Consistency & Comprehensiveness of Pathway Databases

Limsoon Wong

23 February 2010

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

Plan

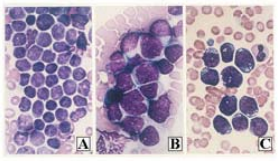
- **Past successes in gene expression analysis**
- **Towards more meaningful genes**
- **Issues on pathway sources**
  - Comprehensiveness
  - Consistency
  - Compatibility
  - Matching pathways in different sources

Past Success: An Example

NUS National University of Singapore

Childhood Acute Lymphoblastic Leukemia

- **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**

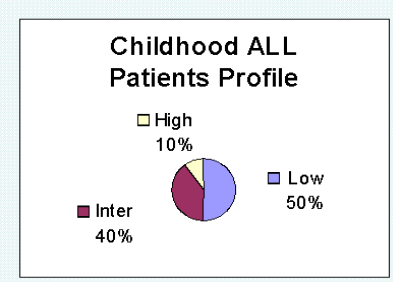


- **Conventional diagnosis**
  - Immunophenotyping
  - Cytogenetics
  - Molecular diagnostics

  ⇒Unavailable in developing countries

5

Patient Profiles & Treatment Costs



- **Treatment for childhood ALL over 2 yrs**
  - Intermediate intensity: US$60k
  - Low intensity: US$36k
  - High intensity: US$72k
- **Treatment for relapse: US$150k**
- **2000 new cases a year in ASEAN countries**
- **Cost for side-effects: Unquantified**

6

Why not high/low intensity to everyone?

- **High-intensity Tx**
  - Over intensive for 90% of patients, thus a lot more side effects
  - US$144m (US$72k * 2000) for high-intensity tx

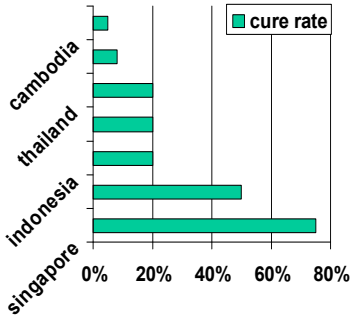
⇒ **Total US$144m/yr plus un-quantified costs for dealing with side effects**

- **Low-intensity Tx**
  - Under intensive for 50% of patients, thus a lot more relapse
  - US$72m (US$36k * 2000) for low-intensity tx
  - US$150m (US$150k * 2000 * 50%) for relapse tx

⇒ **Total US$222m/yr**

Current Situation

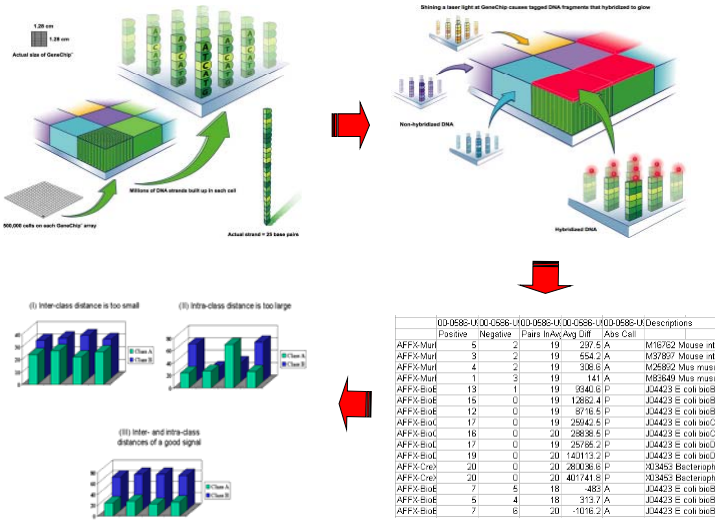
- **Intermediate intensity conventionally applied in ASEAN countries**



- **Over intensive for 50% of patients, thus more side effects**
- **Under intensive for 10% of patients, thus more relapse**

