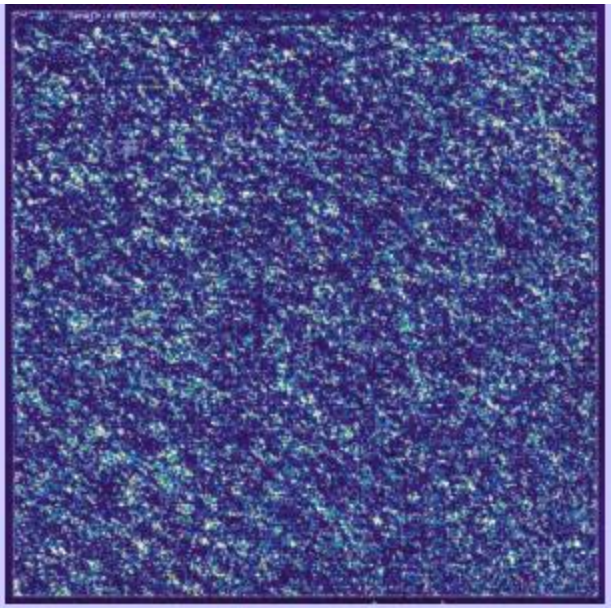
Finding consistent disease subnetworks in data with extremely small sample sizes

Limsoon Wong

Affymetrix GeneChip Array



Source: Affymetrix

Typical Analysis Workflow

- **Gene expression data collection**
- **DE gene selection by, e.g., t-statistic**
- **Classifier training based on selected DE genes**
- **Apply the classifier for diagnosis of future cases**

