# Robust gene expression and proteomic profile analysis in the presence of undeclared heterogeneity

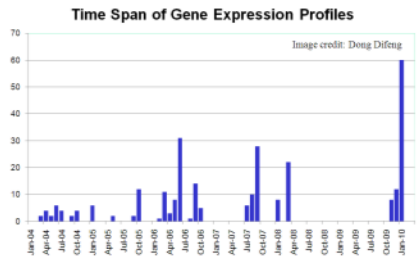
**Limsoon Wong**

**This talk is mainly based on the work of my (ex-)students Abha Belorkar and Wilson Goh**

# Undeclared heterogeneity

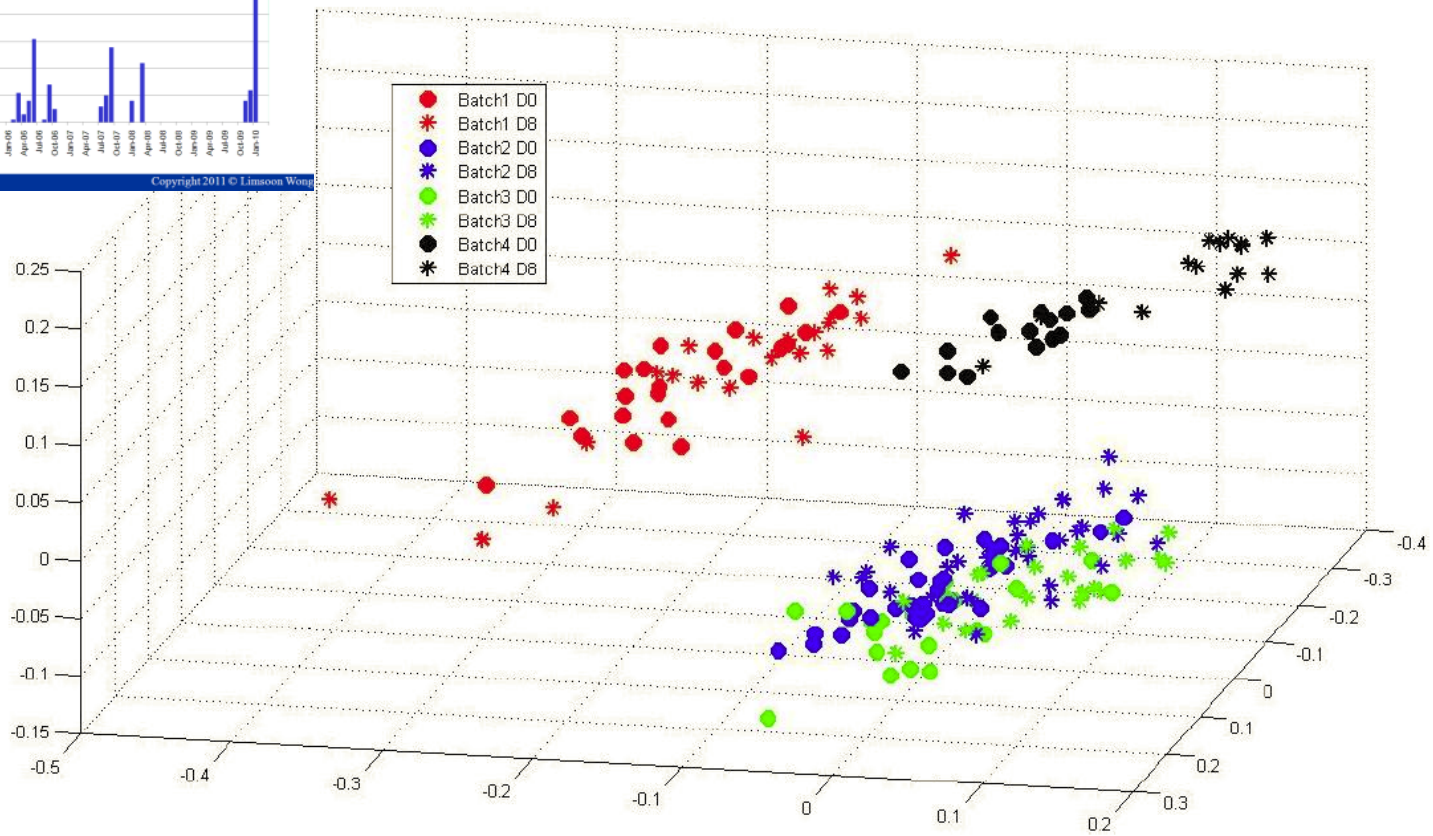
- **Batch effects**
  - Batch effects are unwanted sources of variation caused by different processing date, handling personnel, reagent lots, equipment/machines, etc.
- **Undeclared subtypes**
  - Disease has subtypes but these are not labelled
- **Undeclared heterogeneity is a big challenge faced in biological research, especially towards translational research and precision medicine**

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

# Plan

- **Problems with common normalization methods**
- **A better normalization method: GFS**
- **Batch effect-resistant feature selection built on top of GFS: SNET/FSNET/PFSNET**
- **Subpopulation-sensitive feature selection built on top of PFSNET: SPSNET**

NUS National University of Singapore

# COMMON NORMALIZATION METHODS

# Common normalization methods

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

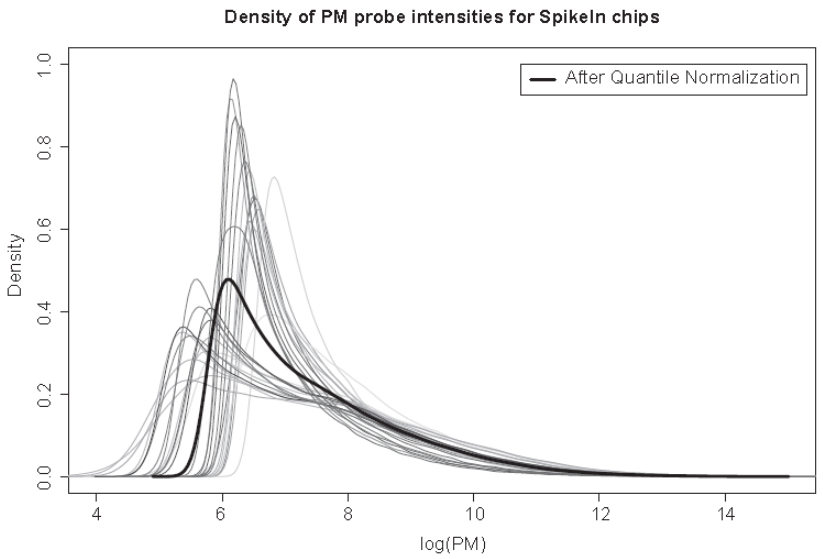
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