- **US$120m (US$60k * 2000) for intermediate intensity tx**
- **US$30m (US$150k * 2000 * 10%) for relapse tx**
- **Total US$150m/yr plus un-quantified costs for dealing with side effects**

Single-Test Platform of Microarray & Machine Learning



	00-0588-U00-0588-U00-0588-U00-0588-U00-0588-U Descriptions					
	Positive	Negative	Pairs In Av	Avg Diff	Abs Call	
AFFX-MurI	5	2	19	297.5	A	M16762 Mouse int
AFFX-MurI	3	2	19	554.2	A	M37897 Mouse int
AFFX-MurI	4	2	19	308.6	A	M26892 Mus musc
AFFX-MurI	1	3	19	141	A	M65849 Mus musc
AFFX-BioB	13	1	19	9340.6	P	J04423 E coli bioB
AFFX-BioB	15	0	19	12882.4	P	J04423 E coli bioB
AFFX-BioB	12	0	19	8716.5	P	J04423 E coli bioB
AFFX-BioC	17	0	19	29042.5	P	J04423 E coli bioC
AFFX-BioC	16	0	20	29836.5	P	J04423 E coli bioC
AFFX-BioD	17	0	19	25785.2	P	J04423 E coli bioD
AFFX-BioD	19	0	20	14011.3.2	P	J04423 E coli bioD
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacteriaph
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacteriaph
AFFX-BioB	7	5	18	485.A		J04423 E coli bioB
AFFX-BioB	5	4	18	313.7	A	J04423 E coli bioB
AFFX-BioB	7	6	20	-1016.2	A	J04423 E coli bioB

Individual Gene Testing

Fold Change

$$FC_{ratioi} = \frac{x_i}{y_i} \qquad FC_{diff_i} = x_i - y_i$$

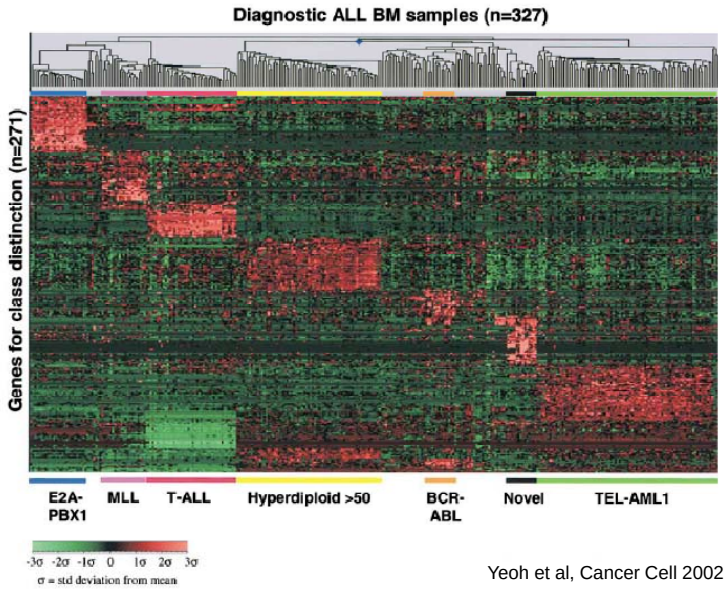
x – Microarray value after drug
y – Microarray value before drug
i – Gene

T-test

$$T_i = \frac{\hat{x}_i - \hat{y}_i}{s_i} \qquad T_i = \frac{\hat{x}_i - \hat{y}_i}{s_i + s_o} \qquad T_i = \frac{\hat{x}_i - \hat{y}_i}{\sqrt{Bs^2 + (1-B)s_i^2}}$$

x – Log2 value of treatment
y – Log2 value of control
s – Standard error
i – Gene

Golub et al, Science, 1999



Yeoh et al, Cancer Cell 2002

Exploit Invariant Gene Expr Profiles

- Low intensity applied to 50% of patients
- Intermediate intensity to 40% of patients
- High intensity to 10% of patients

