A quantum leap in the reproducibility, precision, & sensitivity of gene-expression-profile analysis even when sample size is extremely small

Limsoon Wong

(Based on the work of my student Kevin Lim)

NUS
National University
of Singapore

Why small sample size?

- **Biological constraint**
  - Comparing cell lines
  - Comparing mutants vs wildtype
- **Rare-sample constraint**
- **Population-size constraint**
  - Singapore is small, we often wait a long time for enough patients presenting the desired phenotype
- **Cost constraint**

Outline

- Ideals of a perfect method for gene selection in gene expression profile analysis
- Failure of commonly-used methods
- Reproducible precise & sensitive selection of genes, even when sample size is extremely small
- Reliable accurate cross-batch classification, even when batch effect is severe and sample size is small

THE IDEAL

A perfect method for identifying causal factors of a disease

- **A perfect method should …**
  - – Completeness: Report all causal factors in a dataset
  - – Soundness: Not report any non-factor

⇒ **When it is applied to two representative datasets of the disease, the two sets of identified factors should be the same**

⇒ **When it is applied to a subset of a dataset, the set of identified factors should be a subset of the set of identified factors when it is applied to the whole dataset**

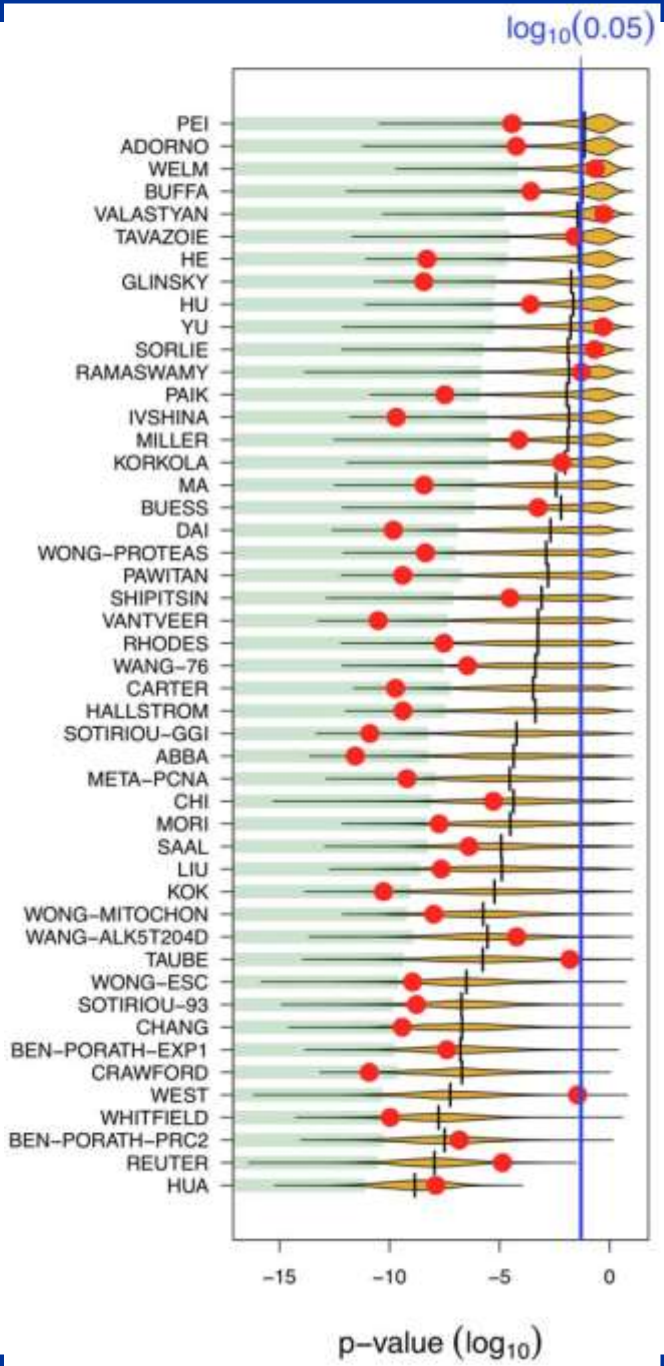
THE REALITY

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

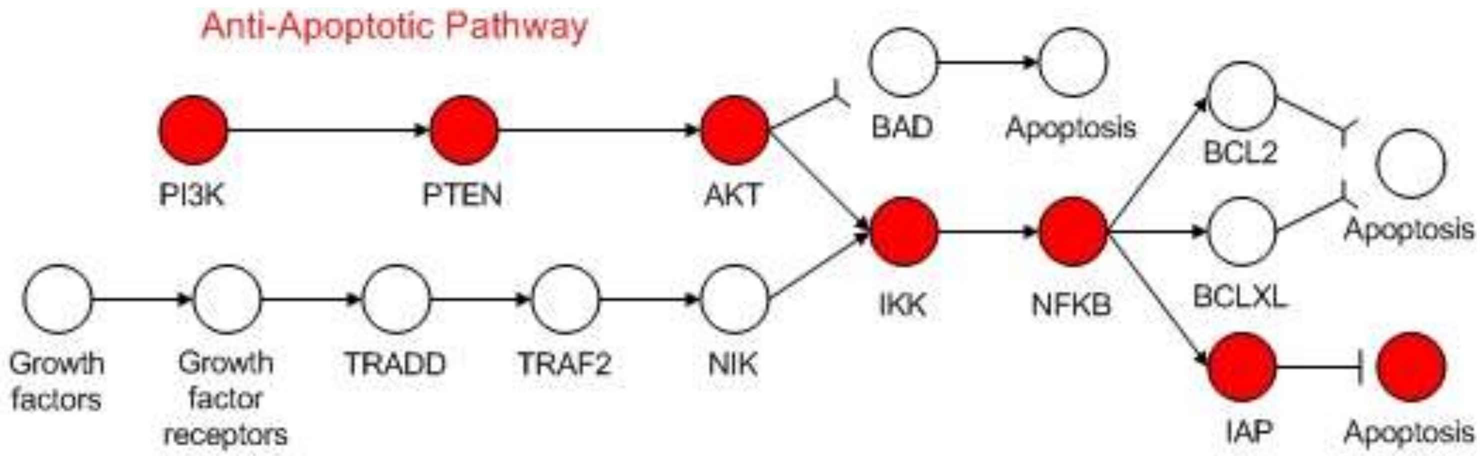
Zhang et al, *Bioinformatics*, 2009



"Most random gene expression signatures are significantly associated with breast cancer outcome"

Venet et al., *PLoS Comput Biol*, 7(10):e1002240, 2011.

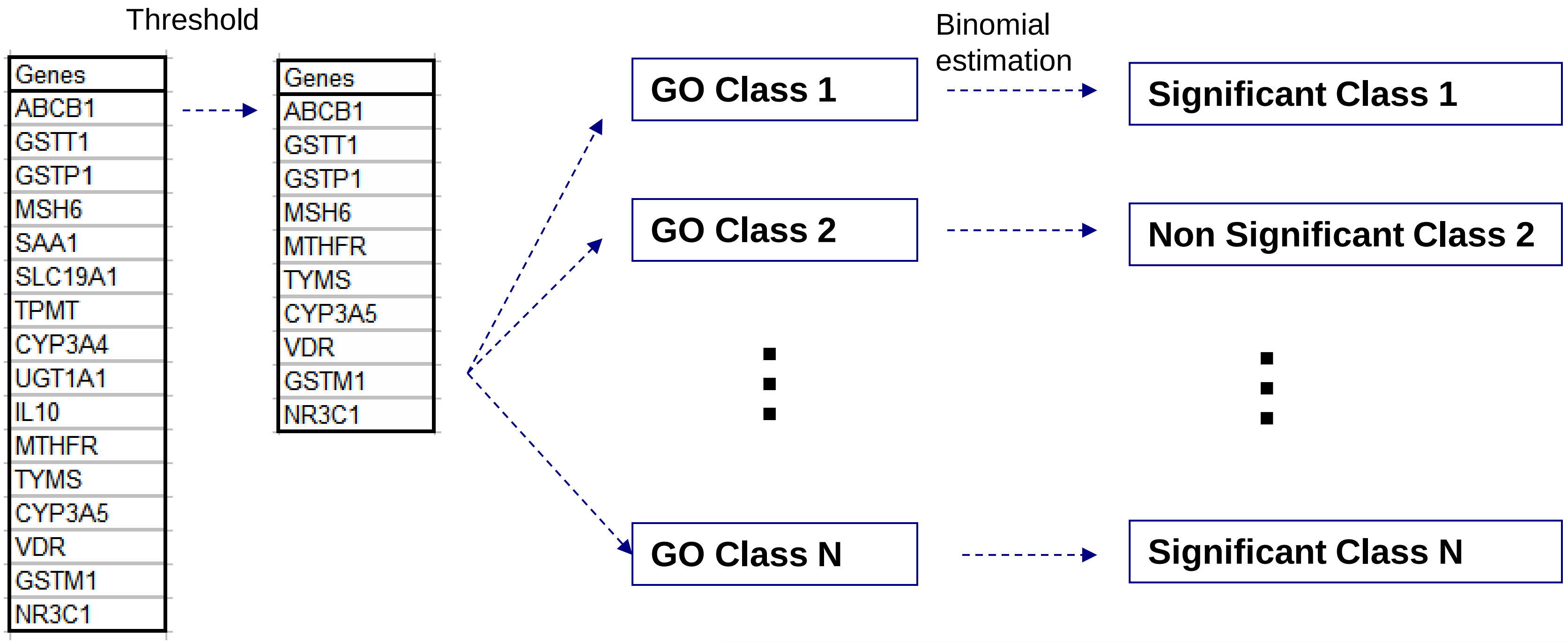
Gene regulatory circuits



- **Each disease has some underlying cause**

- **There is some unifying biological theme for genes that are truly associated with a disease**

Overlap analysis: ORA



Threshold

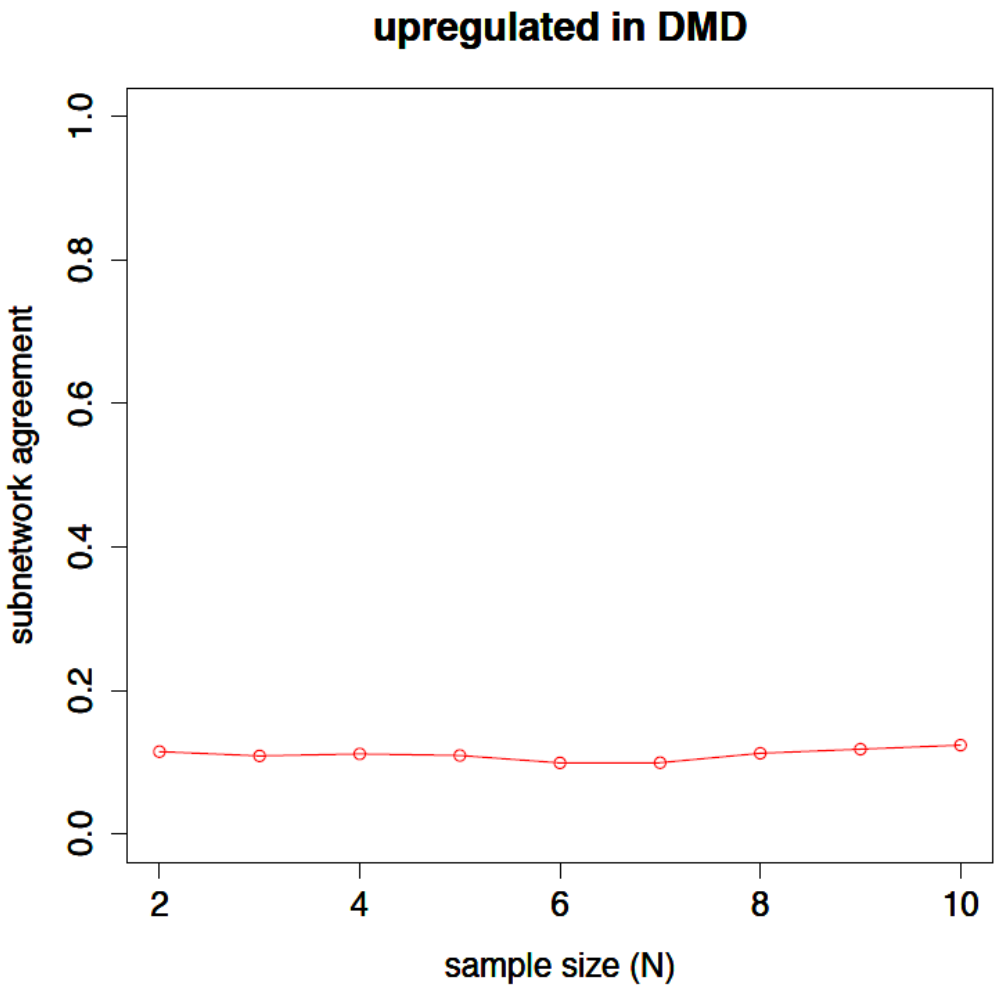
Genes
ABCB1
GSTT1
GSTP1
MSH6
SAA1
SLC19A1
TPMT
CYP3A4
UGT1A1
IL10
MTHFR
TYMS
CYP3A5
VDR
GSTM1
NR3C1