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

- **Quantile normalization**

# 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

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

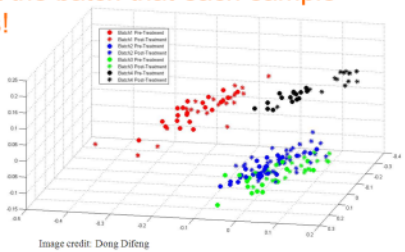


Image credit: Dong Difeng

⇒ Need normalization to correct for batch effect

# After quantile normalization



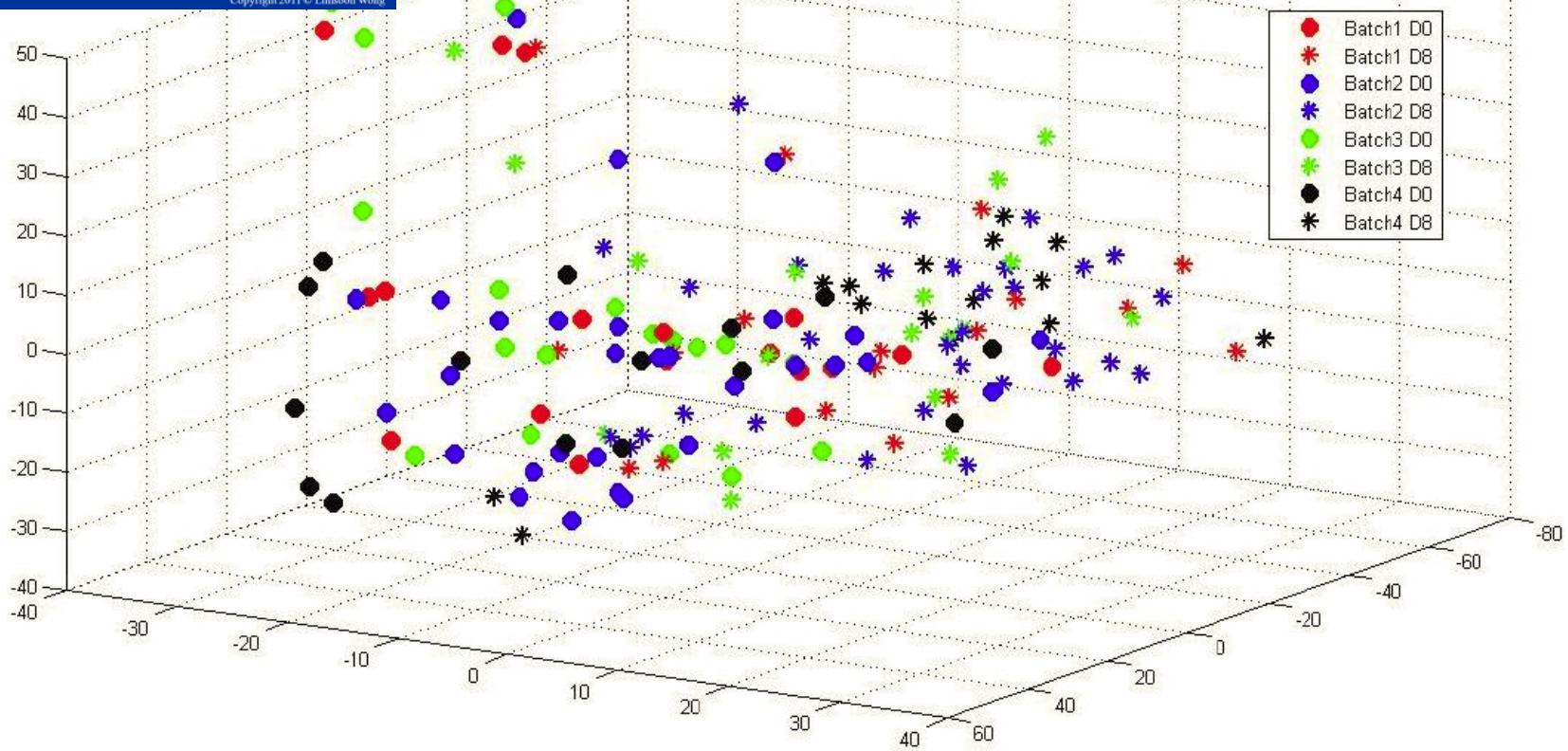
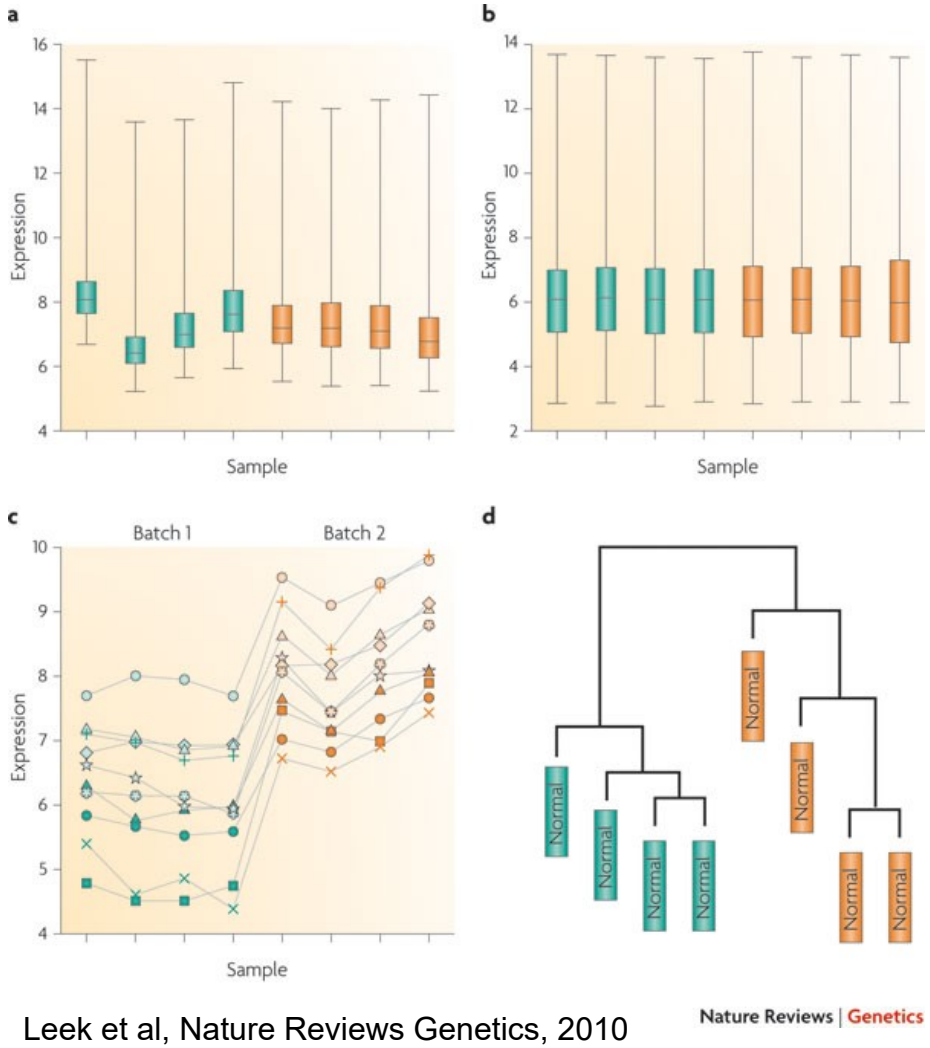



