Big data in bioinformatics: Is more better?

Limsoon Wong

What is Big Data and Why

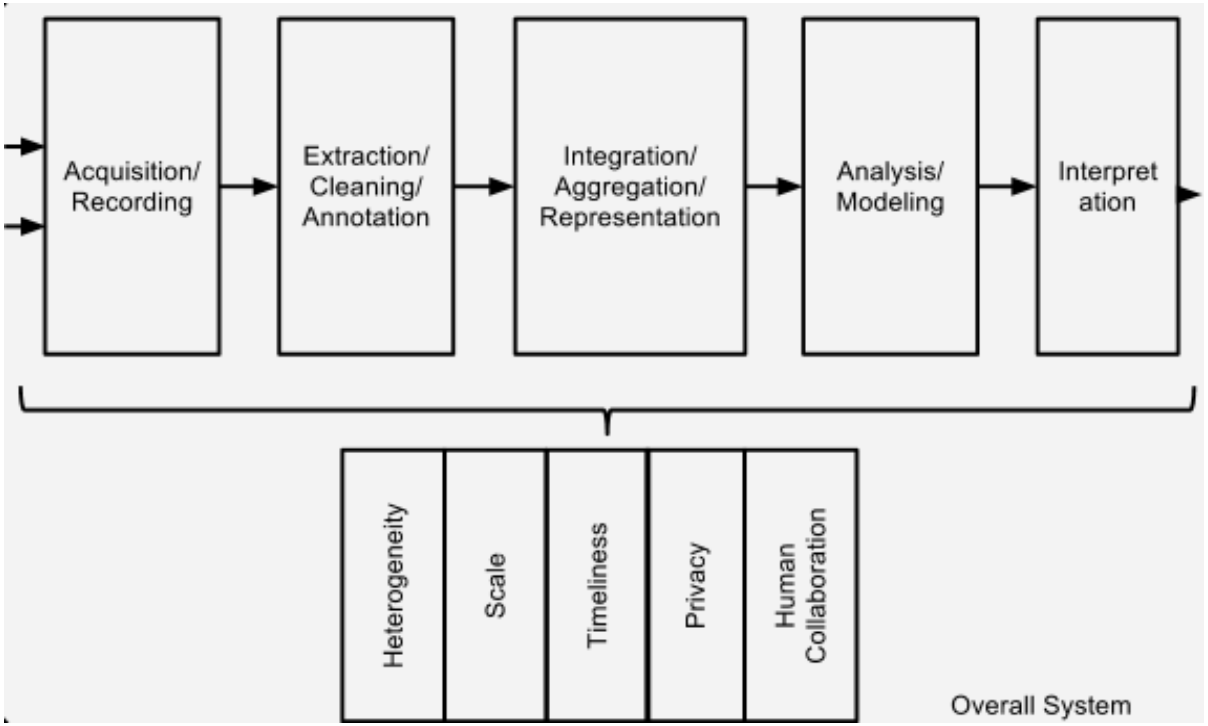
- **Big data *a la* Gartner**
  - Volume, velocity, variety
- **Other characteristics**
  - Veracity, v...

A practical definition

"More than you know how to handle"

- **Why big data?**
  - Can collect cheaply, due to automation
  - Can store cheaply, due to falling media prices
  - Many success stories, where useful predictions were made with the data