Genes
ABCB1
GSTT1
GSTP1
MSH6
MTHFR
TYMS
CYP3A5
VDR
GSTM1
NR3C1

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

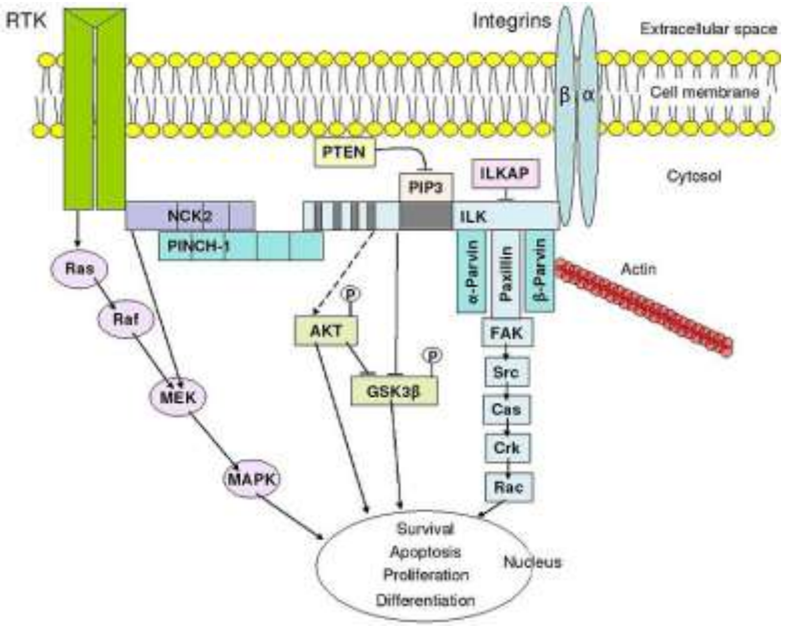
THE REASONS

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 null hypothesis is obviously false**

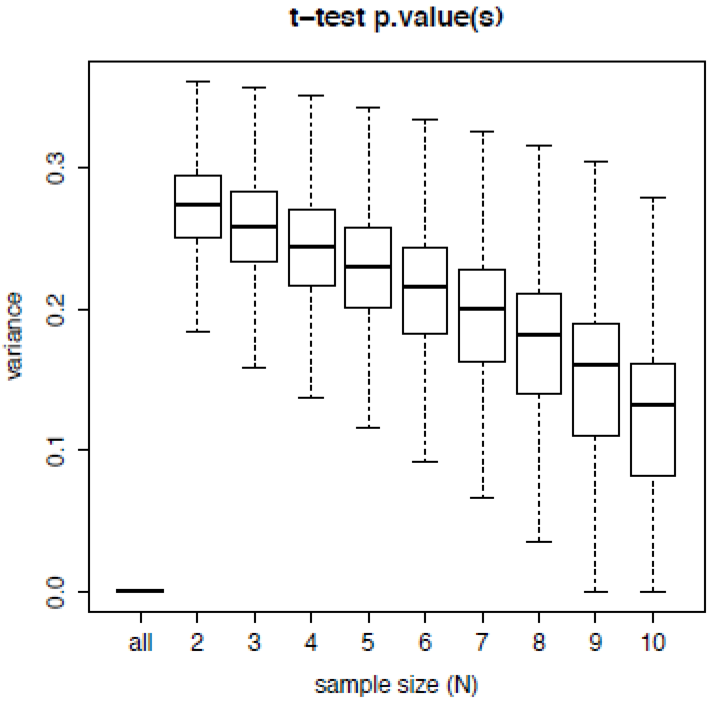
⇒ **Lots of false positives**



- 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

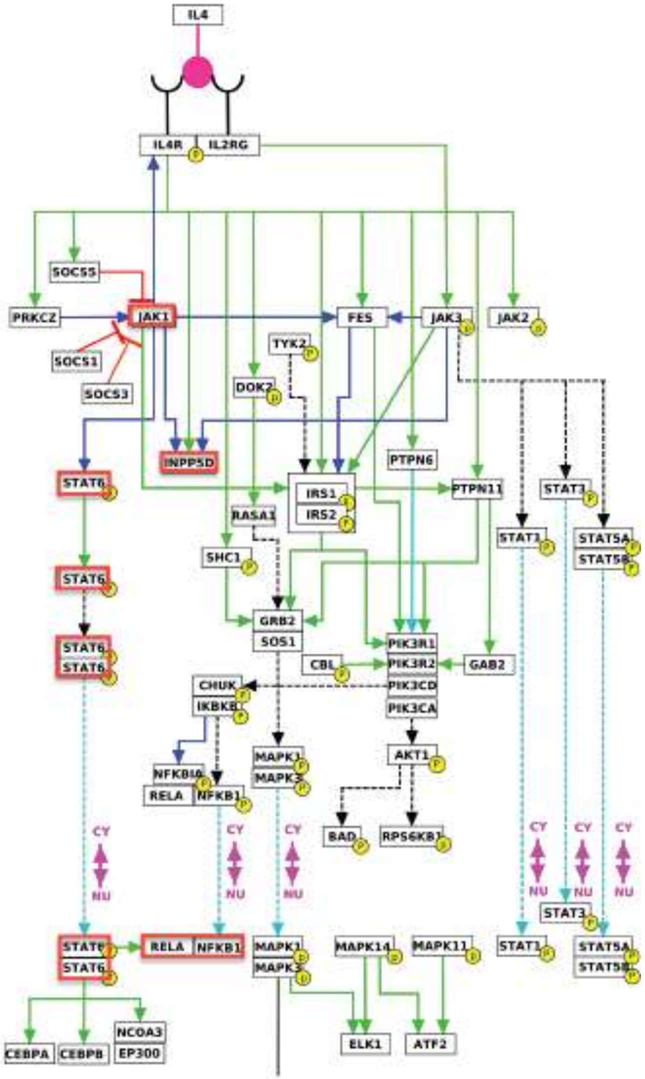
- **It relies on a pre-determined list of DE genes**
- **This list is sensitive to the test statistic used and to the significance threshold used**
- **This list is unstable regardless of the threshold used when sample size is small**



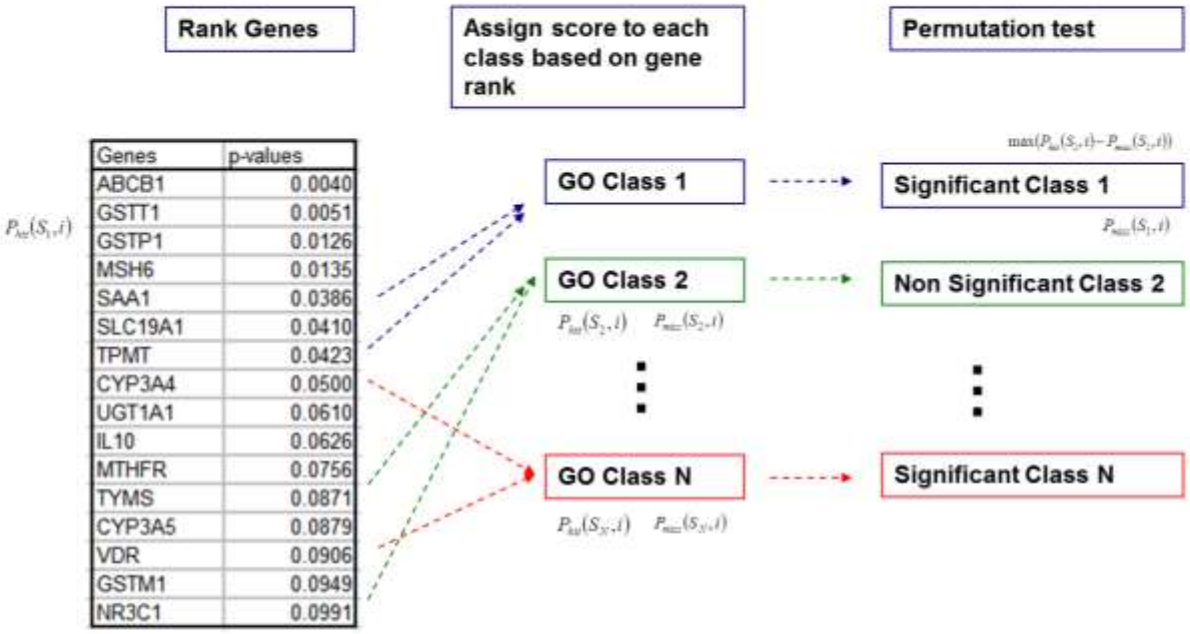
Issue #3 with ORA

- **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**

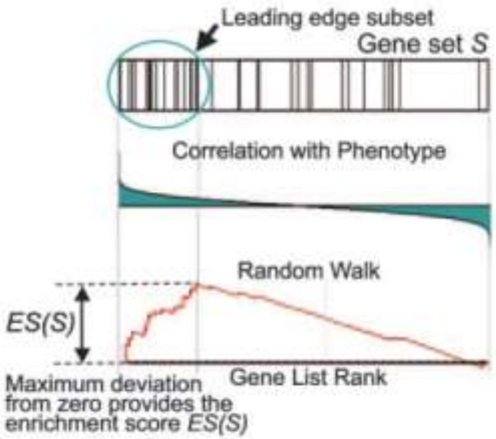


GSEA in gene-permutation mode


Rank Genes
Assign score to each class based on gene rank
Permutation test
$P_{hit}(S_1,i)$
GO Class 1
GO Class 2
$P_{hit}(S_2,i)$ $P_{miss}(S_2,i)$
GO Class N
$P_{hit}(S_N,i)$ $P_{miss}(S_N,i)$
$\max(P_{hit}(S_1,i)-P_{miss}(S_1,i))$
Significant Class 1
$P_{miss}(S_1,i)$
Non Significant Class 2
Significant Class N

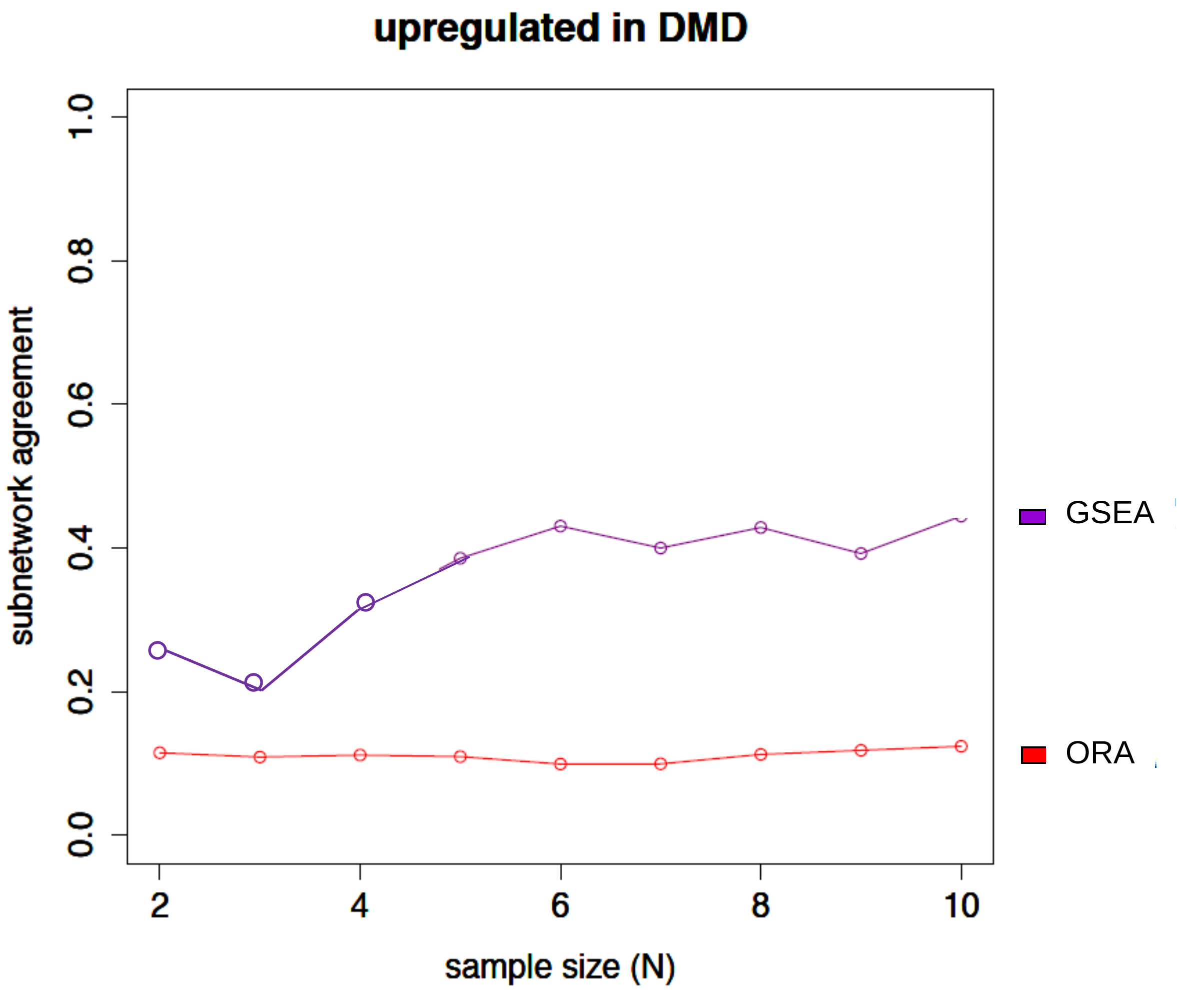
Genes	p-values
ABCB1	0.0040
GSTT1	0.0051
GSTP1	0.0126
MSH6	0.0135
SAA1	0.0386
SLC19A1	0.0410
TPMT	0.0423
CYP3A4	0.0500
UGT1A1	0.0610
IL10	0.0626
MTHFR	0.0756
TYMS	0.0871
CYP3A5	0.0879
VDR	0.0906
GSTM1	0.0949
NR3C1	0.0991