Image credit: Difeng Dong's PhD dissertation, 2011

# 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

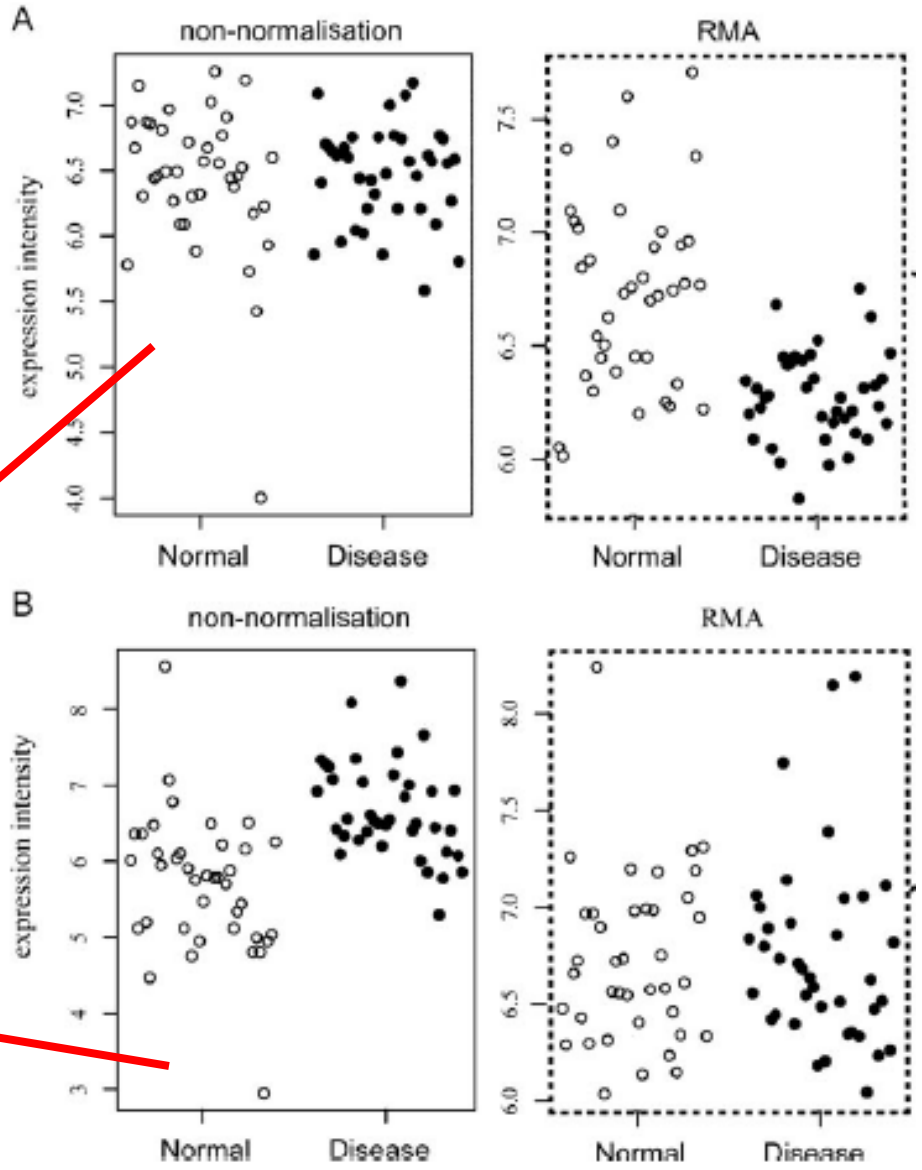
c: Checks on individual genes susceptible to batch effects

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

# Caution: "Over normalized" signals in cancer samples

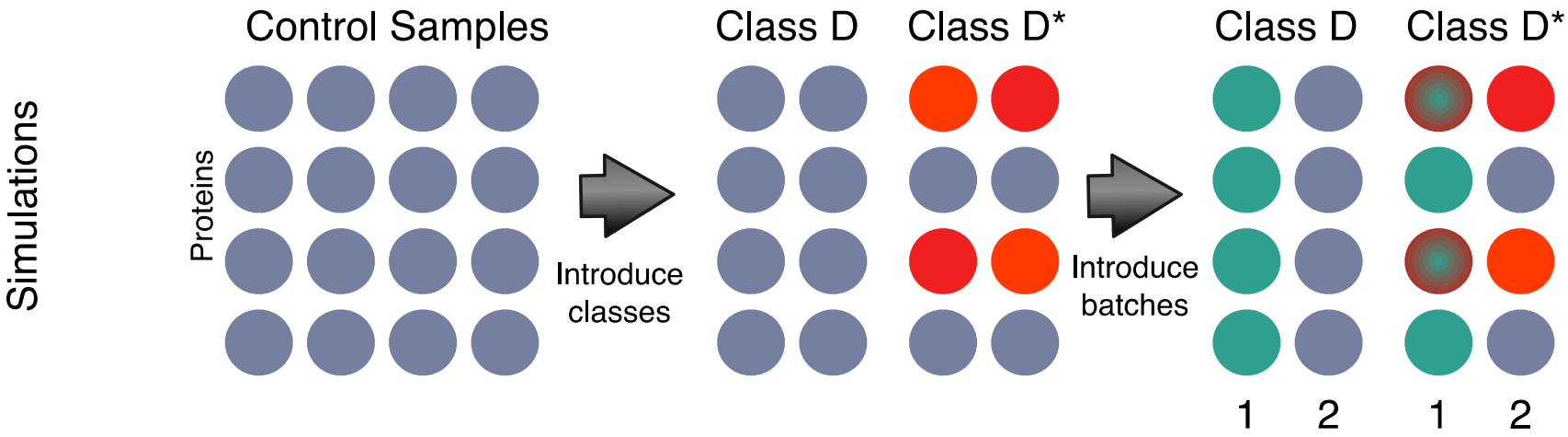
A gene normalized by quantile normalization (RMA) was detected as down-regulated DE gene, but the original probe intensities in cancer samples were not diff from those in normal samples