⇒ Reduced side effects
⇒ Reduced relapse
⇒ 75-80% cure rates

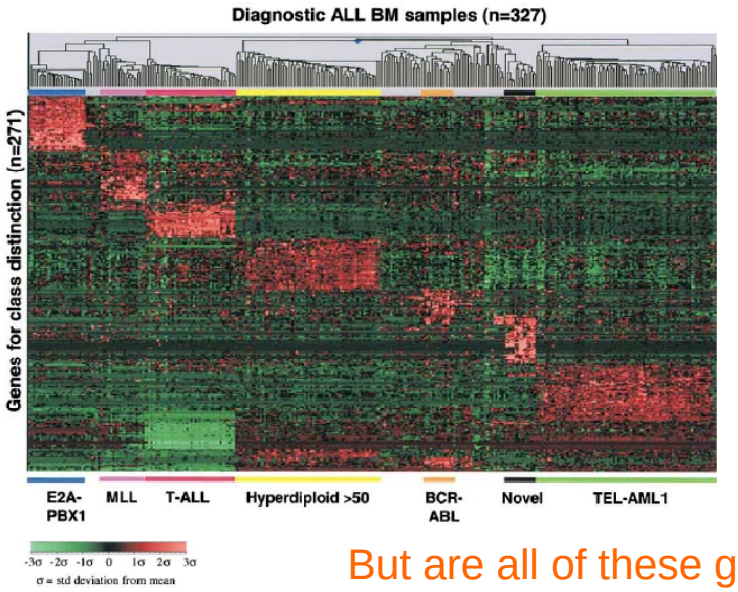
- US$36m (US$36k * 2000 * 50%) for low intensity
- US$48m (US$60k * 2000 * 40%) for intermediate intensity
- US$14.4m (US$72k * 2000 * 10%) for high intensity

- Total US$98.4m/yr

⇒ Save US$51.6m/yr

Yeoh et al, Cancer Cell 2002

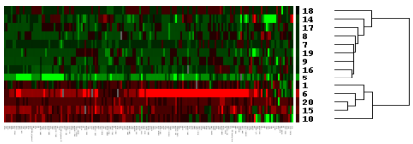
Toward More Meaningful Genes



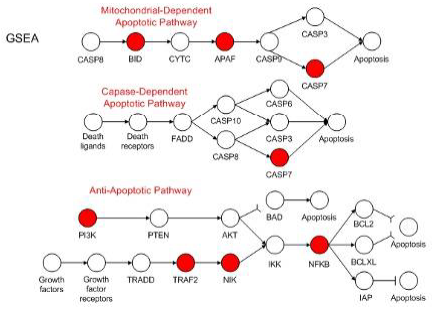
But are all of these genes meaningful?

Nasopharyngeal Carcinoma

- **NPC patients respond differentially to CYC202**



- **Can we identify drug action pathways by more sophisticated methods?**



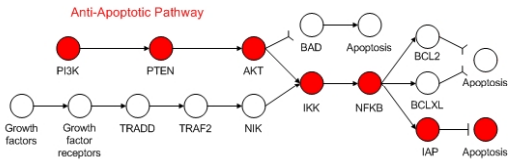
Percentage of Overlapping Genes

- **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



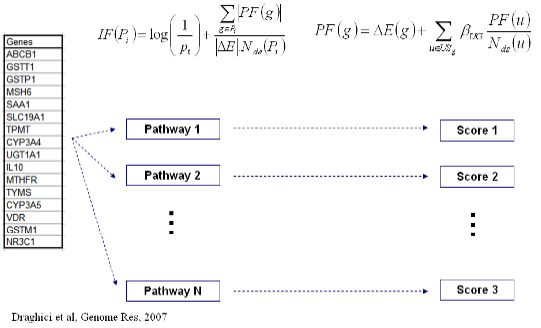
- **Each disease subtype has underlying cause**
- **There is a unifying biological theme for genes that are truly associated with a disease subtype**

- **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



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?

