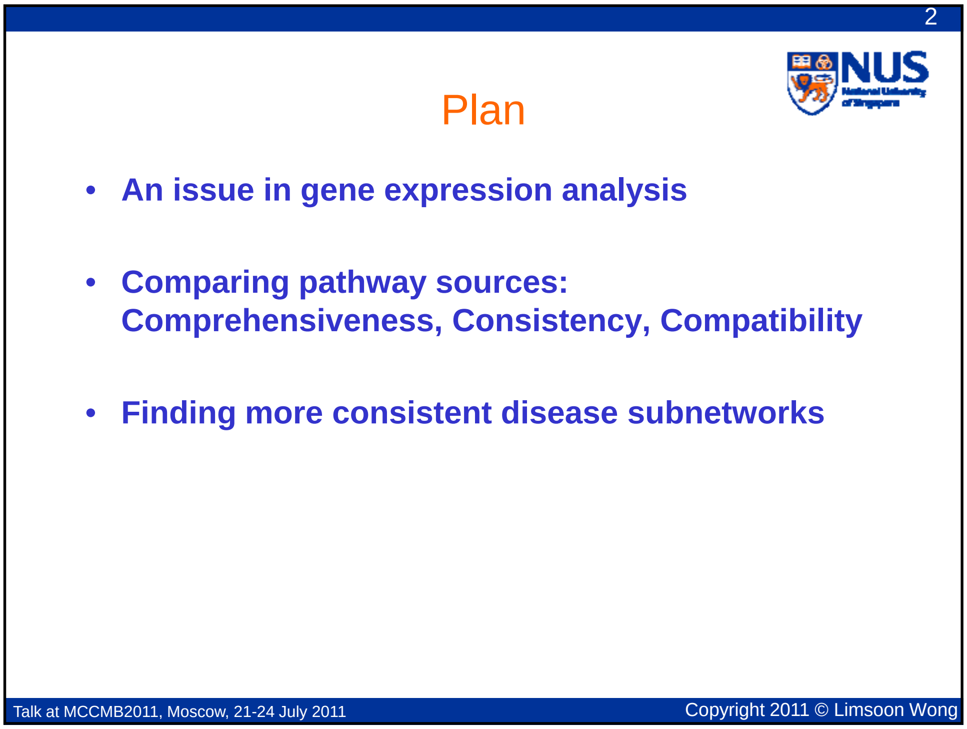


Enabling Reproducible Gene Expression Analysis

Limsoon Wong

25 July 2011

(Joint work with Donny Soh, Difeng Dong, Yike Guo)

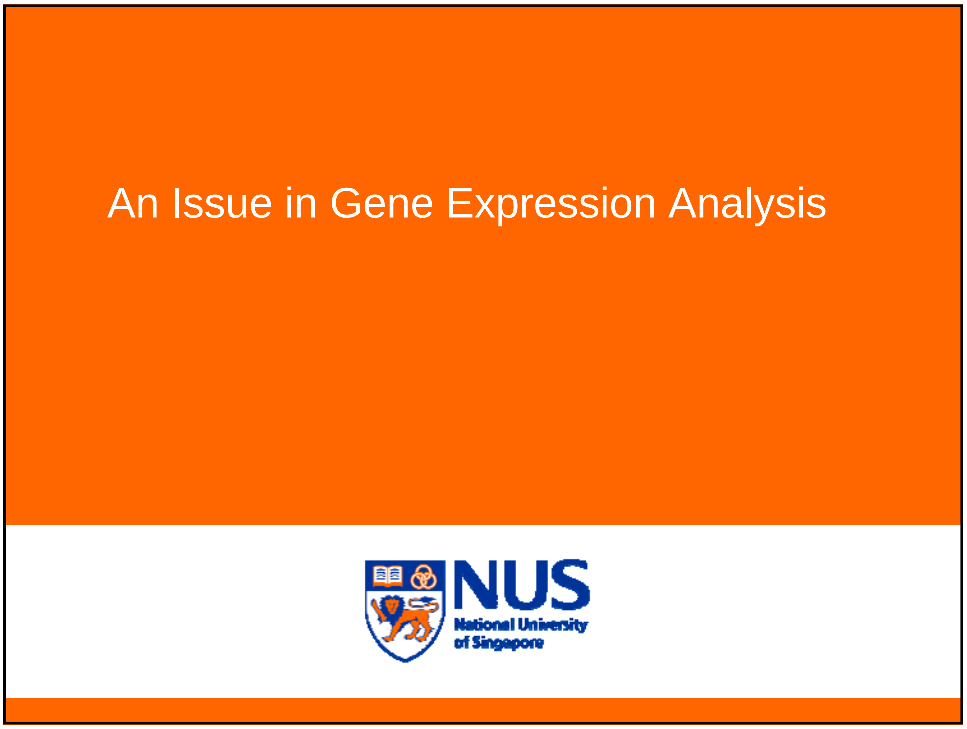


2

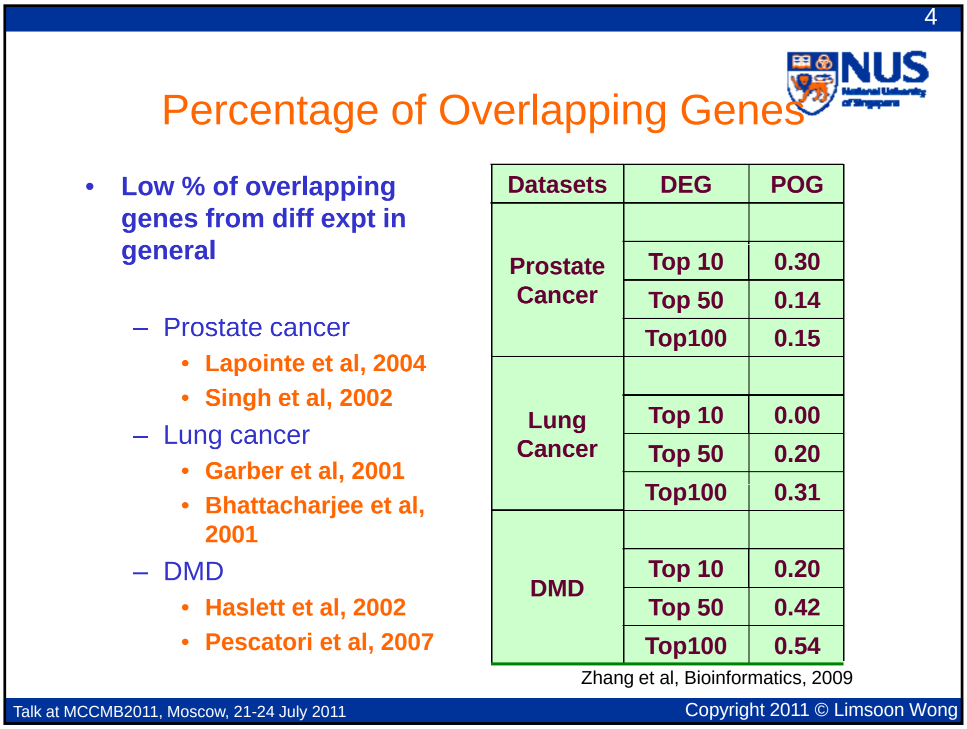
Plan

- An issue in gene expression analysis
- Comparing pathway sources: Comprehensiveness, Consistency, Compatibility
- Finding more consistent disease subnetworks

Talk at MCCMB2011, Moscow, 21-24 July 2011



An Issue in Gene Expression Analysis



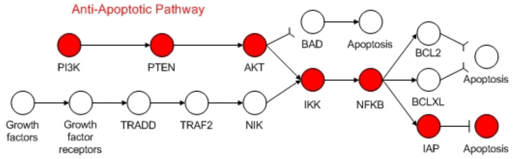
Percentage of Overlapping Genes

- **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

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

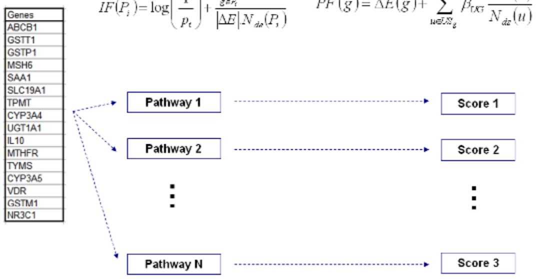
- The unifying biological theme is basis for inferring the underlying cause of disease subtype