Challenges in Big Data

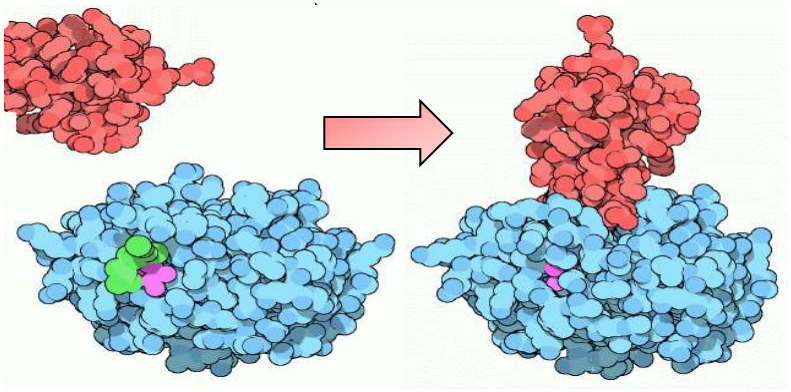


- **Much emphasis is on scaling issues**
- **But there are non-scaling-related issues that affect fundamental assumptions in current bioinformatics and statistical analysis**
  - Big data may break analysis procedures in fundamental ways

A few stories

- **Discovering protein complexes from PPIN**
- **Identifying causal genes**
- **Finding interesting patterns**

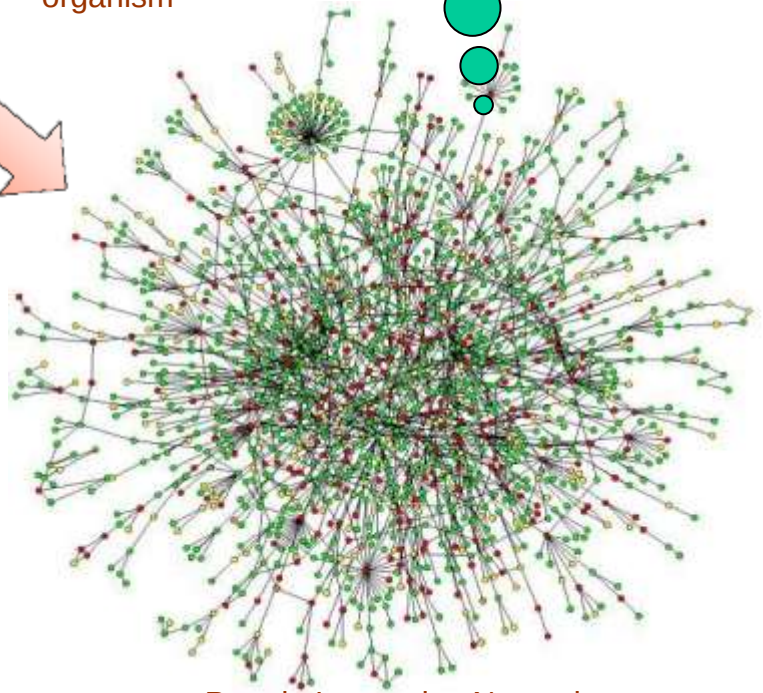
Protein-protein interaction networks



Individual proteins come together and interact

Collection of such interactions in an organism

PPIN
Valuable source of knowledge



Protein Interaction Network

- **Proteins come together & interact**
- **The collection of these interactions form a Protein Interaction Network or PPIN**

Difficulties

- **Cytochrome BC1 complex**
  - Involved in electron-transport chain in mitochondrial inner membrane