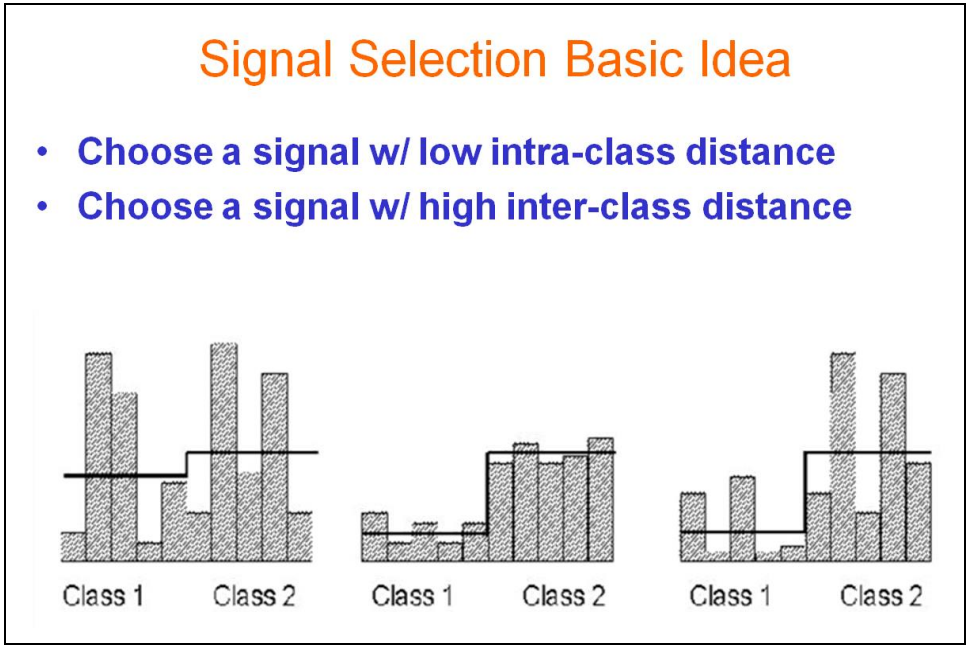


Image credit: Golub et al., *Science*, 286:531–537, 1999

Terminology: DE gene = differentially expressed gene

Hierarchical Clustering

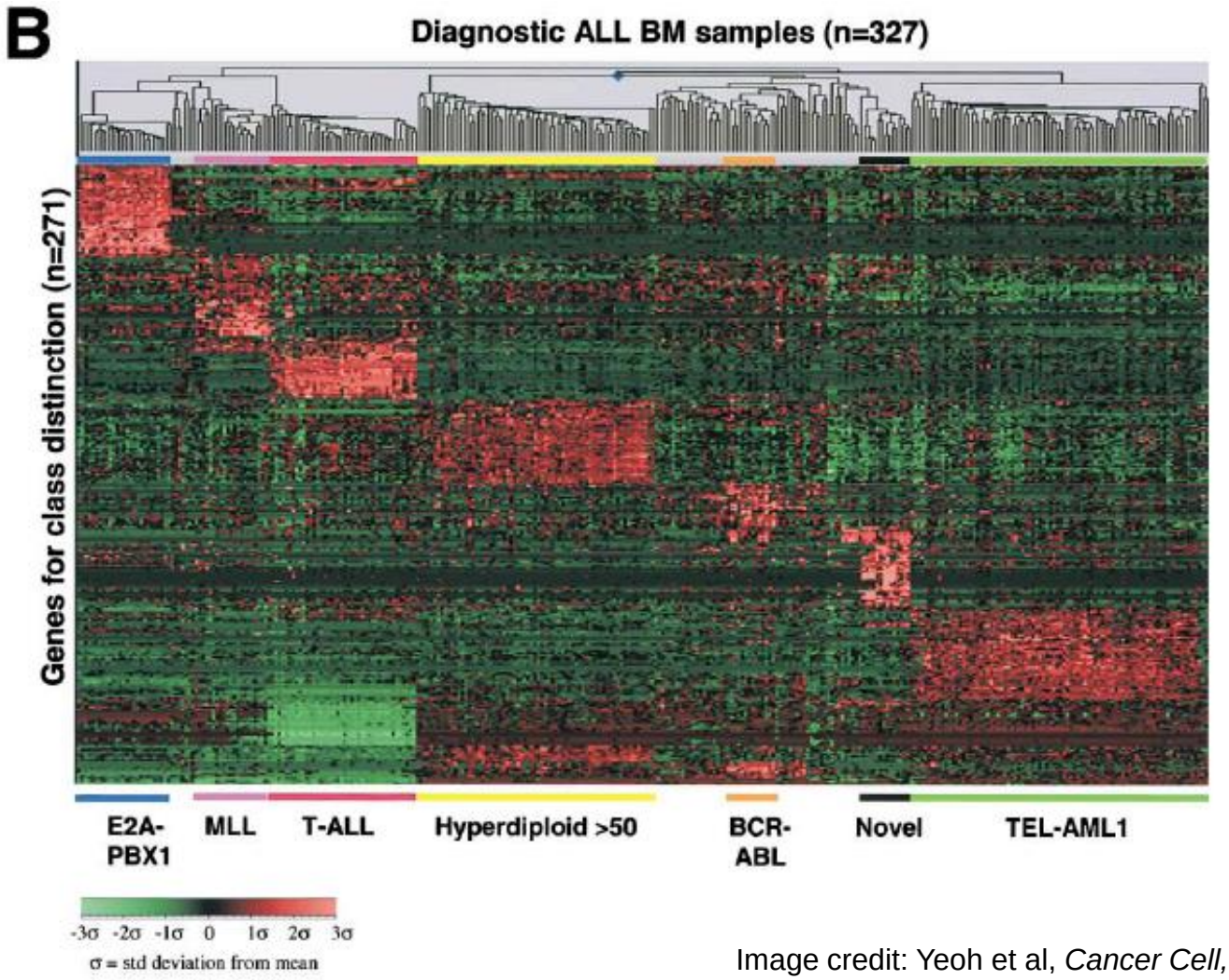


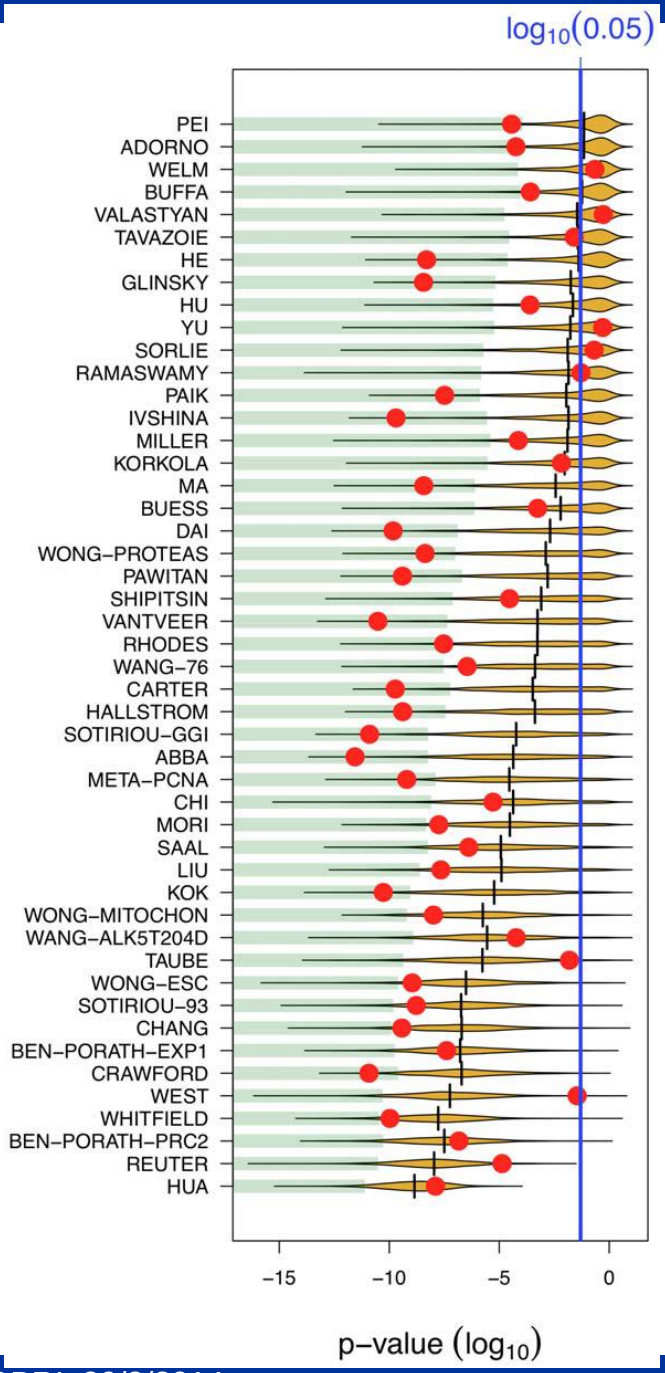
Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

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?**

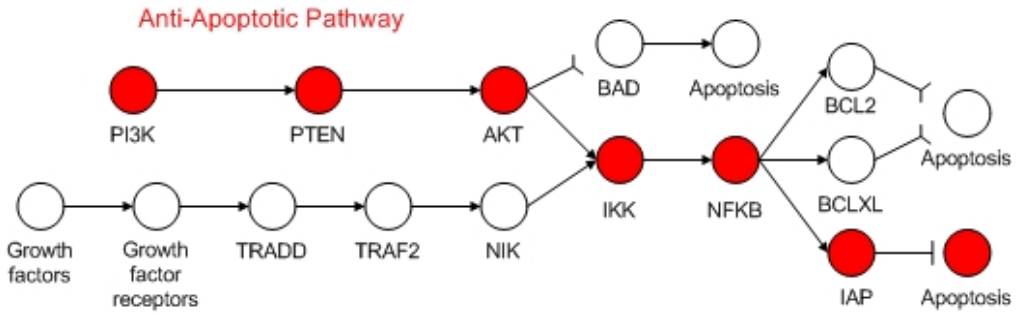
- **Prob(a gene is correlated) = $1/2^6$**
- **# of genes on array = 30,000**