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



- **Issue #2 & #3 solved to different degrees**
  - Does not need pre-determined list of DE genes, but gene ranking (based on t-test p-value) is still unstable for small sample size
  - Irrelevant genes in pathway have only small effect on the ES(S) peak
- **Issues #1 (when sample size is small) is unsolved**

Better performance, but not great



Lim & Wong. *Bioinformatics*, 30(2):189--196, 2014

PFSNet: Exploiting subnetworks

- **Induce subnetworks from pathways by considering only genes highly expressed in majority of patients in any class**

Wt of gene i in +ve class

$$\beta_1^*(g_i) = \sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|} \qquad \beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{fs(e_{g_i,p_j})}{|\neg D|}$$

-ve class wt

Score of subnet S in patient k w/ +ve class wt

$$Score_1^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_1^*(g_i) \qquad Score_2^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_2^*(g_i)$$

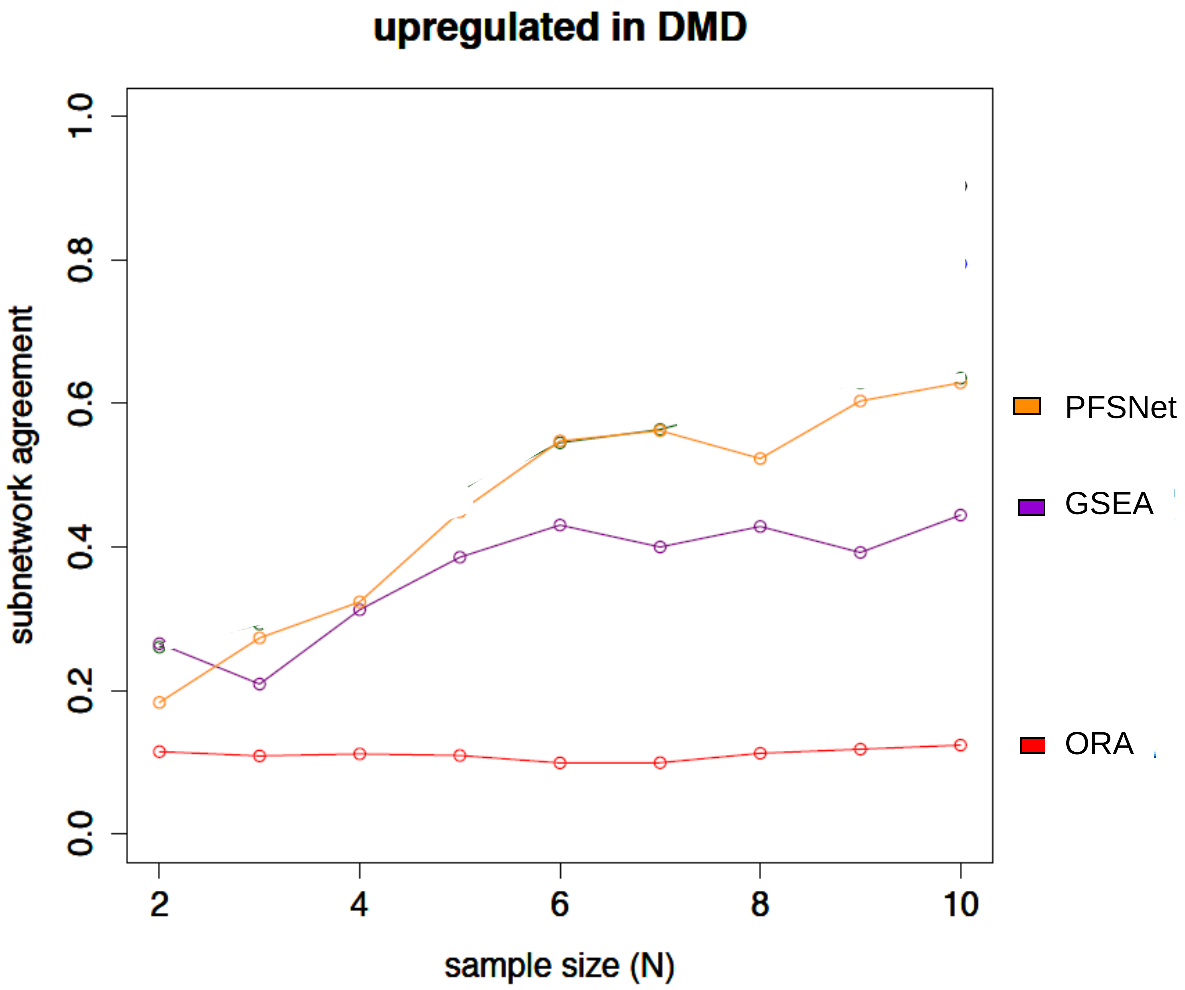
- **For an irrelevant subnetwork S, the two scores above for each patient P_k should be roughly equal, regardless of his class**

- **Do a paired t-test to decide whether S is relevant**
  - Get null distribution by permuting class labels

- β weights become unstable
- Cannot generate null distribution

- **All 3 issues solved, but not when sample size is small**

Much better performance, but still not great



THE QUANTUM LEAP

EVEN WHEN SAMPLE SIZE IS EXTREMELY SMALL

ORA-Paired: Paired test and new null hypothesis

- Let g_i be genes in a given pathway P
- Let p_j be a patient
- Let q_k be a normal

- Let $\Delta_{i,j,k} = \text{Expr}(g_i,p_j) – \text{Expr}(g_i,q_k)$

- Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

- **Issue #1 is solved**
  - Null hypothesis is “Pathway P is irrelevant to the difference between patients and normals, and the genes in P behave similarly in patients and normals”
- **Issue #2 is solved**
  - No longer need a pre-determined list of DE genes
- **Issue #3 is unsolved**
- **Is sample size now larger?**
  - |patients| * |normals| * |genes in P|

Testing the null hypothesis

“Pathway P is irrelevant to the difference between patients and normals and so, the genes in P behave similarly in patients and normals”

- **Method #1**
  - T-test w/ a conservative degree of freedom
    - **E.g., # normals + # patients**
- **Method #2**
  - By the null hypothesis, a dataset and any of its class-label permutations are exchangeable

  ⇒ Get null distribution by class-label permutations
    - **Only for large-size sample**
- **Method #3**
  - Modified null hypothesis
    - **“Pathway P induces gene-gene correlations, and genes in P behave according to these gene-gene correlations;**
    - **P is irrelevant to the diff betw patients and normals and so, genes in P behave similarly in patients and normals”**

  ⇒ Get null distribution using datasets that conserve gene-gene correlations in the original dataset
    - **E.g., array rotation**

Array rotation

- **QR decomposition**

$$\mathbf{X = X_Q \,.\, X_R}$$

Where

- – X is gene expression array of n samples * m genes
- – X_Q is n * r orientation matrix, r is rank of X
- – X_R is sufficient statistics of covariance between the m genes