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



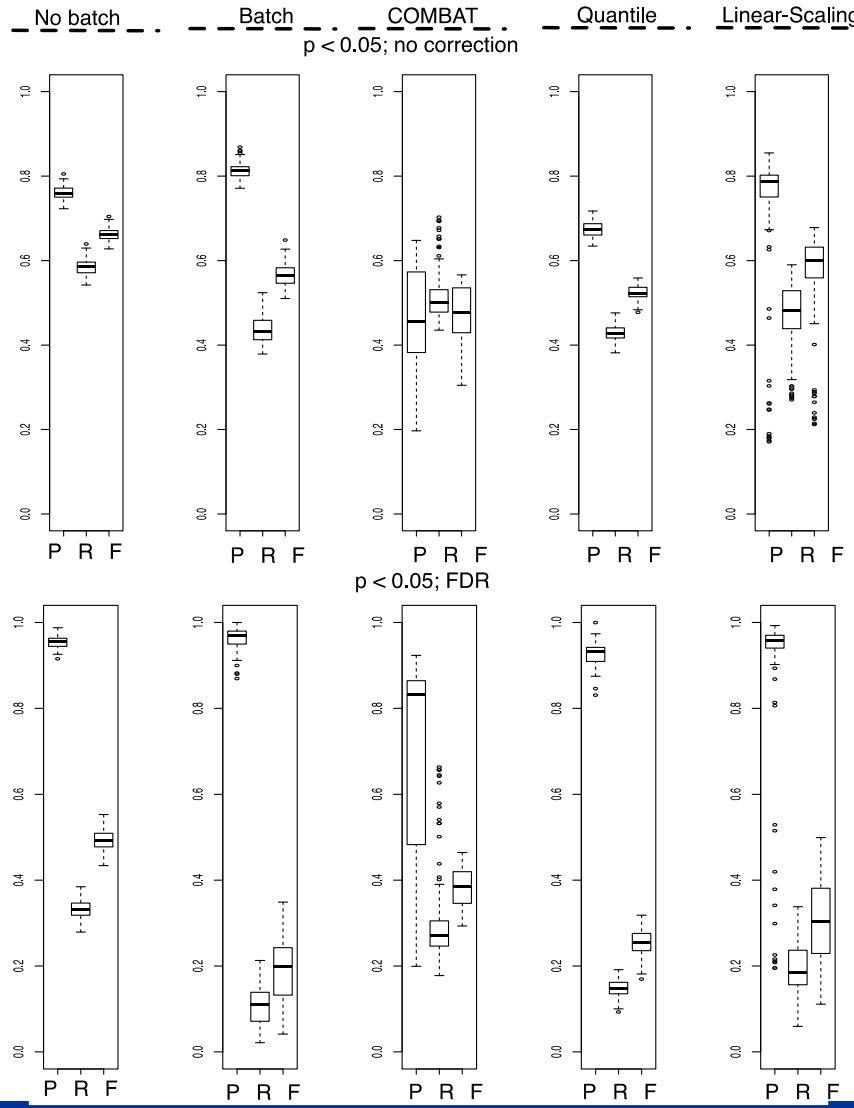

Wang et al. *Molecular Biosystems*, 8:818-827, 2012

# 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

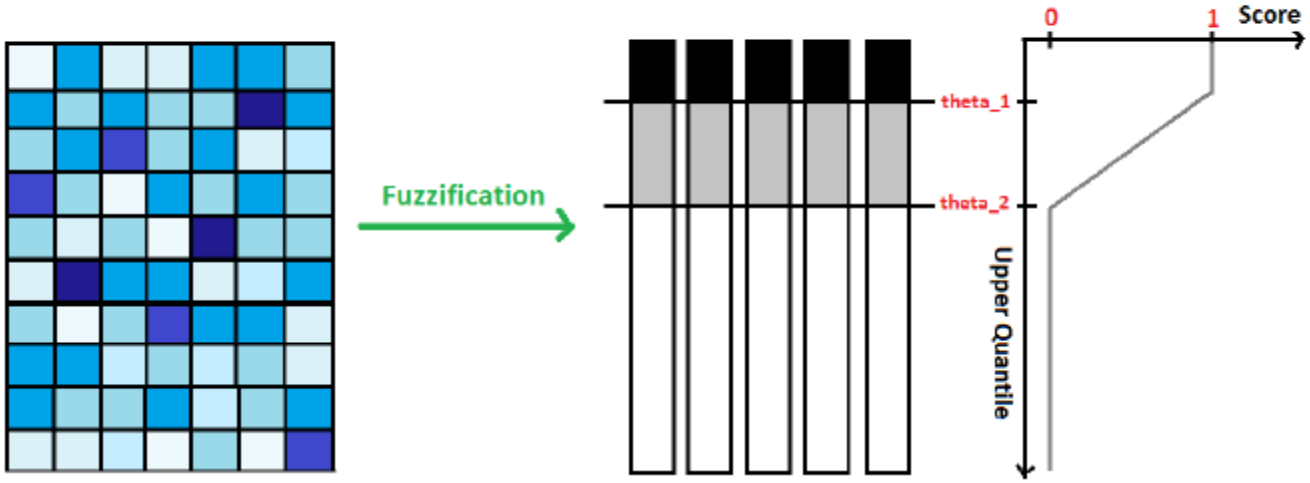
NUS National University of Singapore

# GENE FUZZY SCORE

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

# Gene fuzzy score (GFS)

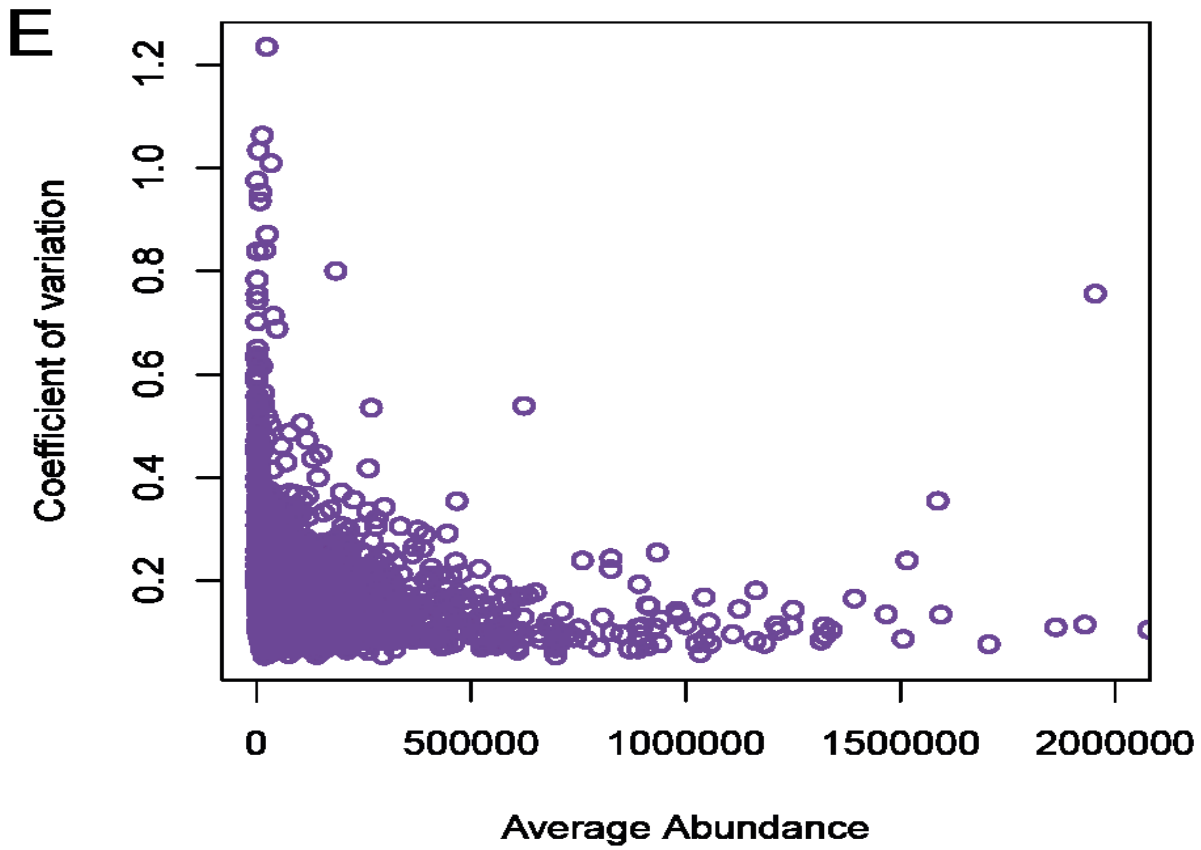
Raw gene expression → gene ranks within microarrays → fuzzified scores




- **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