Towards More Meaningful Genes

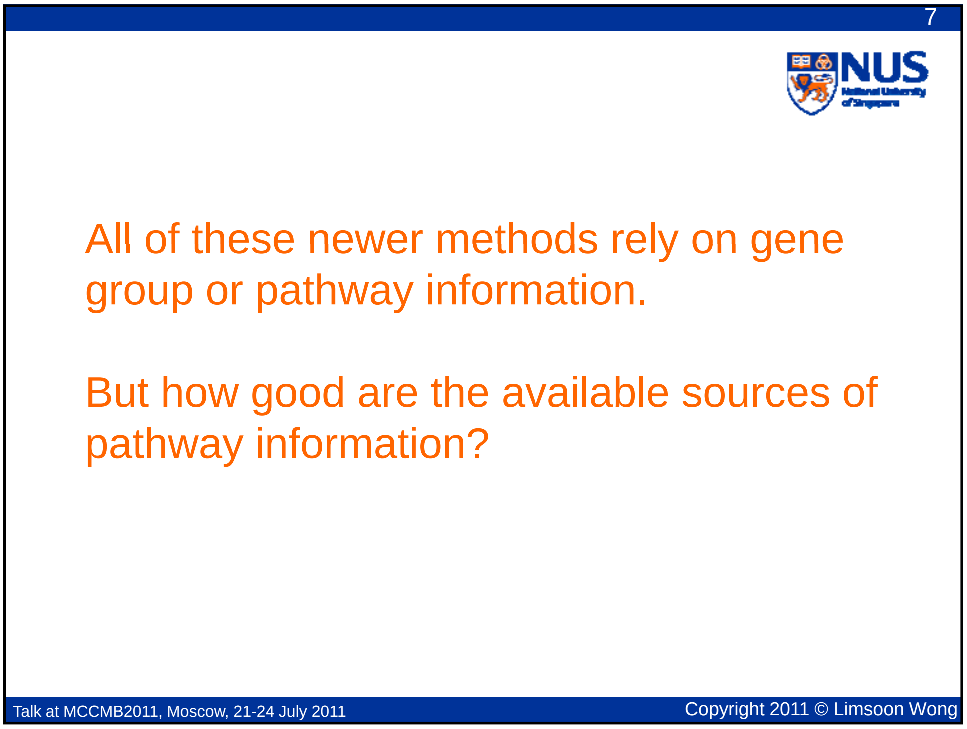
- **ORA**
  - Khatri et al
  - *Genomics*, 2002
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - *PNAS*, 2005
- **Pathway Express**
  - Draghici et al
  - *Genome Res*, 2007

Gene Class Testing: Pathway Express

$$IF(P_i) = \log\left(\frac{1}{p_i}\right) + \frac{\sum_{g \in P_i} |PF(g)|}{|\Delta E| N_{de}(P_i)} \qquad PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \frac{PF(u)}{N_{ds}(u)}$$

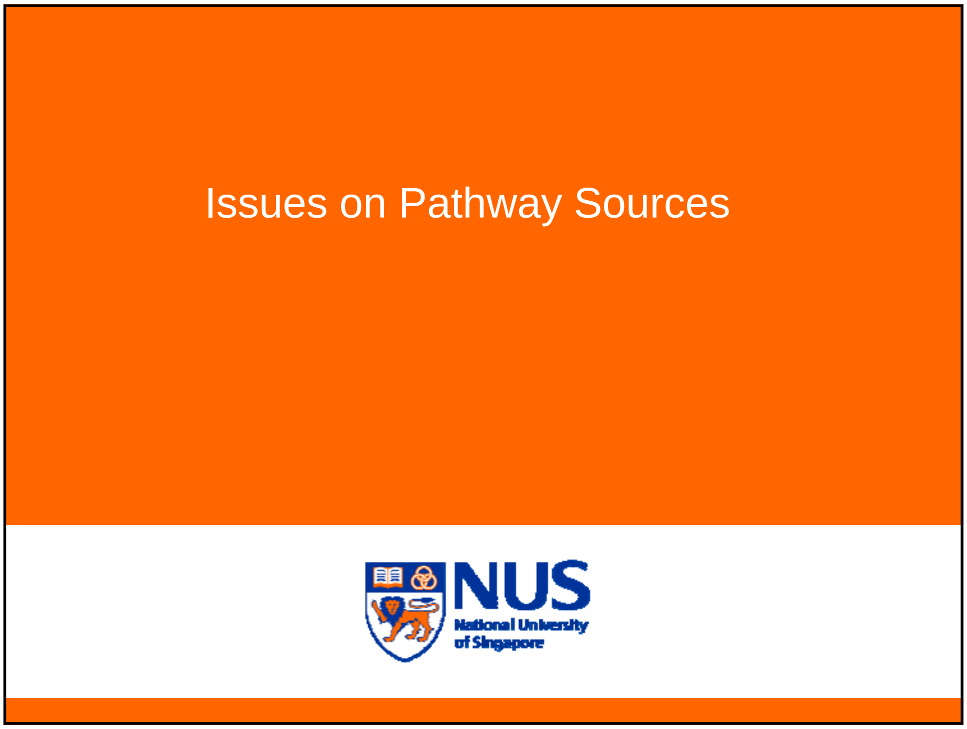


Draghici et al, Genome Res, 2007

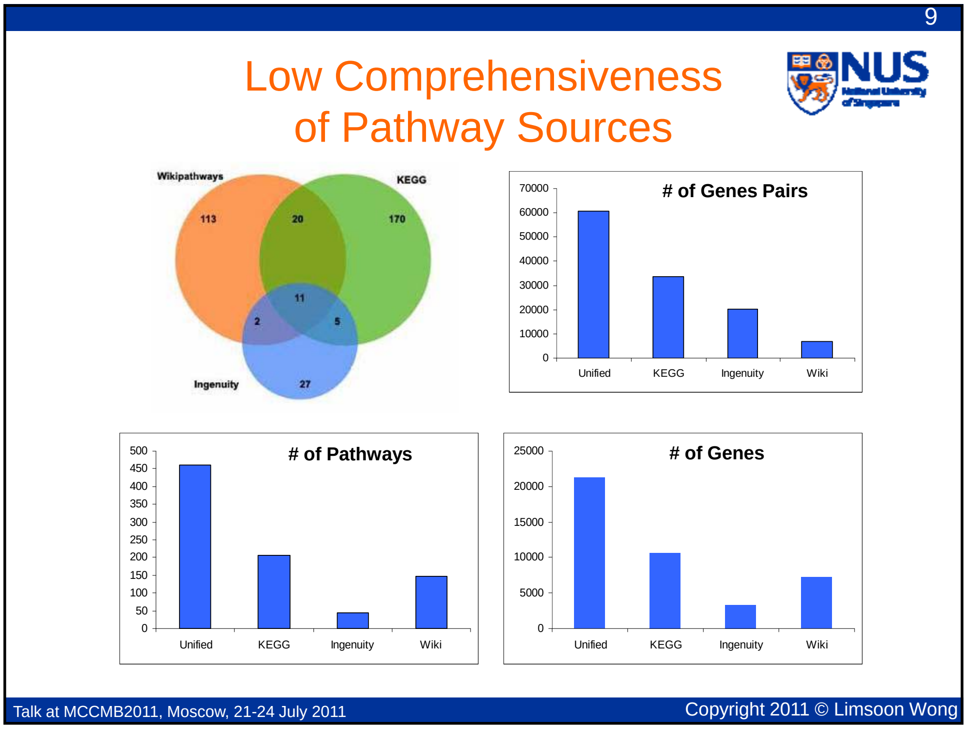


All of these newer methods rely on gene group or pathway information.