- **Rotation**

$$\mathbf{X' = R_Q \,.\, X_Q .\, X_R}$$

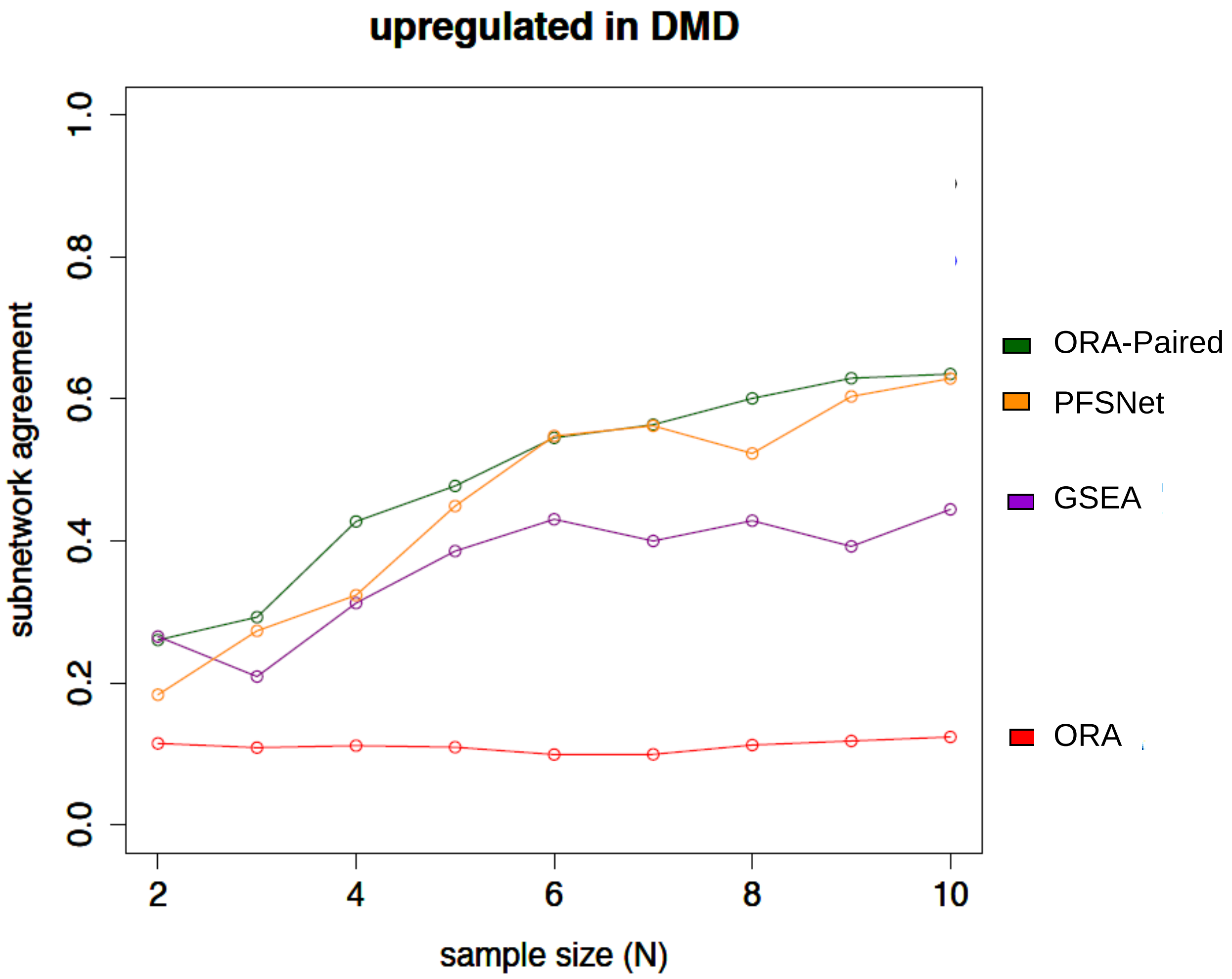
Where

- – R_Q is an n * n rotation operation

⇒ **X' is rotation of X preserving gene-gene correlations**

⇒ **i.e., preserving constraints induced by pathways**

Similar to PFSNet, good but not great

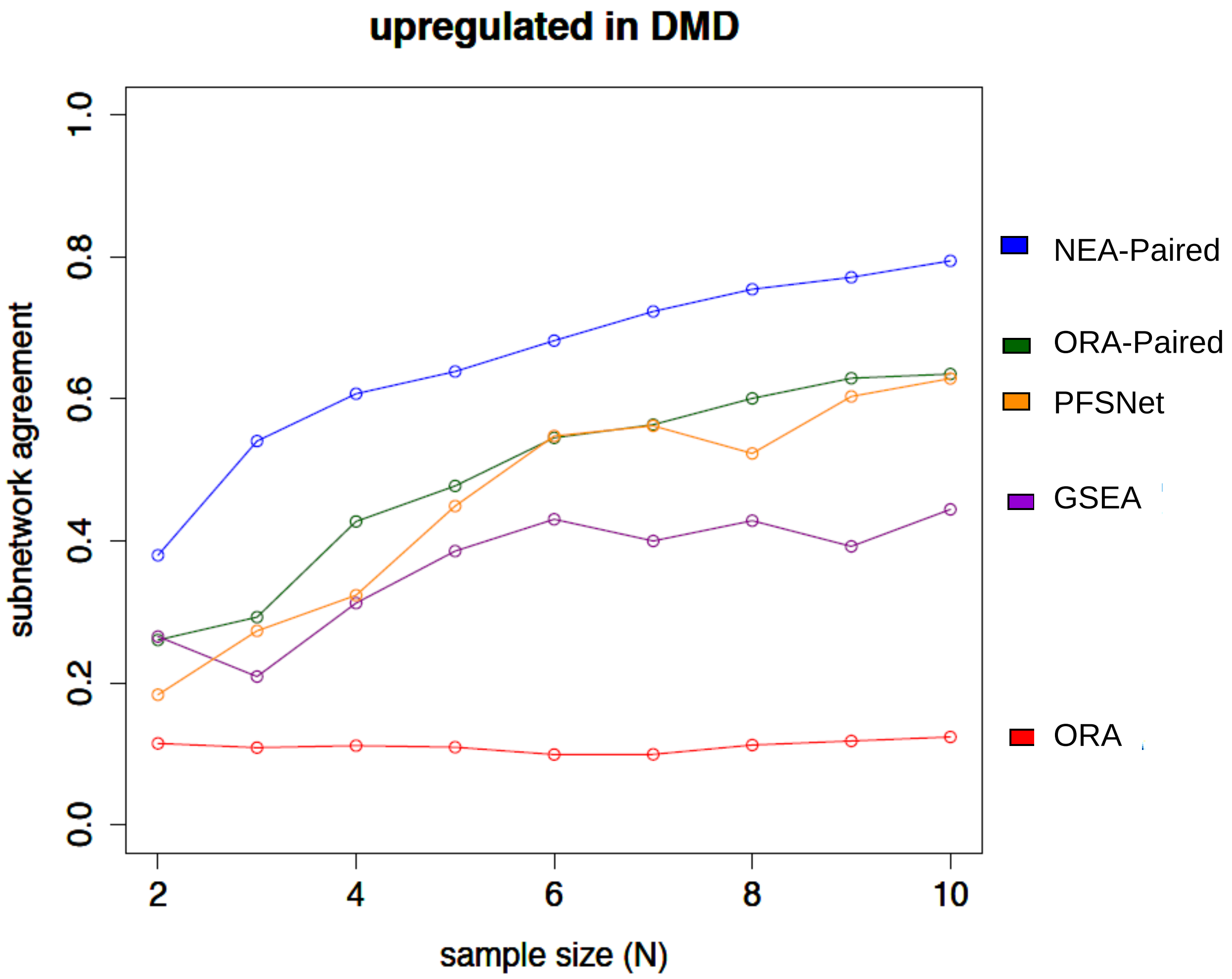


NEA-Paired: Paired test on subnetworks

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

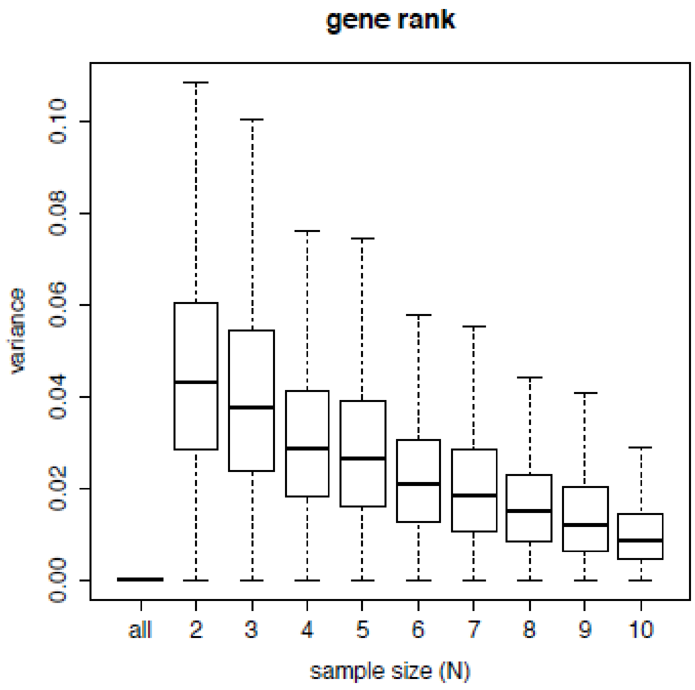
- **Issues #1 & #2 are solved as per ORA-Paired**
- **Issue #3 is partly solved**
  - Testing subnetworks instead of whole pathways
  - But subnetworks derived in a fragmented way