NUS National University of Singapore

# Why low-expression genes are discarded in GFS



# Desirable characteristics of normalization methods

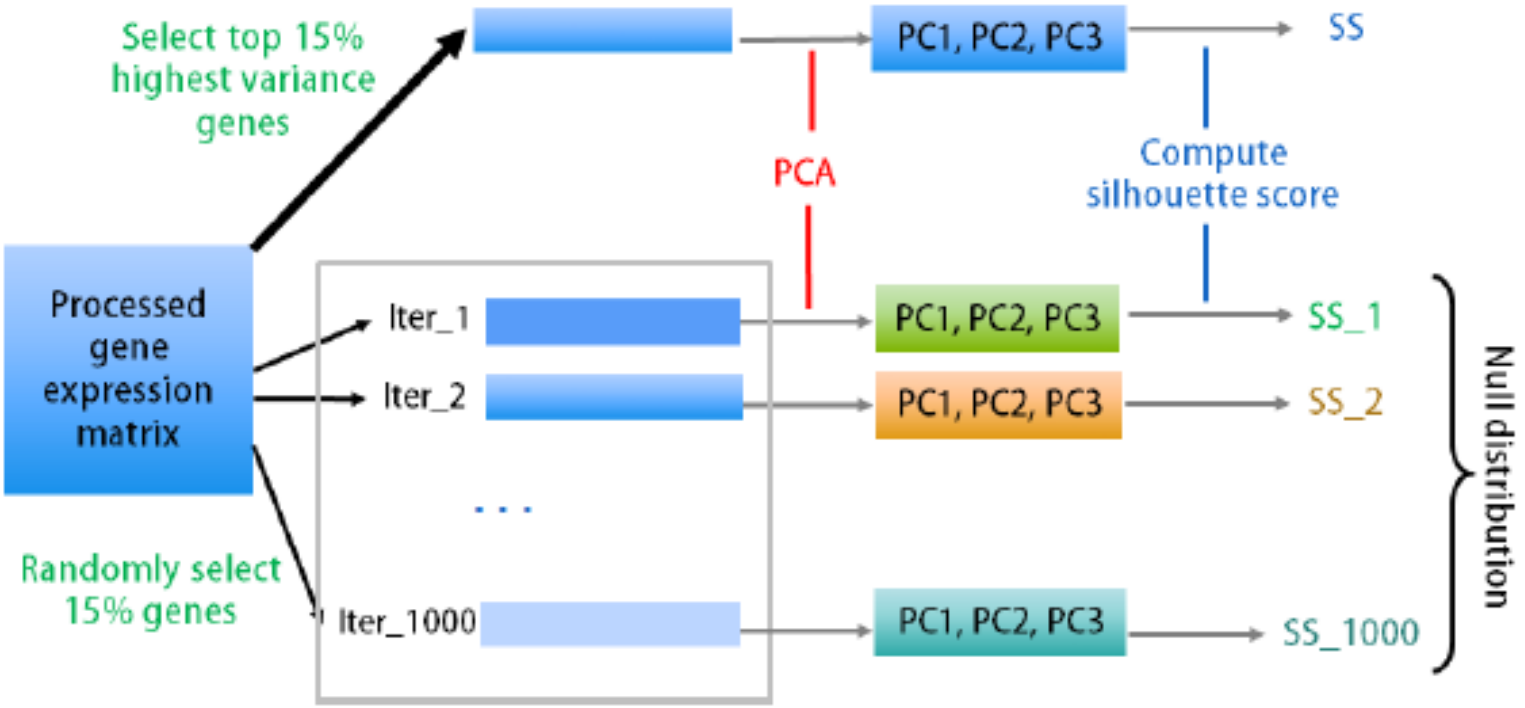
- **High quality**
  - The output of the method is useful in separating samples of different phenotypes from each other
- **High consistency**
  - When applied to any two representative batches of data, the overlap between high-variance features (e.g. genes) is high
- **High biological coherence**
 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