But how good are the available sources of pathway information?



Issues on Pathway Sources

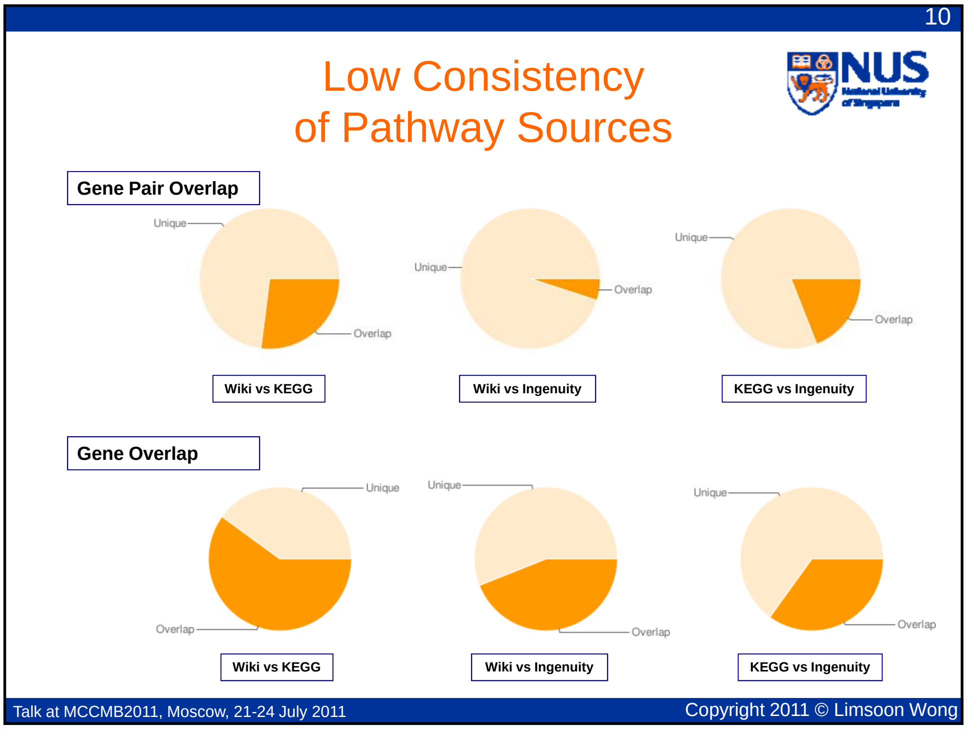


9

Low Comprehensiveness of Pathway Sources

Talk at MCCMB2011, Moscow, 21-24 July 2011

Copyright 2011 © Limsoon Wong



10

Low Consistency of Pathway Sources

Gene Pair Overlap

Wiki vs KEGG

Wiki vs Ingenuity

KEGG vs Ingenuity

Gene Overlap

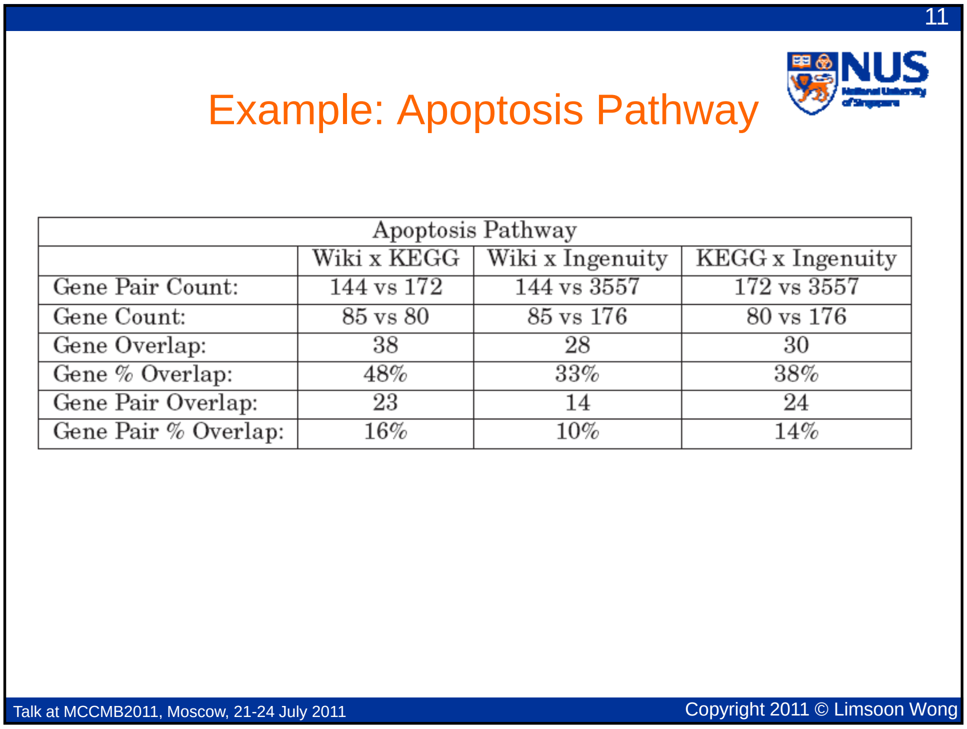
Wiki vs KEGG

Wiki vs Ingenuity

KEGG vs Ingenuity

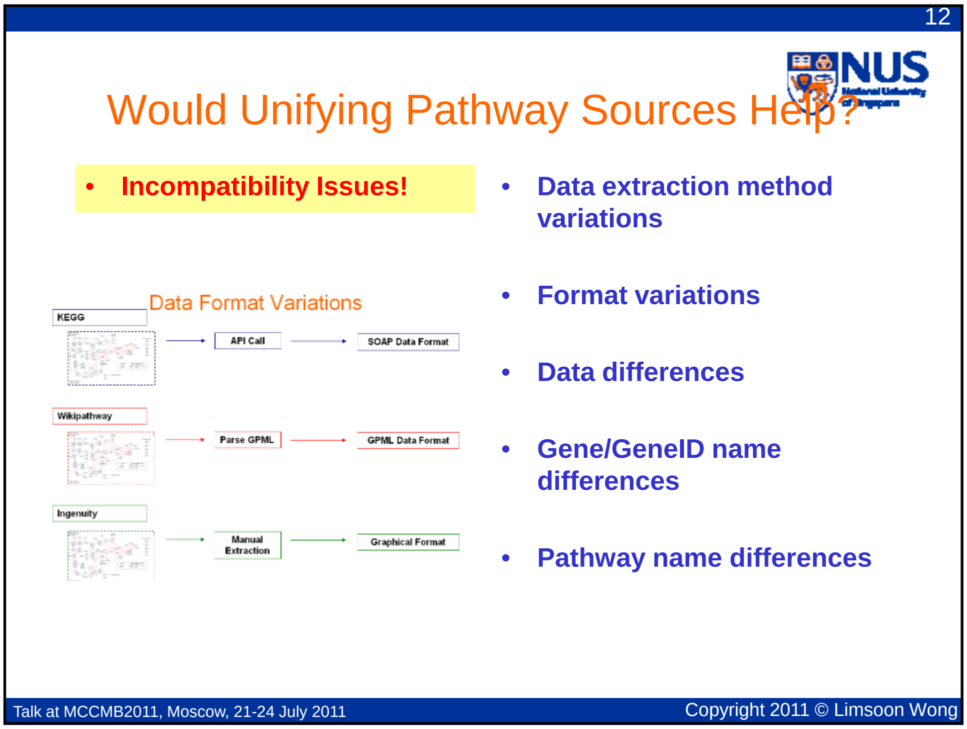
Talk at MCCMB2011, Moscow, 21-24 July 2011

Copyright 2011 © Limsoon Wong



Example: 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%



Would Unifying Pathway Sources Help?

