CS4220: Knowledge Discovery Methods for Bioinformatics

Unit 4: Batch Effects

Wong Limsoon

Plan

- Batch effects
- Visualization
- Normalization
- PC1 removal
- Batch effect-resistant feature selection
- Batch effect-resistant classifiers

What are batch effects?

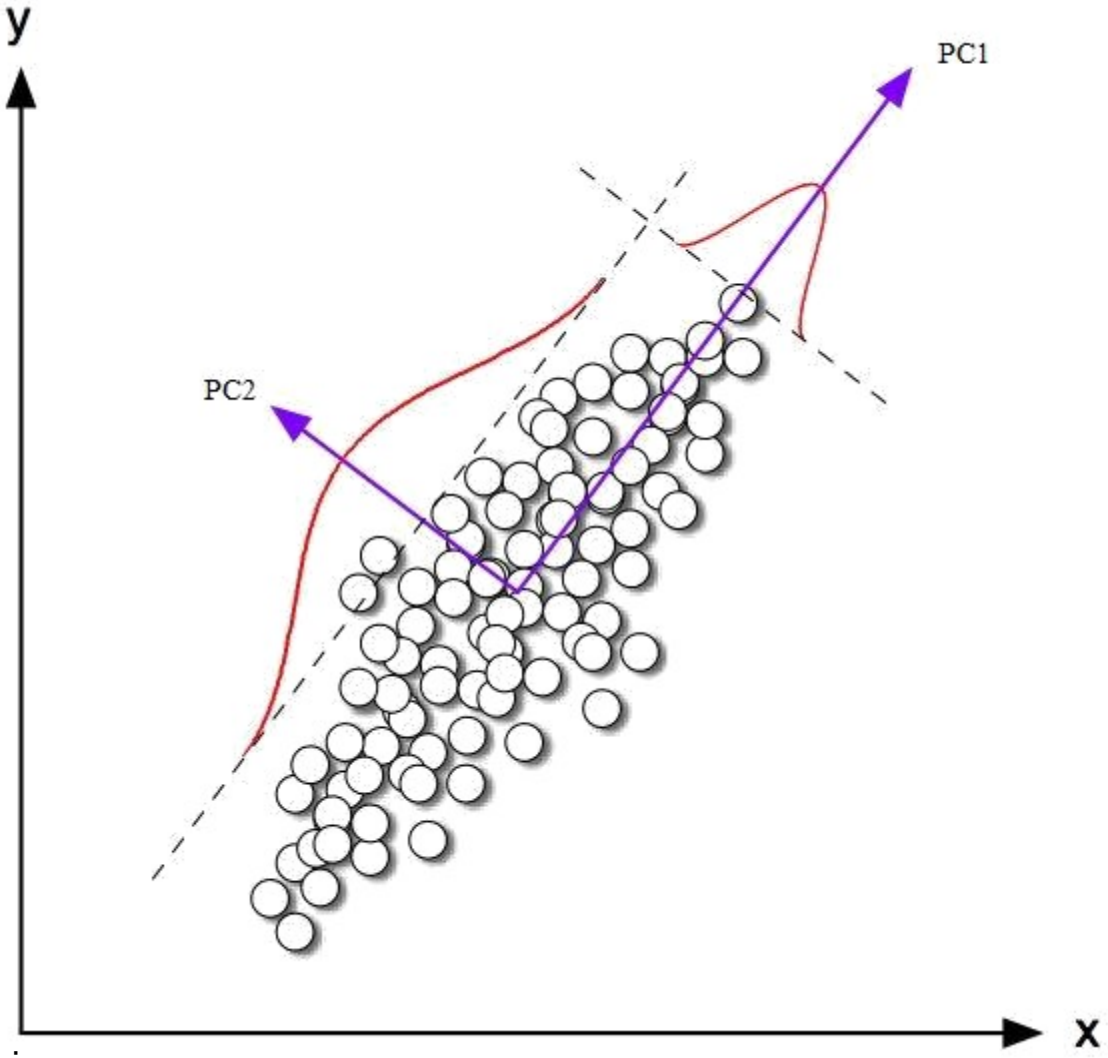
- **Batch effects are unwanted sources of variation caused by different processing date, handling personnel, reagent lots, equipment/machines, etc.**

- **Batch effects is a big challenge faced in biological research, especially towards translational research and precision medicine**

NUS National University of Singapore

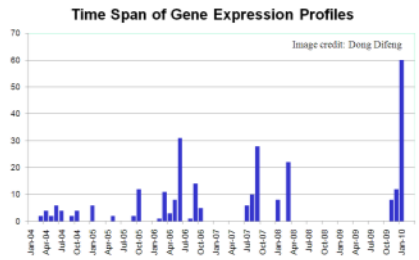
VISUALIZATION

Principal component analysis

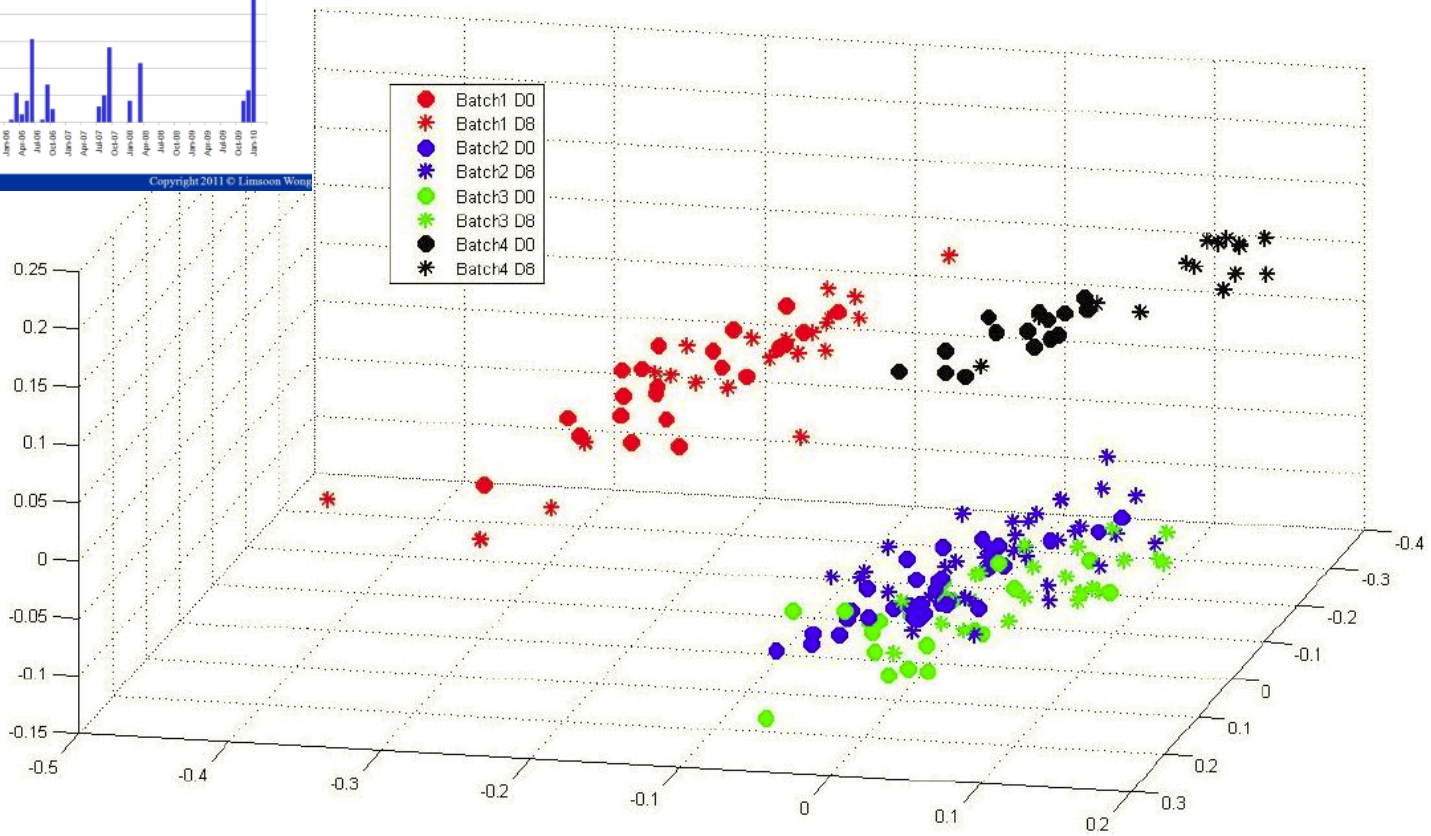


Credit: Alessandro Giuliani

Sometimes, a gene expression study may involve batches of data collected over a long period of time…



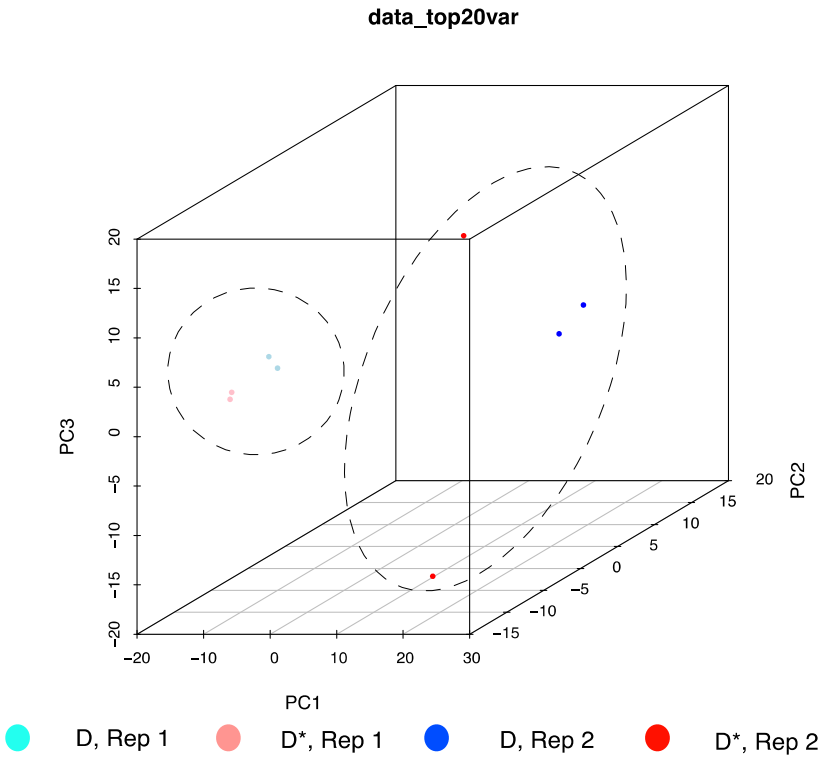
PCA scatter plot



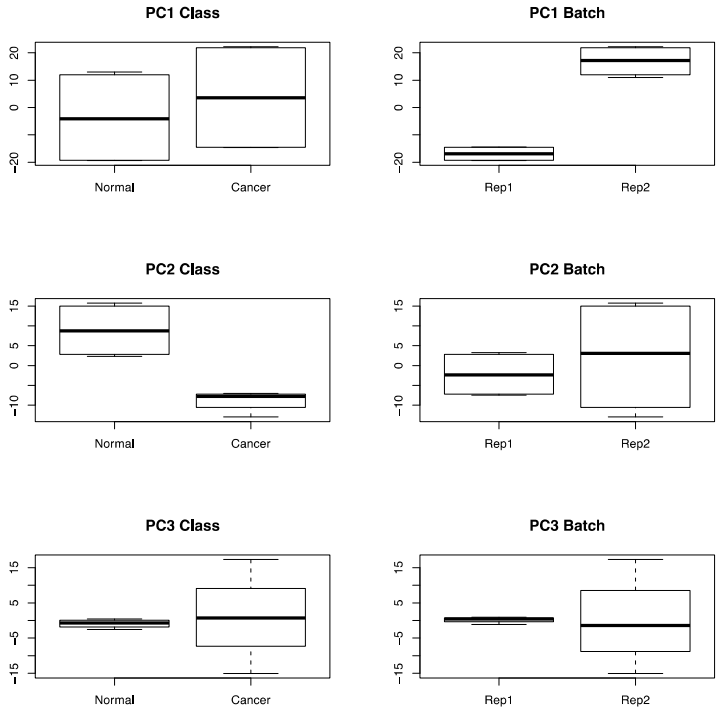
- **Samples from diff batches are grouped together, regardless of subtypes and treatment response**

Image credit: Difeng Dong's PhD dissertation, 2011

Paired boxplots of PCs



Sometime it is not easy to decide which PC is enriched in batch effects using the standard PCA scatter plot



It is easier to see which PC is enriched in batch effects by showing, side by side, the distribution of values of each PC stratified by class and suspected batch variables

NUS National University of Singapore

NORMALIZATION

Approaches to normalization

- **Aim of normalization:**
  Reduce variance w/o increasing bias

- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

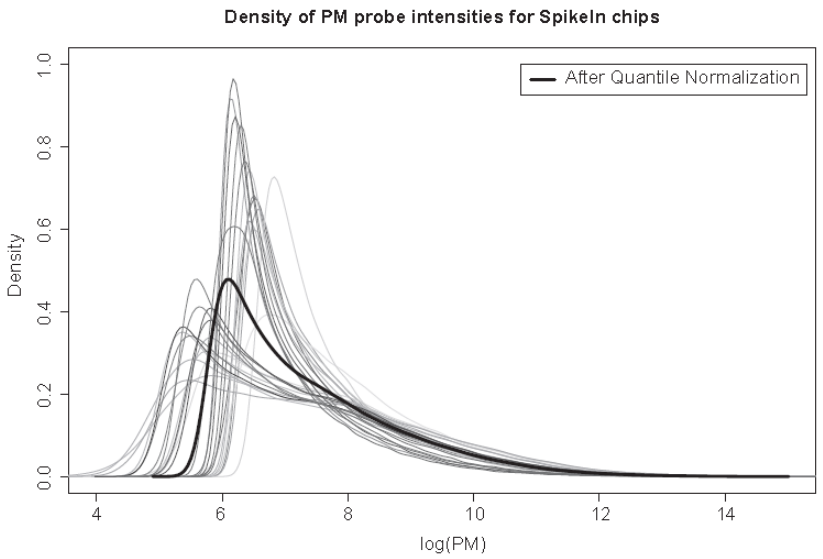
- **Transform data so that distribution of probe intensities is same on all arrays**
  - E.g., $(x - \mu) / \sigma$

- **Quantile normalization**

- **Gene fuzzy score, GFS**

Quantile normalization

- Given *n arrays of length p, form X of size p × n where each array is a column*
- Sort each column of *X to give* X_{sort}
- Take means across rows of X_{sort} *and assign this* mean to each elem in the row to get X'_{sort}
- Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as *X*