Much better performance



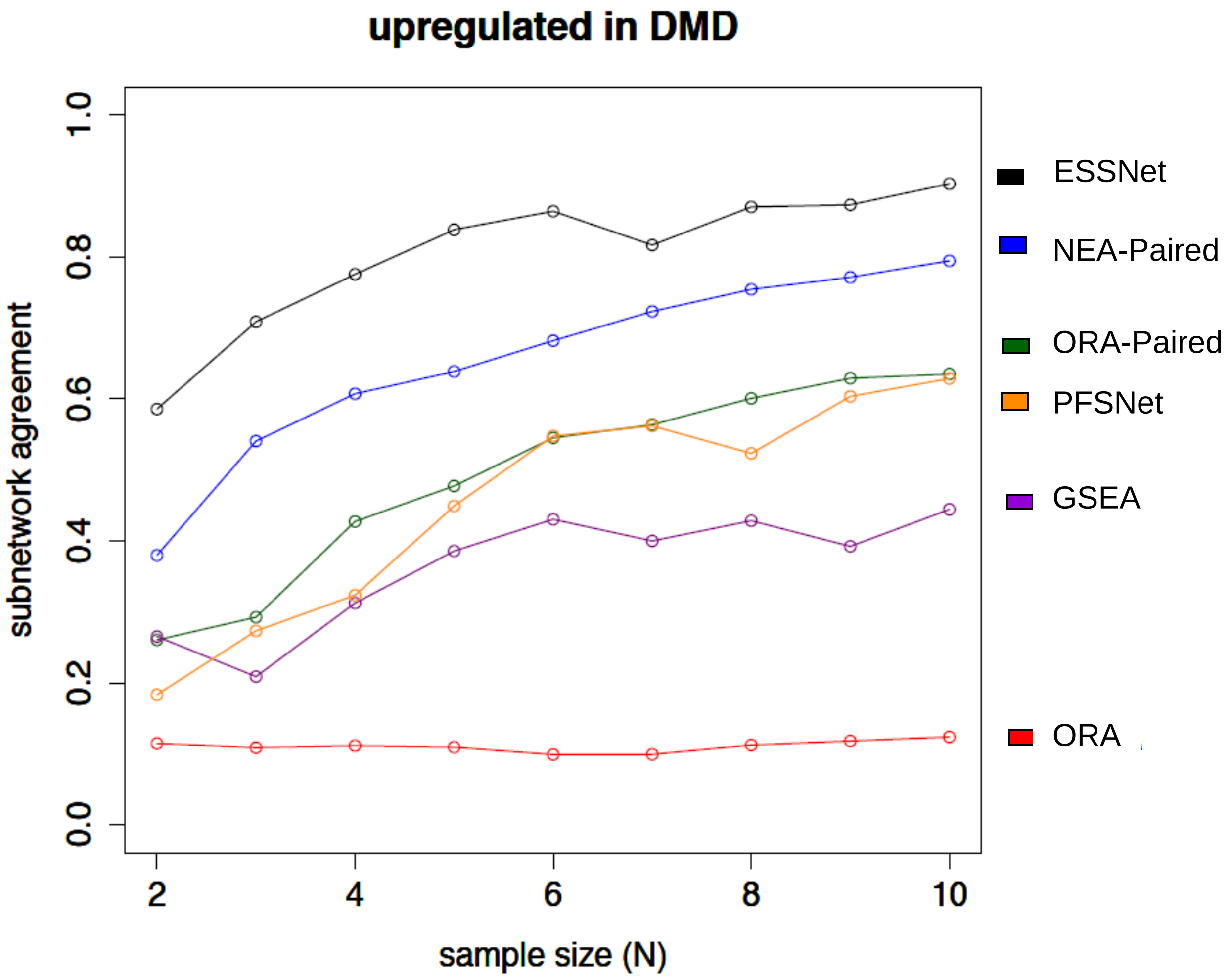
ESSNet: Larger subnetworks

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

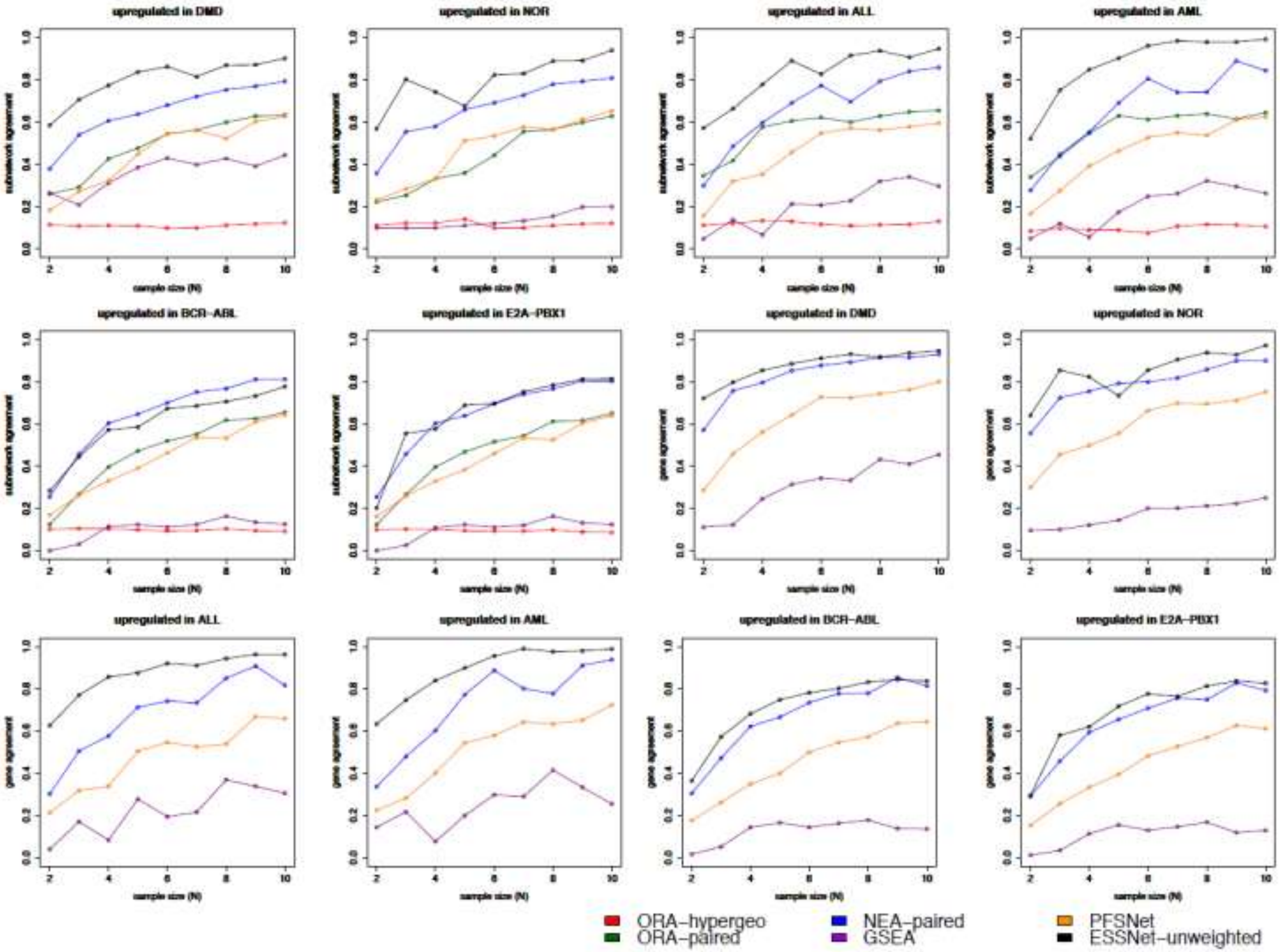


- Gene rank is very stable
- Issues #1 - #3 solved

Fantastic performance



More datasets tested



ESSNet is unlikely to report junk

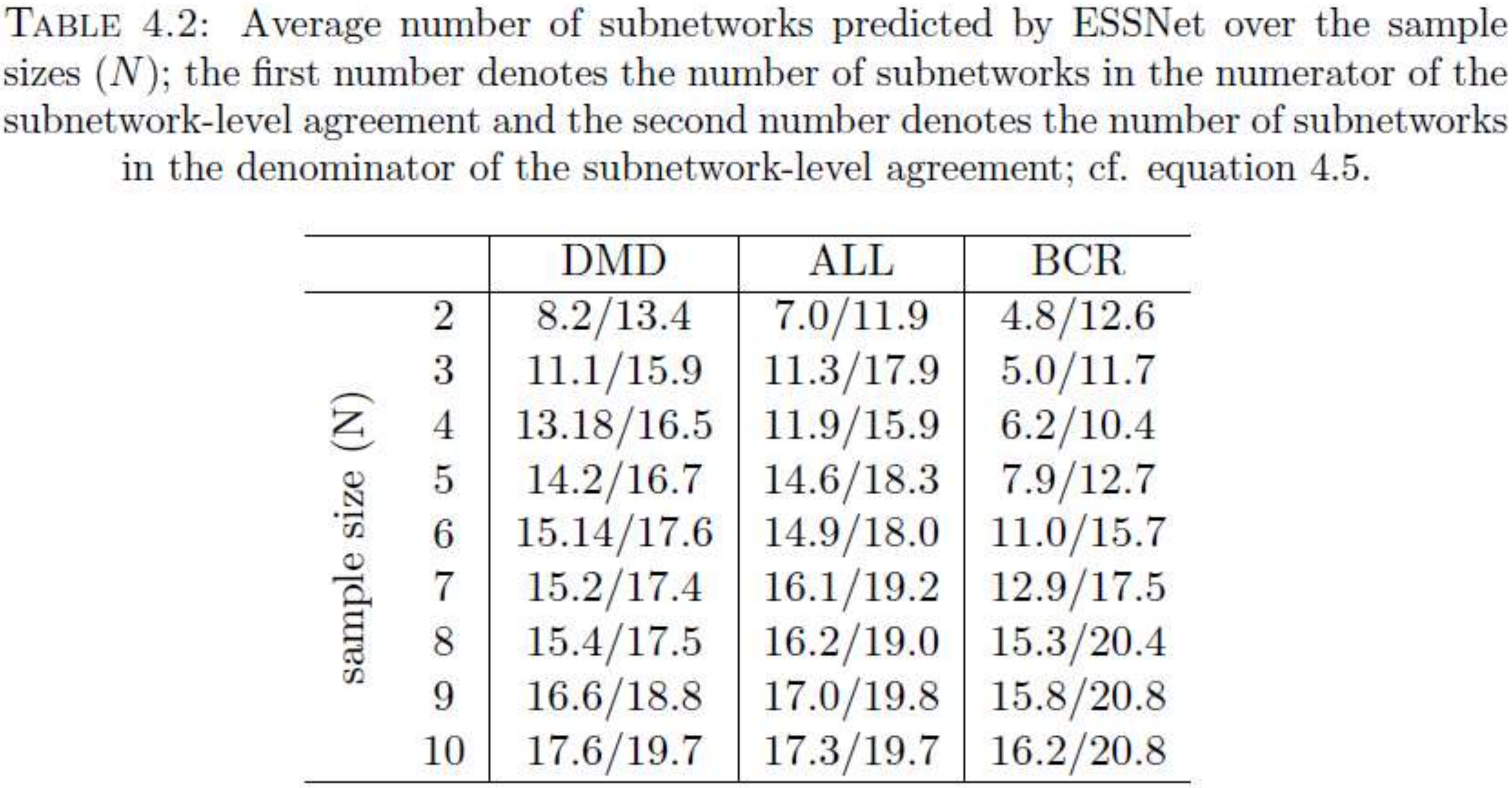
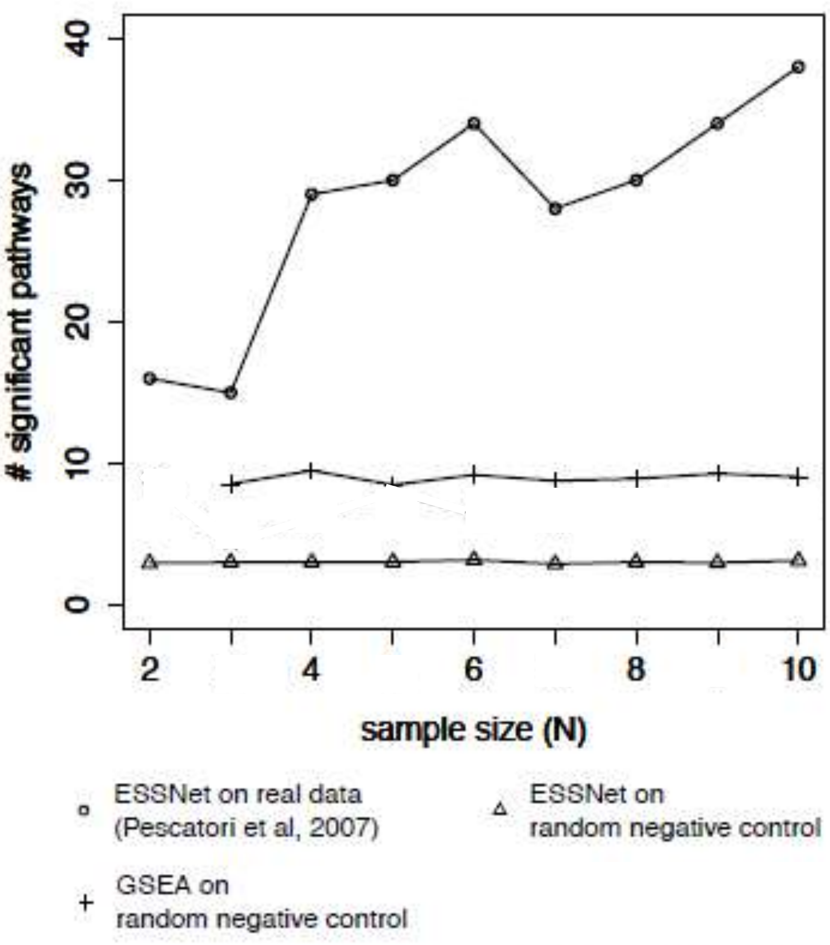


TABLE 4.2: Average number of subnetworks predicted by ESSNet over the sample sizes (N); the first number denotes the number of subnetworks in the numerator of the subnetwork-level agreement and the second number denotes the number of subnetworks in the denominator of the subnetwork-level agreement; cf. equation 4.5.

		DMD	ALL	BCR
sample size (N)	2	8.2/13.4	7.0/11.9	4.8/12.6
	3	11.1/15.9	11.3/17.9	5.0/11.7
	4	13.18/16.5	11.9/15.9	6.2/10.4
	5	14.2/16.7	14.6/18.3	7.9/12.7
	6	15.14/17.6	14.9/18.0	11.0/15.7
	7	15.2/17.4	16.1/19.2	12.9/17.5
	8	15.4/17.5	16.2/19.0	15.3/20.4
	9	16.6/18.8	17.0/19.8	15.8/20.8
	10	17.6/19.7	17.3/19.7	16.2/20.8

A negative-control experiment showing that ESSNet does not report junk



ESSNet also dominates when sample size is large

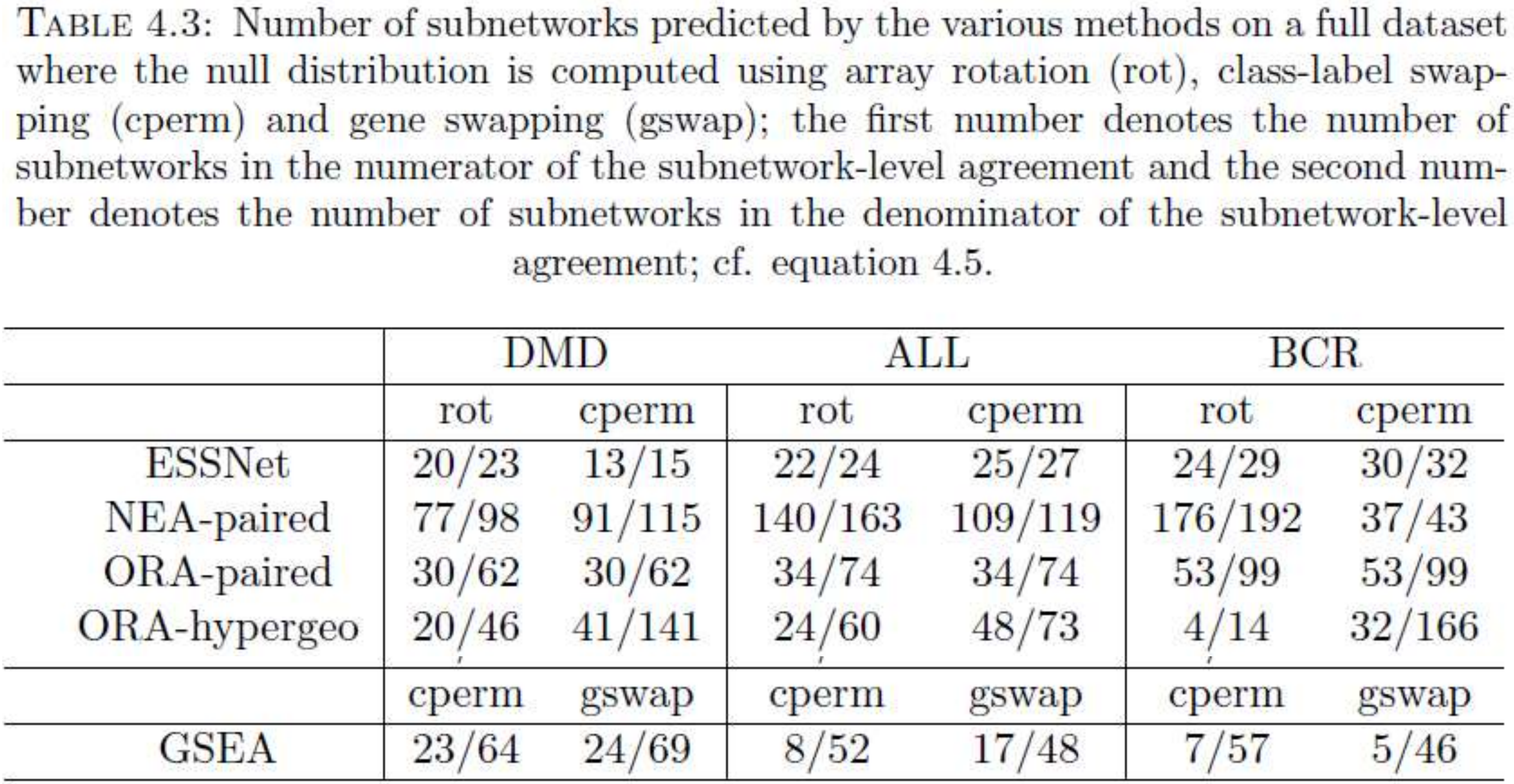
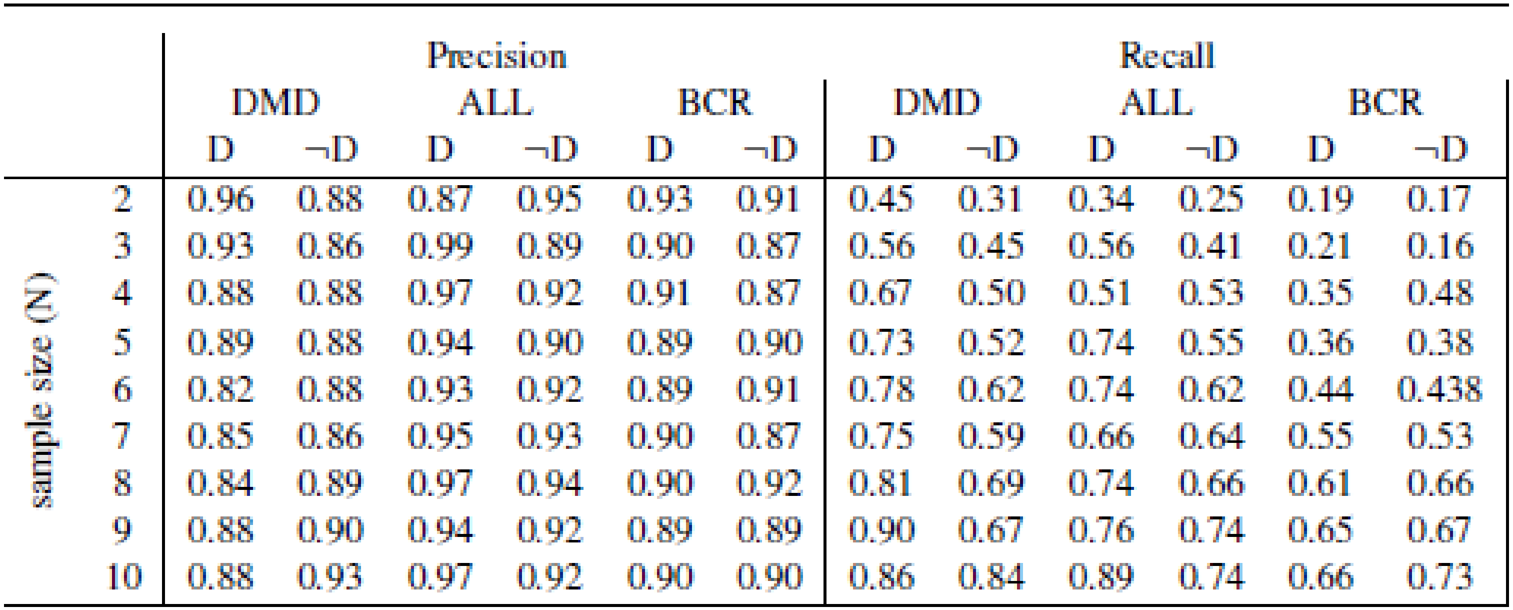