- **Incompatibility Issues!**

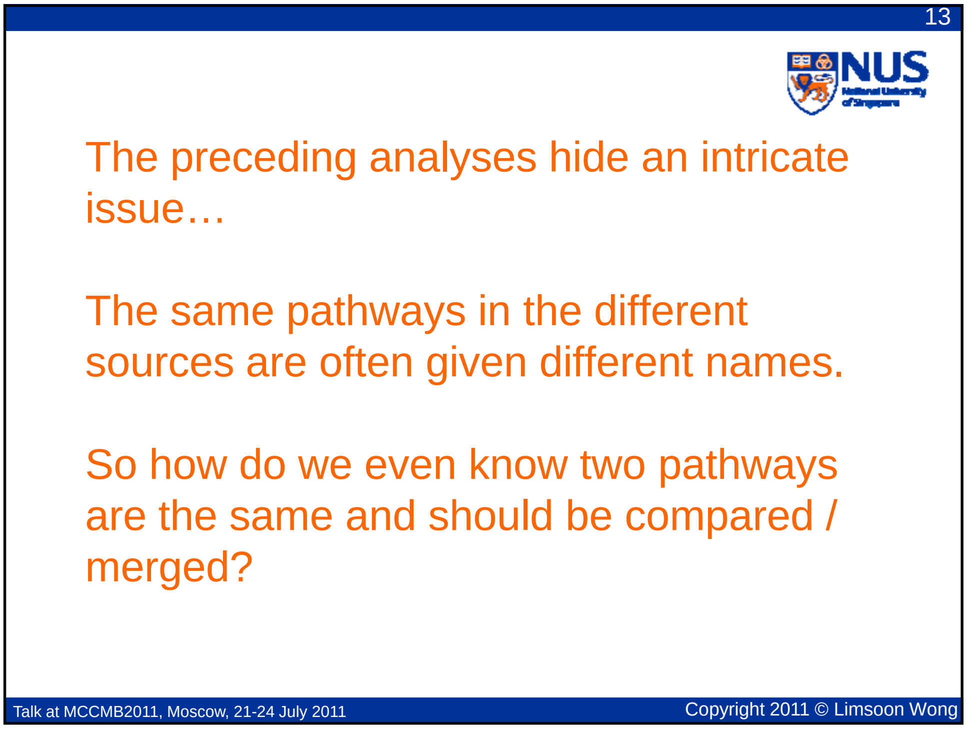
Data Format Variations

KEGG → API Call → SOAP Data Format

Wikipathway → Parse GPML → GPML Data Format

Ingenuity → Manual Extraction → Graphical Format

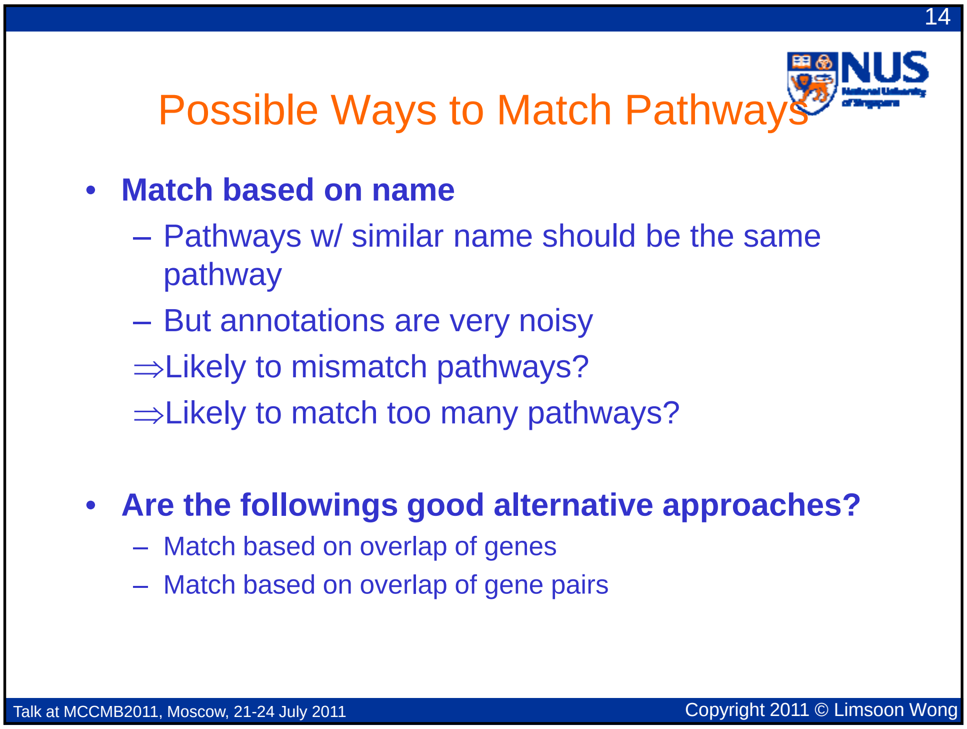
- **Data extraction method variations**
- **Format variations**
- **Data differences**
- **Gene/GeneID name differences**
- **Pathway name differences**



The preceding analyses hide an intricate issue…

The same pathways in the different sources are often given different names.

So how do we even know two pathways are the same and should be compared / merged?



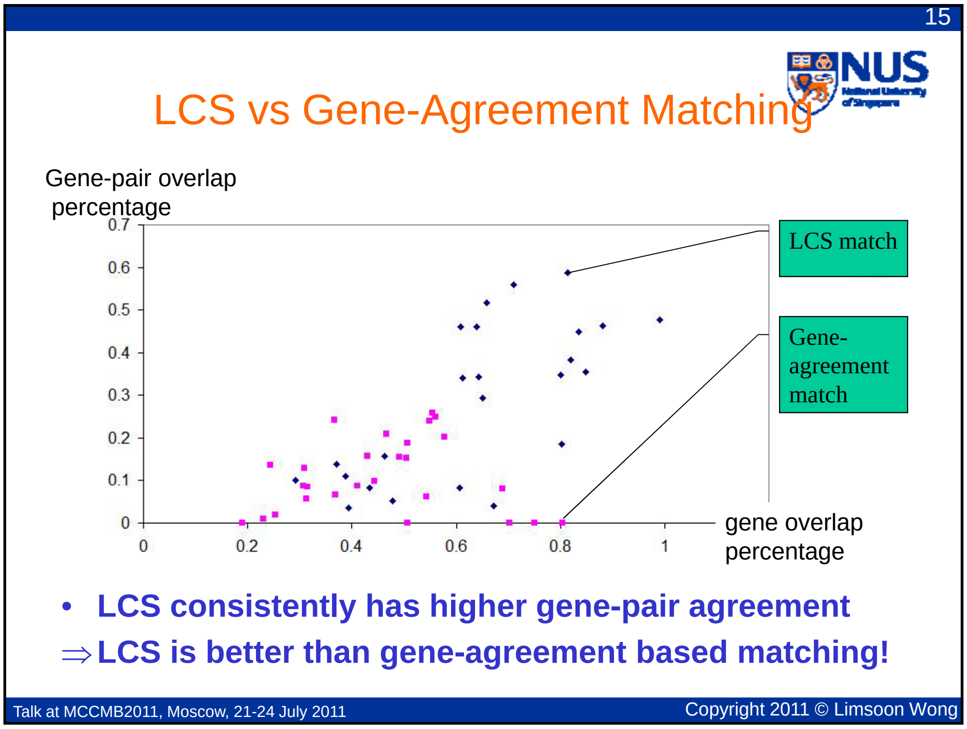
Possible Ways to Match Pathways

- **Match based on name**
  - Pathways w/ similar name should be the same pathway
  - But annotations are very noisy

  ⇒Likely to mismatch pathways?

  ⇒Likely to match too many pathways?