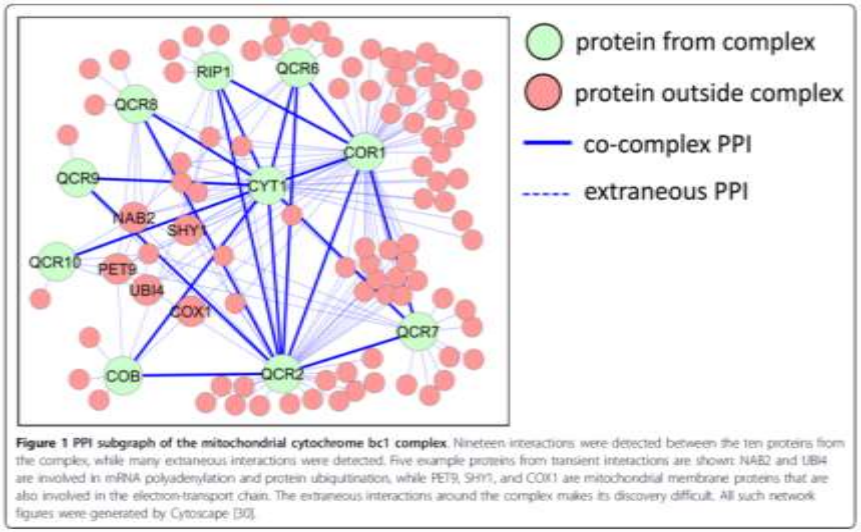


Figure 1 PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation and protein ubiquitination, while PET9, SHY1, and COX1 are mitochondrial membrane proteins that are also involved in the electron-transport chain. The extraneous interactions around the complex makes its discovery difficult. All such network figures were generated by Cytoscape [30].

- **Discovery of BC1 from PPI data is difficult**
  - Sparseness of its PPI subnetwork
    - **Only 19 out of 45 possible interactions were detected between the complex's proteins**
  - Extraneous interactions with other proteins outside the complex
    - **E.g., UBI4 is involved in protein ubiquitination, and binds to many proteins to perform its function**

Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

Perhaps "big data" can help?

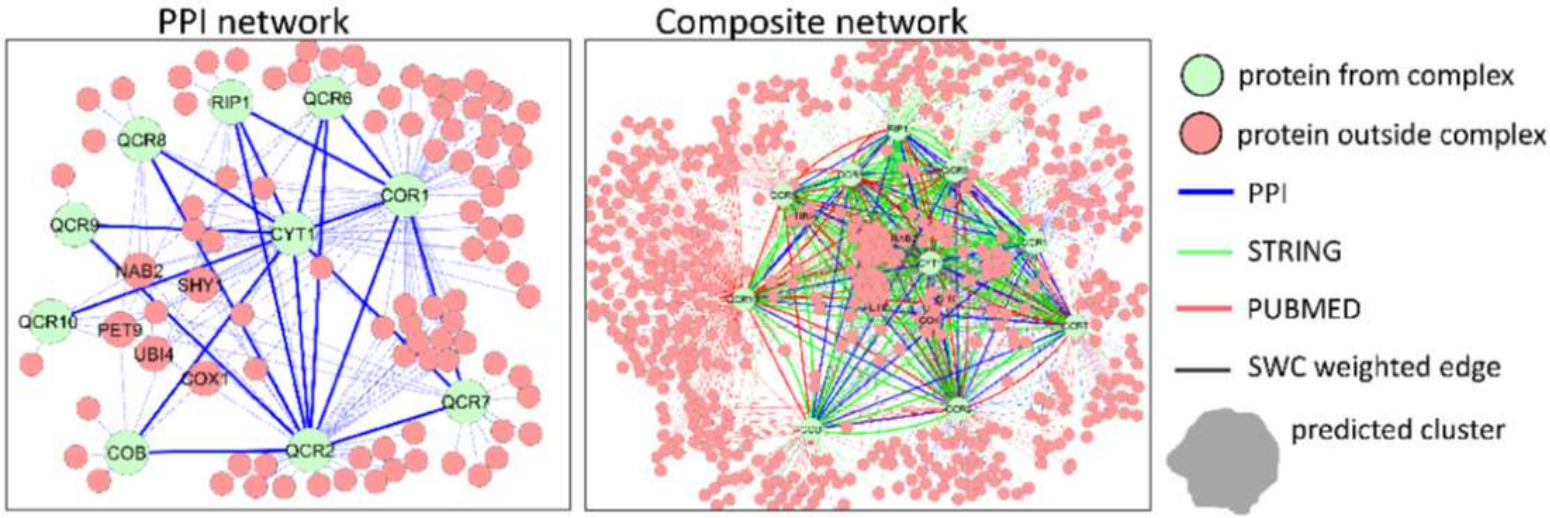
- **Composite network**
  - Vertices represent proteins, edges represent relationships between proteins. Put an edge betw proteins u, v, iff u and v are related according to any of the data sources