- Implemented in some microarray s/w, e.g., EXPANDER

48

In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

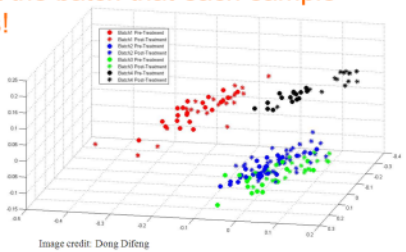


Image credit: Dong Difeng

⇒ Need normalization to correct for batch effect

After quantile normalization

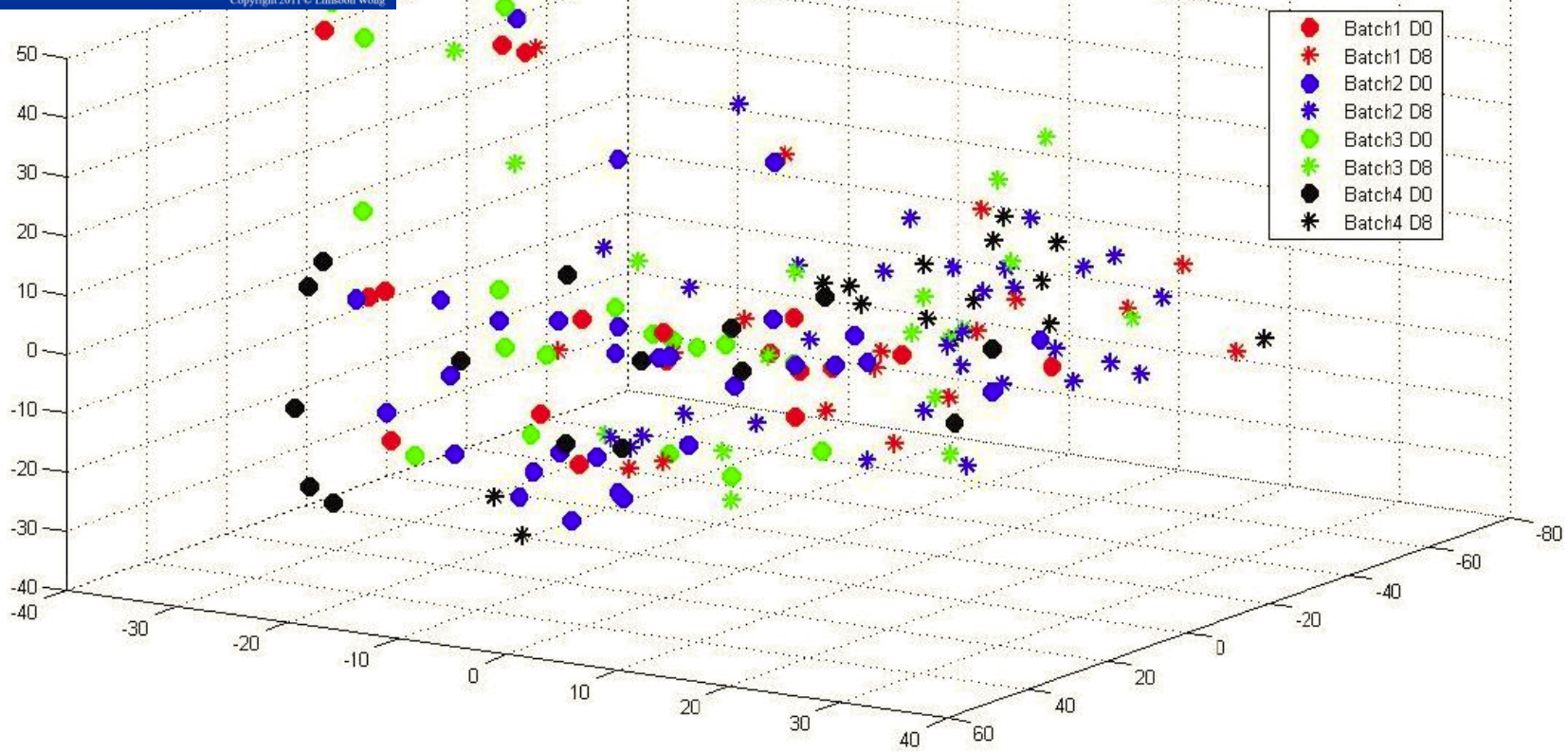
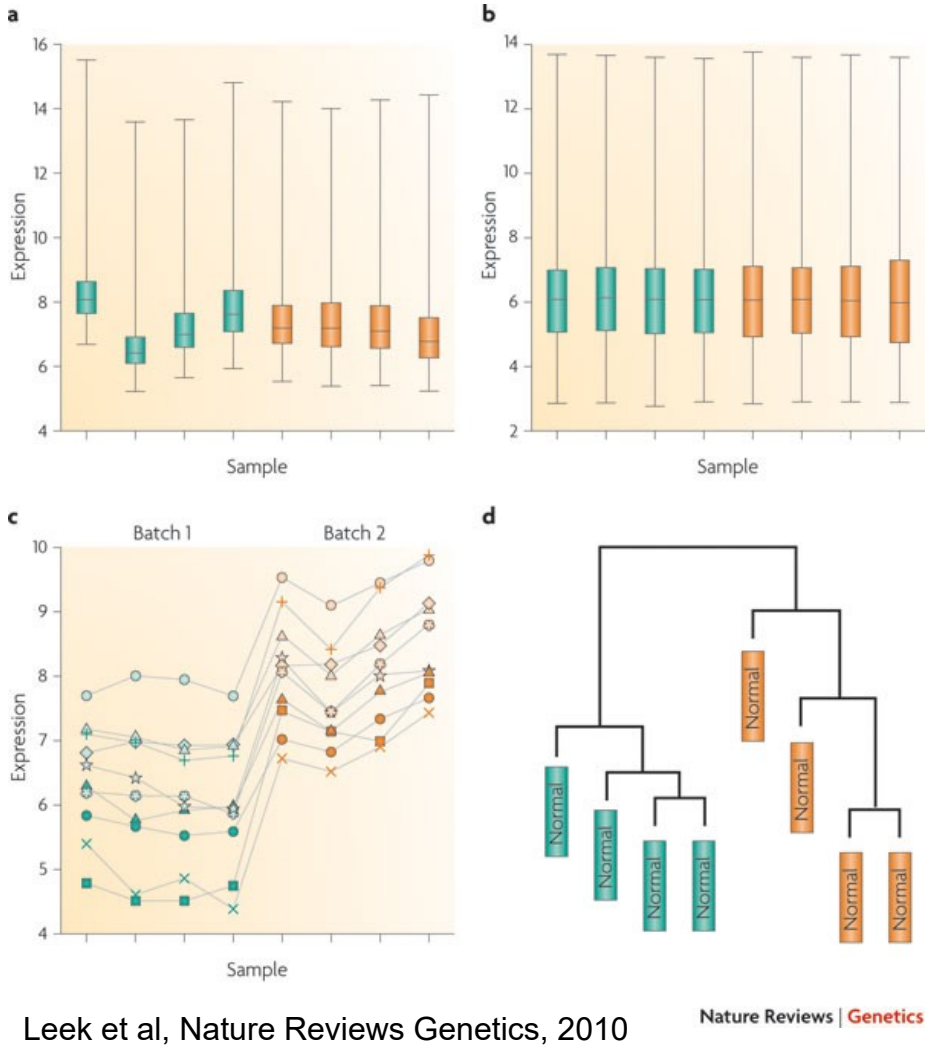


Image credit: Difeng Dong's PhD dissertation, 2011

Caution: It is difficult to eliminate batch effects effectively



Leek et al, Nature Reviews Genetics, 2010

Green and orange are normal samples differing in processing date

a: Before normalization

b: Post normalization

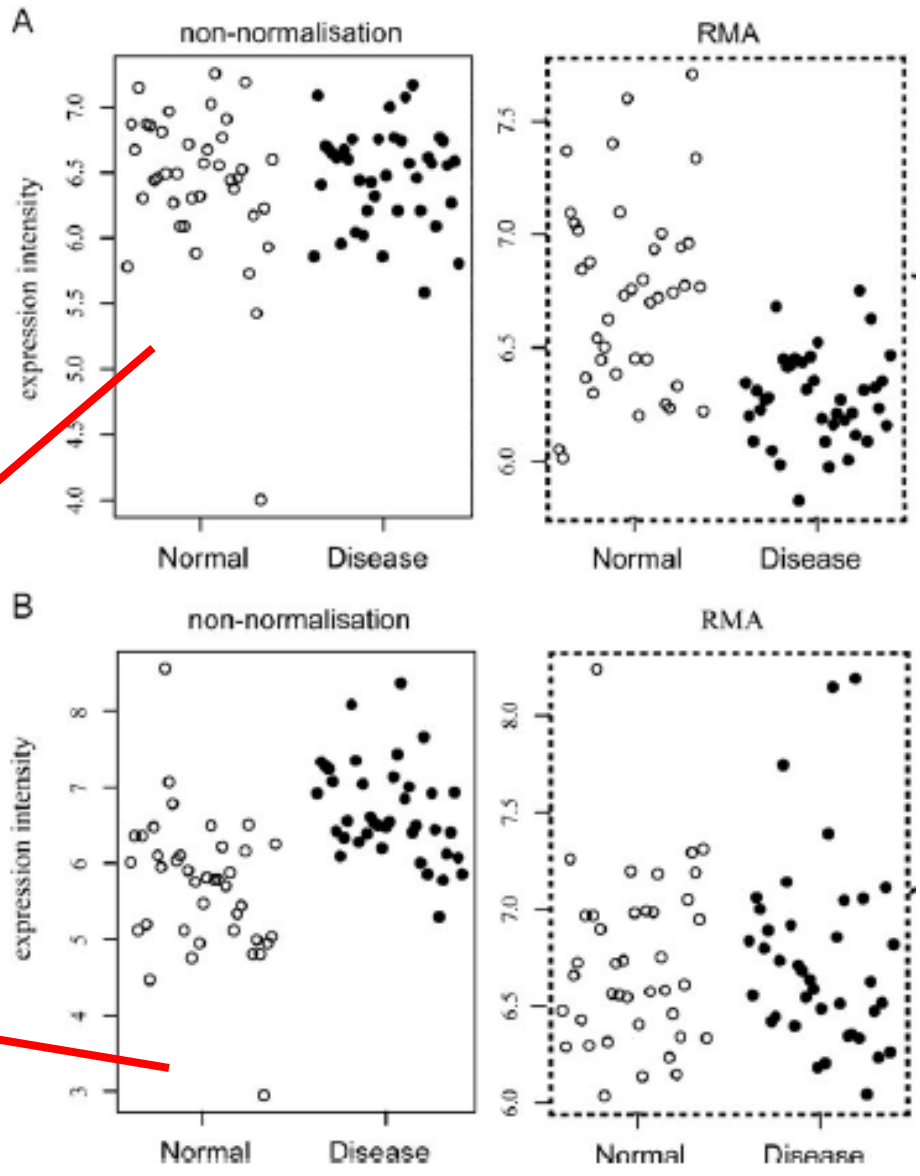
c: Checks on individual genes susceptible to batch effects

d: Clustering after normalization (samples still cluster by processing date)

Caution: "Over normalized" signals in cancer samples

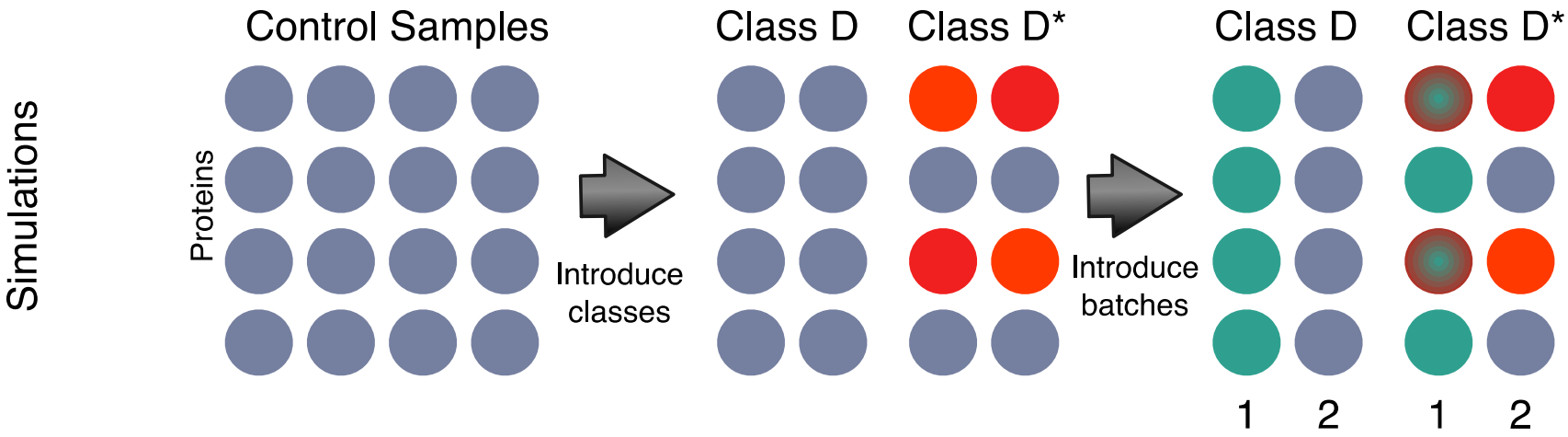
A gene normalized by quantile normalization (RMA) was detected as down-regulated DE gene, but the original probe intensities in cancer samples were not diff from those in normal samples

A gene was detected as an up-regulated DE gene in the non-normalized data, but was not identified as a DE gene in the quantile-normalized data



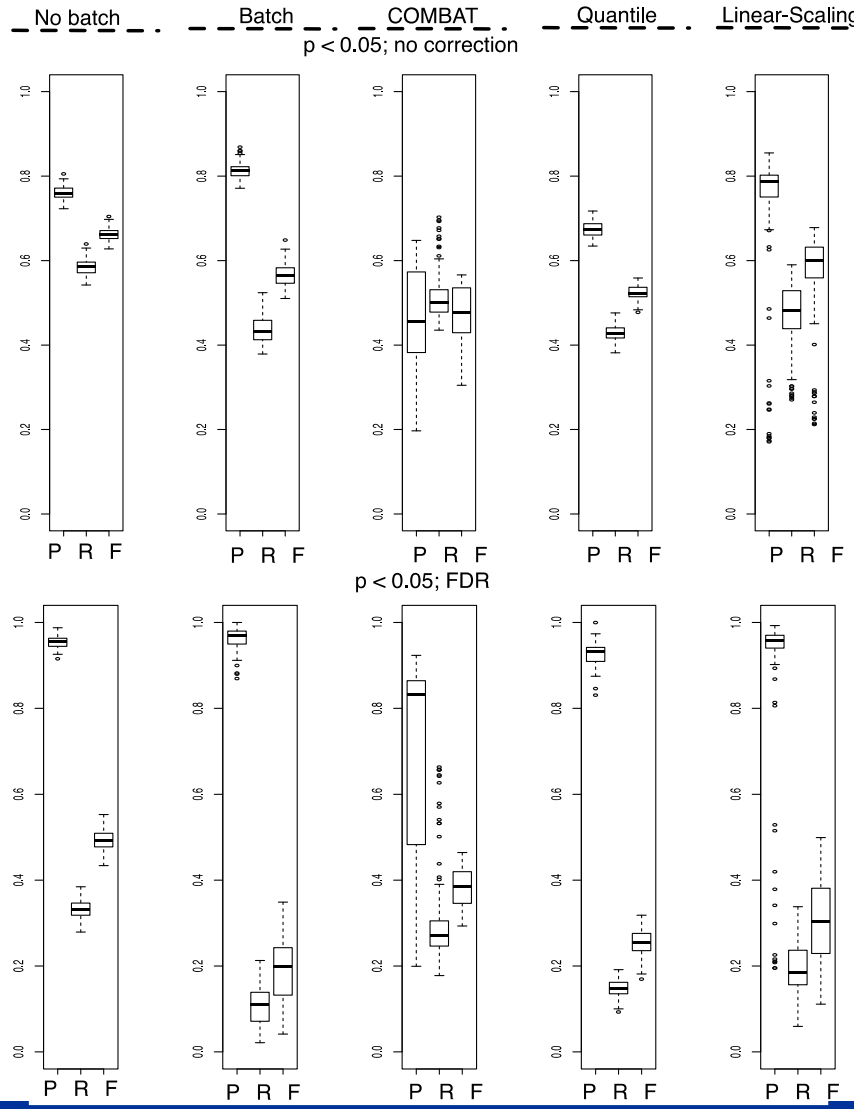
Wang et al. *Molecular Biosystems*, 8:818-827, 2012

Simulated data



- **Real one-class data from a multiplex experiment (no batches); n = 8**
- **Randomly assigned into two phenotype classes D and D*, 100x**
- **20% biological features are assigned as differential, and a randomly selected effect size (20%, 50%, 80%, 100% and 200%) added to D***
- **Half of D and D* are assigned to batch 1, and the other half assigned to batch 2. A randomly selected batch effect (20%, 50%, 80%, 100% and 200%) is added to all features in batch 1**

Batch-effect correction can introduce false positives



P: Precision R: Recall F: F-measure

Feature selection via t-test

Precision is strongly affected by batch correction via COMBAT

This means that false positives are added post-batch correction. Data integrity is affected

Moreover, post-batch correction does not restore performance to where no batch is present

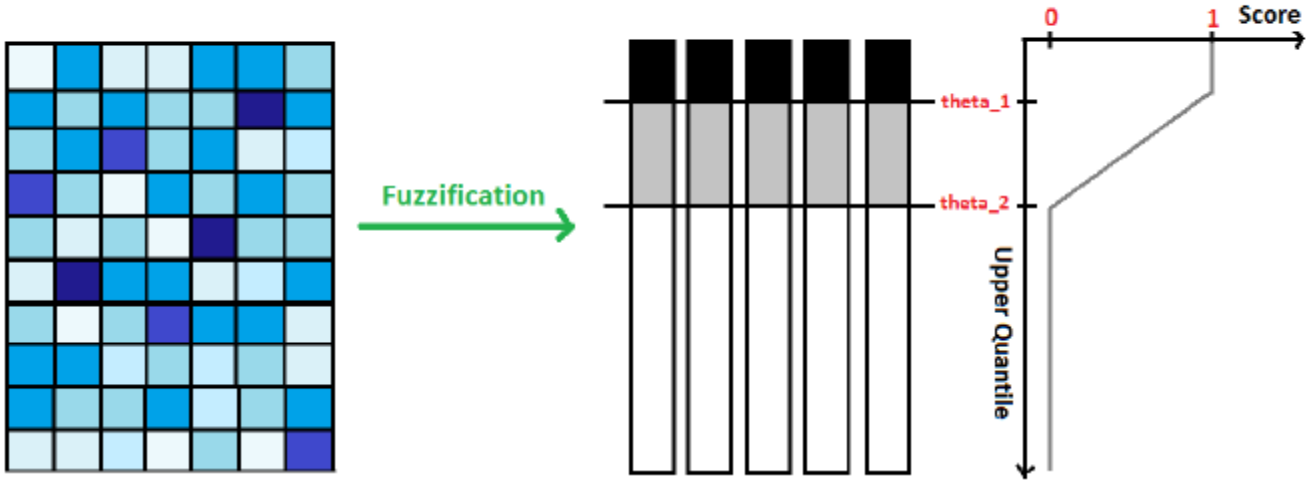
Time for Exercise #1

- **Discuss why normalization methods like mean scaling, z-score, and quantile normalization sometimes do not work well**

Belorkar & Wong, "GFS: Fuzzy preprocessing for effective gene expression analysis", BMC Bioinformatics, 17(S17):1327, 2016

Gene fuzzy score (GFS)

Raw gene expression → gene ranks within microarrays → fuzzified scores



- **Ranks rather than absolute values**
  - No assumption on identical expression distribution
- **Fuzzification**
  - Reduced fluctuations from minor rank differences
  - Noise from rank variation in low-expression genes discarded