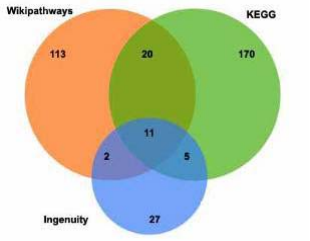
Issues on Pathway Sources

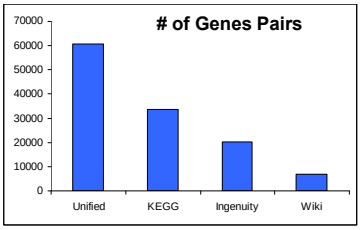
Data Sources

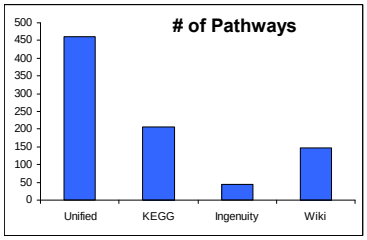
- **KEGG**
  - Curated by a single lab
  - Long famous history
  - Used by many people
- **Wikipathways**
  - Community effort
  - new curation model
- **Ingenuity**
  - Commercial effort
  - Used by many biopharma's

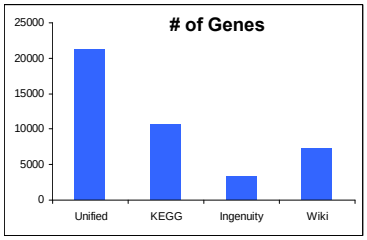
INGENUITY SYSTEMS

Low Comprehensiveness of Pathway Sources



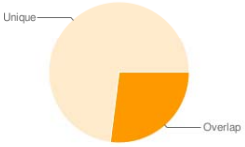




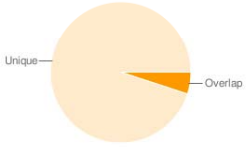


Low Consistency of Pathway Sources

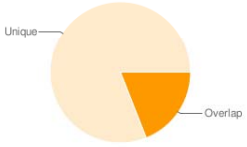
Gene Pair Overlap



Wiki vs KEGG

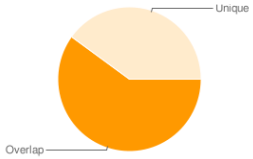


Wiki vs Ingenuity

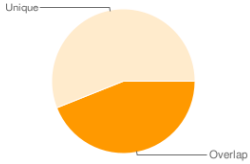


KEGG vs Ingenuity

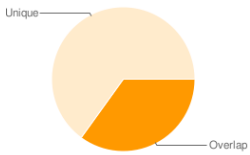
Gene Overlap



Wiki vs KEGG



Wiki vs Ingenuity



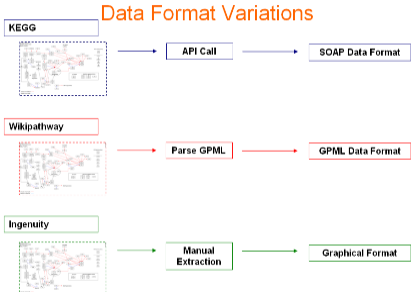
KEGG vs Ingenuity

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



Intricacy of Pathway Matching

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

Matching Pathways by Name

- **LCS procedure**
  - Given pathway X in db A
  - Sort pathways in db B by "longest common substring" with X
  - Manually scan the ranked list to choose closest nomen-clatural match
- **Issue: Accuracy**
  - When LCS says two pathways are the same one, are they really the same?
- **Issue: Completeness**
  - When LCS says two pathways are different, are they really different?