- **Are the followings good alternative approaches?**
  - Match based on overlap of genes
  - Match based on overlap of gene pairs



15

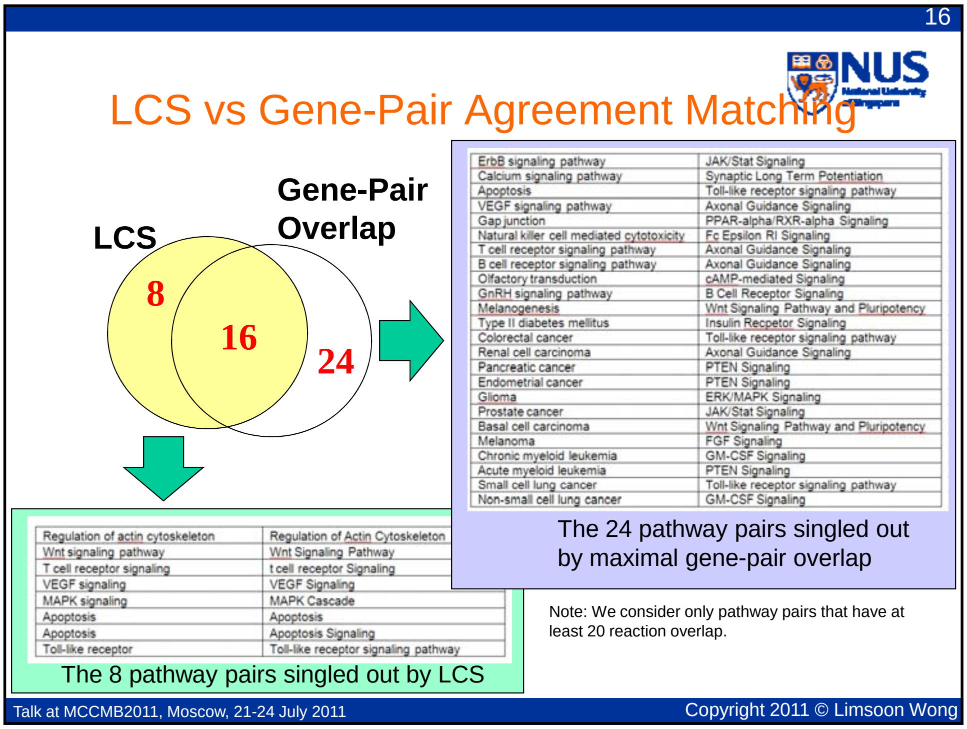
LCS vs Gene-Agreement Matching

- **LCS consistently has higher gene-pair agreement**

⇒ **LCS is better than gene-agreement based matching!**

Talk at MCCMB2011, Moscow, 21-24 July 2011

Copyright 2011 © Limsoon Wong



16

LCS vs Gene-Pair Agreement Matching

ErbB signaling pathway	JAK/Stat Signaling
Calcium signaling pathway	Synaptic Long Term Potentiation
Apoptosis	Toll-like receptor signaling pathway
VEGF signaling pathway	Axonal Guidance Signaling
Gap junction	PPAR-alpha/RXR-alpha Signaling
Natural killer cell mediated cytotoxicity	Fc Epsilon RI Signaling
T cell receptor signaling pathway	Axonal Guidance Signaling
B cell receptor signaling pathway	Axonal Guidance Signaling
Olfactory transduction	cAMP-mediated Signaling
GnRH signaling pathway	B Cell Receptor Signaling
Melanogenesis	Wnt Signaling Pathway and Pluripotency
Type II diabetes mellitus	Insulin Recpetor Signaling
Colorectal cancer	Toll-like receptor signaling pathway
Renal cell carcinoma	Axonal Guidance Signaling
Pancreatic cancer	PTEN Signaling
Endometrial cancer	PTEN Signaling
Glioma	ERK/MAPK Signaling
Prostate cancer	JAK/Stat Signaling
Basal cell carcinoma	Wnt Signaling Pathway and Pluripotency
Melanoma	FGF Signaling
Chronic myeloid leukemia	GM-CSF Signaling
Acute myeloid leukemia	PTEN Signaling
Small cell lung cancer	Toll-like receptor signaling pathway
Non-small cell lung cancer	GM-CSF Signaling

The 24 pathway pairs singled out by maximal gene-pair overlap

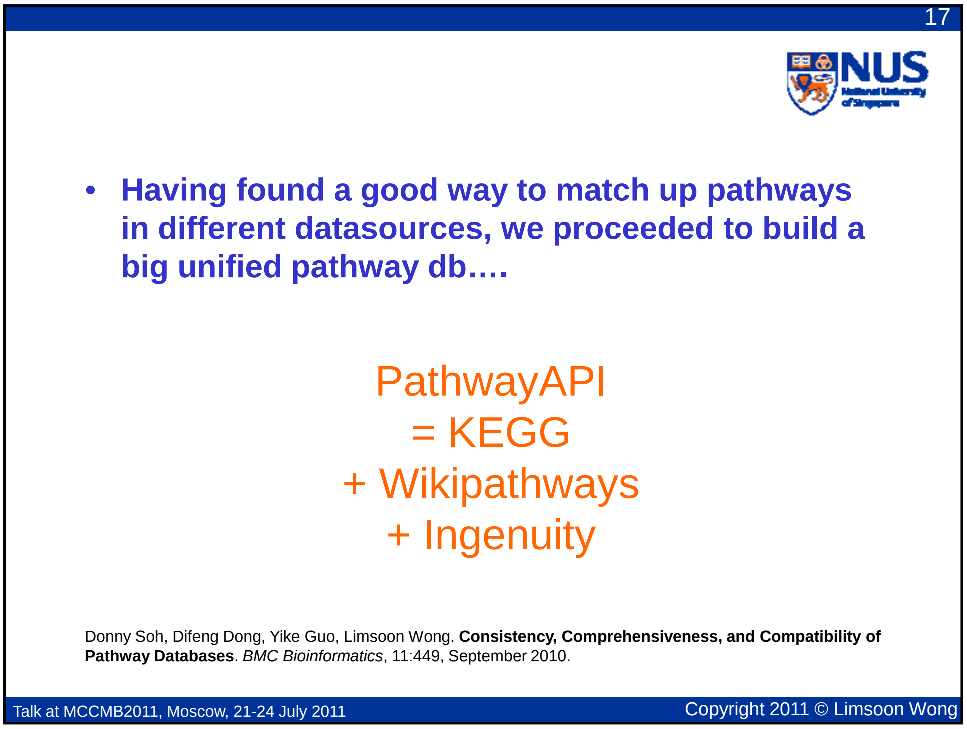
Regulation of actin cytoskeleton	Regulation of Actin Cytoskeleton
Wnt signaling pathway	Wnt Signaling Pathway
T cell receptor signaling	t cell receptor Signaling
VEGF signaling	VEGF Signaling
MAPK signaling	MAPK Cascade
Apoptosis	Apoptosis
Apoptosis	Apoptosis Signaling
Toll-like receptor	Toll-like receptor signaling pathway

The 8 pathway pairs singled out by LCS

Note: We consider only pathway pairs that have at least 20 reaction overlap.