Desirable characteristics of normalization methods

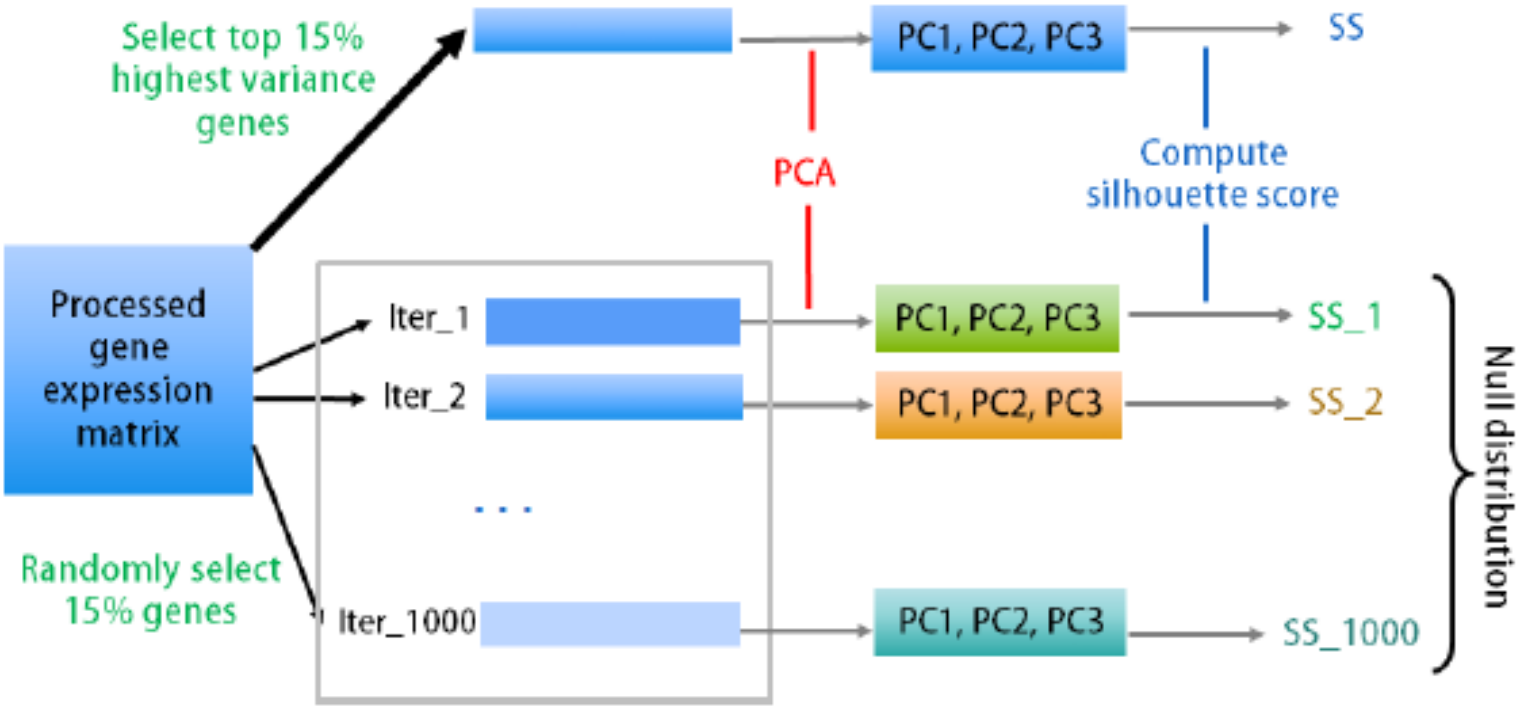
- **High quality**
  - The output of the method is useful in separating samples of different phenotypes from each other
- **High consistency**
  - When applied to any two representative batches of data, the overlap between high-variance features (e.g. genes) is high
- **High biological coherence**
  - E.g. high-variance genes in the normalized output induce large subnetworks on known pathways

Datasets used in evaluating GFS

Disease type	Source	Affy GeneChip	Dataset composition
DMD	Haslett et al.	HG-U95Av2	12 DMD, 12 controls
	Pescatori et al.	HG-U133A	22 DMD, 14 controls
Leukemia	Golub et al.	HU-6800	47 ALL, 25 AML
	Armstrong et al.	HG-U95Av2	24 ALL, 24 AML
ALL	Yeoh et al.	HG-U95Av2	15 BCR-ABL, 27 E2A-PBX1
	Ross et al.	HG-U133A	15 BCR-ABL, 18 E2A-PBX1
ALL	Yeoh et al.	HG-U95Av2	6 Normal, 26 TEL-AML1, 22 Hyperdip>50, 15 T-ALL, 10 Pseudodip, 6 BCR-ABL, 7 MLL, 8 Hyperdip47-50 9 E2A-PBX1, 3 Hypodip

- Haslett, et al. *PNAS*, 99(23):15000-15005, 2002.
- Pescatori et al. *FASEB Journal,* 21(4):1210-1226, 2007
- Golub et al. *Science*, 286(5439):531-537, 1999
- Armstrong et al. *Nature Genetics,* 30(1):41-47, 2002
- Yeoh, et al. *Cancer Cell,* 1(2):133-143, 2002.
- Ross, Mary E., et al. *Blood* ,104(12):3679-3687, 2004

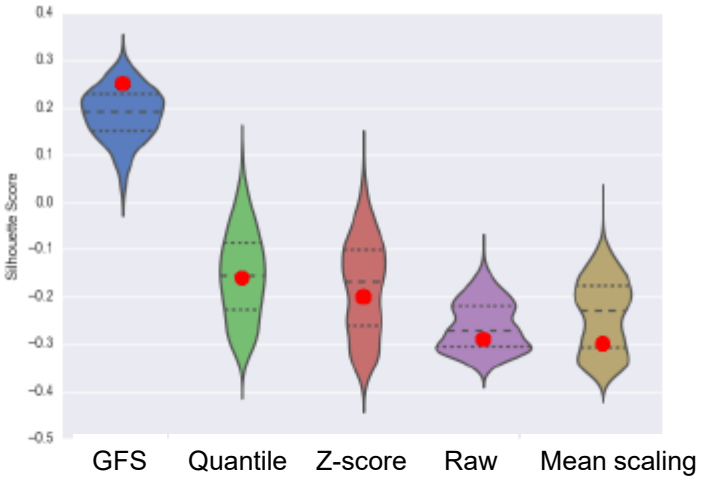
Evaluating quality



- **An ideal normalization method should produce a silhoutte score distribution that is high and stable**

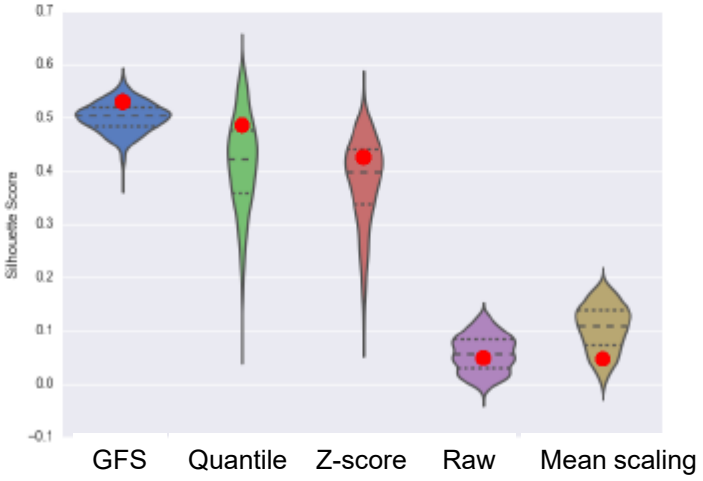
Observations

- **The GFS null distribution is stable at high clustering index**



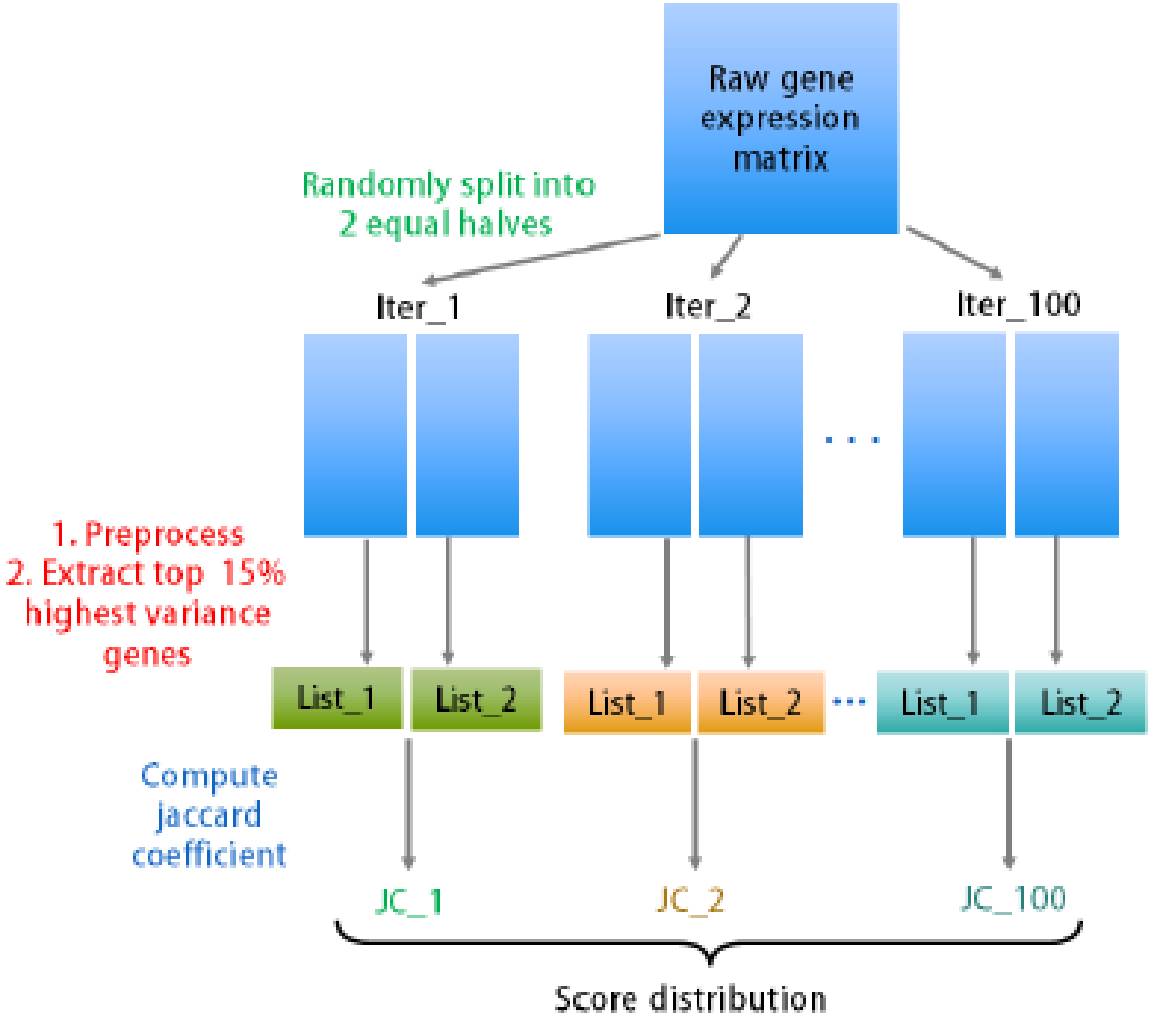
(a) Acute Lymphoblastic Leukemia (ALL)

- **For GFS, the score obtained from the to| 15% highest varianc| genes is always greater than the scor from the 5th percentil of the null distributio**



(b) Duschenne Muscular Dystrophy (DMD)

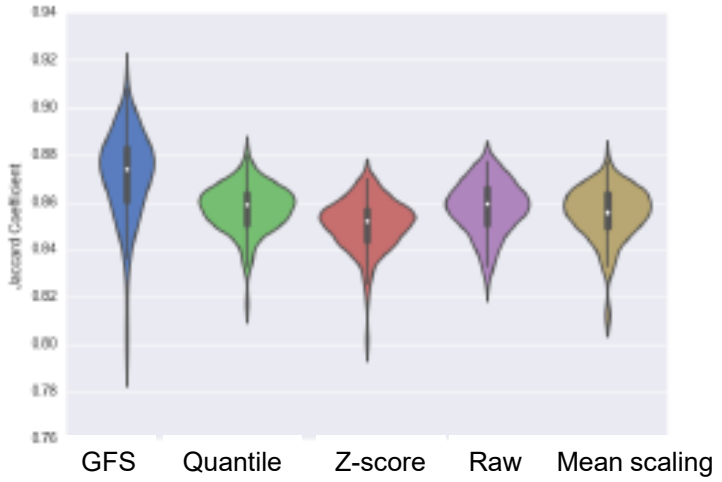
Evaluating consistency



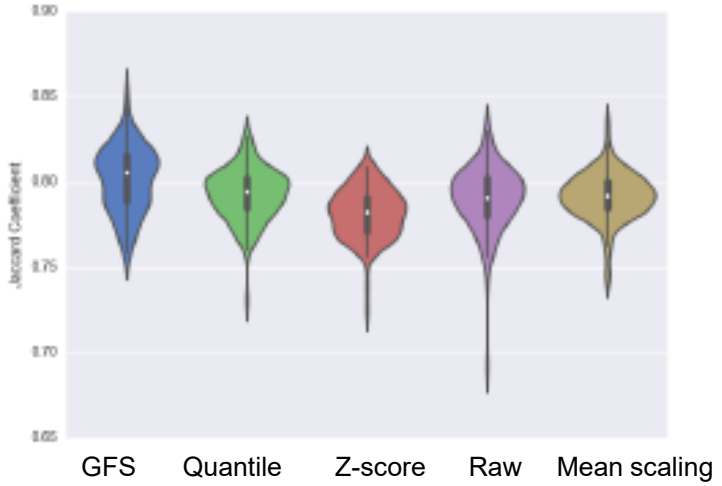
- **An idea method should produce a Jaccard coefficient distribution that is high and stable**

Observations

- **The Jaccard coefficient of GFS over all subsamplings is stable at a coefficient equal to or higher than other methods**

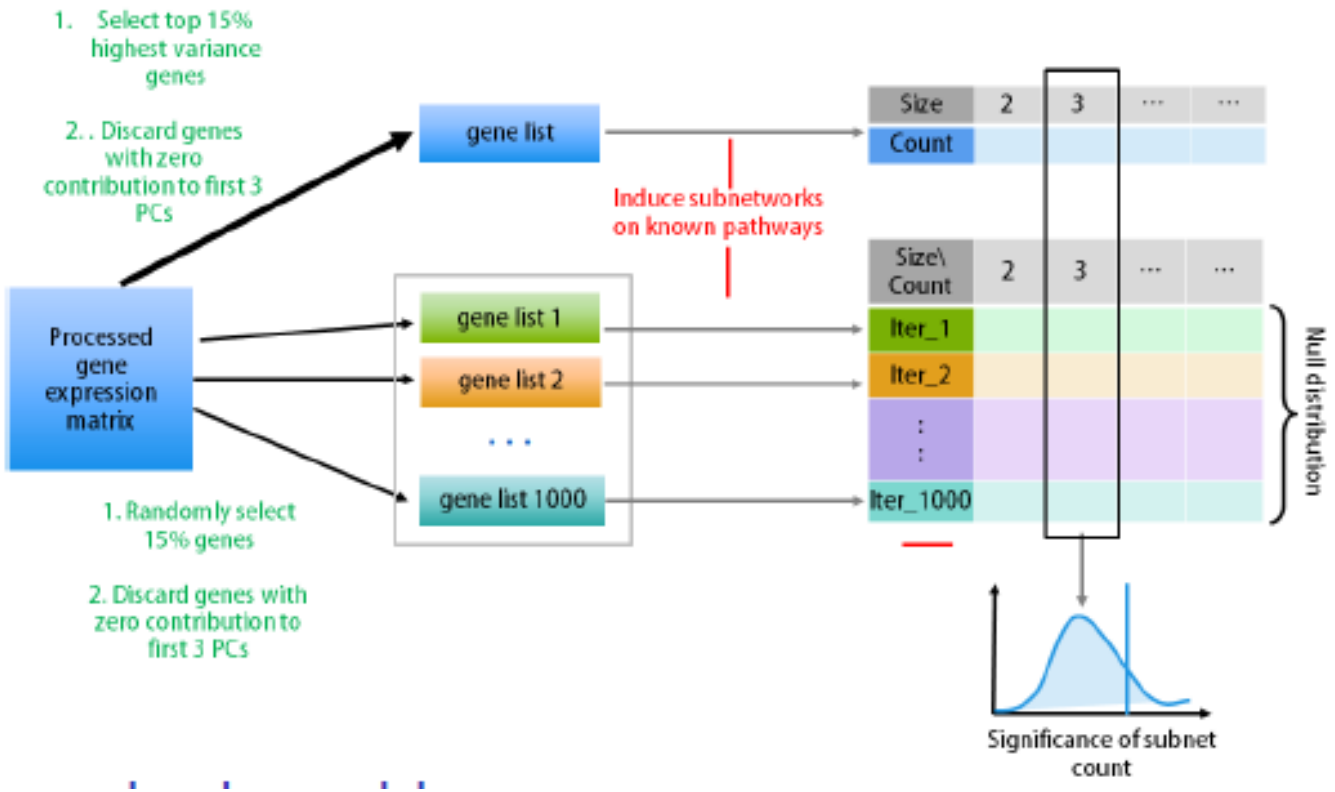


(a) Acute Lymphoblastic Leukemia (ALL)



(b) Duschenne Muscular Dystrophy (DMD)

Evaluating biological coherence



- **An ideal method should produce high-variance genes that induce larger and more significant subnetworks**

Observations

- **High-variance genes from methods other than GFS induce subnetworks that are generally not very different from those produced by random genes**

	Raw		Scaled		Z-score		Quantile		GFS	
size	freq	p	freq	p	freq	p	freq	p	freq	p
2	87	0.672	77	0.861	76	0.876	87	0.672	80	0.071
3	44	0.621	46	0.545	41	0.722	45	0.577	67	0.000
4	24	0.483	24	0.483	24	0.483	23	0.546	39	0.000
5	18	0.105	18	0.105	18	0.105	18	0.105	16	0.001
6	3	0.890	2	0.958	4	0.804	2	0.958	11	0.000
7	9	0.025	4	0.408	3	0.588	9	0.025	4	0.029
8	2	0.492	3	0.289	4	0.144	3	0.289	4	0.013
9	5	0.017	6	0.004	4	0.057	5	0.017	1	0.170
10	3	0.062	3	0.062	4	0.021	2	0.165	5	0.000
:	:	:	:	:	:	:	:	:	:	:
21	-	-	-	-	1	0.038	-	-	1	0.000

(a) Acute Lymphoid Leukemia (ALL)