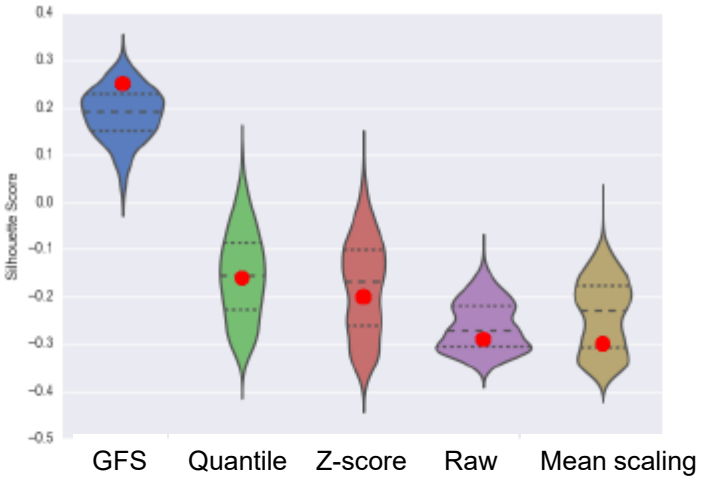
# Evaluating quality




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

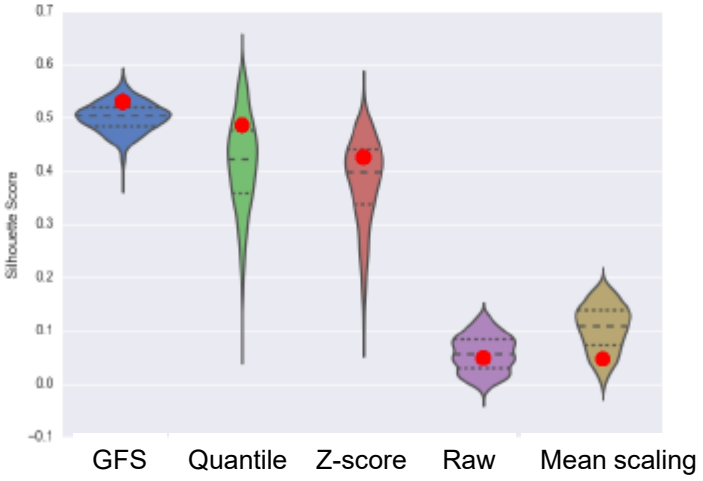
# Observations

- **The GFS null distribution is stable and has high silhouette score**



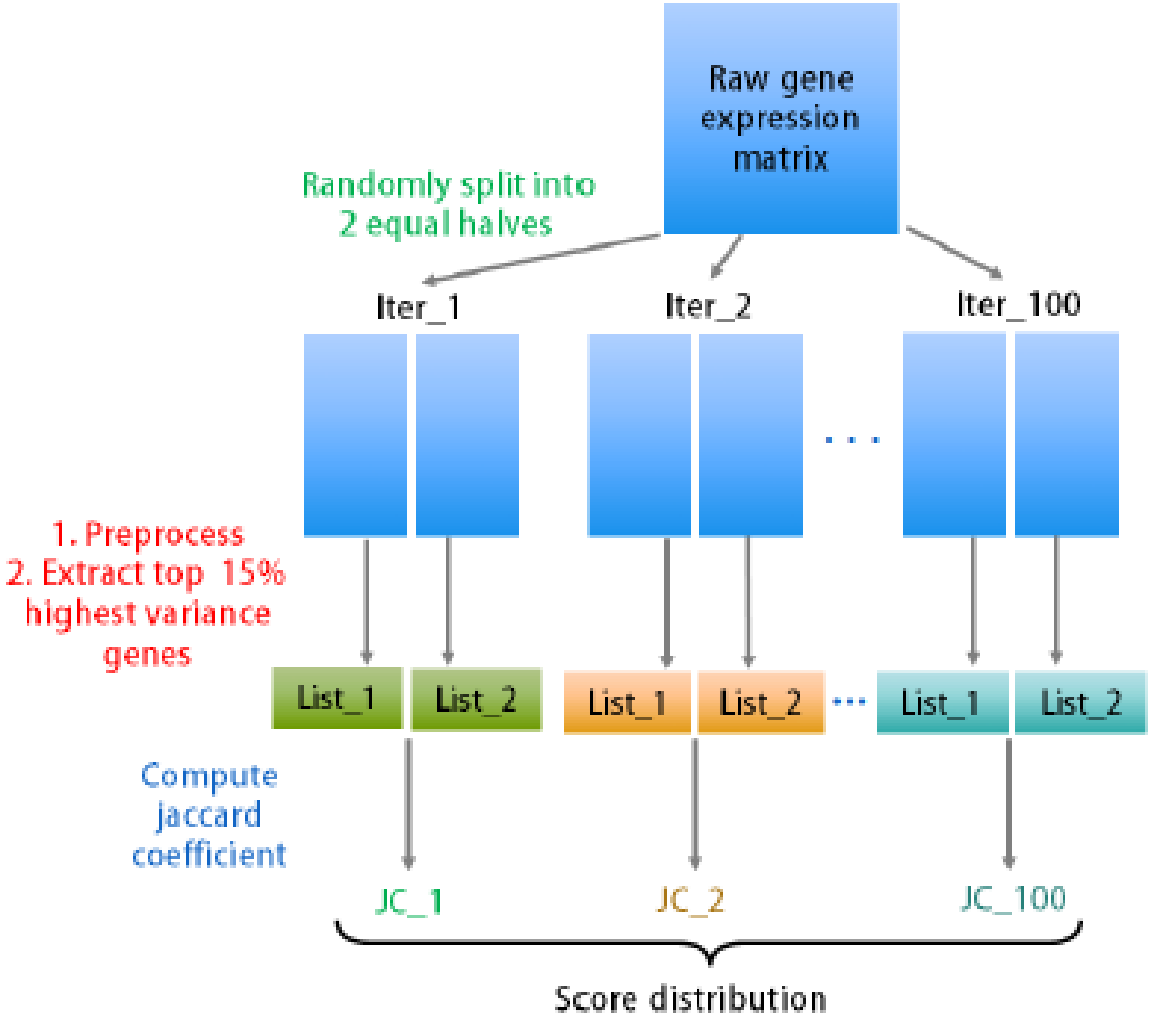

(a) Acute Lymphoblastic Leukemia (ALL)