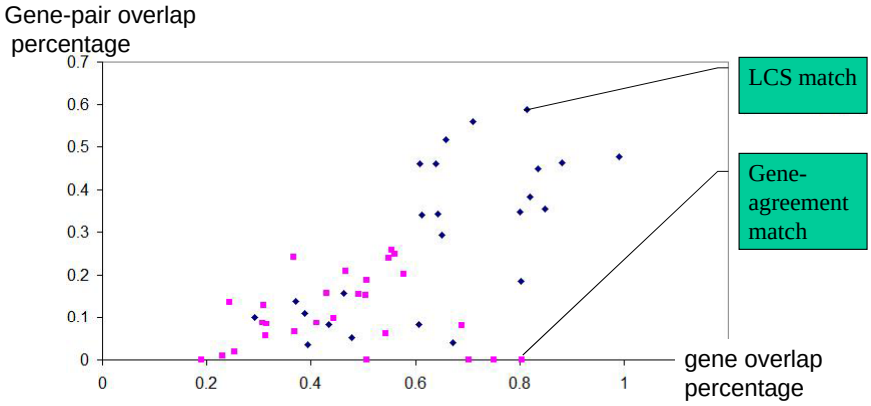
LCS vs Gene-Agreement Matching

- **Accuracy**
  - 94% of LCS matches are in top 3 gene agreement matches
  - 6% of LCS matches not in top 3 of gene agreement matches; but their gene-pair agreement levels are higher
- **Completeness**
  - Let Pi be pathway in db A LCS cannot find match in db B
  - Let Qi be pathway in db B with highest gene agreement to Pi
  - Gene-pair agreement of Pi-Qi is much lower than pathway pairs matched by LCS

LCS is better than gene-agreement based matching!

29

LCS vs Gene-Agreement Matching

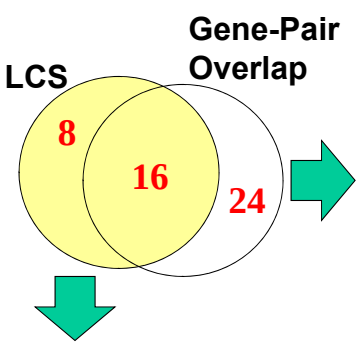


- **LCS consistently has higher gene-pair agreement**

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

30

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

31

LCS vs Gene-Pair Agreement Matching

- **Gene-pair agreement match will miss when**
  - Pathway P in db A has few overlap with pathway P in db B due to incompleteness of db, even if pathway name matches perfectly!
  - Example: wnt signaling pathway, VEGF signaling pathway, MAPK signaling pathway, etc. in KEGG don't have largest gene-pair overlap w/ corresponding pathways in Wikipathways & Ingenuity

⇒ **Bad for getting a more complete unified pathway P**

32

LCS vs Gene-Pair Agreement Matching

- **Pathways having large gene-pair overlap are not necessarily the same pathways**
- **Examples**
  - "Synaptic Long Term Potentiation" in Ingenuity vs "calcium signalling" in KEGG
  - "PPAR-alpha/RXR-alpha Signaling" in Ingenuity vs "TGF-beta signaling pathway" in KEGG

⇒ **Difficult to set correct gene-pair overlap threshold to balance against false positive matches**

Bad Gene-Pair Agreement Matches

- **"Synaptic Long Term Potentiation" in Ingenuity vs "calcium signalling" in KEGG**
  - **Calcium signaling pathway in KEGG describes general mechanism of external calcium signal transduction into cells**
  - **Calcium signal transduction can activate multiple downstream pathways, LTP is one of them**
  - ⇒ **LTP in Ingenuity is only a downstream event of the calcium pathway in KEGG**
- **"PPAR-α/RXR- α signaling" in Ingenuity vs "TGF-β signaling" in KEGG**
  - **PPARα/RXRα plays essential roles in the regulation of cellular differentiation, development, metabolism, and tumorigenesis**
  - **TGF-β acts as antiproliferative factor in normal cells at early stages of oncogenesis**
  - ⇒ **They are independent. The reason they are paired is that they have a mutual inhibition**

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%
- **Significant lack of comprehensiveness**
  - 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**

Acknowledgements


Donny Soh


Difeng Dong


Yike Guo

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

Agency for Science, Technology and Research
SINGAPORE