	Raw		Scaled		Z-score		Quantile		GFS	
size	freq	p	freq	p	freq	p	freq	p	freq	p
2	74	0.903	970	0.415	57	0.995	104	0.278	85	0.009
3	83	0.007	44	0.644	23	0.999	40	0.777	81	0.000
4	19	0.799	22	0.643	17	0.894	18	0.861	28	0.004
5	15	0.324	11	0.665	12	0.586	13	0.485	18	0.000
6	7	0.521	11	0.145	7	0.521	10	0.206	11	0.000
7	8	0.106	12	0.005	4	0.519	10	0.022	9	0.000
8	7	0.018	6	0.045	3	0.392	6	0.045	3	0.011
9	1	0.615	5	0.031	3	0.148	7	0.008	4	0.002
10	2	0.229	1	0.467	3	0.084	2	0.229	2	0.007
:	:	:	:	:	:	:	:	:	:	:
20	-	-	-	-	-	-	-	-	1	0.000

(b) Duchenne Muscular Dystrophy (DMD)

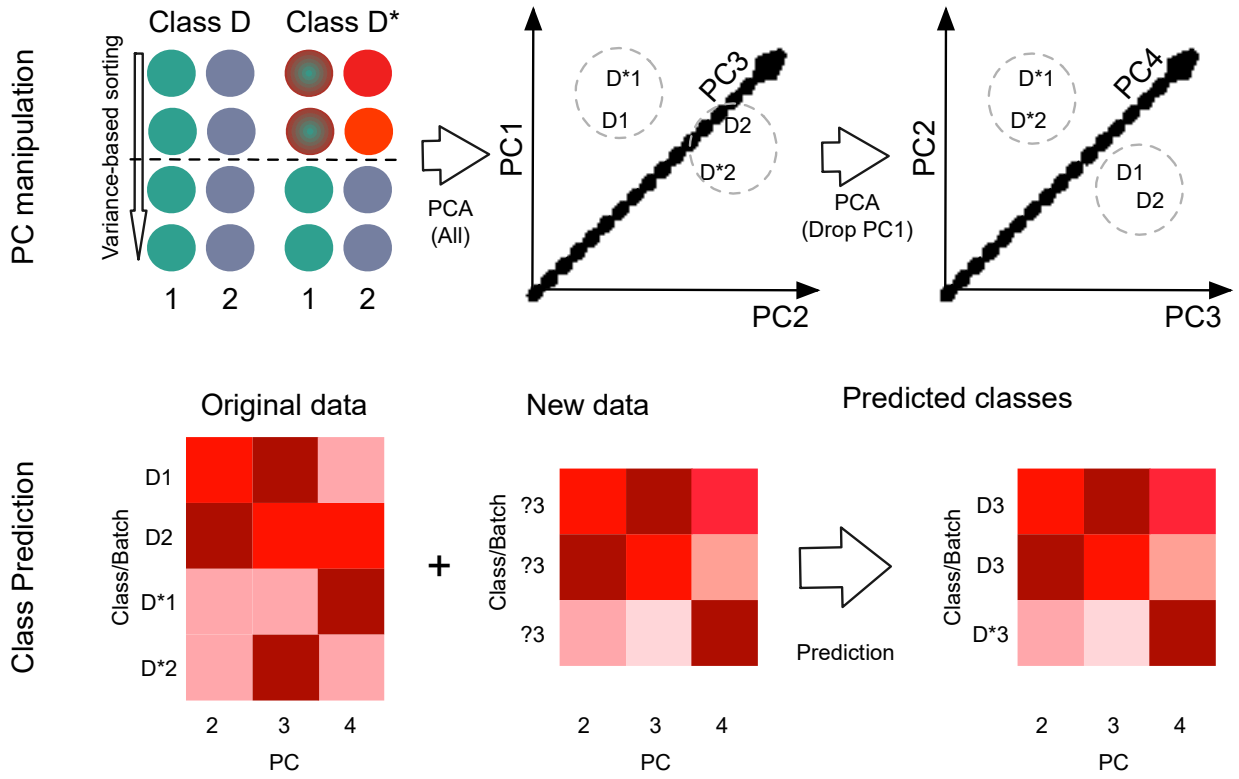
Time for Exercise #2

- **An ideal normalization method should not degrade suddenly when sample size is small. Discuss how you can check this**

PC1 Removal

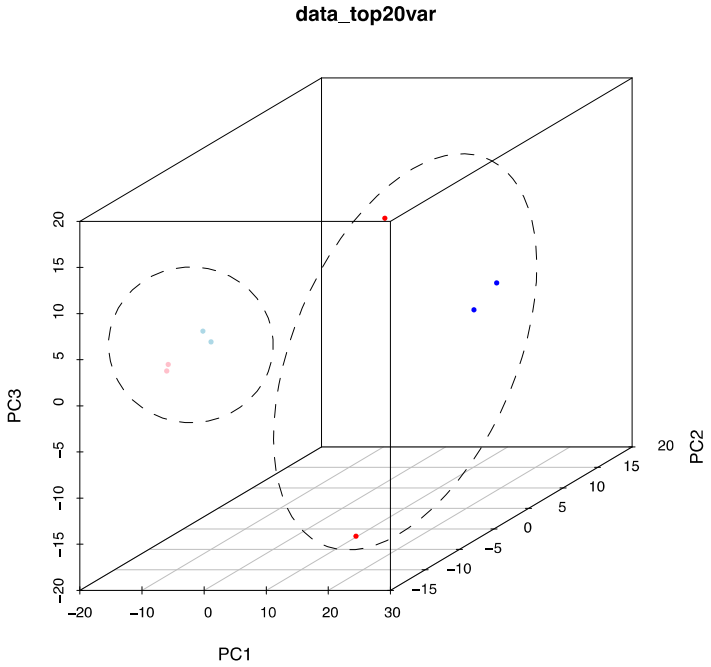
Using PCA for batch-effect correction

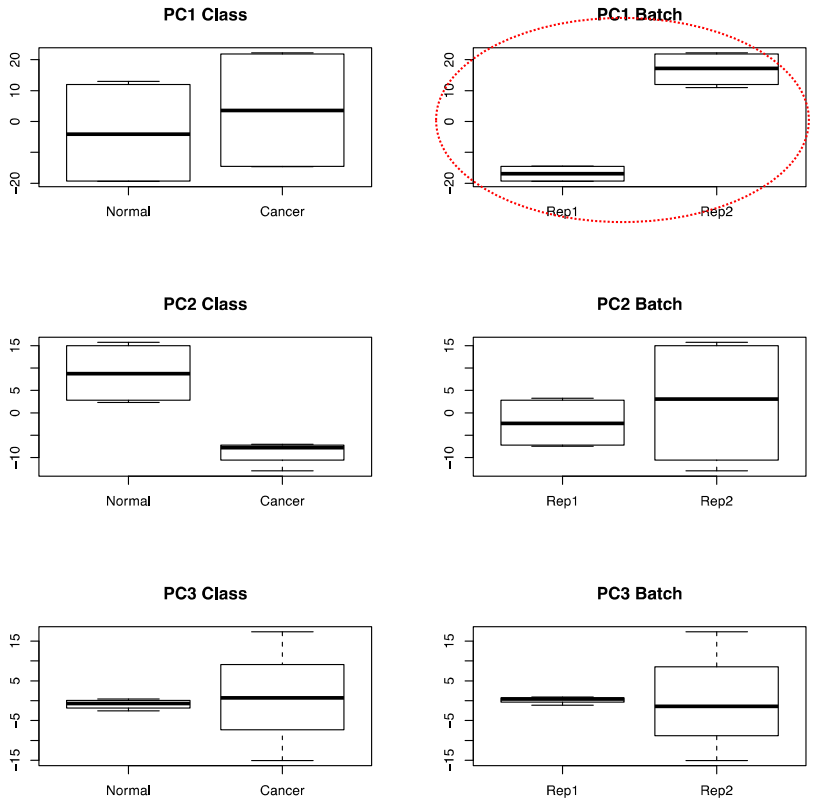
- When a batch effect is observed, it is common practice to apply a batch effect-removal or -correction method. However, this does not necessarily work well in practice. Moreover, if the data does not fit the correction method's assumptions, it may lead to false positives. Instead, we may opt for a more direct strategy by simply removing PCs (usually PC1) enriched in batch effects, and deploying the remaining PCs as features for analysis



Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects”, *BMC Genomics*, in press

PC1 is often associated with batch

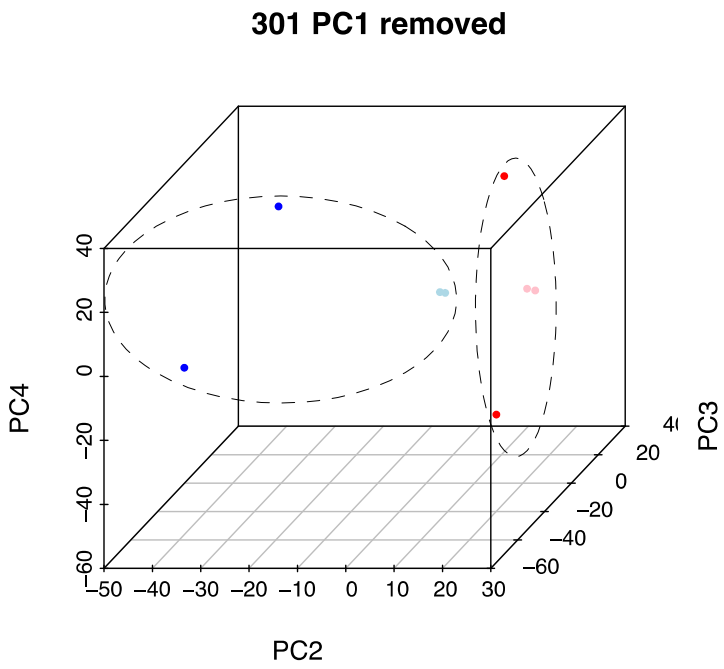


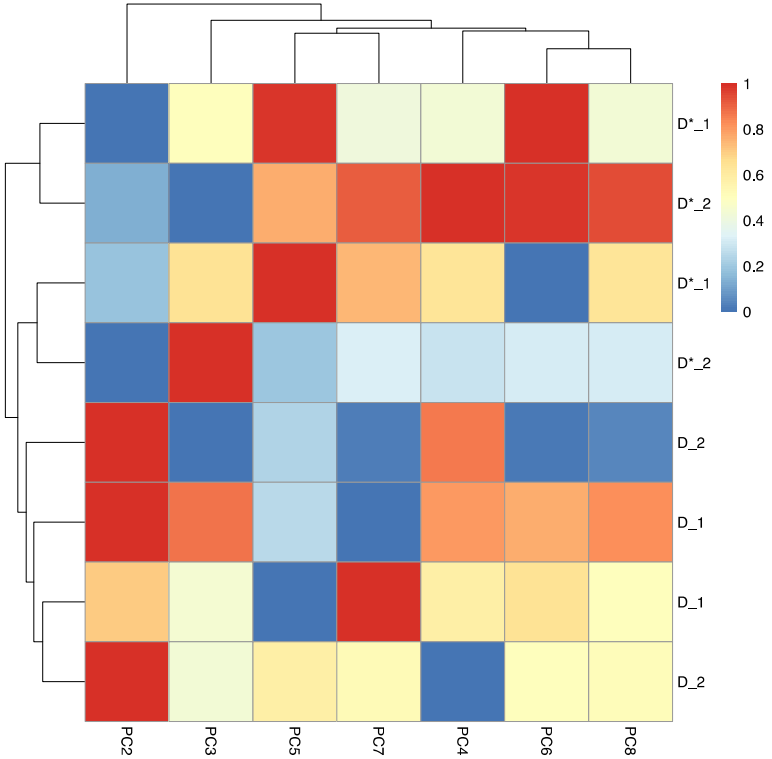


D, Rep 1 D*, Rep 1 D, Rep 2 D*, Rep 2

Batch effects dominate in PC1

Removal of PC1 removes most batch effects



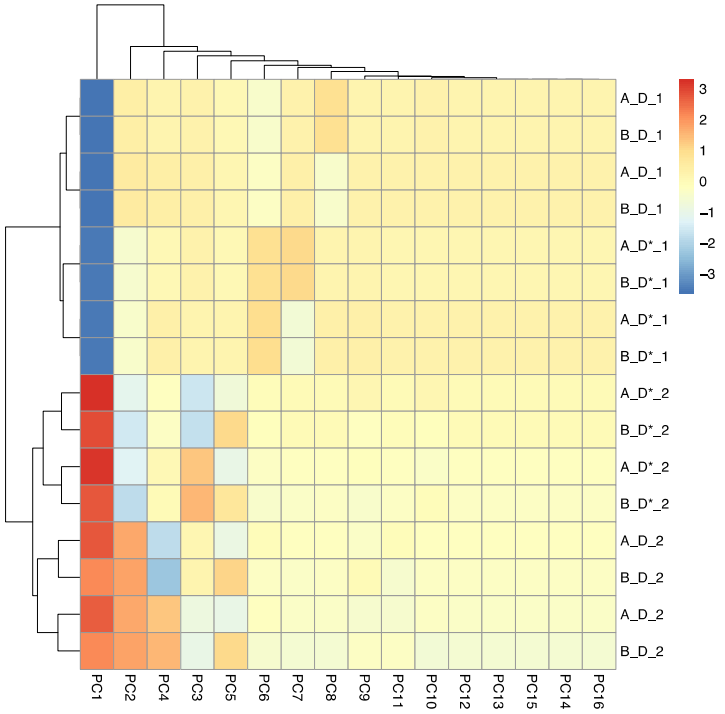


D, Rep 1 D*, Rep 1 D, Rep 2 D*, Rep 2

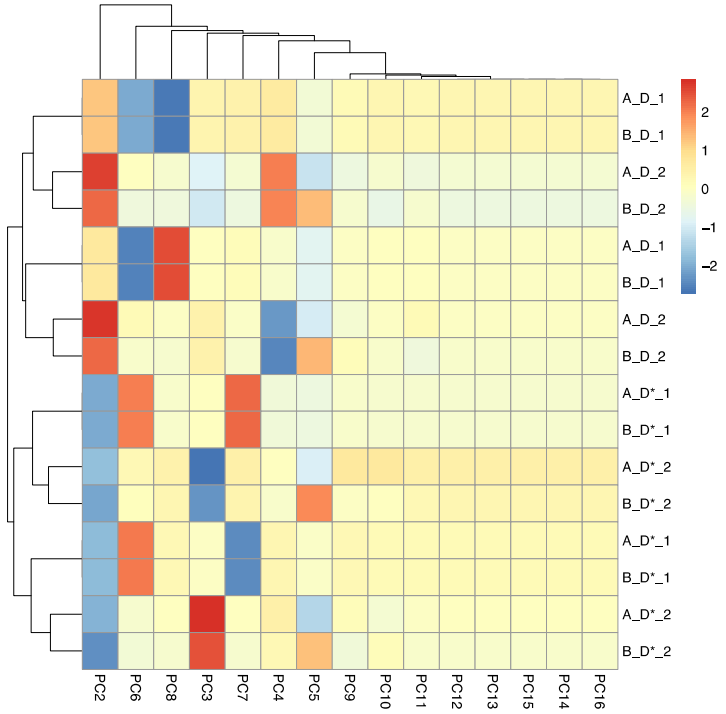
Samples segregate perfectly by class. No batch-associated subgrouping

Post PC1 removal permits data integration

A and B are different datasets with different batch effects inserted



Batch effects dominate



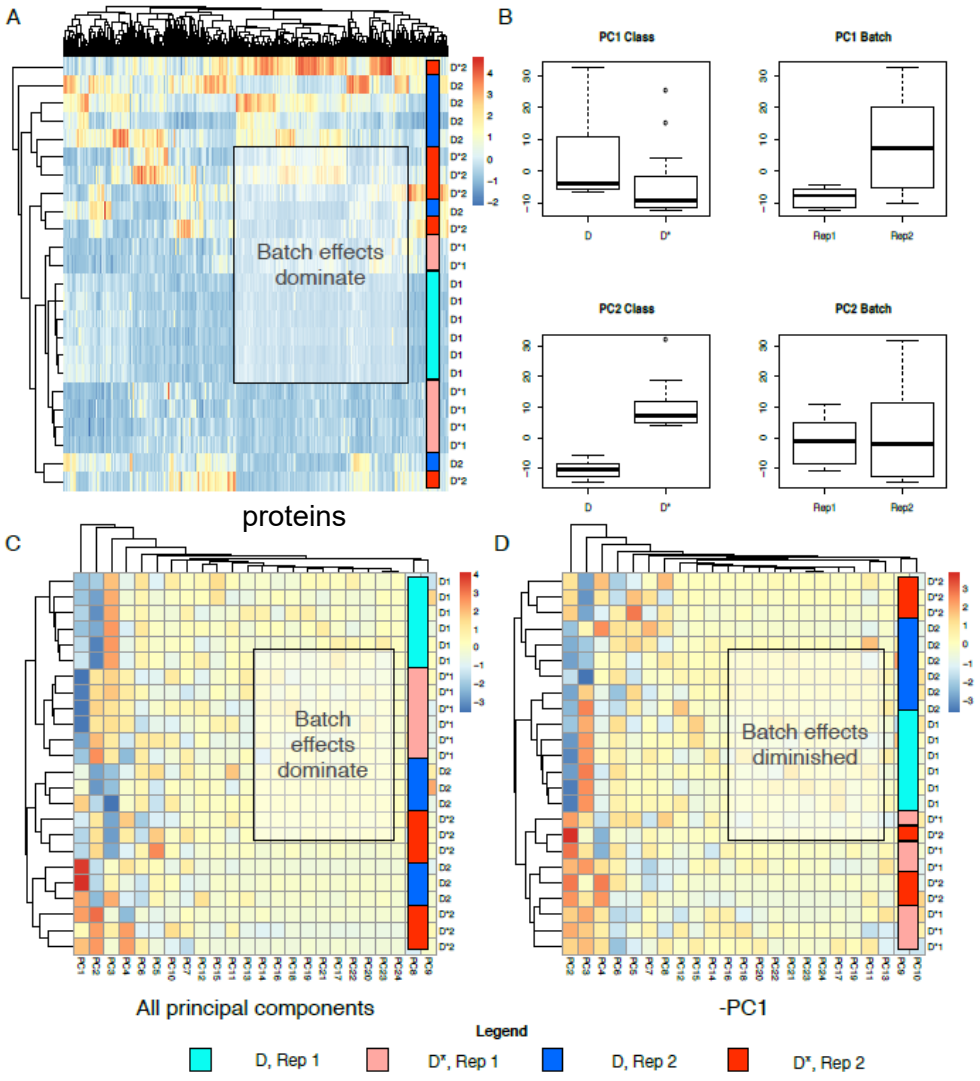
Class-effect discrimination recovered

(Notation: A/B_D/D*_1/2 refers to the dataset, class and batches respectively)

Test using real biological data

- **Proteomics data used: Renal cancer (RC)**
  - Guo et al. *Nature Medicine*, 21(4):407-413, 2015
  - 6 pairs of normal vs cancer ccRCC tissues
  - 2 technical replicates on which we can evaluate batch effects

PC1 elimination also works on real biological data



A: Batch effects dominate. Clustering is based on all protein expression. No feature-selection was performed prior

B: PC1 is associated strongly with batch effects although there is also some association with class effects, though this is not seen in PC2

C: Batch effects dominate in real data (RC). Note that these batch effects are inserted into RC rep2 samples

D: Removal of PC1 diminishes batch effects while also improving class discrimination

NUS National University of Singapore

BATCH EFFECT-RESISTANT FEATURE SELECTION

What if class and batch effects are strongly confounded?