⇒ **E(# of correlated genes) = 469**

⇒ **Many false positives**

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

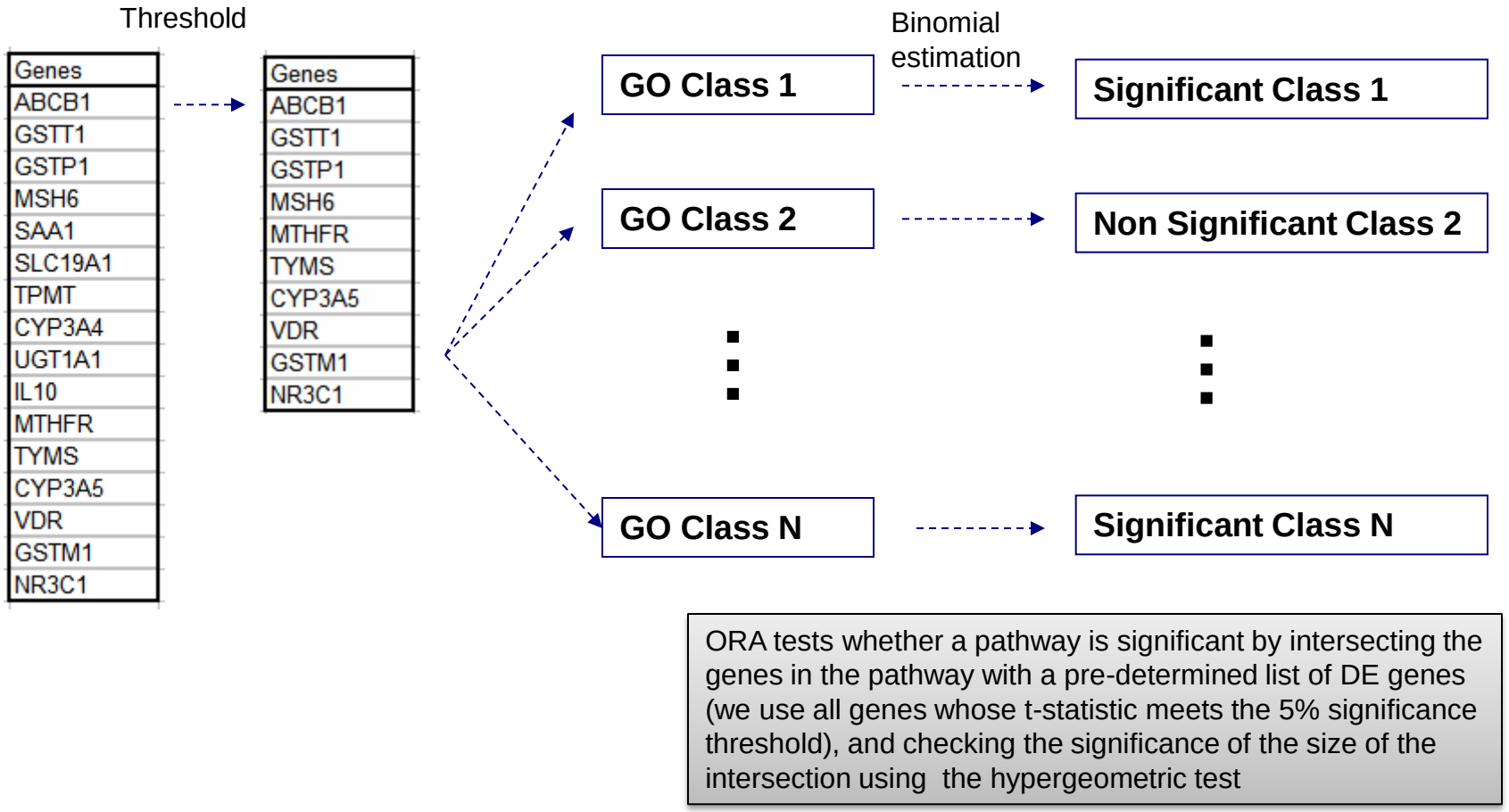
Gene Regulatory Circuits



- **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**

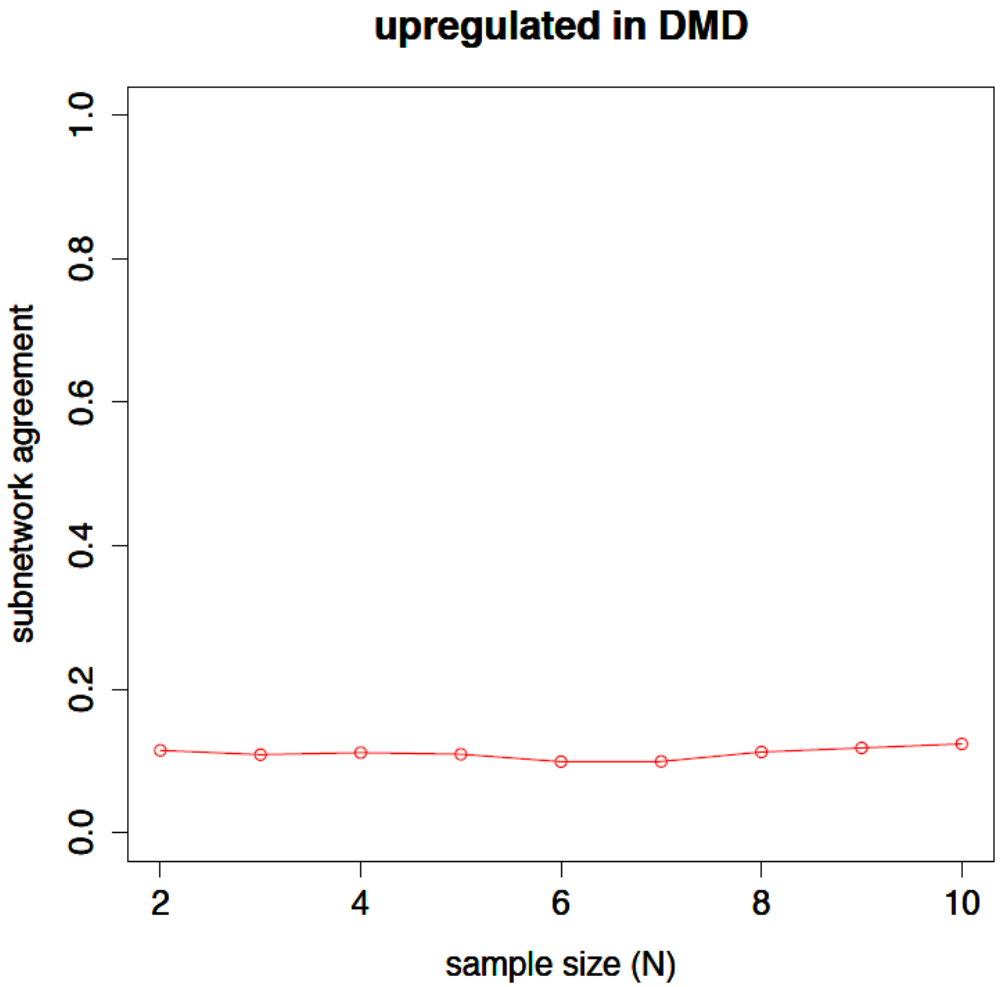
Overlap Analysis: ORA



ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (we use all genes whose t-statistic meets the 5% significance threshold), and checking the significance of the size of the intersection using the hypergeometric test

S Draghici et al. “Global functional profiling of gene expression”. *Genomics*, 81(2):98-104, 2003.

Disappointing Performance



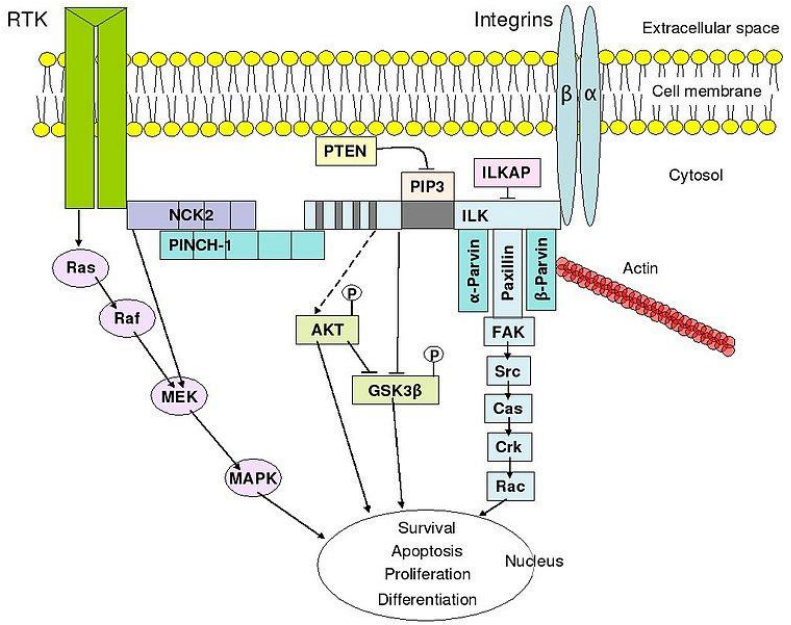
DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010

Issue #1 with ORA

- **Its null hypothesis basically says "Genes in the given pathway behaves no differently from randomly chosen gene sets of the same size"**