Talk at MCCMB2011, Moscow, 21-24 July 2011

Copyright 2011 © Limsoon Wong



17

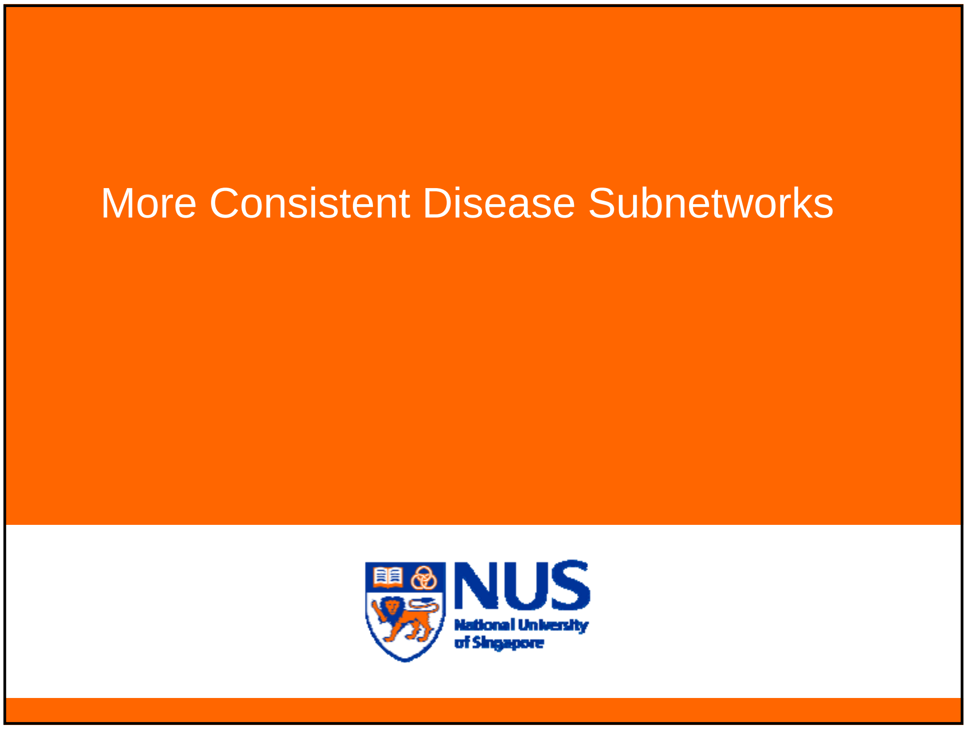
- **Having found a good way to match up pathways in different datasources, we proceeded to build a big unified pathway db….**

PathwayAPI
= KEGG
+ Wikipathways
+ Ingenuity

Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases**. *BMC Bioinformatics*, 11:449, September 2010.

Talk at MCCMB2011, Moscow, 21-24 July 2011

Copyright 2011 © Limsoon Wong

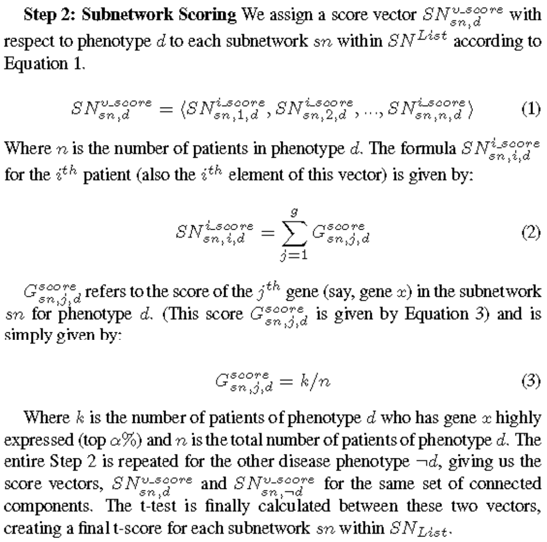


More Consistent Disease Subnetworks

The SNet Method

- **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 $SN^{v\_score}_{sn,d}$ with respect to phenotype d to each subnetwork sn within SN^{List} according to Equation 1.

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

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:

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

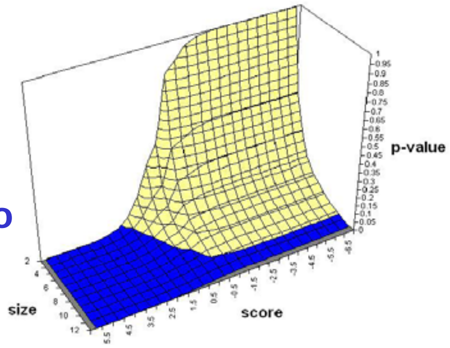
$G^{score}_{sn,j,d}$ refers to the score of the j^{th} gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G^{score}_{sn,j,d}$ is given by Equation 3) and is simply given by:

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

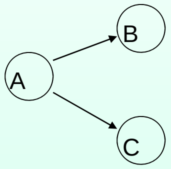
Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d. The entire Step 2 is repeated for the other disease phenotype $\neg d$, giving us the score vectors, $SN^{v\_score}_{sn,d}$ and $SN^{v\_score}_{sn,\neg d}$ 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

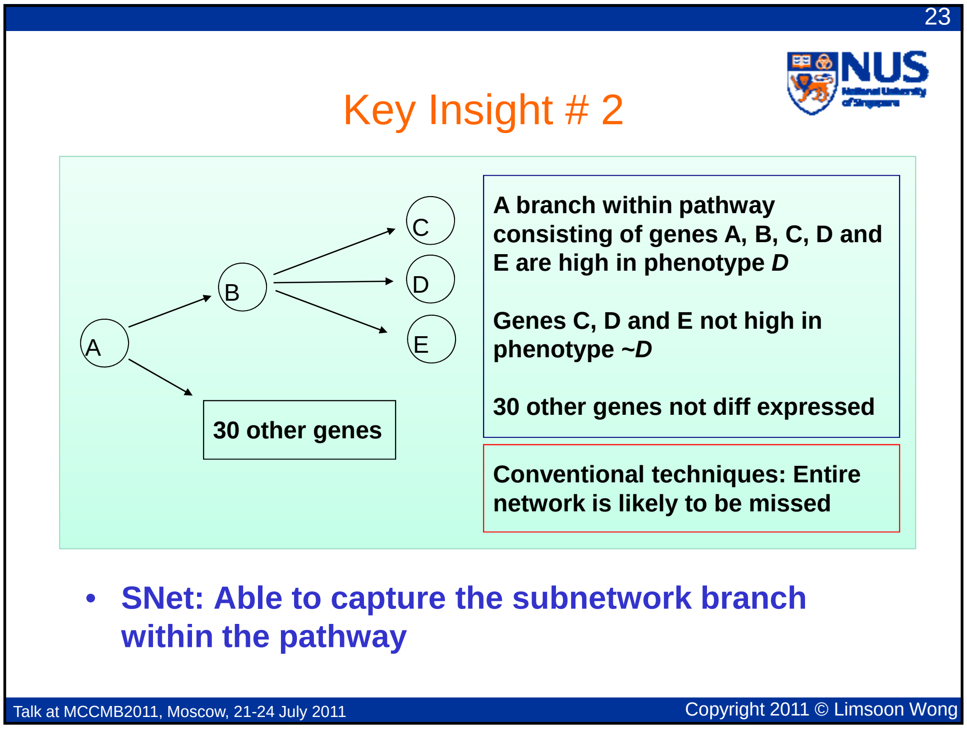


Genes A, B, C are high in phenotype *D*

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



23

Key Insight # 2

30 other genes

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*

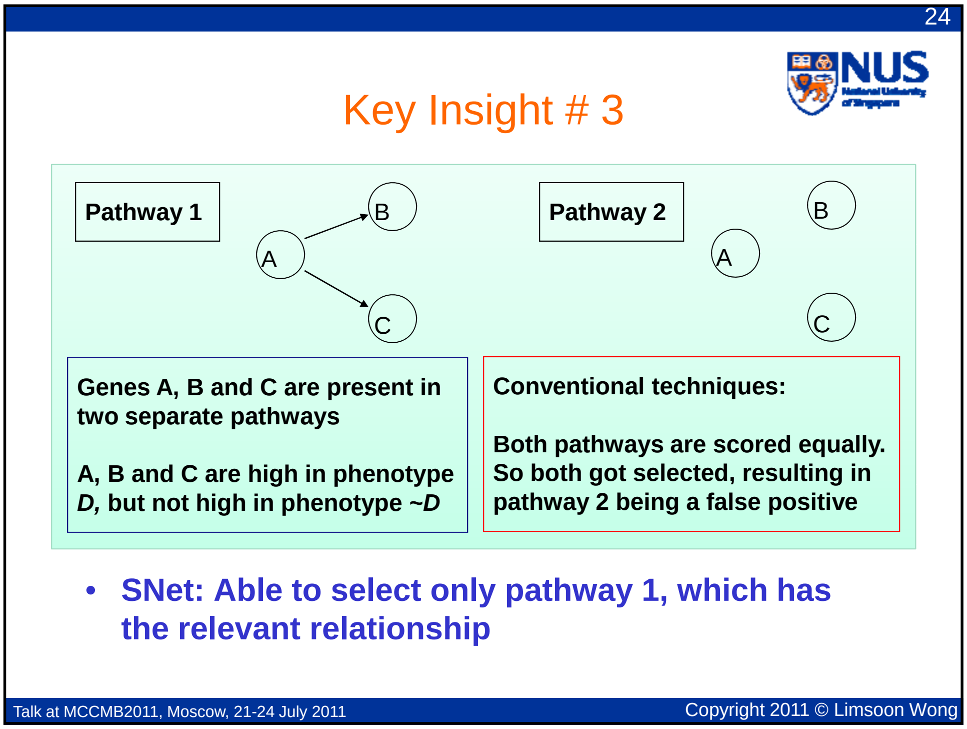
30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