- **Neither batch correction nor PCA work well**
- **We also do not want to inadvertently lose information on disease subpopulations (which look like batch effects but are meaningful)**

$\Rightarrow$ **Consider batch-resistant methods instead of batch removal**

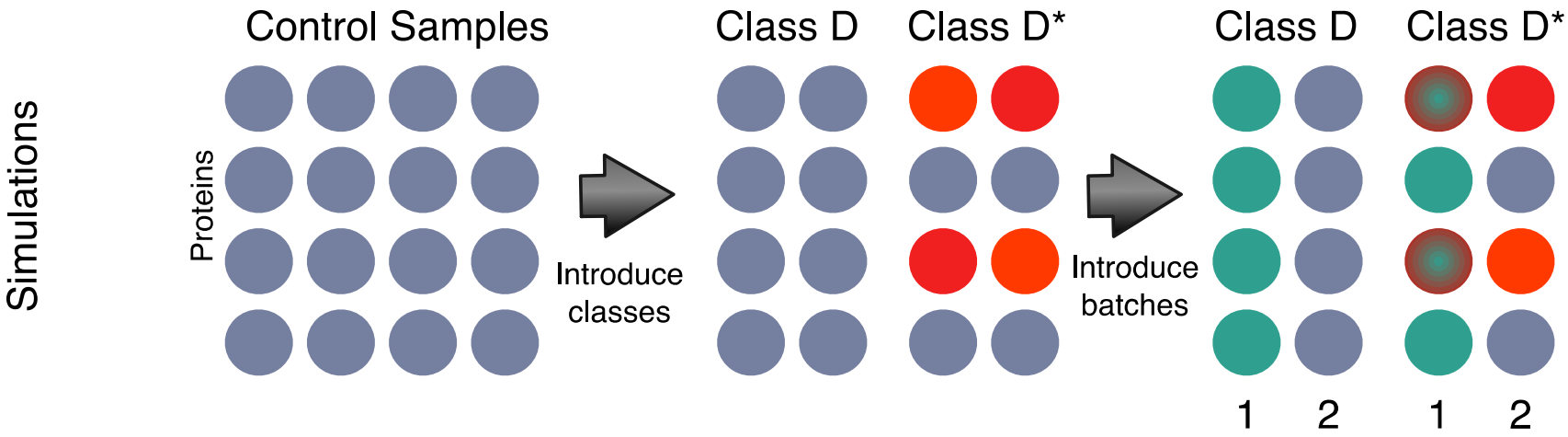
- **Protein complex- / network-based feature selection methods (SNET, FSNET, etc.) exhibit strong reproducibility with high phenotype specificity, maybe they are batch resistant?**

Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects”, *BMC Genomics*, in press

SNET and FSNET

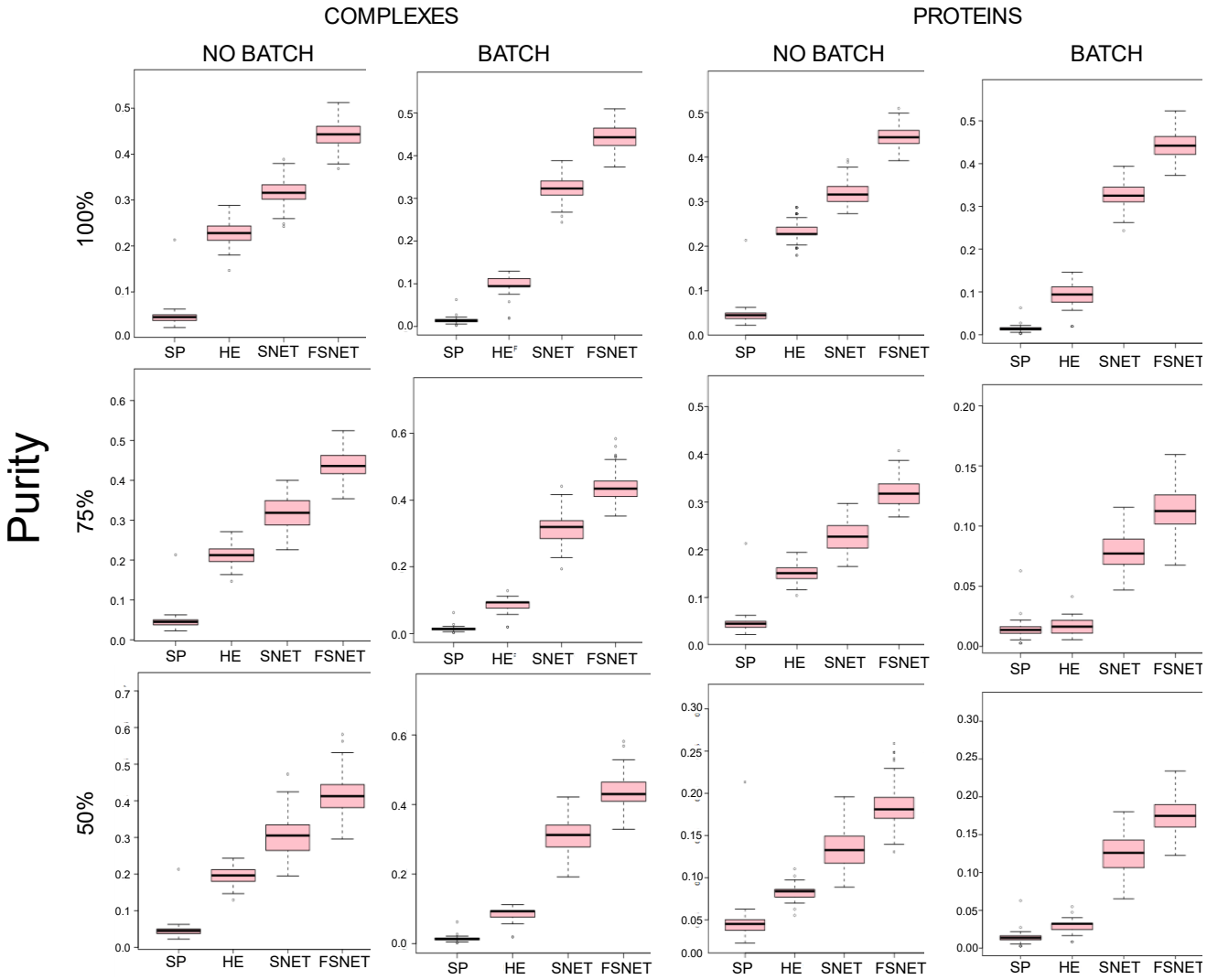
- **SNET and FSNET --- two protein complex-based feature-selection methods. Use expression rank-based weighting method (viz. GFS) on individual proteins, followed by intra-class-proportion weighting**
- SP is the protein-based two-sample t-test and HE is a two-step procedure deploying SP first, followed by the Fisher’s exact test on networks
- Significant artificial complexes are constructed with various level of purity (i.e. proportion of significant proteins in the complex). Equal # of non-significant complexes are constructed as well

Simulated data



- **Real one-class data from a multiplex experiment (no batches); n = 8**
- **Randomly assigned into two phenotype classes D and D*, 100x**
- **20% biological features are assigned as differential, and a randomly selected effect size (20%, 50%, 80%, 100% and 200%) added to D***
- **Half of D and D* are assigned to batch 1, and the other half assigned to batch 2. A randomly selected batch effect (20%, 50%, 80%, 100% and 200%) is added to all features in batch 1**

Batch resistance (Simulated data)

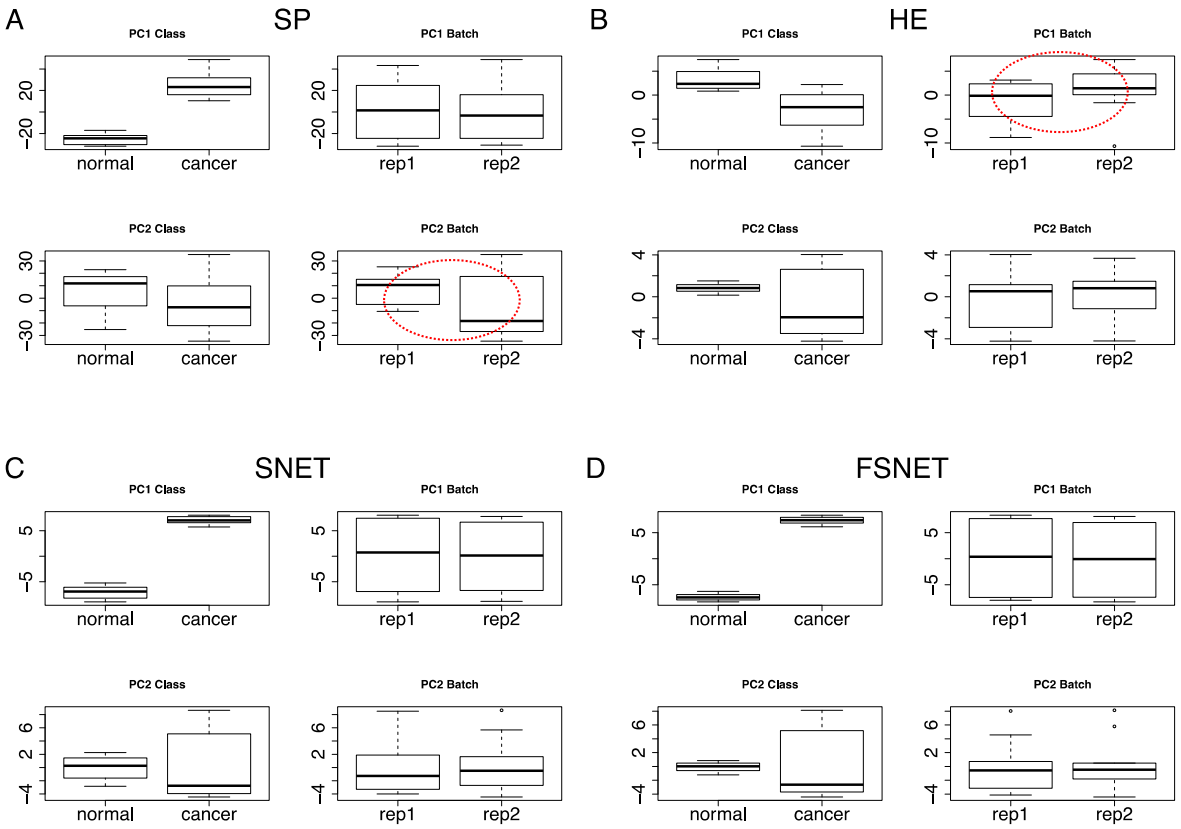


F-score distributions SNET and FSNET is robust against batch effects relative to traditional methods e.g. SP and HE

As a fairer comparison, we consider both complex and constituent protein scenarios (SP does not use complexes)

But how does it look on real data?

Network-based methods are enriched for class-related variation (Real data)

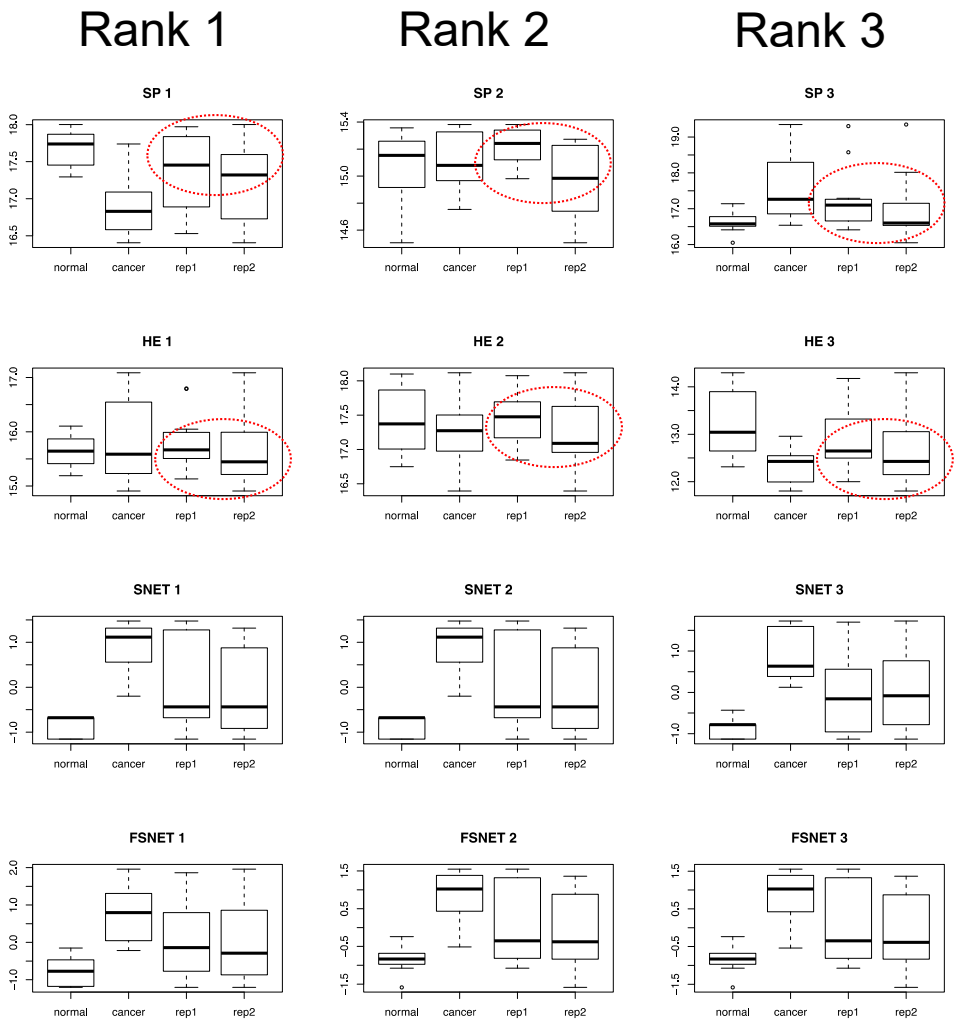


Protein complexes used as reference

Side-by-side boxplots stratified by class and batch tested on real data

SNET and FSNET are robust against batch effects, and only seems to capture variation stemming from class effects

Top complex-based features are strongly associated with class, not batch



SNET and FSNET can capture the class effects while being robust against batch effects

In contrast, both class and batch variability are present in the top variables selected by SP and HE

Time for Exercise #3

- SNET/FSNET are resistant to batch effects. They analyze GFS-processed proteomic profiles in the context of protein complexes instead of individual proteins
- So their batch-effect resistance could be due to the use of GFS rather than their protein complex-level analysis
- Discuss how you can show that their batch-effect resistance is not due solely to GFS

BATCH EFFECT-RESISTANT CLASSIFIERS: EMBRACING-NOISE APPROACH

NUS National University of Singapore

The Pharmacogenomics Journal (2010) 10, 278–291

www.nature.com/tpj

Open

ORIGINAL ARTICLE

A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data

- Study how various batch effect removal algorithm influence cross-batch prediction performance

Results

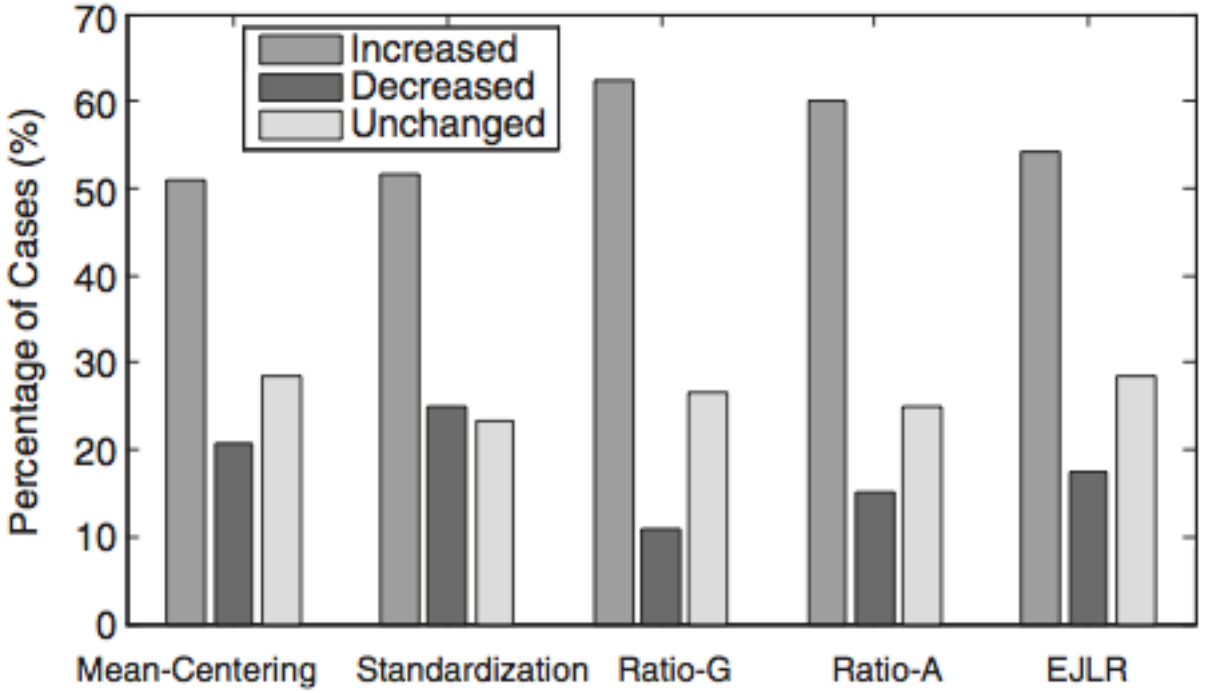


Figure 10 Percentages of increased, decreased and unchanged cases in prediction performance after applying different batch effect removal methods. The total number of cases explored is 120.