TABLE 4.3: Number of subnetworks predicted by the various methods on a full dataset where the null distribution is computed using array rotation (rot), class-label swapping (cperm) and gene swapping (gswap); the first number denotes the number of subnetworks in the numerator of the subnetwork-level agreement and the second number denotes the number of subnetworks in the denominator of the subnetwork-level agreement; cf. equation 4.5.

	DMD		ALL		BCR	
	rot	cperm	rot	cperm	rot	cperm
ESSNet	20/23	13/15	22/24	25/27	24/29	30/32
NEA-paired	77/98	91/115	140/163	109/119	176/192	37/43
ORA-paired	30/62	30/62	34/74	34/74	53/99	53/99
ORA-hypergeo	20/46	41/141	24/60	48/73	4/14	32/166
	cperm	gswap	cperm	gswap	cperm	gswap
GSEA	23/64	24/69	8/52	17/48	7/57	5/46

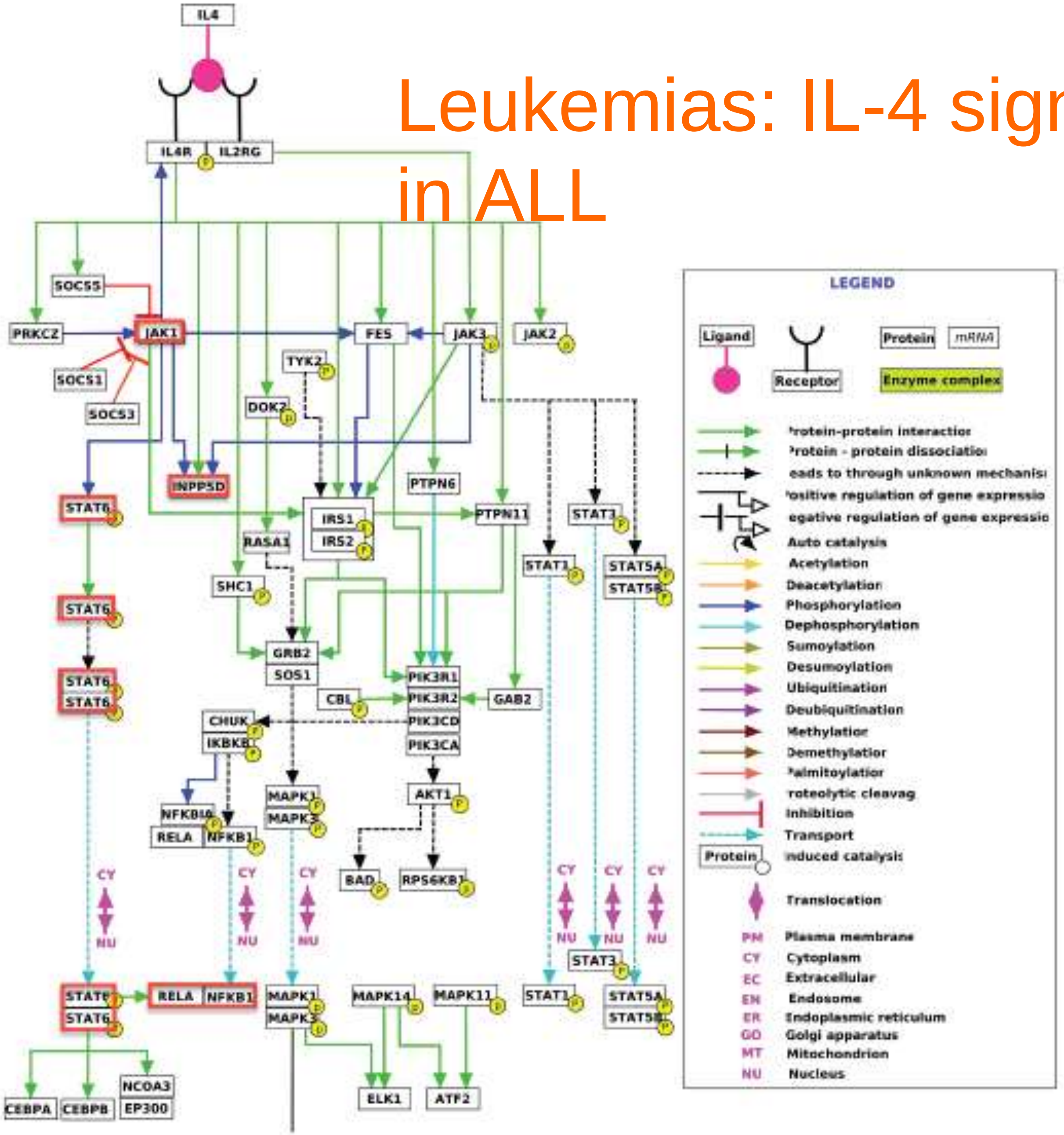
Do ESSNet results agree on small datasets vs big datasets?



		Precision						Recall					
		DMD		ALL		BCR		DMD		ALL		BCR	
		D	¬D	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D
sample size (N)	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

- **Use 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.

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**

A caveat?

- **The complaint:**
  - Genes and subnets of lower expression levels are ignored by ESSNet

- **A caution against the complaint:**
  - These genes have higher variance

  ⇒ More false positives

- **If you really wish to insist on this:**
  - Consider also significant lower-expression subnets of NEA-paired
    - **NEA-paired uses the same scoring method as ESSNet, but scores every node & their immediate neighbourhood**

DIFFICULTY OF CROSS-BATCH CLASSIFICATION

Batch effects

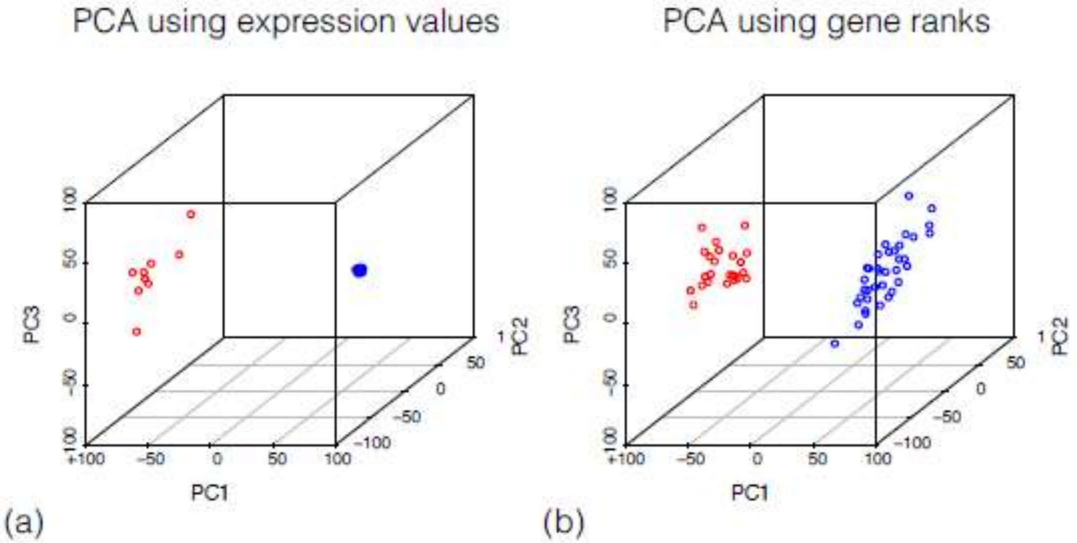


FIGURE 5.1: Batch effects in the DMD/NOR datasets, the blue and red color denote different data batches. (a) Scatterplot on the first 3 components using gene-expression values. (b) Scatterplot on the first 3 components using gene ranks.

- **Batch effects are common**
- **Batch effects cannot always be removed using common normalization methods**

Gene-feature-based classifiers do badly when there are batch effects, even after normalization