Data source	Database	Scoring method
PPI	BioGRID, IntACT, MINT	Iterative AdjustCD.
L2-PPI (indirect PPI)	BioGRID, IntACT, MINT	Iterative AdjustCD
Functional association	STRING	STRING
Literature co-occurrence	PubMed	Jaccard coefficient

	Yeast			Human		
	# Pairs	% co-complex	coverage	# Pairs	% co-complex	coverage
PPI	106328	**5.8%**	**55%**	48098	10%	14%
L2-PPI	181175	1.1%	18%	131705	5.5%	20%
STRING	175712	5.7%	89%	311435	3.1%	27%
PubMed	161213	4.9%	70%	91751	4.3%	11%
All	531800	**2.1%**	**98%**	522668	3.4%	49%

Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

More is not always better, unless..



While proteins in BC1 become fully connected in the composite network, there is a blow-up in extraneous proteins. So clustering won't discover the complex, unless you know how to remove the extraneous proteins

A few stories

- Discovering protein complexes from PPIN
- **Identifying causal genes**
- Finding interesting patterns

NUS National University of Singapore

Gene Expression Profiling

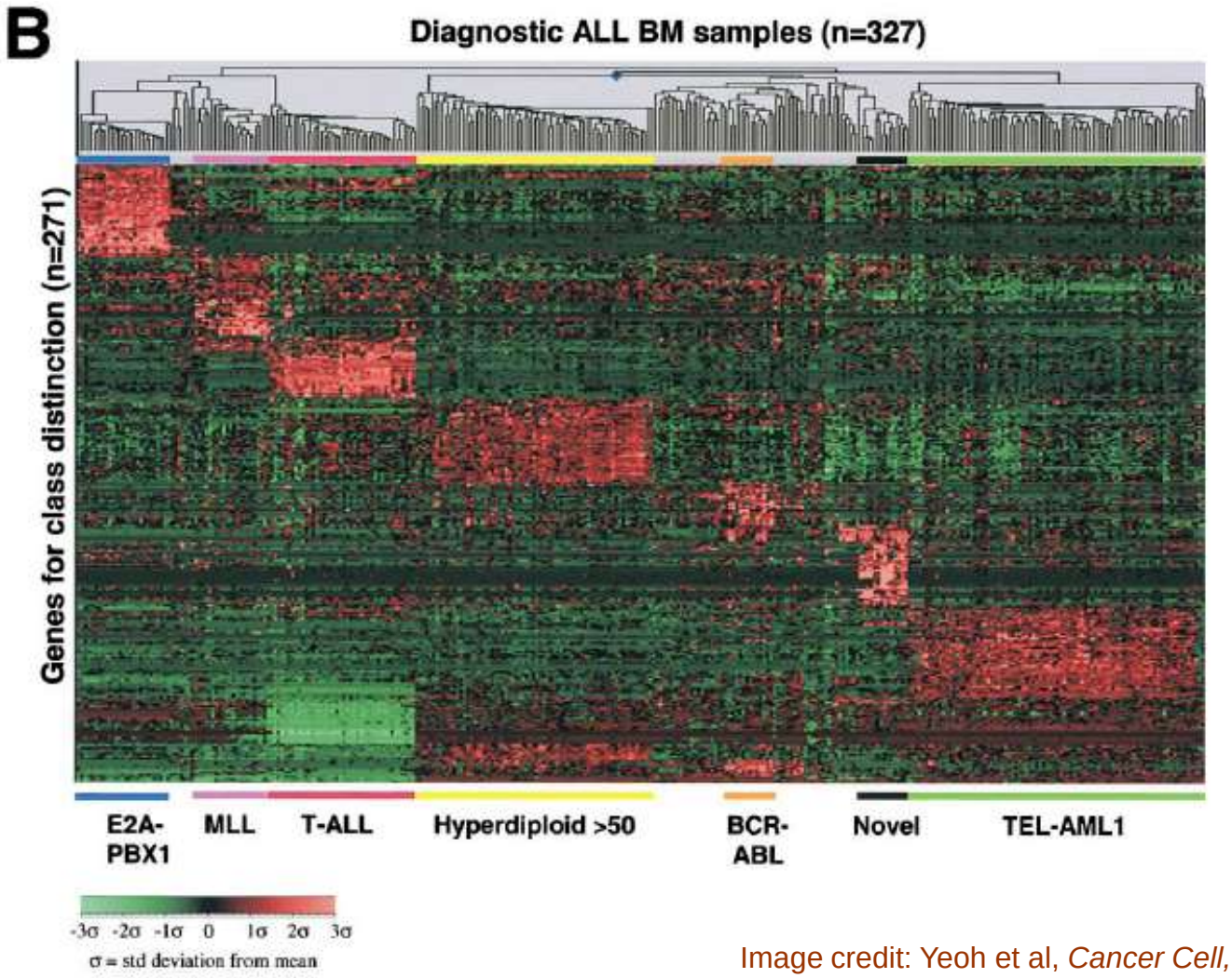


Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

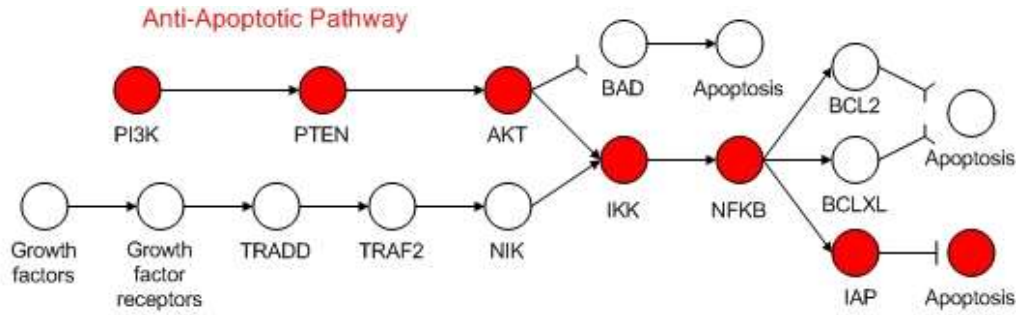
Difficulties

- **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

Biology to the rescue: Gene Regulatory Circuits



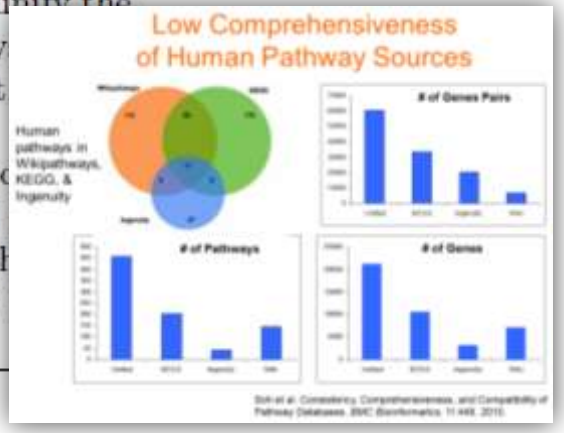
- **Each disease phenotype has some underlying cause**
- **There is some unifying biological theme for genes that are truly associated with a disease subtype**

- **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**

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 pathway 564 organisms. The data is returned in BioPax format
PathwayAPI	PathwayAPI (`http://www.pathwayapi.com`) contains unified human pathways obtained from a merge of WikiPathways and Ingenuity® Knowledge Base (Soh 2010). Data is downloadable as a SQL dump or as a and is also interfaceable in JSON format.

Goh, et al. *Proteomics*, 12(4-5):550-563, 2012.