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

Talk at MCCMB2011, Moscow, 21-24 July 2011

Copyright 2011 © Limsoon Wong



24

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

Talk at MCCMB2011, Moscow, 21-24 July 2011

Copyright 2011 © Limsoon Wong

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

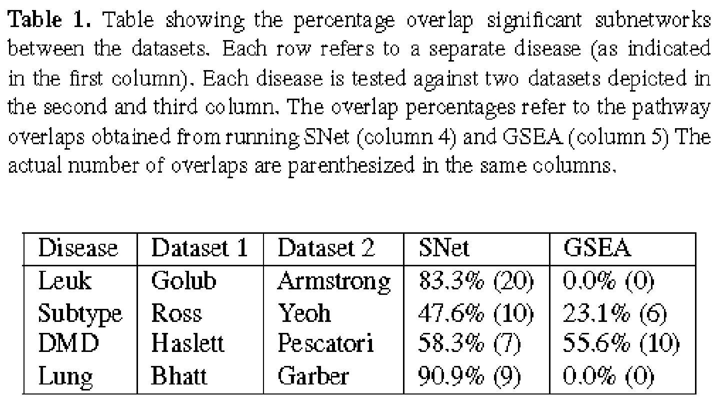
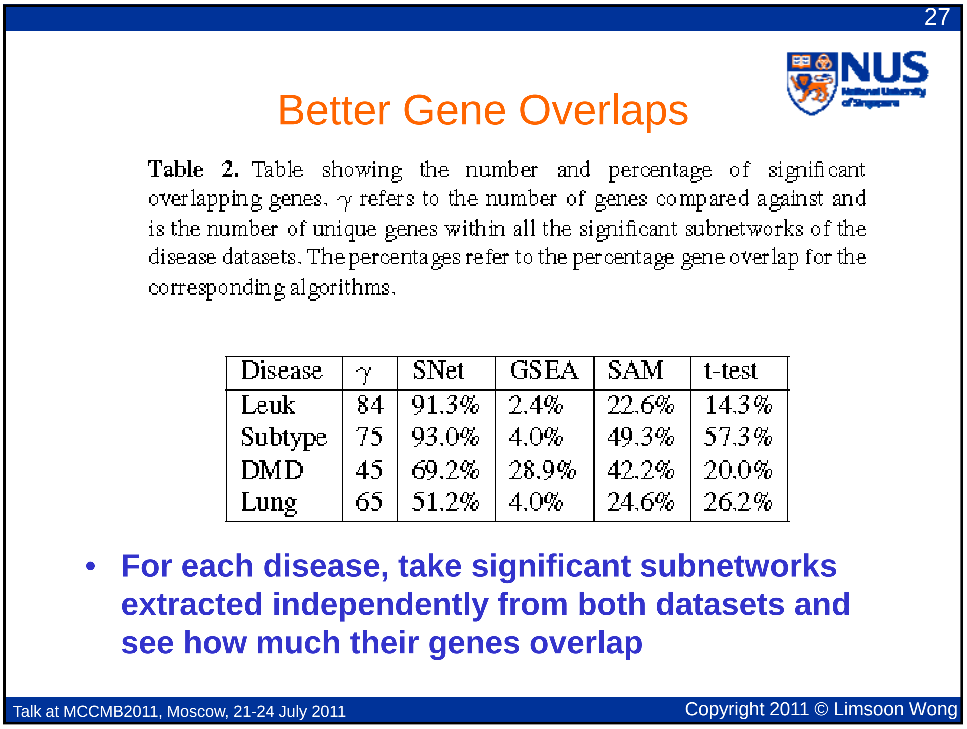


Table 1. Table showing the percentage overlap significant subnetworks between the datasets. Each row refers to a separate disease (as indicated in the first column). Each disease is tested against two datasets depicted in the second and third column. The overlap percentages refer to the pathway overlaps 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**



27

Better Gene Overlaps

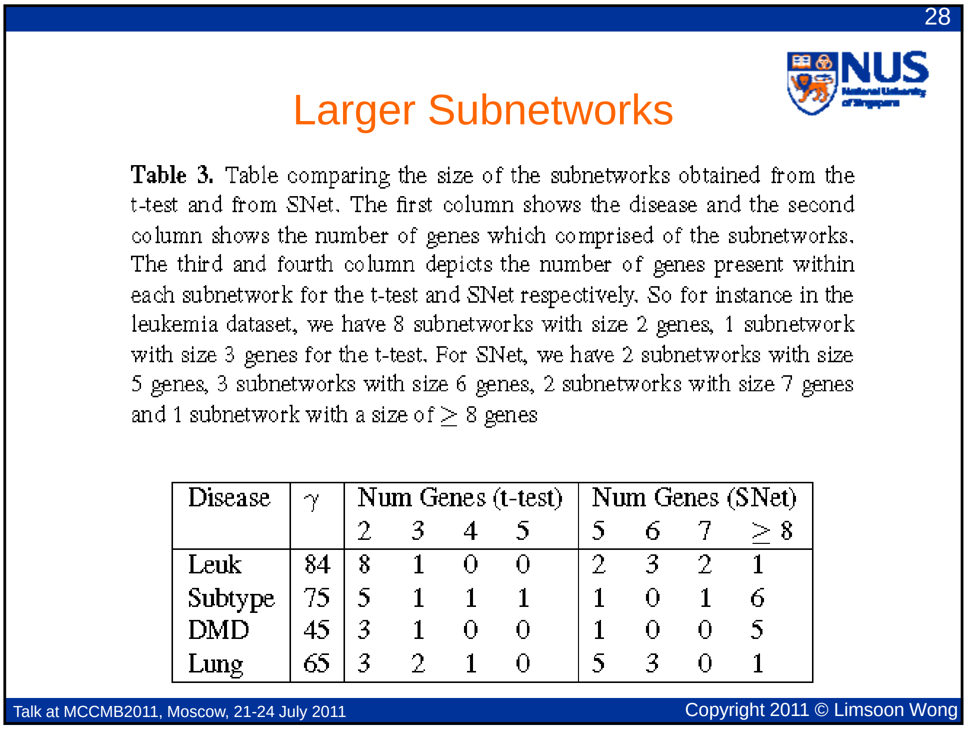
Table 2. Table showing the number and percentage of significant overlapping genes. γ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding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