Increased: Difference in MCC with and without batch removal > 0.05
Decreased: Difference in MCC with and without batch removal < -0.05
Unchanged: Difference in MCC with and without batch removal ≤ 0.05 & ≥ -0.05

Findings

- **Around 10-20% of the times, doing batch effect removal actually reduces prediction power**
- **Batch removal is not practical in real situations**

constructed predictive models to future data sets. It is desirable to have a large sample size or good quality data in each batch, so that the characteristics of each batch can be summarized more accurately and batch effects can be removed more effectively. If the sample sizes of the training and the test set are too small, it is difficult to draw a conclusive inference due to the large uncertainty. In the context of implementing an array-based diagnostic test in a clinical setting, it should be appreciated that batches may, in practice, be composed of a single clinical sample. In this regard, the use of reference samples for the purpose of calibrating batch effects may be of paramount importance.

Typical approaches

- **Typical**
  - Attempt to accurately estimate the batch effects
  - Then remove them
  - Therefore large sample sizes are often required for each batch and a balanced class ratio is often desired

- **A new approach**
  - "Embracing noise"

Koh & Wong, "Embracing noise to improve cross-batch prediction accuracy", *BMC Systems Biology*, 6(Suppl 2):S3, 2012

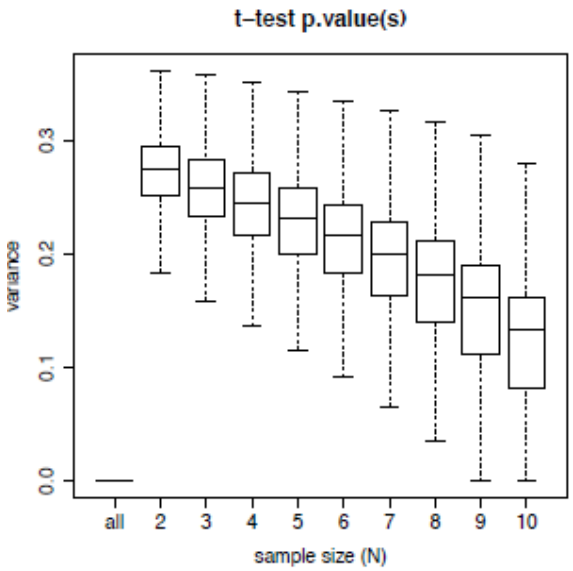
"Embracing-noise" approach

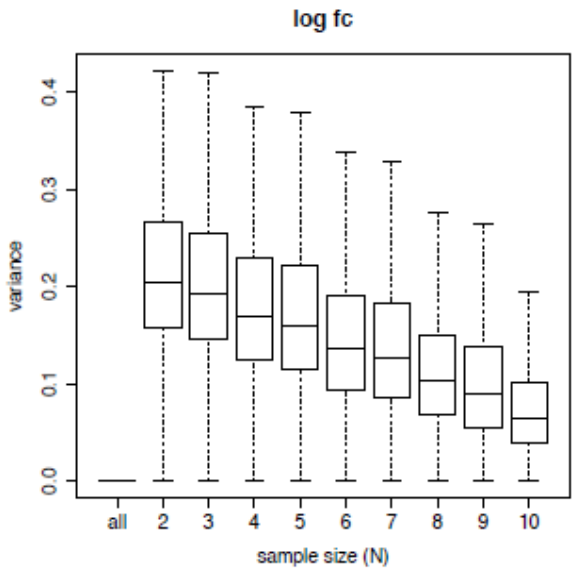
- **Ranking values (c.f. quantile normalization)**
  - Instead of absolute values
  - Inspired by MAQC project
    - **"Absolute values may be different (among batches) but relative values are conserved between different platforms"**

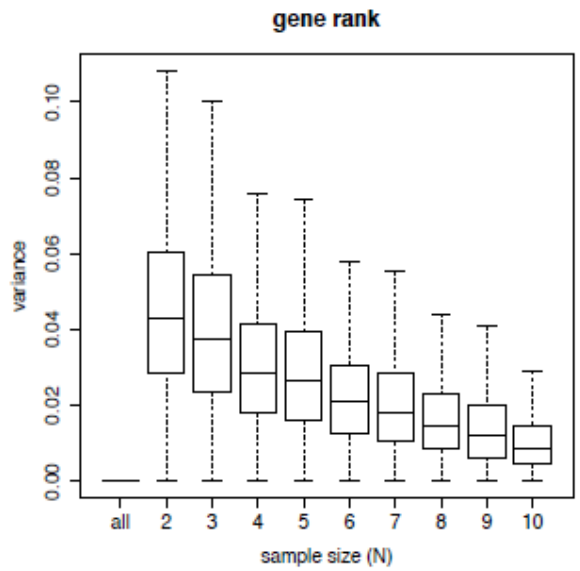
- **Stochastic sampling with replacement**
  - Bootstrapping suppresses noise
    - **Training clones produced are likely to be enriched with more "clean" samples**

Ranking values

- **Findings from MAQC project**
  - Median coefficient of variations
    - **Within Lab: 5-15% for different platforms**
    - **Inter Lab: 10-20% for different platforms**
  - High correlation between the ranks of log ratios between different platforms
    - **Absolute values may be different but relative values are conserved between different platforms**
  - Conclusion
    - **Microarray are still reproducible (ranking values) despite being noisy**







Ranking values are stable
even when sample size is small

Bootstrapping suppresses noise

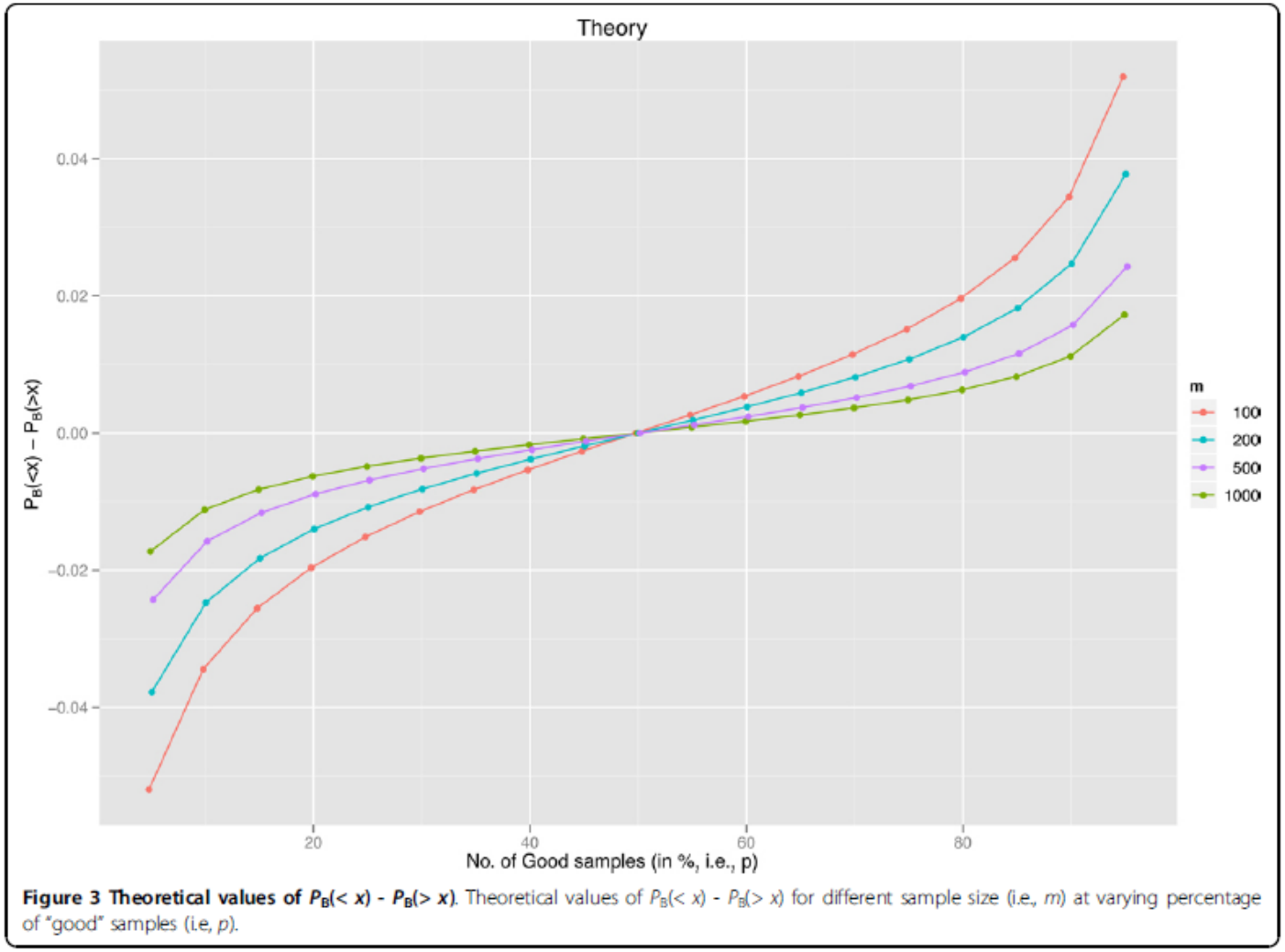


Figure 3 Theoretical values of $P_B(< x)$ - $P_B(> x)$. Theoretical values of $P_B(< x)$ - $P_B(> x)$ for different sample size (i.e., m) at varying percentage of "good" samples (i.e, p).

Dynamic bagging

- **Integrates bootstrapping with sequential hypothesis testing**
- **Removes the need to a priori and arbitrarily fixing the number of bootstrap replicates (N)**
- **N is minimum for each test instance with statistical guarantees on the error rates**
  - An error is define as the difference in decision with the minimum N and infinite N

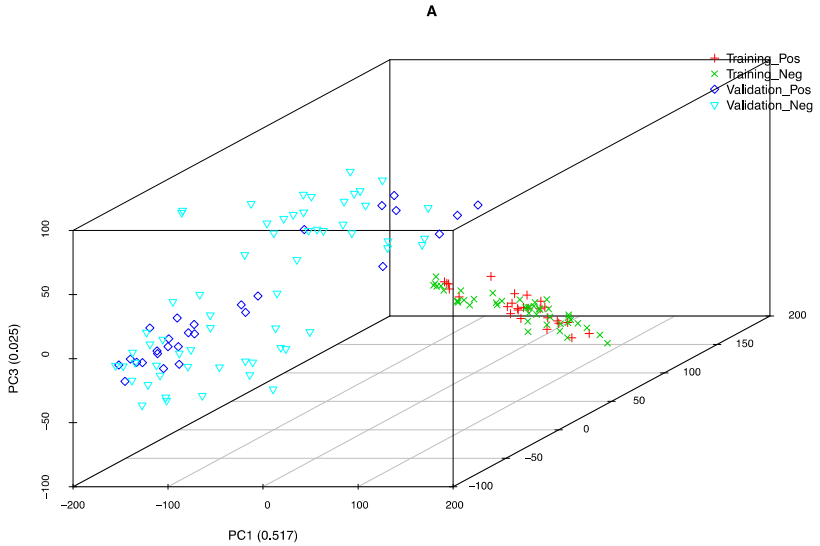
Koh, et al. **Improved statistical model checking methods for pathways analysis**. *BMC Bioinformatics*, 13(Suppl 17):S15, 2012

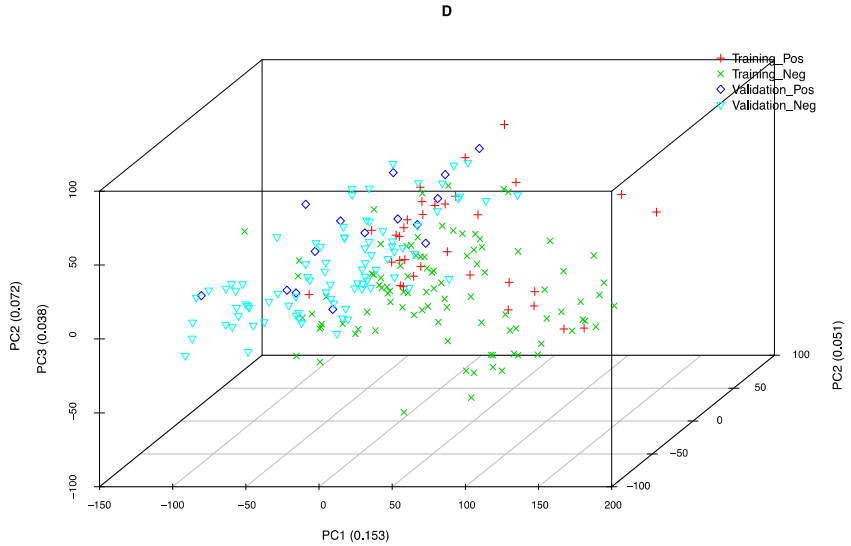
Datasets

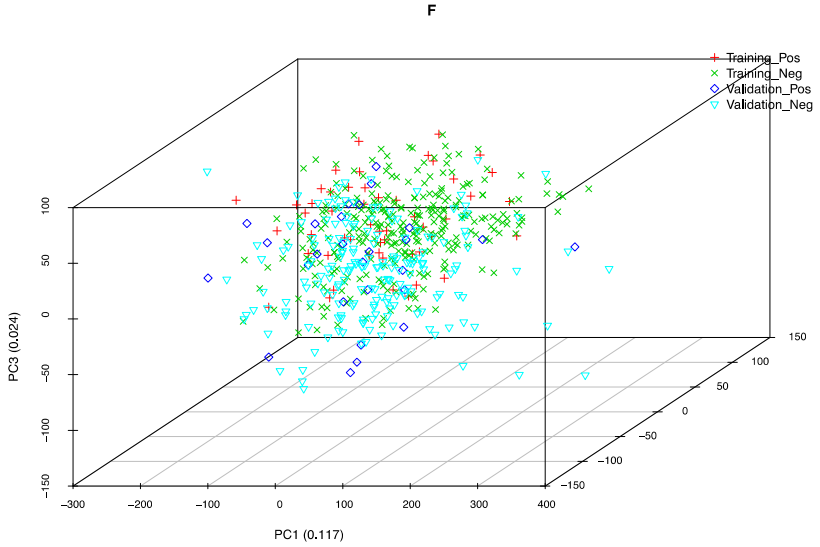
Data set code	Data set description	Training set			Validation set		
		Number of samples	Positives	Negatives	Number of Samples	Positives	Negatives
A	Lung tumorigen vs. non-tumorigen (Mouse)	70	26	44	88	28	60
D	Breast cancer pre-operative treatment response (pathologic complete response)	130	33	97	100	15	85
F	Multiple myeloma overall survival milestone outcome	340	51	289	214	27	187
I	Same as data set F but class labels are randomly assigned	340	200	140	214	122	92

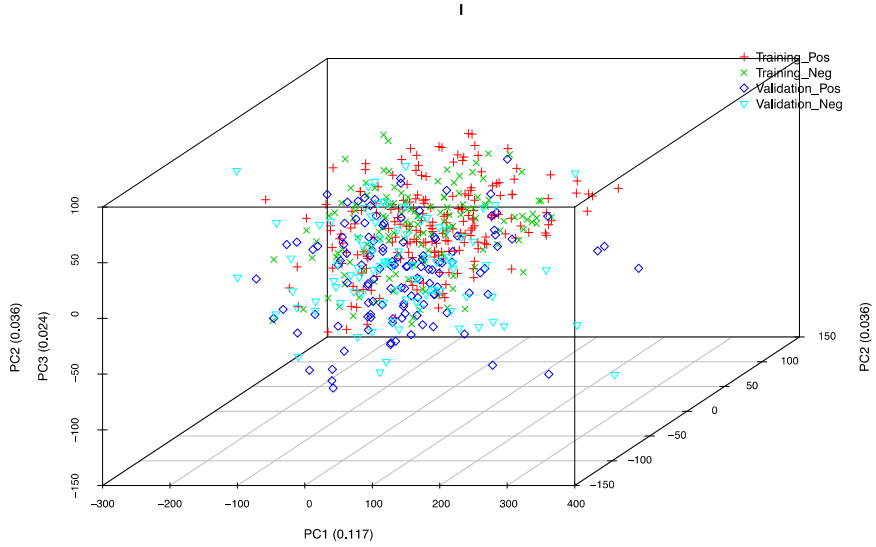
- Two additional data sets of size 25% or 50% of the above (with same class ratio)
- Total of 12 training sets and 12 validation sets