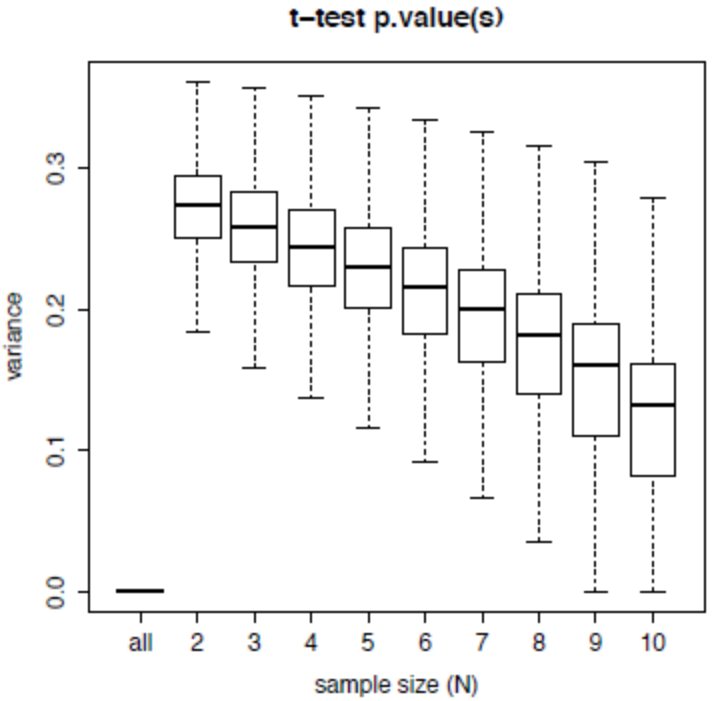
- **This may lead to lots of false positives**



- A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Thus necessarily the behavour of genes in a pathway is more coordinated than random ones

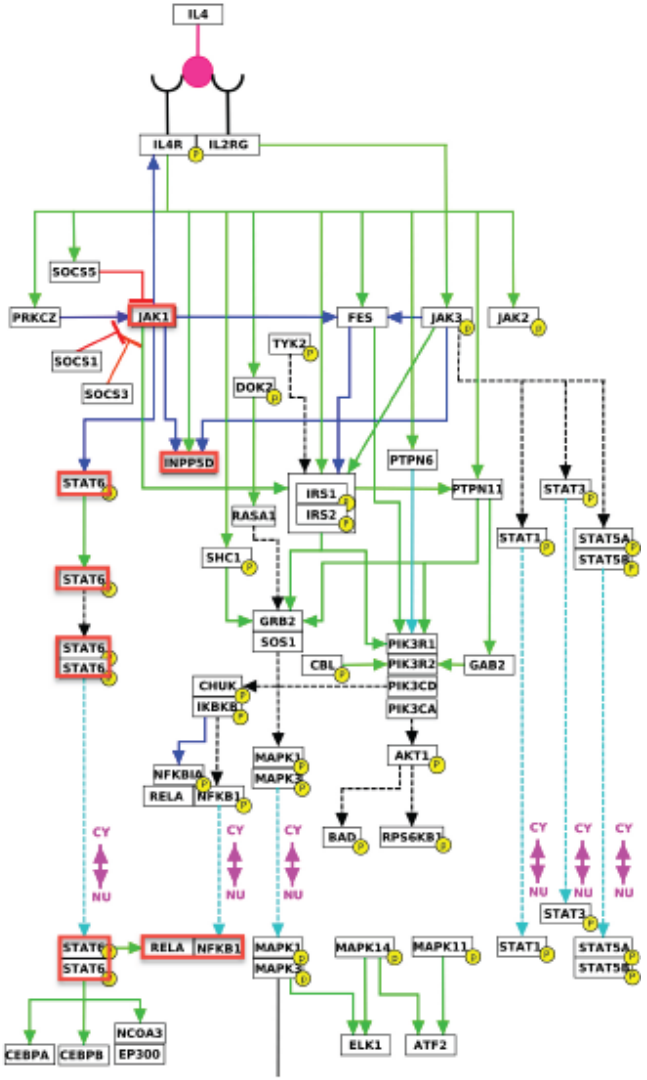
Issue #2 with ORA

- **It relies on a pre-determined list of DE genes**
- **This list is sensitive to the test statistic used and to the significance threshold used**
- **This list is unstable regardless of the threshold used when sample size is small**

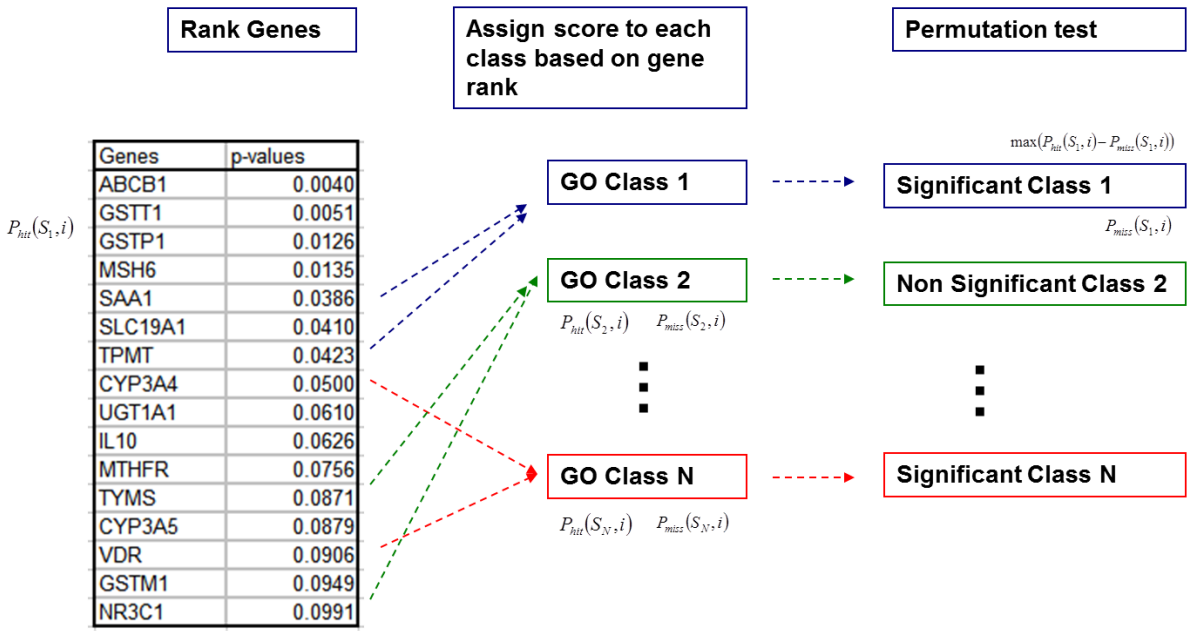


Issue #3 with ORA

- **It tests whether the entire pathway is significantly differentially expressed**
- **If only a branch of the pathway is relevant to the phenotypes, the noise from the large irrelevant part of the pathways can dilute the signal from that branch**