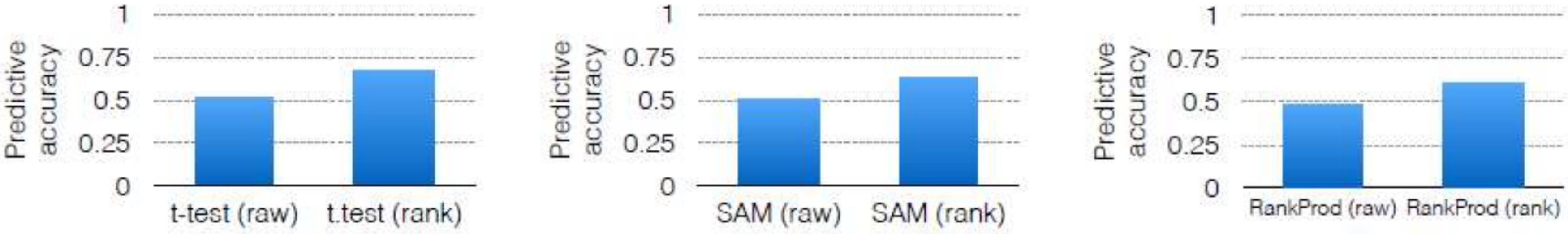
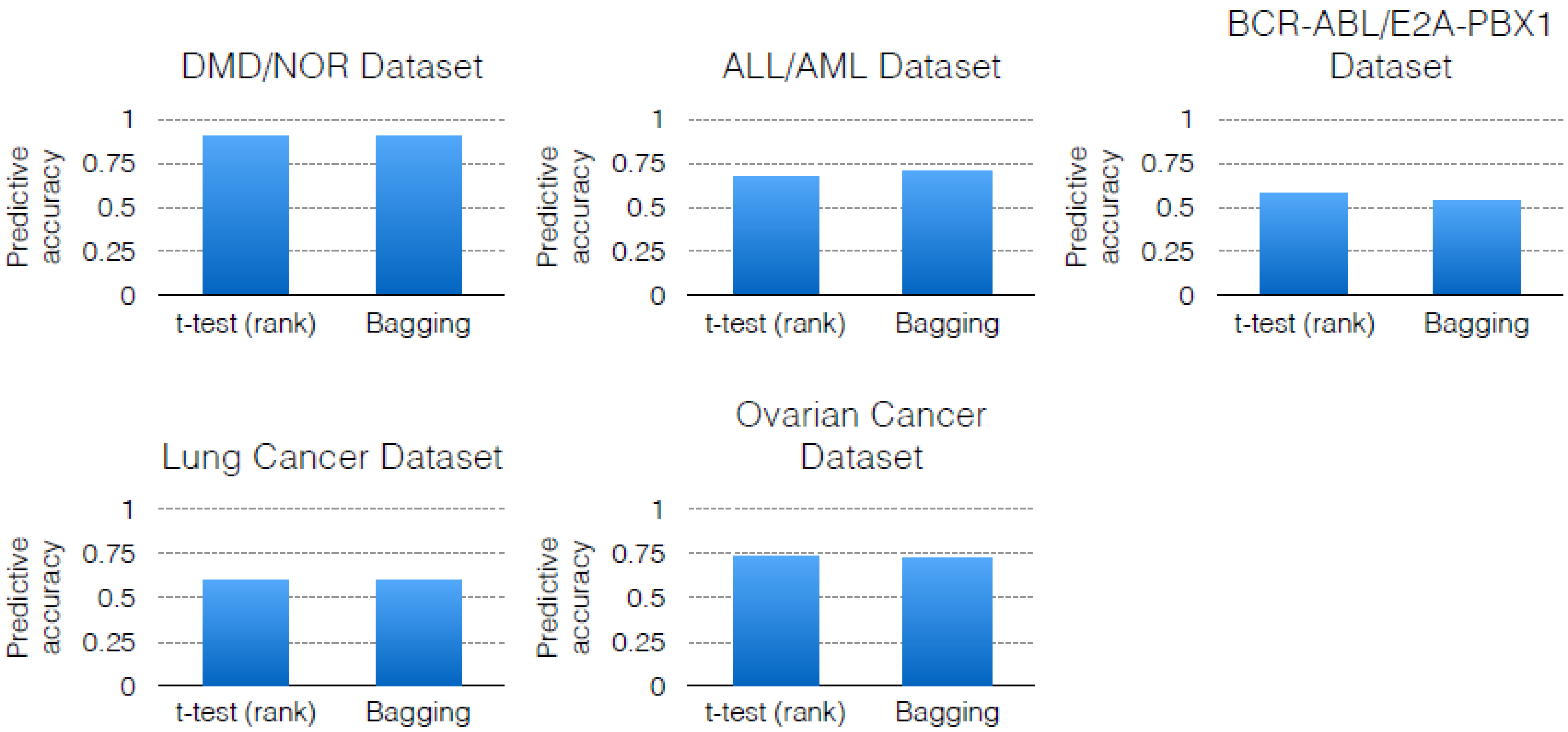


FIGURE 5.8: Predictive accuracy of gene-feature-based classifiers with and without rank normalization in the DMD/NOR dataset.

Gene selection by t-test, SAM, or rank product. Classifier by naïve Bayes

Ensemble classifiers can't always improve results of gene-feature-based classifiers with normalization


DMD/NOR Dataset

ALL/AML Dataset

BCR-ABL/E2A-PBX1 Dataset

Lung Cancer Dataset

Ovarian Cancer Dataset

Genes from subnetworks produced by ESSNet can't help gene-feature-based classifiers

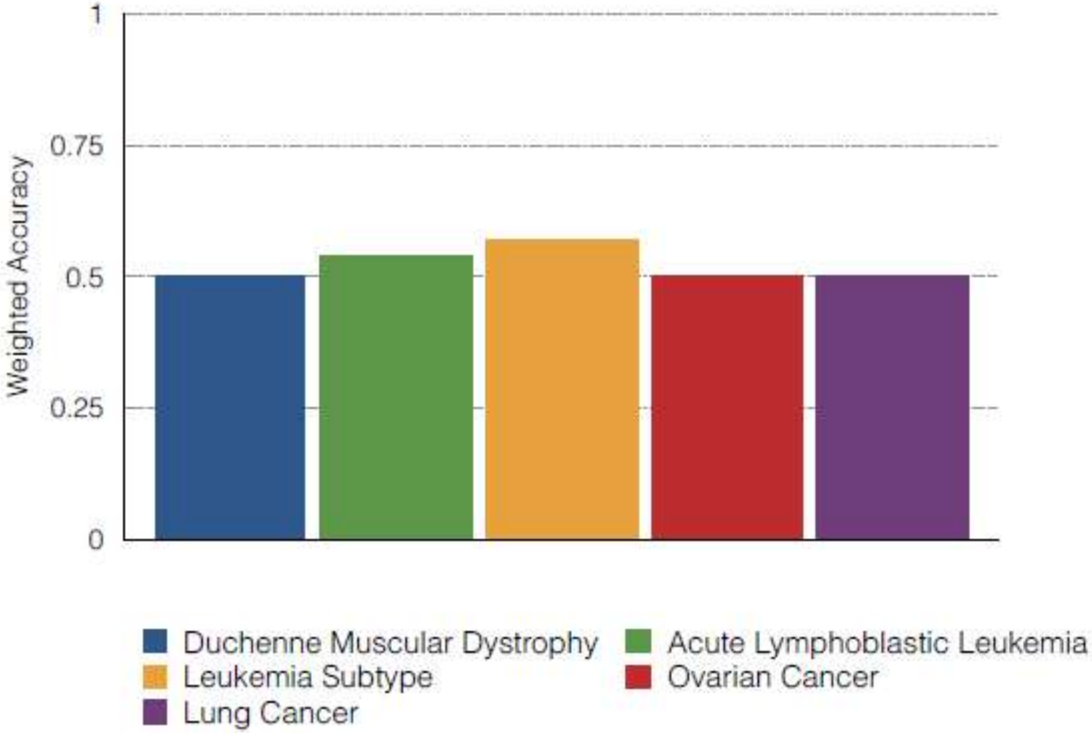


FIGURE 5.15: Predictive accuracy of gene-feature-based classifier using genes extracted from subnetworks in ESSNet; demonstrating that genes in the subnetworks by themselves are not a good discriminator for classification

SUCCESSFUL CROSS-BATCH CLASSIFICATION

WHEN SAMPLE SIZE IS LARGE

PFSNet-based features

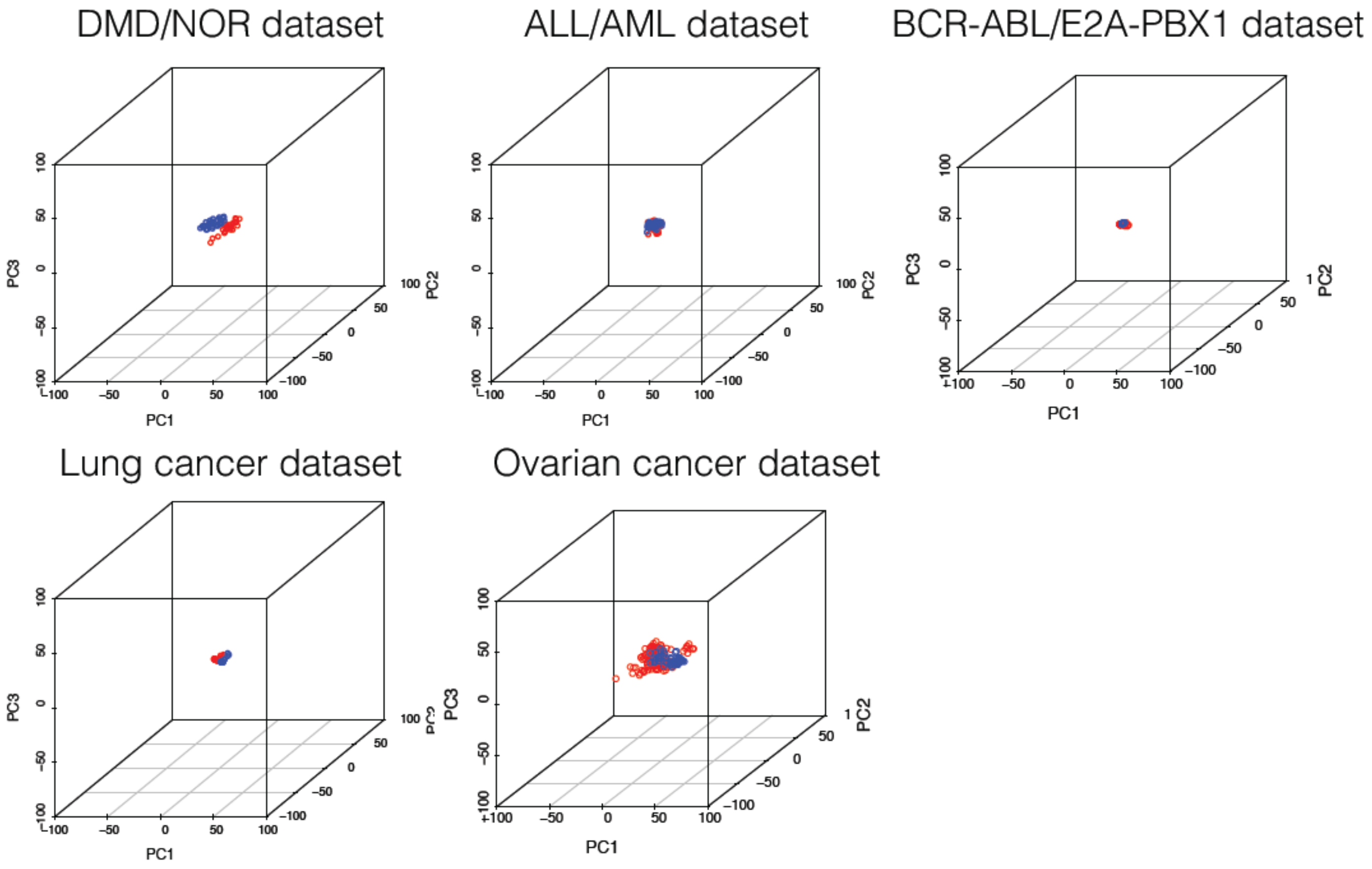
- **PFSNet**
  - Induce subnetworks from pathways by considering only genes highly expressed in majority of patients in any class
  - For each subnetwork S and each patient P$_k$, compute a pair of scores:

$$\beta_1^*(g_i) = \sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|} \qquad \beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{fs(e_{g_i,p_j})}{|\neg D|}$$

$$Score_1^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_1^*(g_i) \quad Score_2^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_2^*(g_i)$$

- **Straightforward to use these scores as features**

Successfully reducing batch effects


DMD/NOR dataset

ALL/AML dataset

BCR-ABL/E2A-PBX1 dataset

Lung cancer dataset

Ovarian cancer dataset

FIGURE 5.6: A figure showing that the batch effects are reduced by PFSNet subnetwork features. The colors red and blue represent different batches.

Successful cross-batch classification

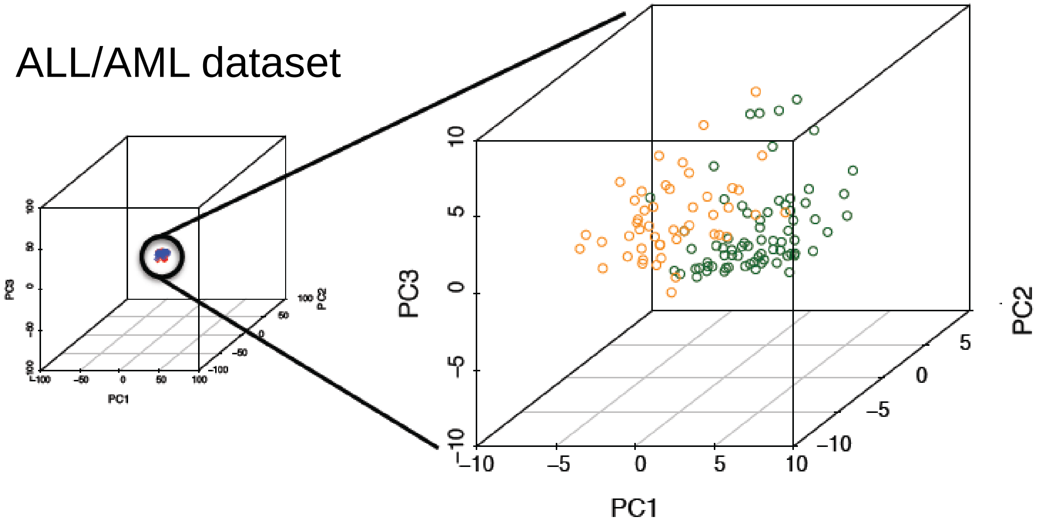

ALL/AML dataset

FIGURE 5.7: A figure showing that data points are separated by class labels instead of batch when PFSNet features are used. The colors green and orange represent different classes.