PCA of these datasets



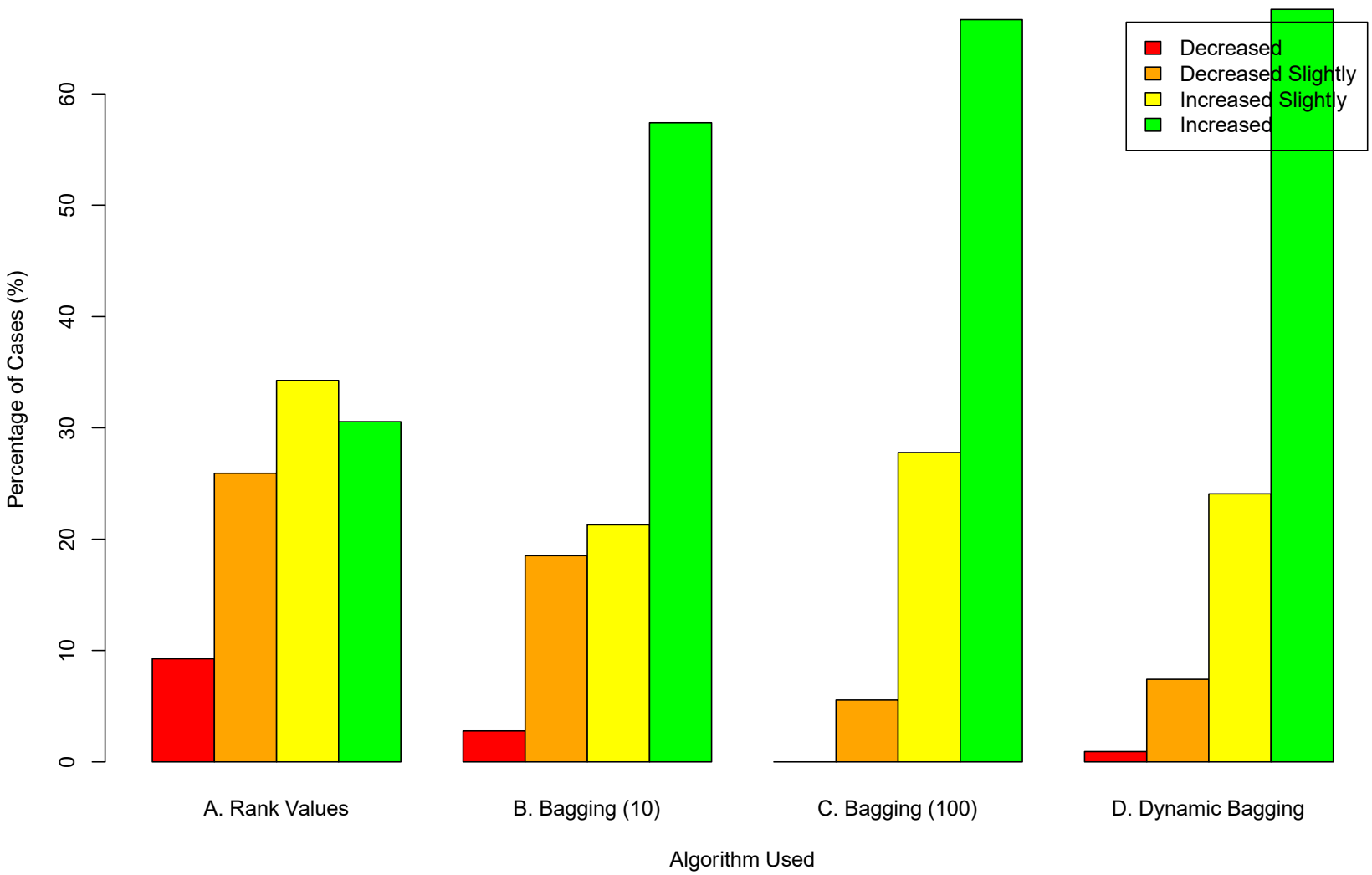


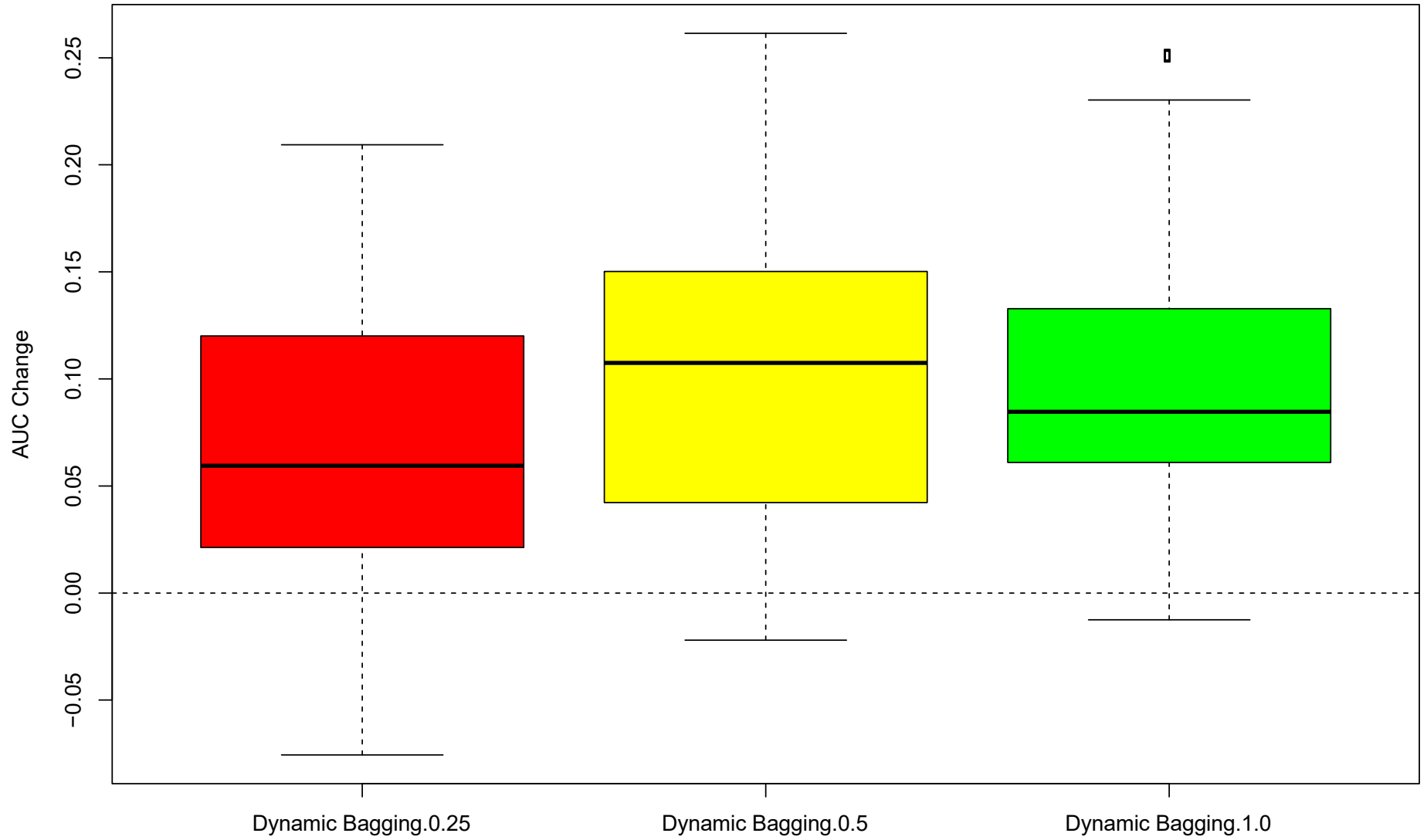




Experiments

- **Feature selection**
  - t-Test (Parametric)
  - Wilcoxon Rank Sum Test (Non-Parametric)
- **Classification algorithms**
  - C4.5 (Tree)
  - Support Vector Machine (Linear)
  - Nearest Neighbor (Instance-based)
- **Performance metric**
  - Area Under Curve





Conclusion

- **An unconventional yet simple approach**
  - Ranking values
  - Dynamic bagging
- **Great performance**
  - Shows improvements in most cases
- **Practically applicable**
  - Works on small training data sets
  - Independent of the sample size of the test data

NUS National University of Singapore

BATCH EFFECT-RESISTANT CLASSIFIERS: USING NETWORK-BASED FEATURES

Batch effects

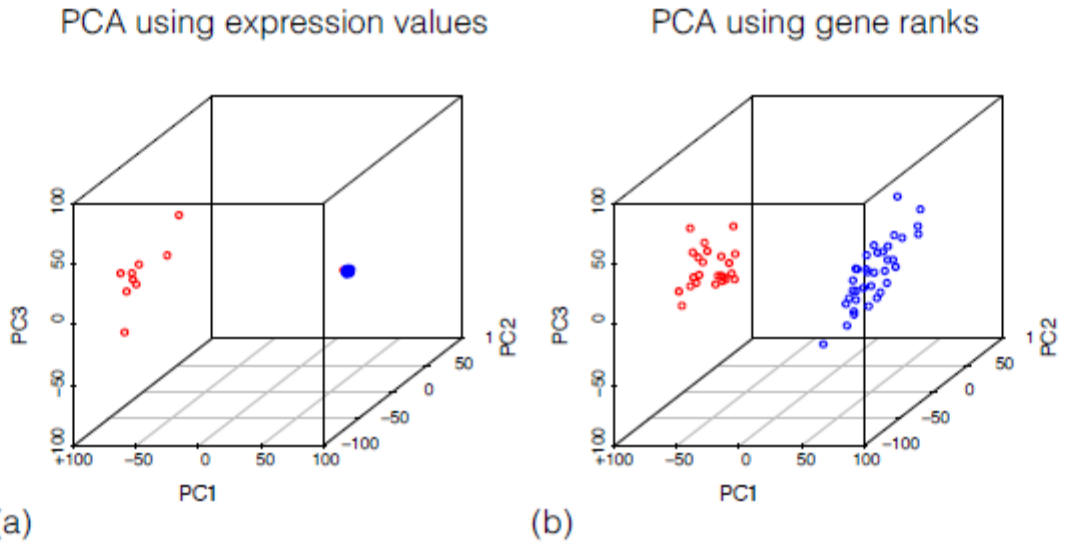


FIGURE 5.1: Batch effects in the DMD/NOR datasets, the blue and red color denote different data batches. (a) Scatterplot on the first 3 components using gene-expression values. (b) Scatterplot on the first 3 components using gene ranks.

- **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

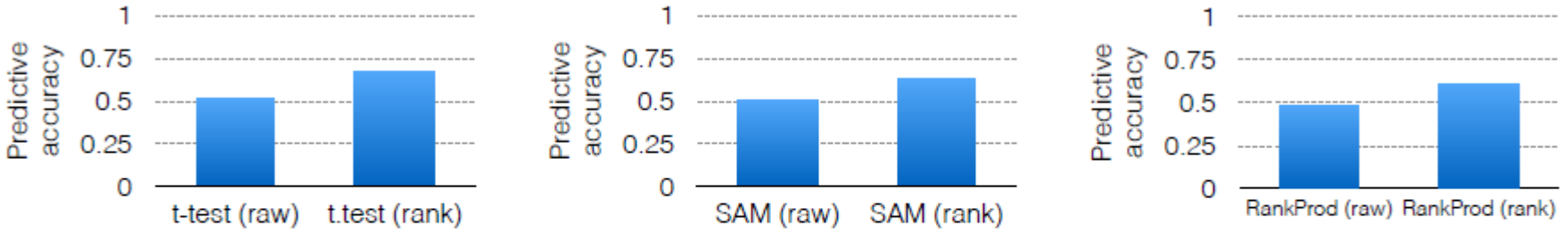
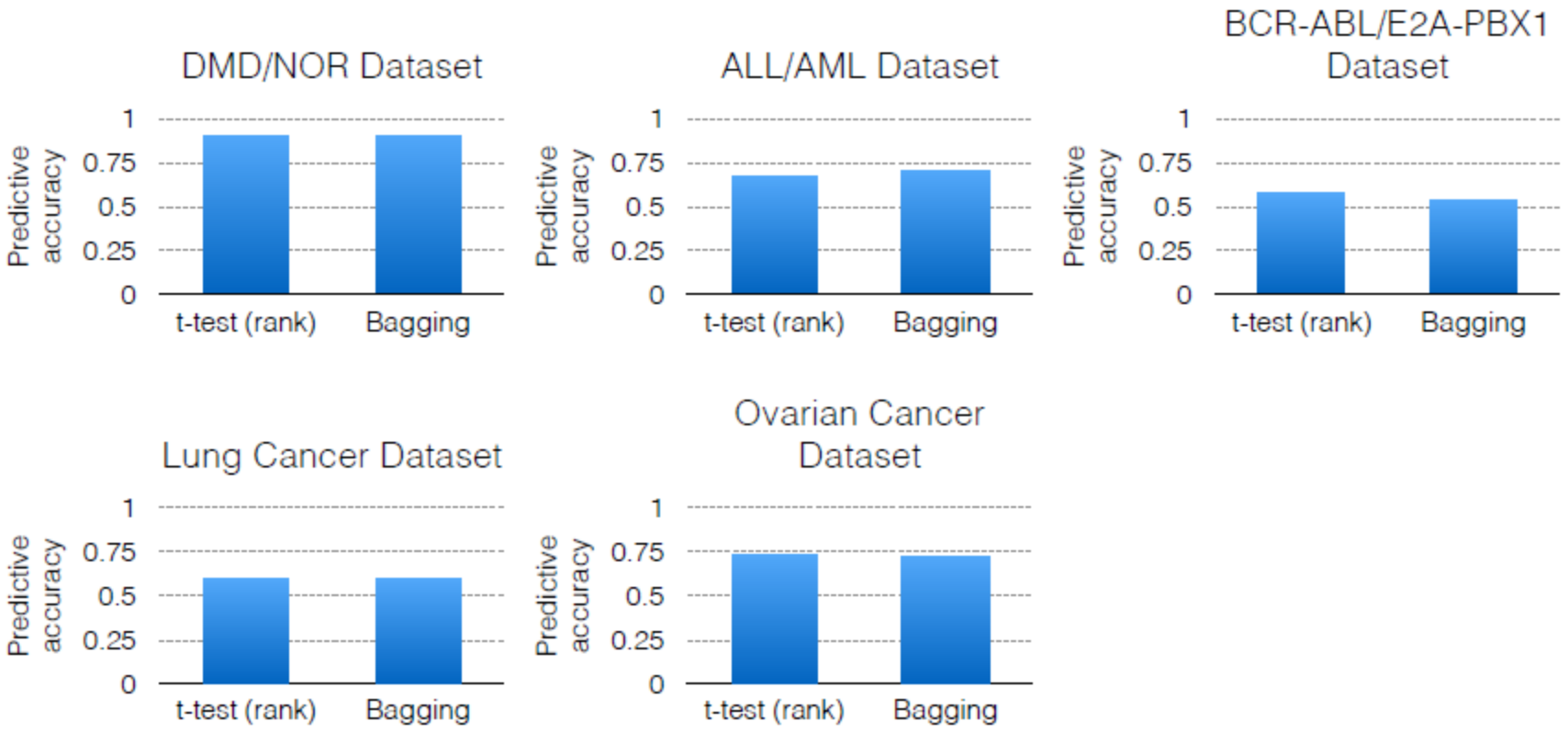


FIGURE 5.8: Predictive accuracy of gene-feature-based classifiers with and without rank normalization in the DMD/NOR dataset.