GSEA in Gene Permutation Mode

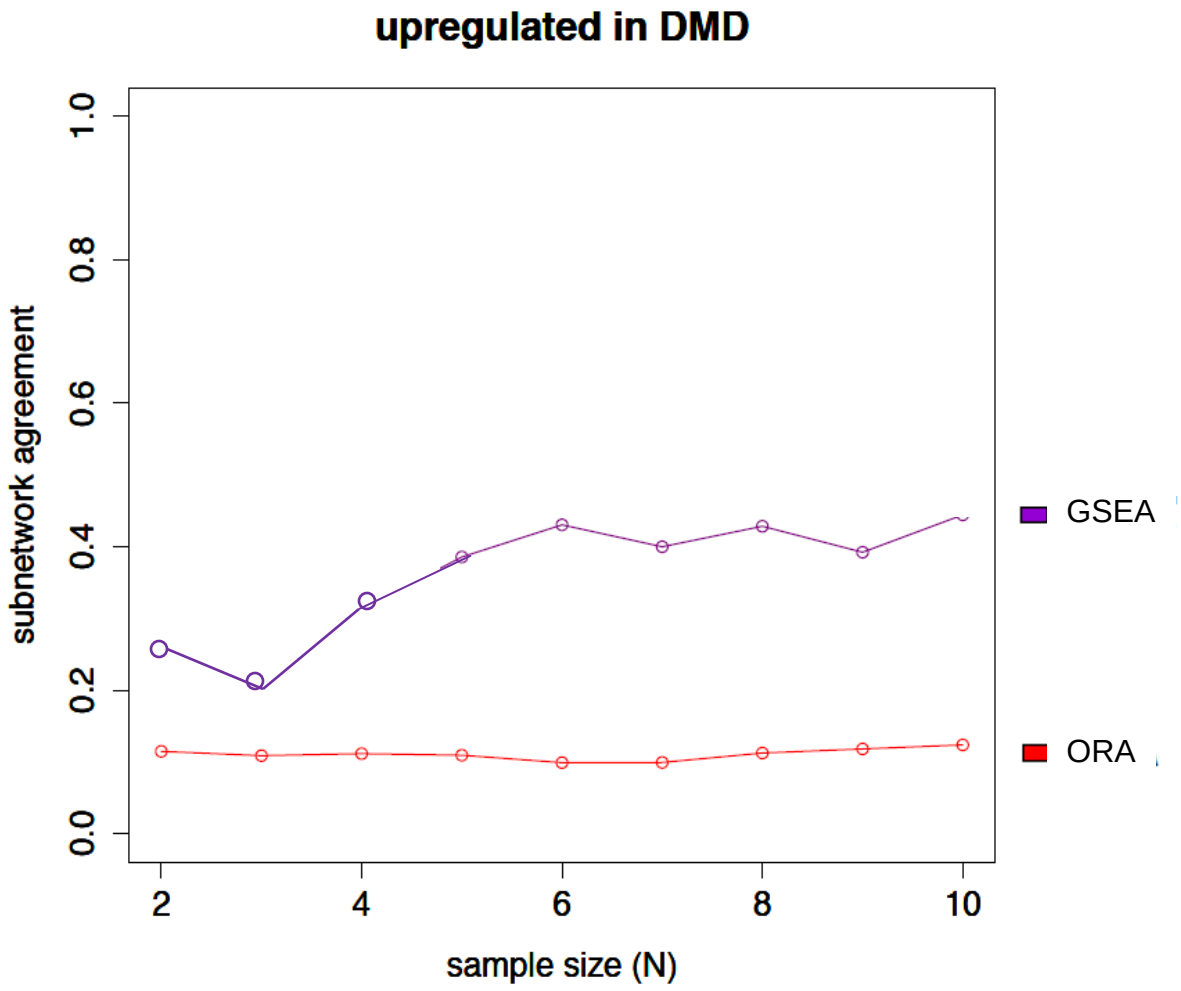


A Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome wide expression profiles". *PNAS*, 102(43):15545-15550, 2005

Note: Class label permutation mode cannot be used when sample size is small

- **Issue #2 is mostly solved**
  - Does not need pre-determined list of DE genes
  - But gene ranking (based on t-test p-value) is still unstable when sample size is small

Better Performance

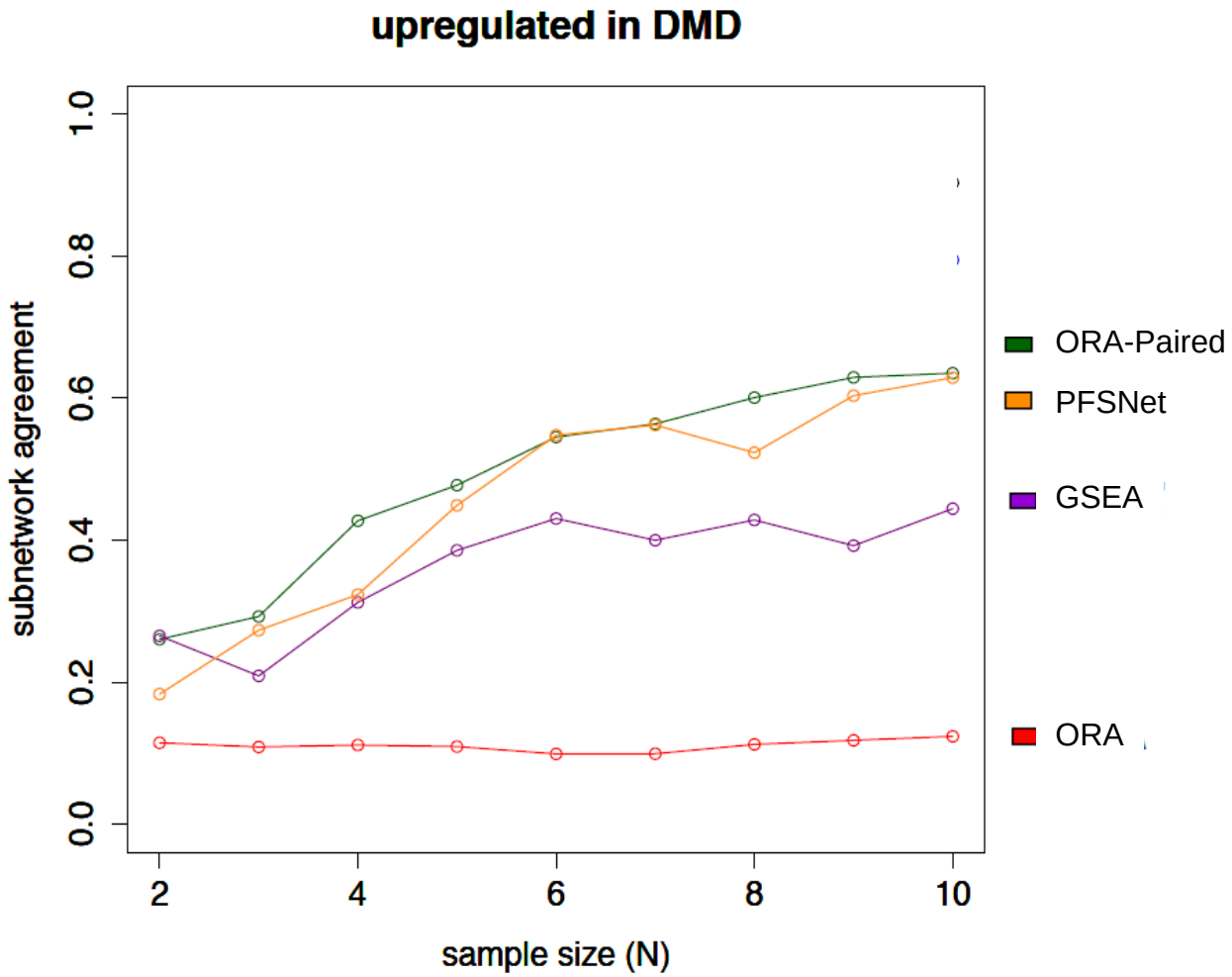


ORA-Paired: Paired Test and New Null Hypothesis

- Let g_i be genes in a given pathway P
- Let p_j be patients
- Let q_k be normals
- Let $\Delta_{i,j,k} = Expr(g_i,p_j) – Expr(g_i,q_k)$
- Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