Talk at MCCMB2011, Moscow, 21-24 July 2011

Copyright 2011 © Limsoon Wong



28

Larger Subnetworks

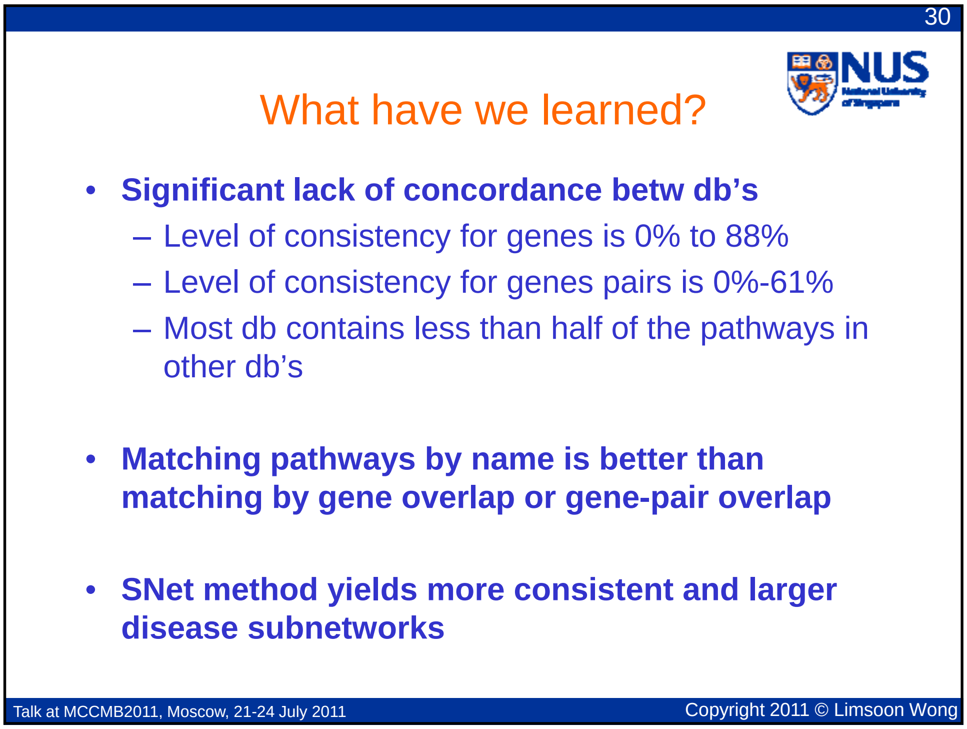
Table 3. Table comparing the size of the subnetworks obtained from the t-test and from SNet. The first column shows the disease and the second column shows the number of genes which comprised of the subnetworks. The third and fourth column depicts the number of genes present within each subnetwork for 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	≥ 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

Talk at MCCMB2011, Moscow, 21-24 July 2011

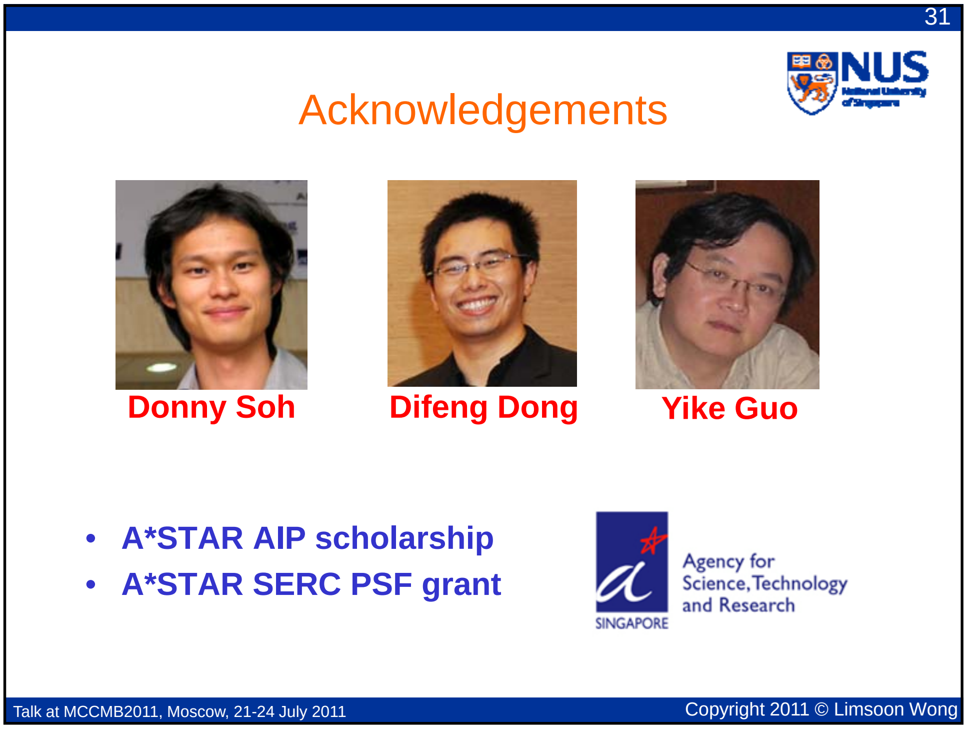
Copyright 2011 © Limsoon Wong

Remarks



What have we learned?

- **Significant lack of concordance betw db's**
  - Level of consistency for genes is 0% to 88%
  - Level of consistency for genes pairs is 0%-61%
  - Most db contains less than half of the pathways in other db's
- **Matching pathways by name is better than matching by gene overlap or gene-pair overlap**
- **SNet method yields more consistent and larger disease subnetworks**

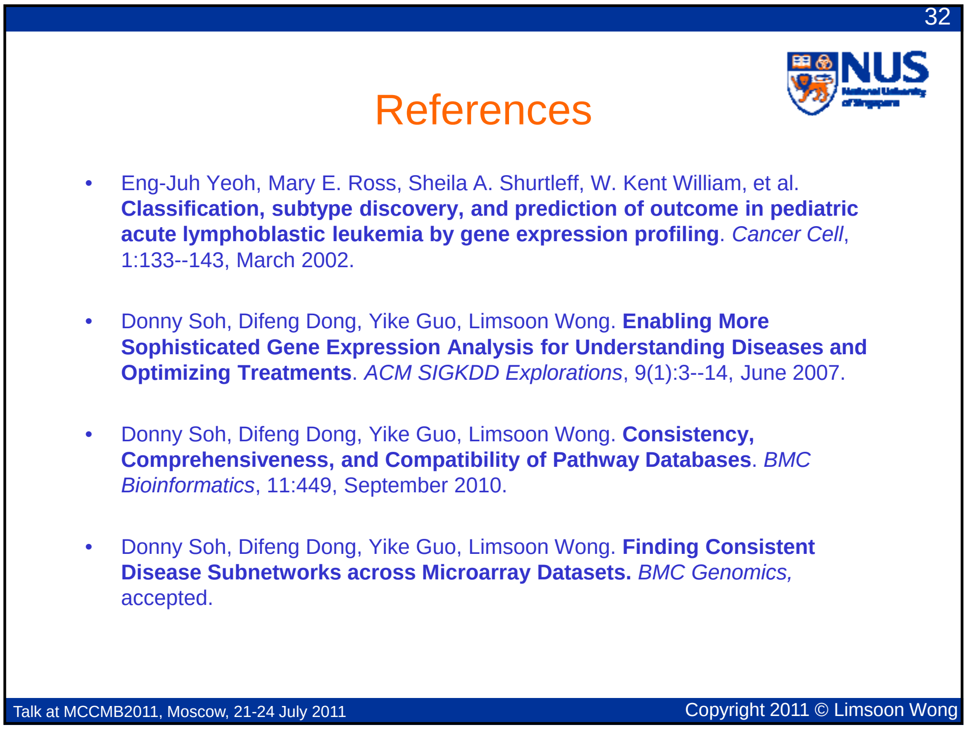


Acknowledgements

Donny Soh   Difeng Dong   Yike Guo

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

Agency for Science, Technology and Research
SINGAPORE



References

- Eng-Juh Yeoh, Mary E. Ross, Sheila A. Shurtleff, W. Kent William, et al. **Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling**. *Cancer Cell*, 1:133--143, March 2002.

- Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Enabling More Sophisticated Gene Expression Analysis for Understanding Diseases and Optimizing Treatments**. *ACM SIGKDD Explorations*, 9(1):3--14, June 2007.

- Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases**. *BMC Bioinformatics*, 11:449, September 2010.

- Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Finding Consistent Disease Subnetworks across Microarray Datasets.** *BMC Genomics,* accepted.