Big data of biological pathways



Pathway-Based Methods

- **ORA**
  - Khatri et al, *Genomics*, 2002

- **FCS**
  - Pavlidis & Noble, *PSB* 2002
- **GSEA**
  - Subramanian et al, *PNAS*, 2005

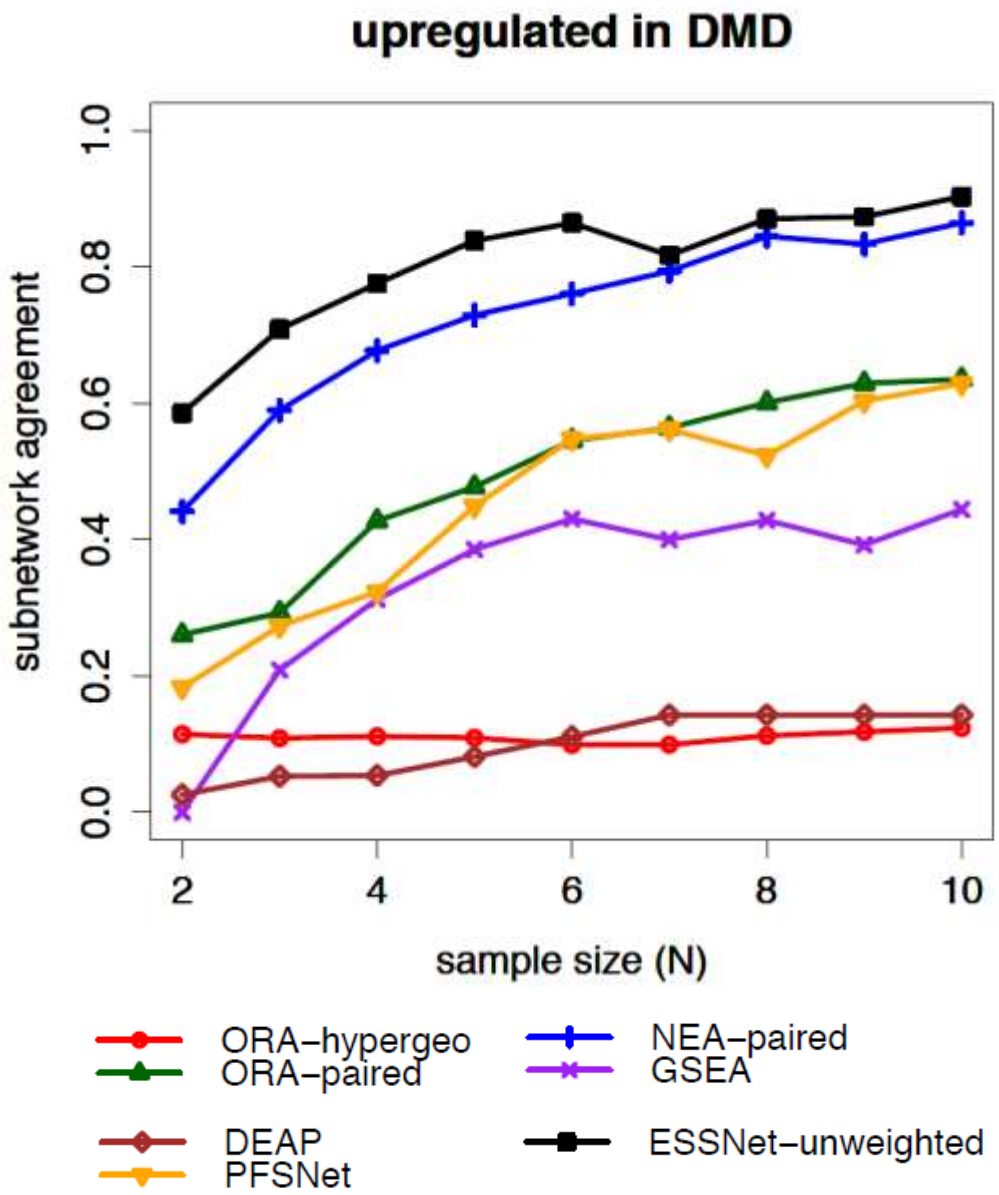
- **DEAP**
  - Winston et al., PLoS Comp Biol, 2013
- **SNet, PFSNet, ESSNet**
  - Soh et al., *BMC Genomics*, 2011
  - Lim & Wong, *Bioinformatics*, 2014