- **Issue #1 is solved**
  - The null hypothesis is now “If a pathway P is irrelevant to the difference between patients and normals, then the genes in P are expected to behave similarly in patients and normals”
- **Issue #2 is solved**
  - No longer need a pre-determined list of DE genes
  - Sample size is now much larger
    - ~~# patients + # normals~~
    - # patients * # normals * # genes in P

Much Better Performance

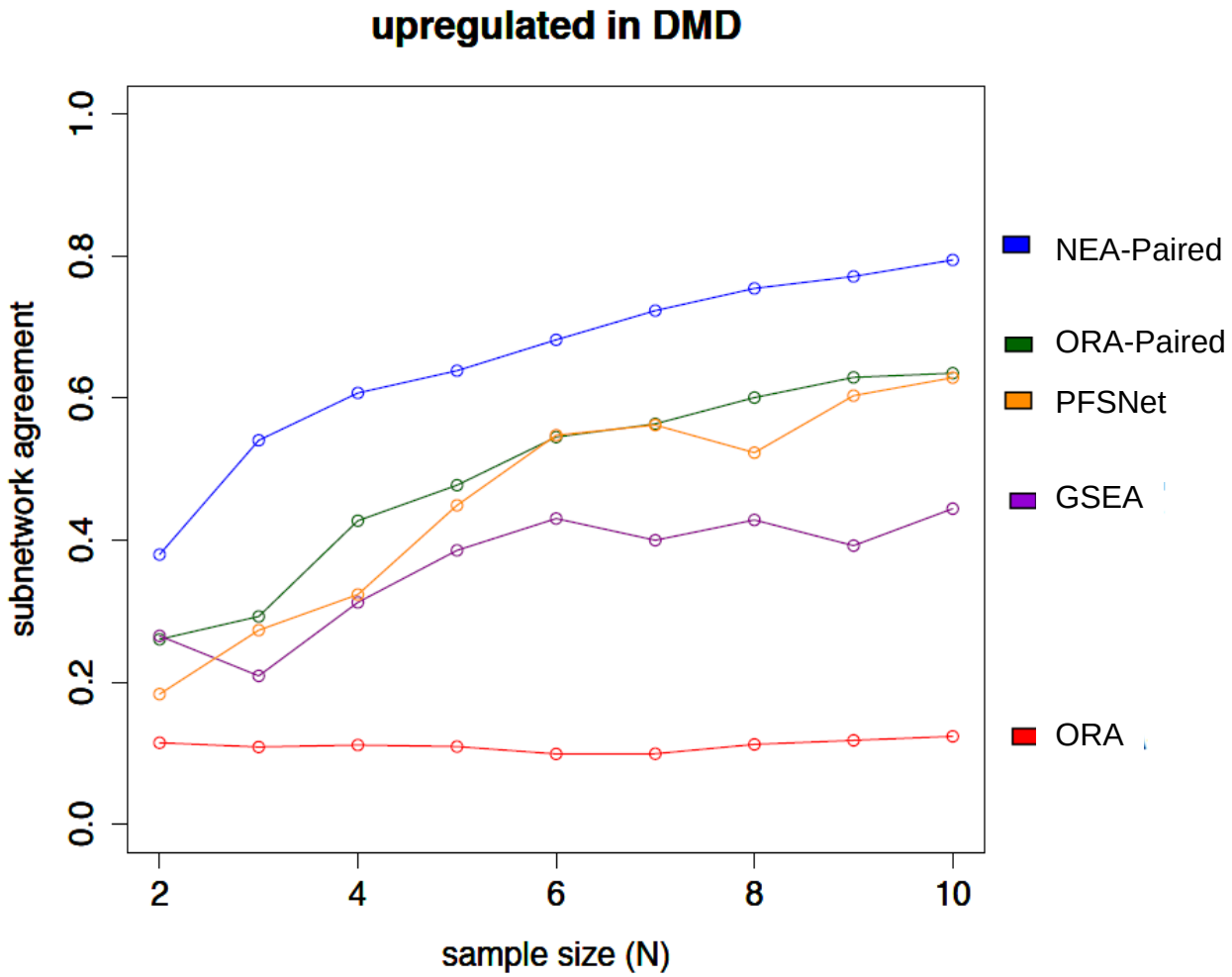


NEA-Paired: Paired Test on Subnetworks

- **Given a pathway P**
- **Let each node and its immediate neighbourhood in P be a subnetwork**
- **Apply ORA-Paired on each subnetwork individually**

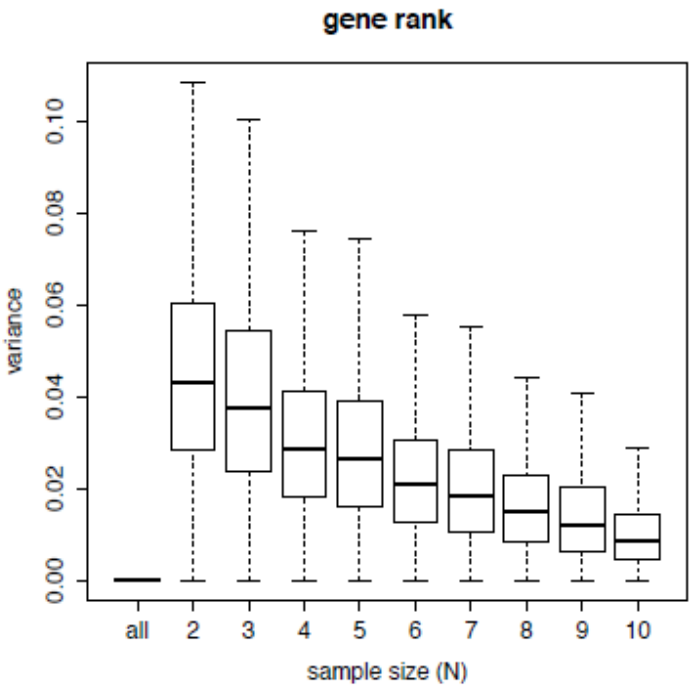
- **Issues #1 & #2 are solved as per ORA-Paired**
- **Issue #3 is partly solved**
  - Testing subnetworks instead of whole pathways
  - But subnetworks derived in fragmented way