SUCCESSFUL CROSS-BATCH CLASSIFICATION

EVEN WHEN SAMPLE SIZE IS SMALL

ESSNet

- Induce subnetworks using genes highly expressed in majority of samples in any class

- Let g_i be genes in a given subnetwork S
- Let p_j be patients
- Let q_k be normals

- Let $\Delta_{i,j,k} = \text{Expr}(g_i,p_j) - \text{Expr}(g_i,q_k)$

- Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

ESSNet scores subnetworks but not patients.

How to produce feature vectors for patients?

ESSNet-based features

- **The idea is to see whether the pairwise differences of genes with a subnetwork betw a given sample p_x and the two separate classes (D and ¬D) have a distribution around 0**

$$\Delta_{(D)}(S, p_x) = \{e_{g_i,p_x} - e_{g_i,p'} \mid g_i \in S \text{ and } p' \in D\}$$

$$\Delta_{(\neg D)}(S, p_x) = \{e_{g_i,p_x} - e_{g_i,p'} \mid g_i \in S \text{ and } p' \in \neg D\}$$

- **We expect Δ(D)(S,P_x) and Δ(¬D)(S,P_x) to have +ve or –ve median for patients in one of the classes iff subnetwork S is useful for classification**
  - The median and ±2 std dev of Δ(D)(S,P_x) and Δ(¬D)(S,P_x) give 6 features for P_x

ESSNet-based features

- **We also obtain pairwise differences of genes within a subnetwork among all possible pairs of patients in D and ¬D**

$$\Delta_{(D-\neg D)}(S) = \{e_{g_i,p'} - e_{g_i,p''} \mid g_i \in S \text{ and } p' \in D \text{ and } p'' \in \neg D\}$$

Similarly for $\Delta_{(\neg D - \neg D)}$(S), $\Delta_{(\neg D - D)}$(S), $\Delta_{(D - D)}$(S)

- **This gives 4 more features**

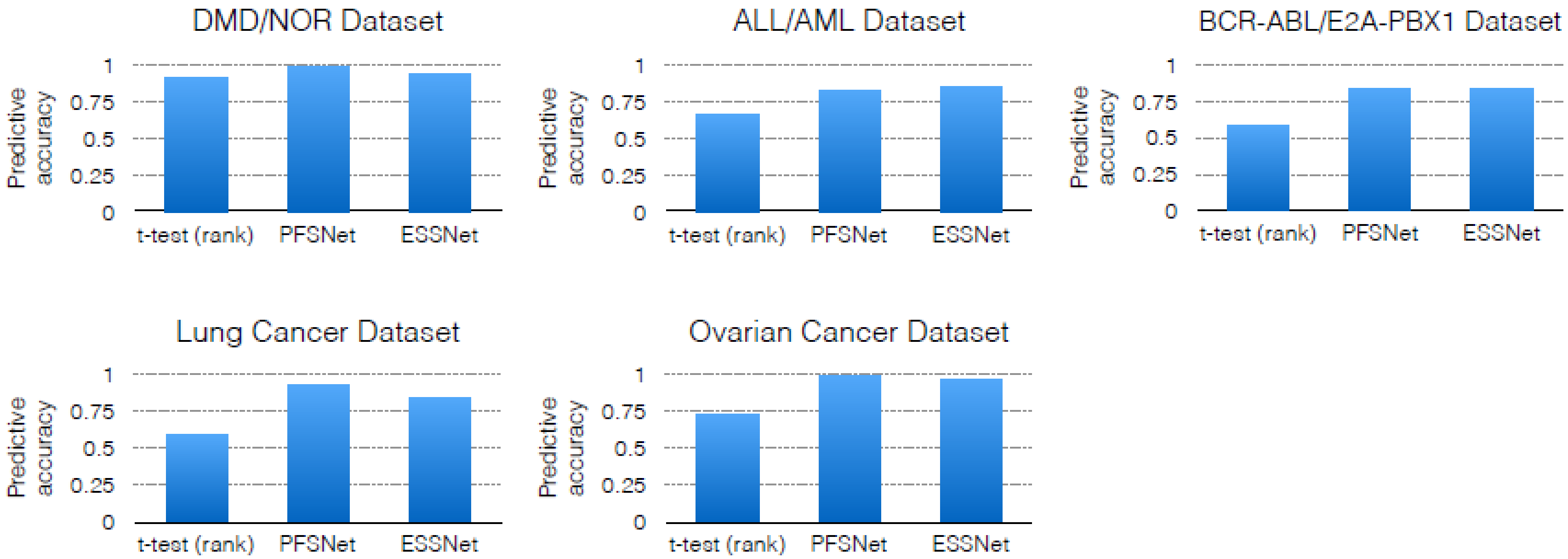
$$ESSNet\_feature_7^{p_x,S} = T\_statistic(\Delta_{(\neg D)}(S,p_x), \Delta_{(D-\neg D)}(S))$$

$$ESSNet\_feature_8^{p_x,S} = T\_statistic(\Delta_{(\neg D)}(S,p_x), \Delta_{(\neg D-\neg D)}(S))$$

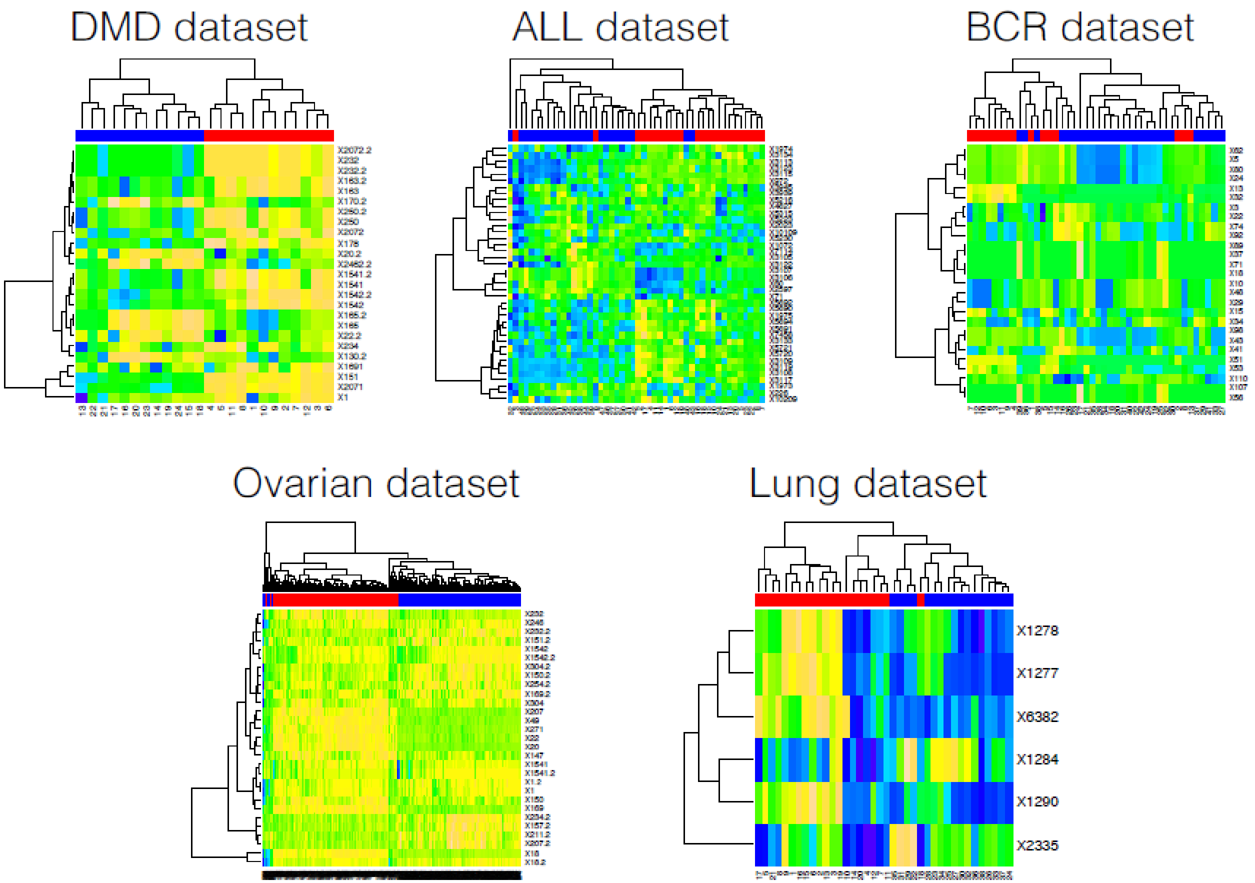
$$ESSNet\_feature_9^{p_x,S} = T\_statistic(\Delta_{(D)}(S,p_x), \Delta_{(D-D)}(S))$$

$$ESSNet\_feature_{10}^{p_x,S} = T\_statistic(\Delta_{(D)}(S,p_x), \Delta_{(\neg D-D)}(S))$$

ESSNet-based features lead to high cross-batch classification accuracy



ESSNet-based cross-batch hierachical clusterings



DMD dataset

ALL dataset

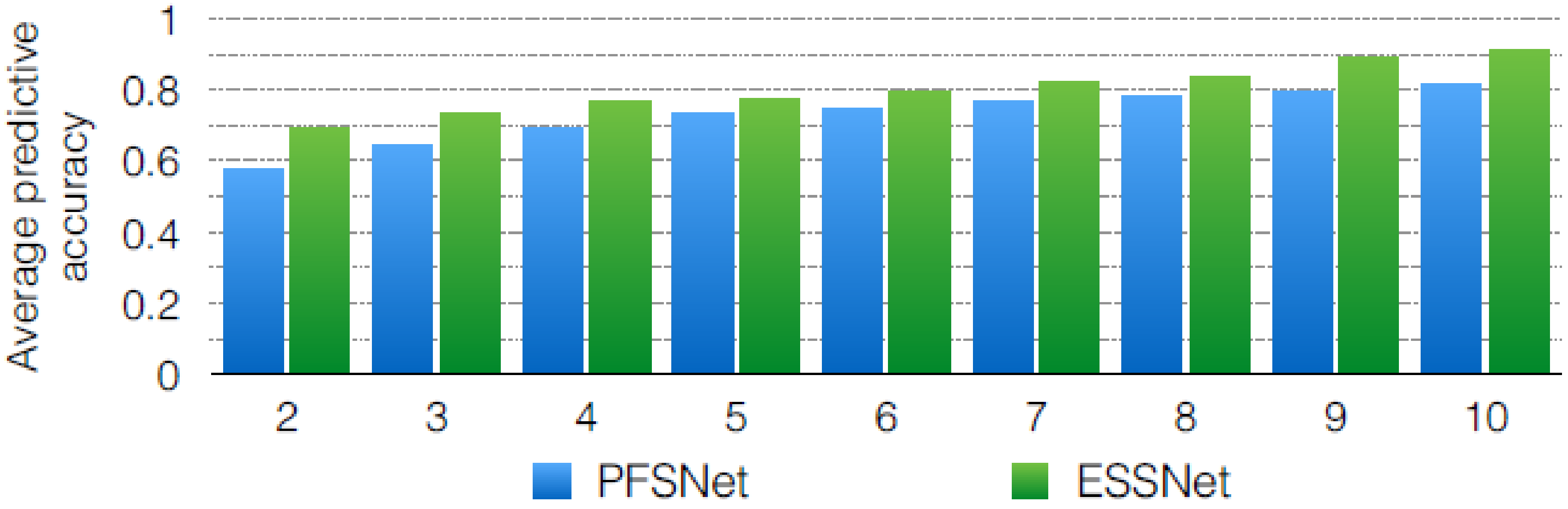
BCR dataset

Ovarian dataset

Lung dataset

FIGURE 5.17: A figure depicting heirarchical clustering performed on the patient's subnetwork scores.

ESSNet-based features retain high cross-batch classification accuracy even when training-sample size is small



Remarks

- **Traditional methods of classifying gene expression profiles often have difficulty predicting outcome of new batches of patients**
  - Normalization does not always help

- **ESSNet-based features are much more robust even when training-sample size is small**
  - Subnetworks found by ESSNet are reproducible and gave high cross-batch classification accuracy

⇒ **ESSNet is successful in isolating disease-relevant subnetworks from pathways**

Acknowledgements

- **My students**
  - Donny Soh
  - Dong Difeng
  - Kevin Lim
- **& collaborator**
  - Choi Kwok Pui
  - Li Zhenhua
- **Singapore Ministry of Education**

- 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**. Submitted
- Kevin Lim. **Using biological networks and gene-expression profiles for the analysis of diseases**. *PhD dissertation*, NUS, November 2014