- **For GFS, the score obtained from the top 15% highest variance genes is always greater than the score from the 95$^{th}$ percentile of the null distribution**




(b) Duschenne Muscular Dystrophy (DMD)

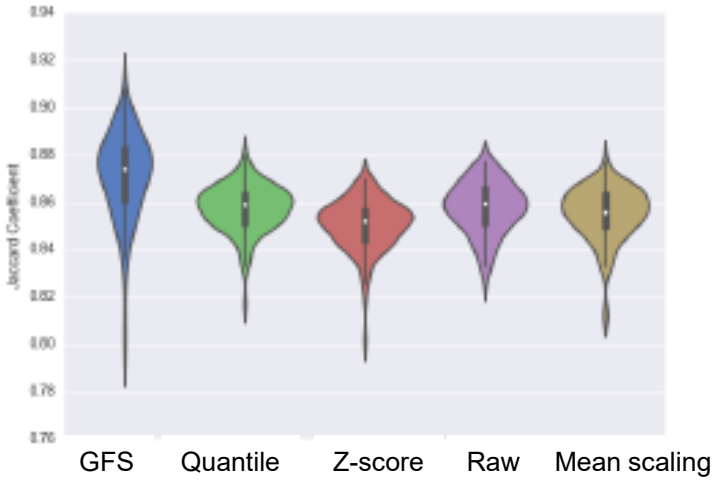
# Evaluating consistency



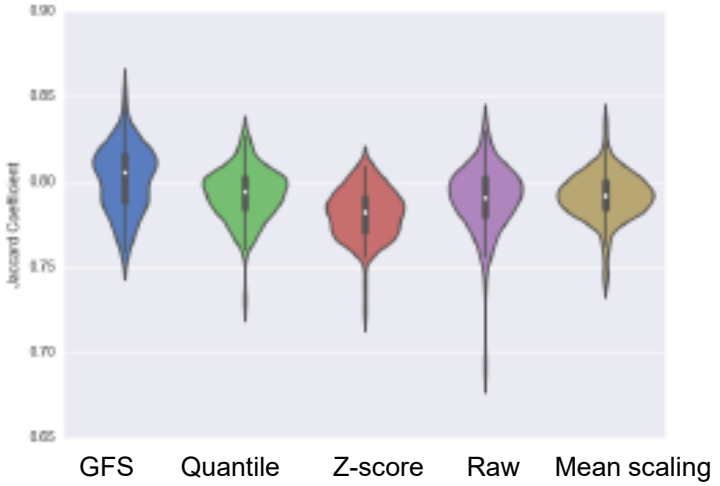

- **An idea method should produce a Jaccard coefficient distribution that is high and stable**

# Observations

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

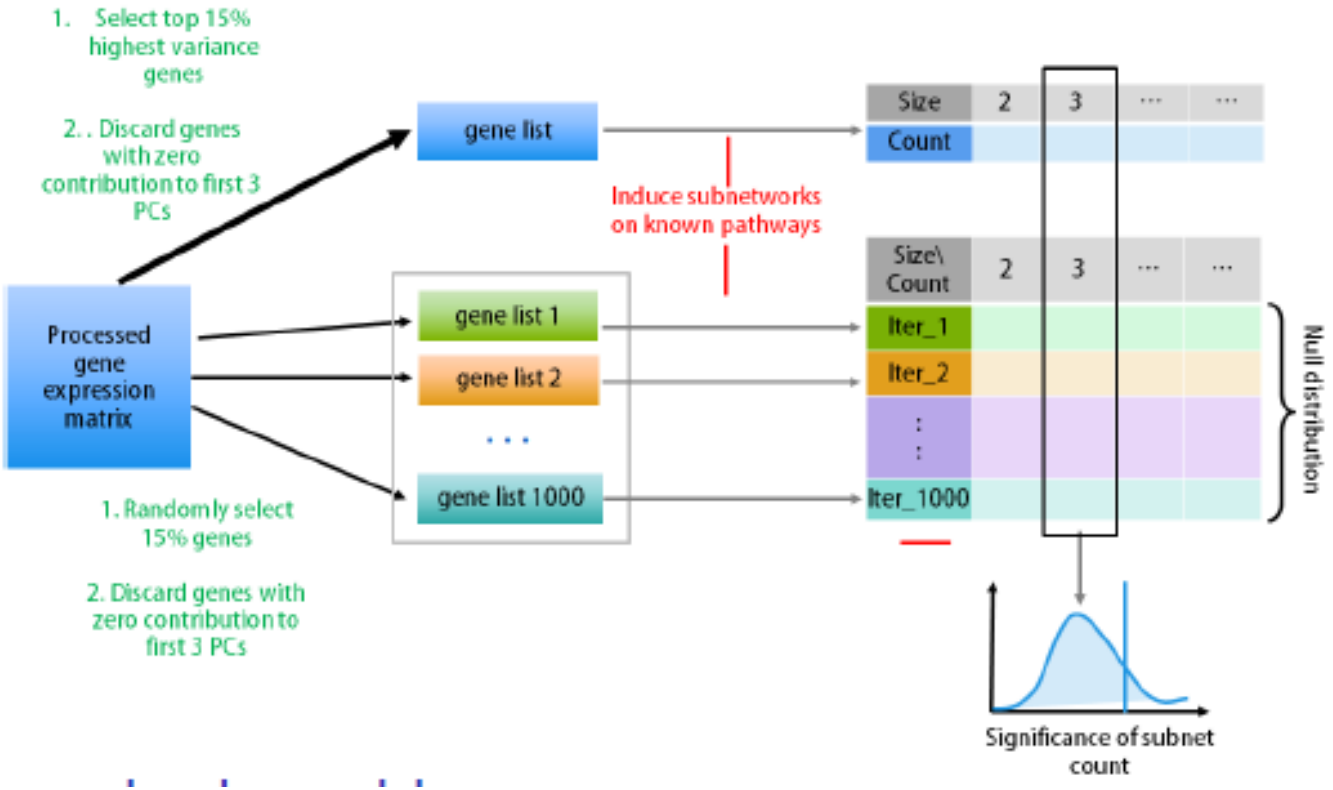



(a) Acute Lymphoblastic Leukemia (ALL)