Even Better Performance



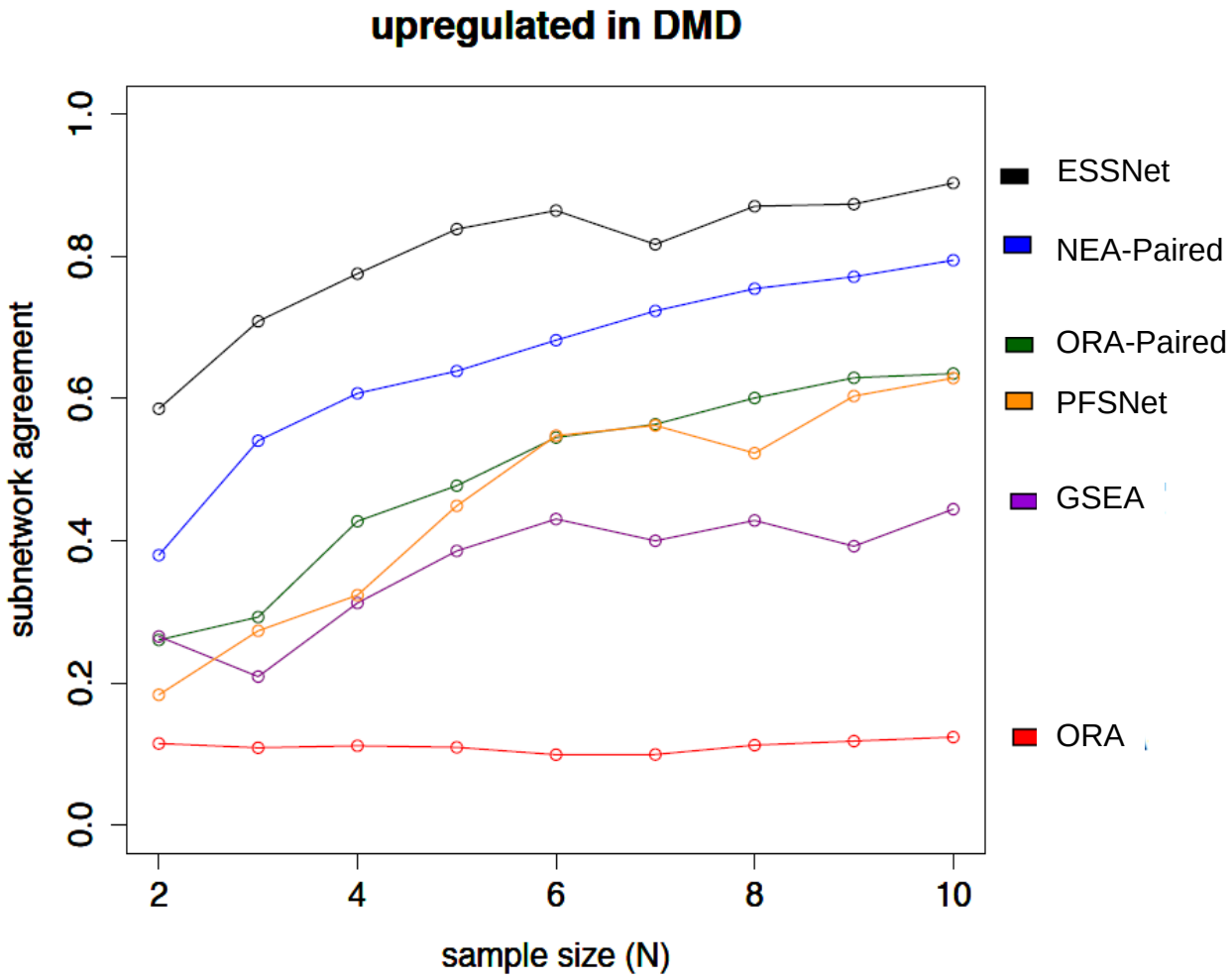
ESSNet: Larger Subnetworks

- Compute the average rank of a gene based on its expression level in patients
- Use the top α% to extract large connected components in pathways
- Test each component using ORA-Paired



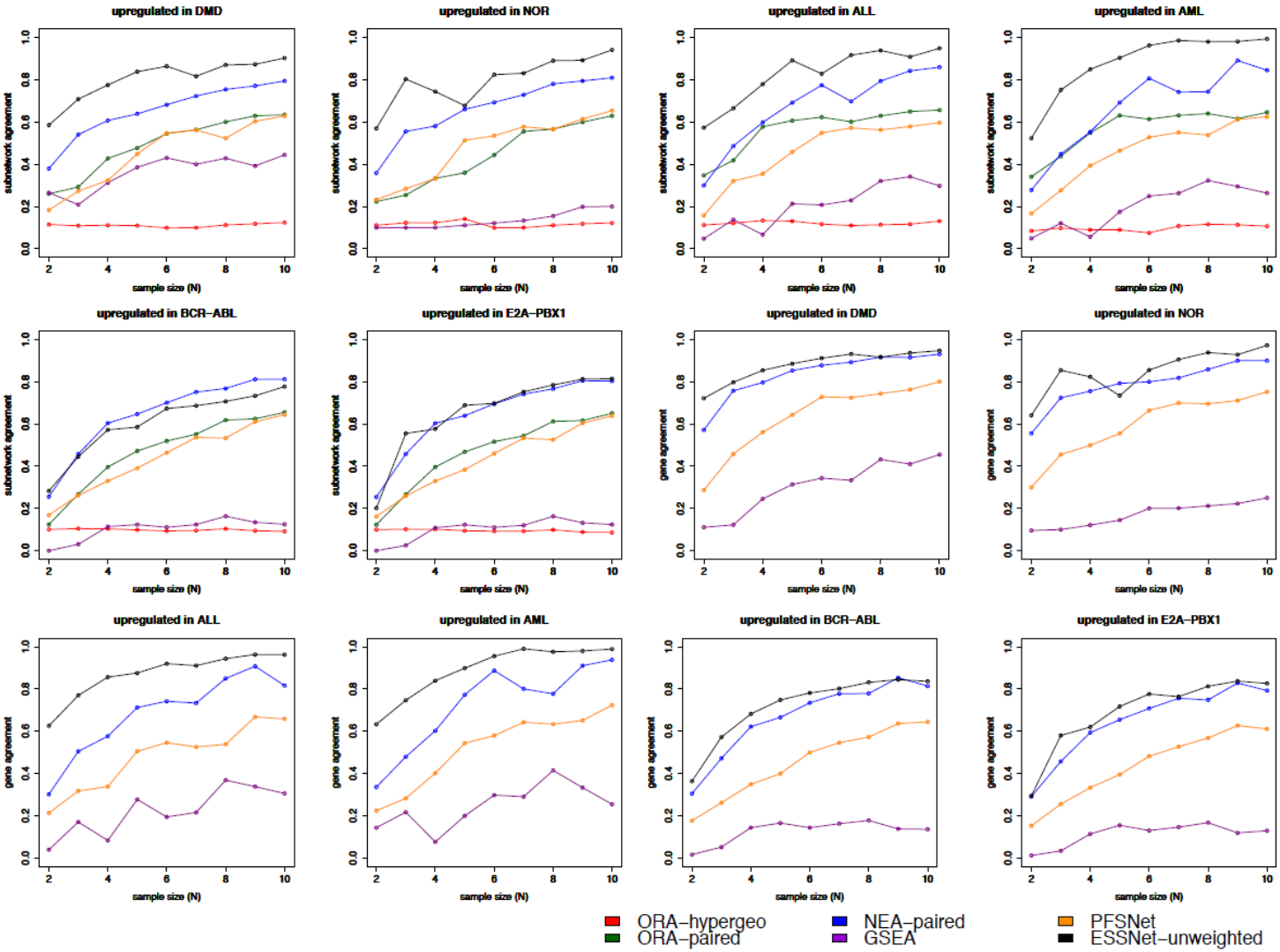
- Gene rank is very stable
- Issues #1 - #3 solved