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



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

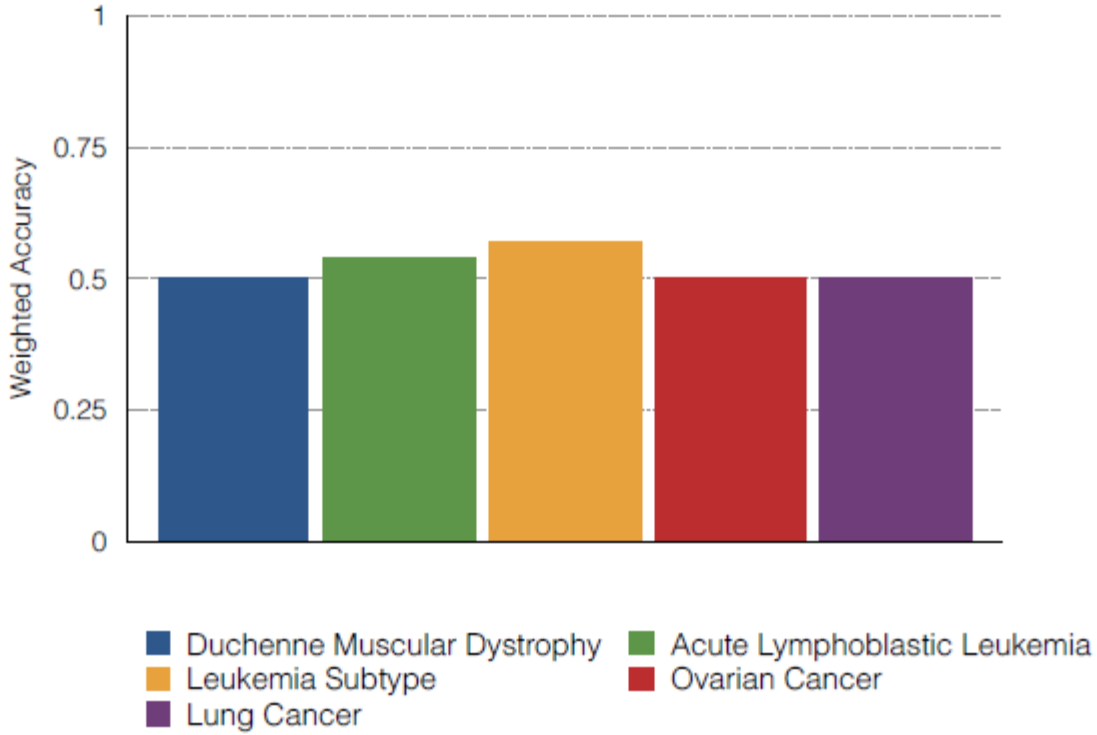


FIGURE 5.15: Predictive accuracy of gene-feature-based classifier using genes extracted from subnetworks in ESSNet; demonstrating that genes in the subnetworks by themselves are not a good discriminator for classification

NUS National University of Singapore

So new ideas to better use subnetwork-based features for successful cross-batch classification is needed…

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:

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

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

- **Straightforward to use these scores (and their paired difference) as features**

Successfully reducing batch effects

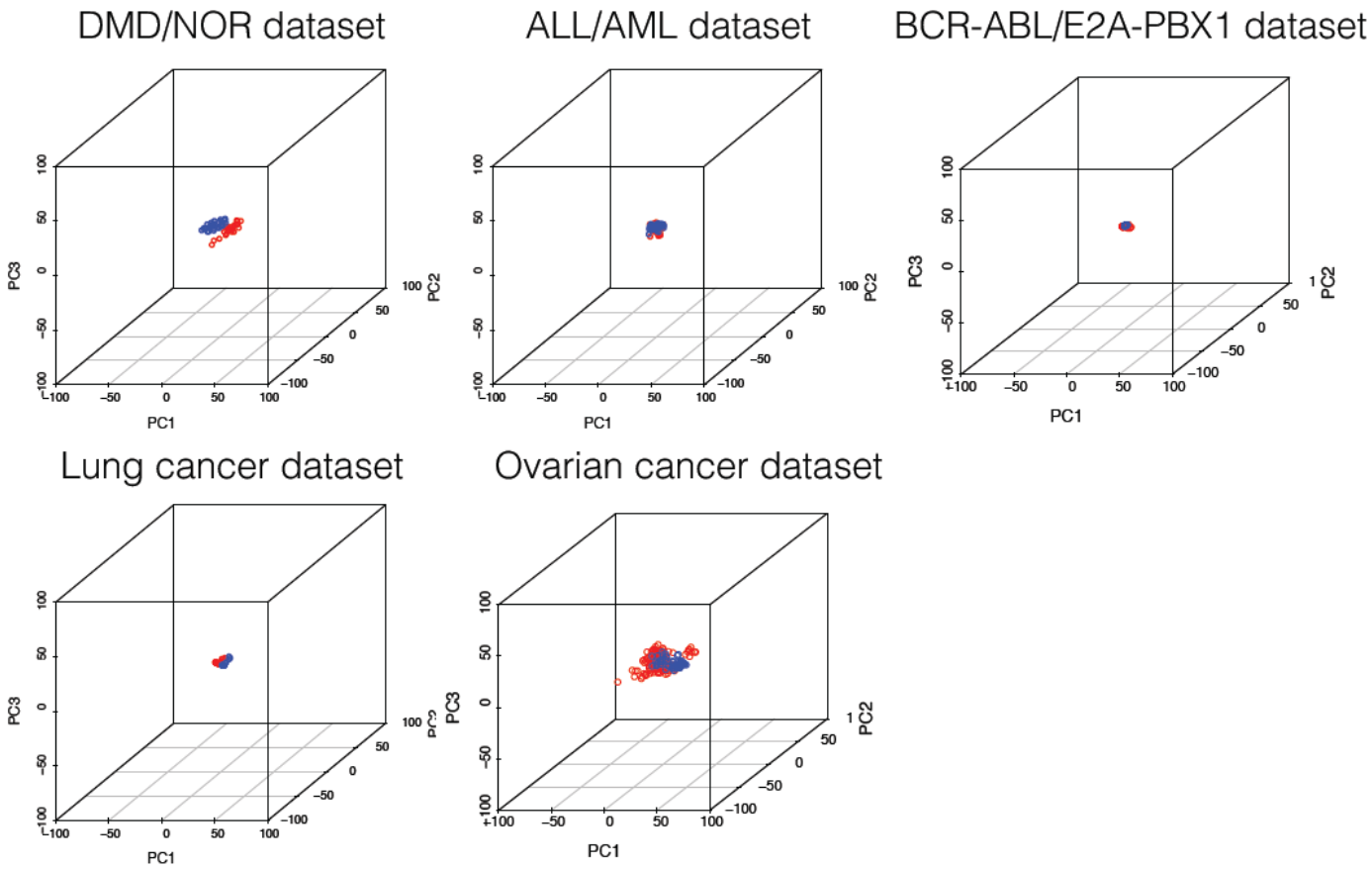


FIGURE 5.6: A figure showing that the batch effects are reduced by PFSNet subnetwork features. The colors red and blue represent different batches.

Successful cross-batch classification

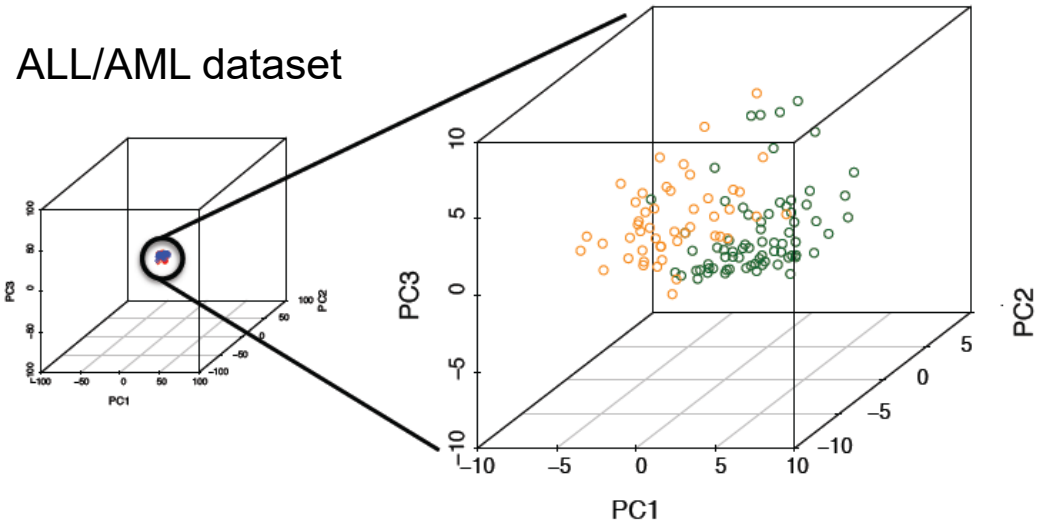


FIGURE 5.7: A figure showing that data points are separated by class labels instead of batch when PFSNet features are used. The colors green and orange represent different classes.

NUS National University of Singapore

How about cross-batch classification 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 $\Delta_{i,j,k} = Expr(g_i,p_j) – Expr(g_i,q_k)$
- 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 $\Delta(D)(S,P_x)$ and $\Delta(\neg D)(S,P_x)$ 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 $\Delta(D)(S,P_x)$ and $\Delta(\neg D)(S,P_x)$ 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**

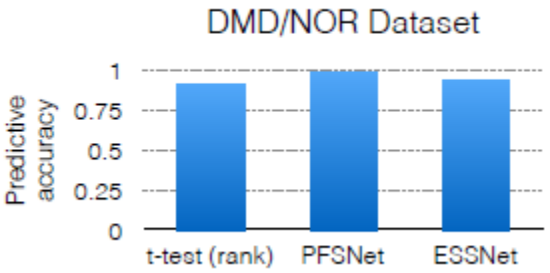
$$\Delta_{(D--\neg D)}(S) = \{e_{g_i,p'} - e_{g_i,p''} \mid g_i \in S \text{ and } p' \in D \text{ and } p'' \in \neg D\}$$

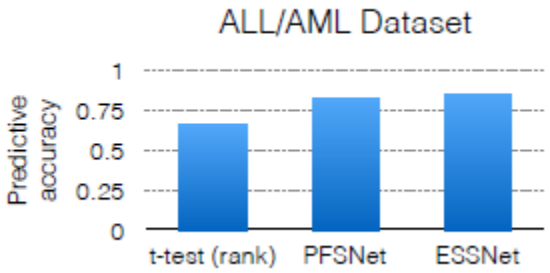
Similarly for $\Delta_{(\neg D - \neg D)}(S)$, $\Delta_{(\neg D - D)}(S)$, $\Delta_{(D - D)}(S)$

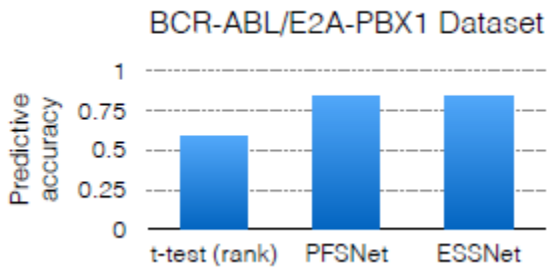
- **This gives 4 more features**

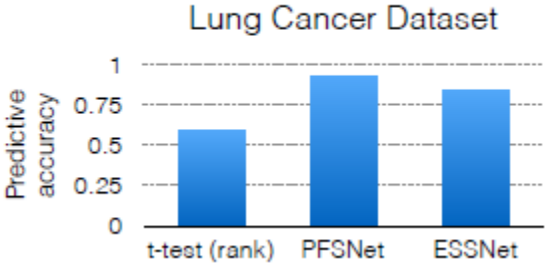
$$ESSNet\_feature_7^{p_x,S} = T\_statistic(\Delta_{(\neg D)}(S,p_x), \Delta_{(D--\neg D)}(S))$$
$$ESSNet\_feature_8^{p_x,S} = T\_statistic(\Delta_{(\neg D)}(S,p_x), \Delta_{(\neg D--\neg D)}(S))$$
$$ESSNet\_feature_9^{p_x,S} = T\_statistic(\Delta_{(D)}(S,p_x), \Delta_{(D-D)}(S))$$
$$ESSNet\_feature_{10}^{p_x,S} = T\_statistic(\Delta_{(D)}(S,p_x), \Delta_{(\neg D-D)}(S))$$

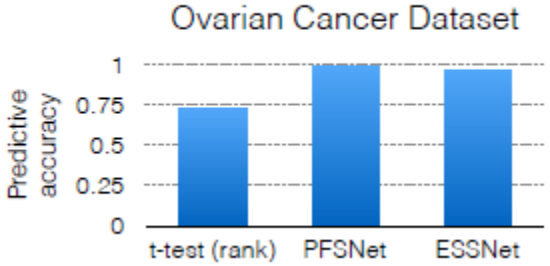
ESSNet-based features lead to high cross-batch classification accuracy











ESSNet-based cross-batch hierarchical clusterings

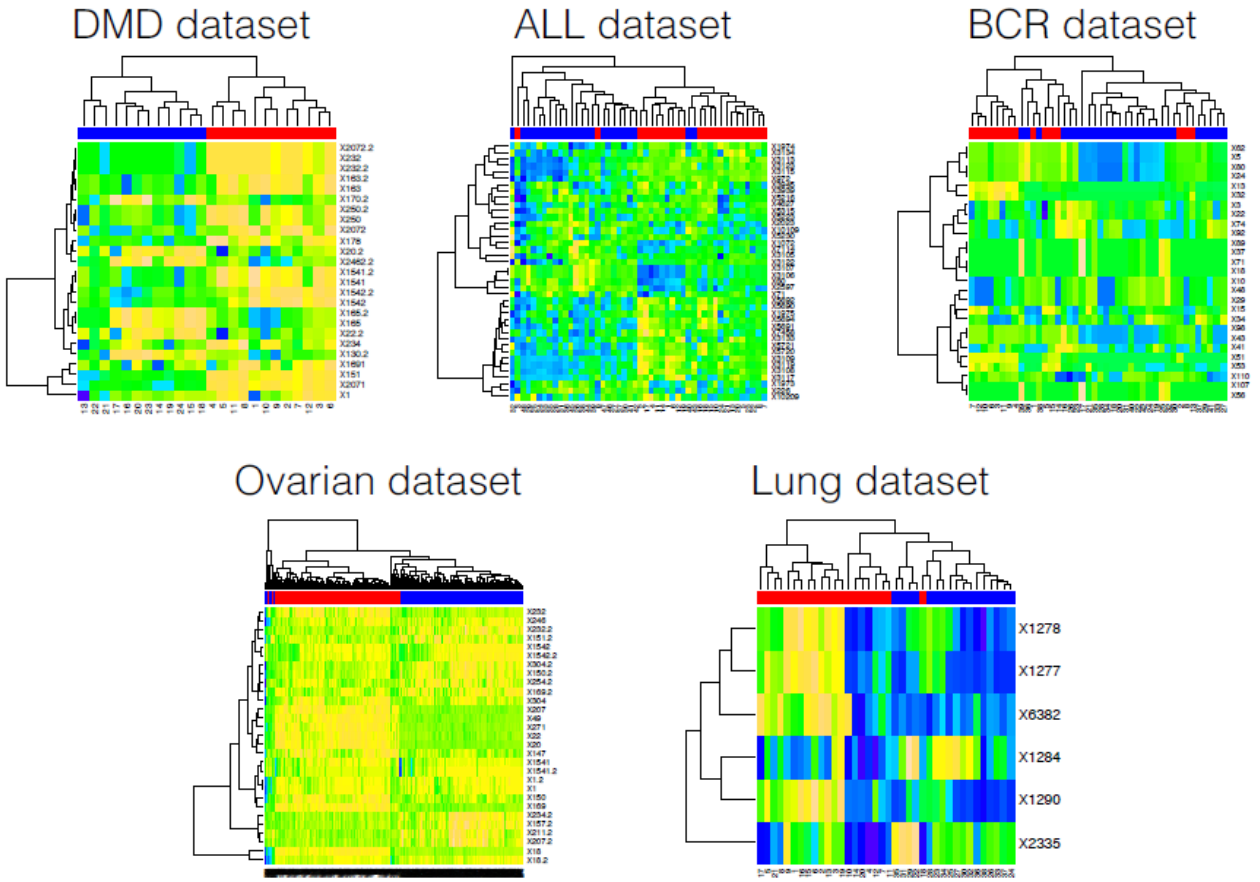
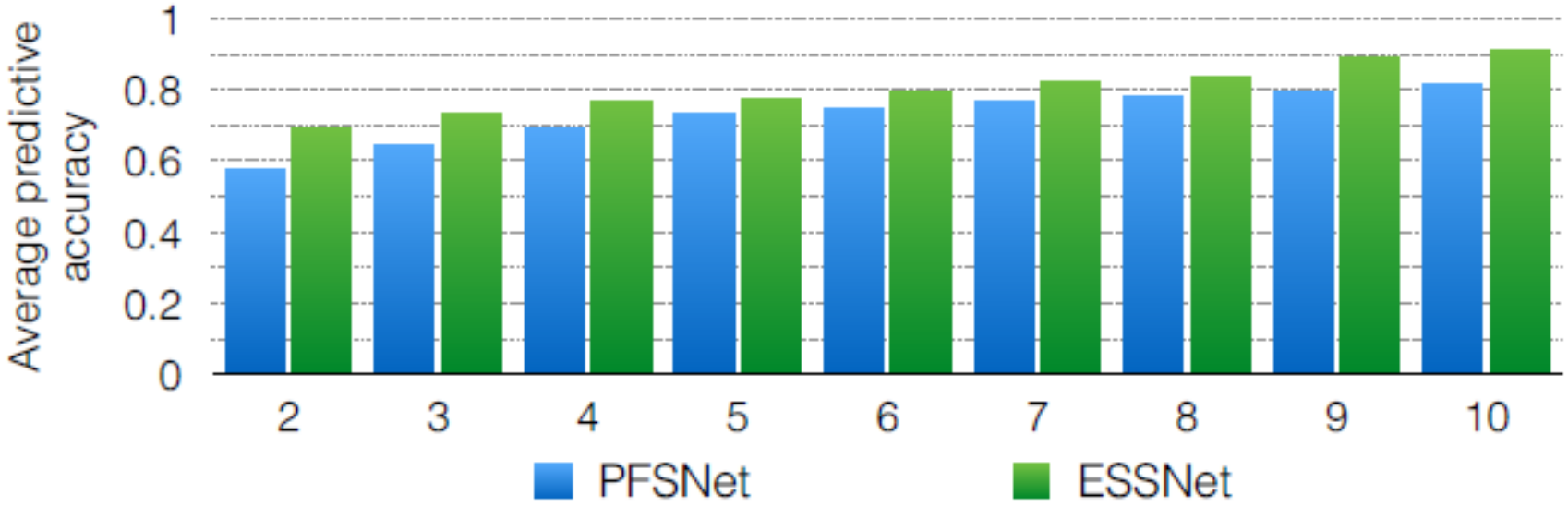


FIGURE 5.17: A figure depicting heirarchical clustering performed on the patient's subnetwork scores.

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



NUS National University of Singapore

CONCLUDING REMARKS

What have we learned?

- **Batch correction can introduce false effects into data; Use with care**
- **Rank fuzzification is a useful normalization method**
- **PCA is not just a visualization tool; it can also be used for dealing with batch effects**
- **Protein complex-based feature-selection is batch-resistant; can deal with batch-related issues without requiring batch correction**

Must read

- Goh et al., "Why batch effects matter in omics data, and how to avoid them", *Trends in Biotechnology*, 35(6):498-507, 2017
- Belorkar & Wong, "GFS: Fuzzy preprocessing for effective gene expression analysis", *BMC Bioinformatics*, 17(S17):1327, 2016
- Koh & Wong, "Embracing noise to improve cross-batch prediction accuracy", *BMC Systems Biology*, 6(Suppl 2):S3, 2012
- Goh & Wong, "Protein complex-based analysis is resistant to the obfuscating consequences of batch effects", *BMC Genomics*, 18(Suppl 2):142, 2017

Acknowledgements

Ah Fu

Kevin

Wilson

Abha

- **Much of this lecture is based on the works of my past/current students**
  - Koh Chuan Hock (Ah Fu)
  - Kevin Lim
  - Wilson Goh
  - Abha Belorkar

Remember this chart from the proteomics lecture, slide #54