Overlap Analysis

Direct-Group Analysis

Network-Based Analysis

Unfortunately, most pathway-based gene expression analysis methods don't always work, especially when sample size is small

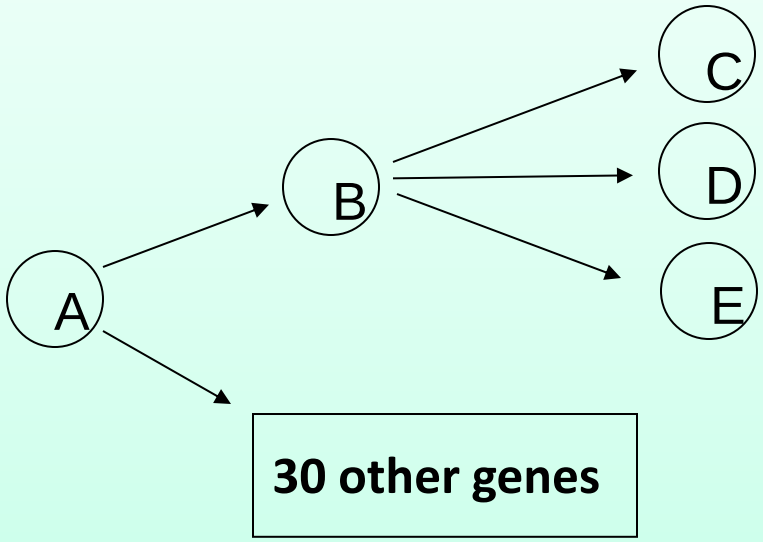


Lim et al. **ESSNet, finding consistent disease subnetworks in data with extremely small sample sizes**. *Bioinformatics*, submitted.

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

NUS National University of Singapore

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**

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

PFSNet vs GSEA & GGEA: Pathway Agreement

Testing subnets from PFSNet using GSEA & GGEA

	PFSNet
Leukemia (GSEA)	0.50
Leukemia (GGEA)	0.67
ALL subtype (GSEA)	1.00
ALL subtype (GGEA)	1.00
DMD (GSEA)	0.90
DMD (GGEA)	0.54

Lim & Wong. **Finding consistent disease subnetworks using PFSNet**. *Bioinformatics*, 30(2):189--196, 2014

A few stories

- Discovering protein complexes from PPIN
- Identifying causal genes
- **Finding interesting patterns**

Statistics lies, unless …

Overall

	A	B
lived	60	65
died	100	165

Looks like treatment A is better

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

Looks like treatment B is better

History of heart disease

	A	B
lived	10	5
died	70	50

No history of heart disease

	A	B
lived	10	45
died	10	110

Looks like treatment A is better

Big data may break the i.i.d. assumption underlying most statistical tests

Rules are just rules, unless …

Data mining sensor & telemetry data in a factory may give you rules like …

Fuse blow → Robot stop

··· a thousand other rules ···

Circuit overload → Fuse blow

··· a thousand other rules ···

Insufficient lubrication → Circuit overload

··· a thousand other rules ···

Oil pump clogged → Insufficient lubrication

··· a thousand other rules ···

Metal shavings → Oil pump clogged

Asking "why" 5 levels deep, and getting to the root cause

What have we learned?

- **More data can offer a more complete picture, fill in gaps, etc.**
- **More data can also introduce noise into an analysis**
- **Unless you know how to tame this noise, more data may not lead to a better analysis**
- **Mechanical application of statistical and data mining techniques often does not work**
- **Must understand statistical and data mining tools & the problem domain**
  - Must know how to logically exploit both