Fantastic Performance



NUS National University of Singapore

More Datasets Tested

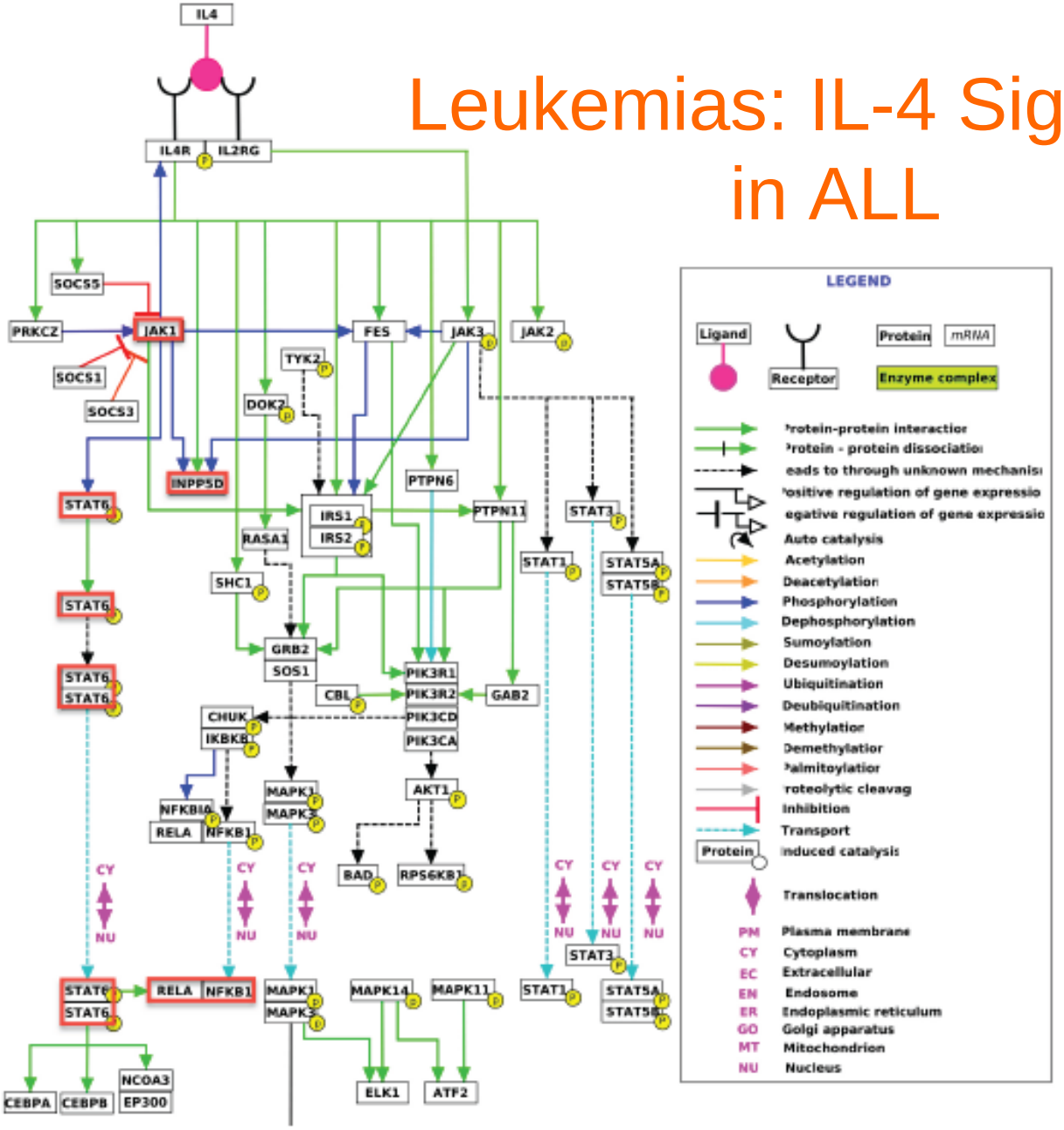


Do ESSNet results agree on small datasets vs big datasets?

sample size (N)	Precision						Recall					
	DMD		ALL		BCR		DMD		ALL		BCR	
	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D
2	0.96	0.88	0.87	0.95	0.93	0.91	0.45	0.31	0.34	0.25	0.19	0.17
3	0.93	0.86	0.99	0.89	0.90	0.87	0.56	0.45	0.56	0.41	0.21	0.16
4	0.88	0.88	0.97	0.92	0.91	0.87	0.67	0.50	0.51	0.53	0.35	0.48
5	0.89	0.88	0.94	0.90	0.89	0.90	0.73	0.52	0.74	0.55	0.36	0.38
6	0.82	0.88	0.93	0.92	0.89	0.91	0.78	0.62	0.74	0.62	0.44	0.438
7	0.85	0.86	0.95	0.93	0.90	0.87	0.75	0.59	0.66	0.64	0.55	0.53
8	0.84	0.89	0.97	0.94	0.90	0.92	0.81	0.69	0.74	0.66	0.61	0.66
9	0.88	0.90	0.94	0.92	0.89	0.89	0.90	0.67	0.76	0.74	0.65	0.67
10	0.88	0.93	0.97	0.92	0.90	0.90	0.86	0.84	0.89	0.74	0.66	0.73

- **The table above uses ESSNet's results on entire datasets as the benchmark to evaluate ESSNet's results on small subsets of the datasets**
- **The precision (i.e., agreement) is superb, though some subnetworks are missed when smaller datasets are analysed**

Leukemias: IL-4 Signaling in ALL



For the Leukemia dataset (in which patients are either classified to have acute lymphoblastic leukemia or acute myeloid leukemia), one of the significant subnetworks that is biologically relevant is part of the Interleukin-4 signaling pathway; see figure 6b (supplementary material). The binding of Interleukin-4 to its receptor (Cardoso *et al.*, 2008) causes a cascade of protein activation involving JAK1 and STAT6 phosphorylation. STAT6 dimerizes upon activation and is transported to the nucleus and interacts with the RELA/NFKB1 transcription factors, known to promote the proliferation of T-cells (Rayet and Gelinas, 1999). In contrast, acute myeloid leukemia does not have genes in this subnetwork up-regulated and are known to be unrelated to lymphocytes.

Concluding Remarks

- **Consistent successful gene expression profile analysis needs deep integration of background knowledge**
- **Most gene expression profile analysis methods fail to give reproducible results when sample size is small (and some even fail when sample size is quite large)**
- **Logical analysis to identify key issues and simple logical solution to the issues can give fantastic results**

Acknowledgements

- **My students**
  - Donny Soh
  - Dong Difeng
  - Kevin Lim
  - Li Zhenhua
- **& collaborator**
  - Choi Kwok Pui
- **MOE**

- Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Finding Consistent Disease Subnetworks Across Microarray Datasets**. *BMC Genomics*, 12(Suppl. 13):S15, November 2011
- Kevin Lim, Limsoon Wong. **Finding consistent disease subnetworks using PFSNet**. *Bioinformatics*, 30(2):189--196, January 2014
- Kevin Lim, Zhenhua Li, Kwok Pui Choi, Limsoon Wong. **ESSNet: Finding consistent disease subnetworks in data with extremely small sample sizes**. In preparation