(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)

NUS National University of Singapore

# BATCH EFFECT-RESISTANT FEATURE SELECTION

Goh & Wong, “Protein complex-based analysis is resistant to the obfuscating consequences of batch effects”, *BMC Genomics*, 18(Suppl 2):142, 2017

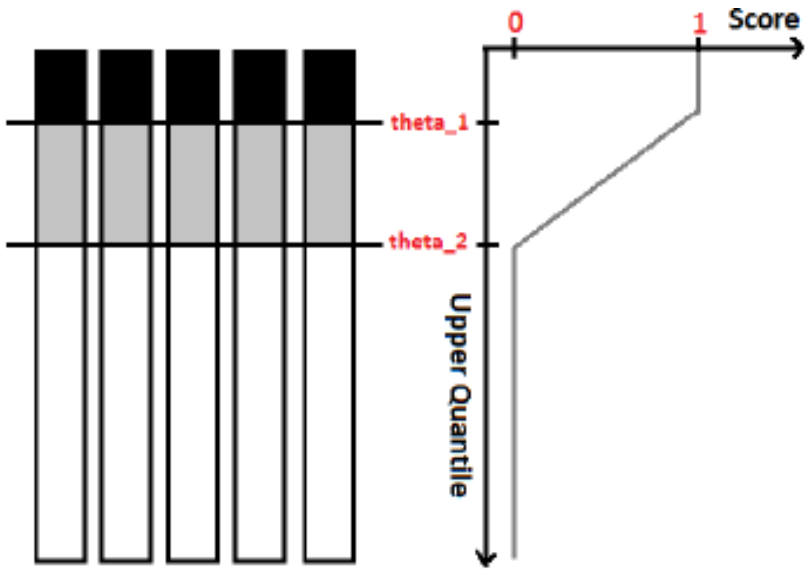
# What if class and batch effects are strongly confounded?

- **Batch-effect correction does not work well**
- **Inadvertently lose info on disease subpopulations (which look like batch effects but are meaningful)**

⇒ **Consider batch effect-resistant methods instead of batch removal**

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

# FSNET




- **β(g,C)**
  - Proportion of tissues in class C that have protein g among their most-abundant proteins
- **Score(S,p,C)**
  - Score of protein complex S and tissue p weighted based on class C
- **$f_{SNET}$(S,X,Y,C)**
  - Complex S is differentially high in sample set X and low in sample set Y, weighted based on class C, when $f_{SNET}$(S,X,Y,C) is at largest extreme of t-distribution

$$\beta(g_i, C_j) = \sum_{pk \epsilon c_j} \frac{fs(g_i, p_k)}{|C_j|}$$

$$\text{score}(S, p_k, C_j) = \sum_{g_i \in S} fs(g_i, p_k) * \beta(g_i, C_j)$$

$$f_{\text{SNET}}(S, X, Y, C_j) = \frac{\text{mean}(S, X, C_j) - \text{mean}(S, Y, C_j)}{\sqrt{\frac{\text{var}(S,X,C_j)}{|X|} + \frac{\text{var}(S,Y,C_j)}{|Y|}}}$$

# SNET and PFSNET

- **SNET**
  - Predecessor of FSNET
  - fs(g,p) is set to 1 if protein g is in top theta1% most abundant proteins in tissue p

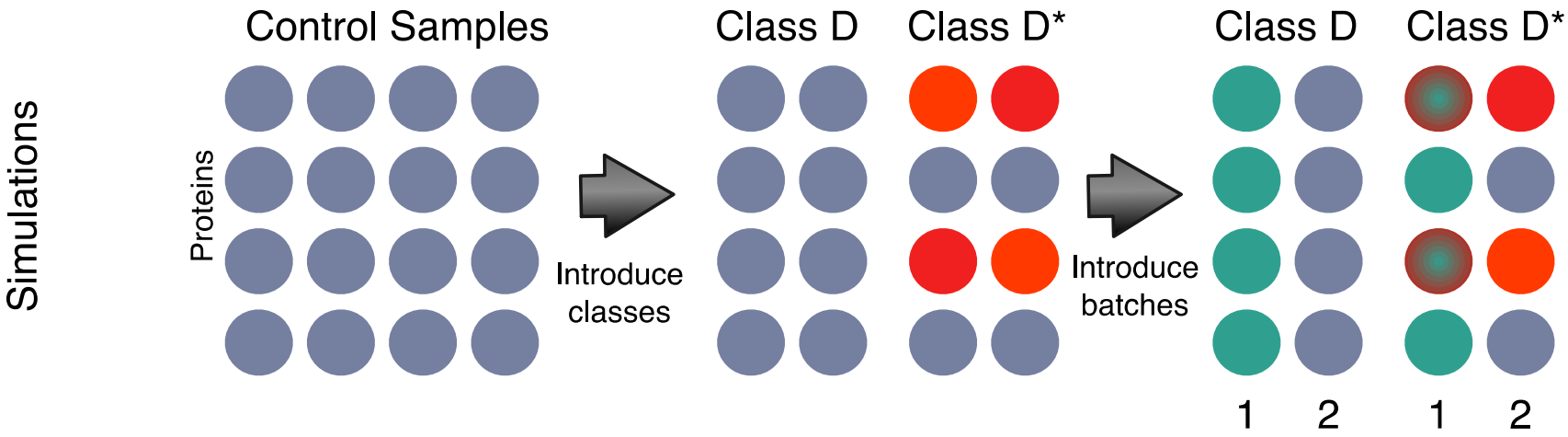
- **PFSNET**
  - Successor of FSNET
  - delta(S, p, X, Y) = score(S, p, X) –score(S, p, Y)
  - $$f_{\mathrm{PFSENT}}(S,X,Y,Z) = \frac{\mathrm{mean}(S,X,Y,Z)}{\mathrm{se}(S,X,Y,Z)}, \quad (7)$$

where $\mathrm{mean}(S,X,Y,Z)$ and $\mathrm{se}(S,X,Y,Z)$ are respectively the mean and standard error of the list $\{\mathrm{delta}(S,p_k,X,Y)|p_k$ is a tissue in $Z\}$. The complex $S$ is considered significantly consistently highly abundant in $X$ but not in $Y$ if $f_{\mathrm{PFSNet}}(S,X,Y,X\cup Y)$ is at the largest 5% extreme of the Student $t$-distribution.

# Comparison with popular feature-selection methods

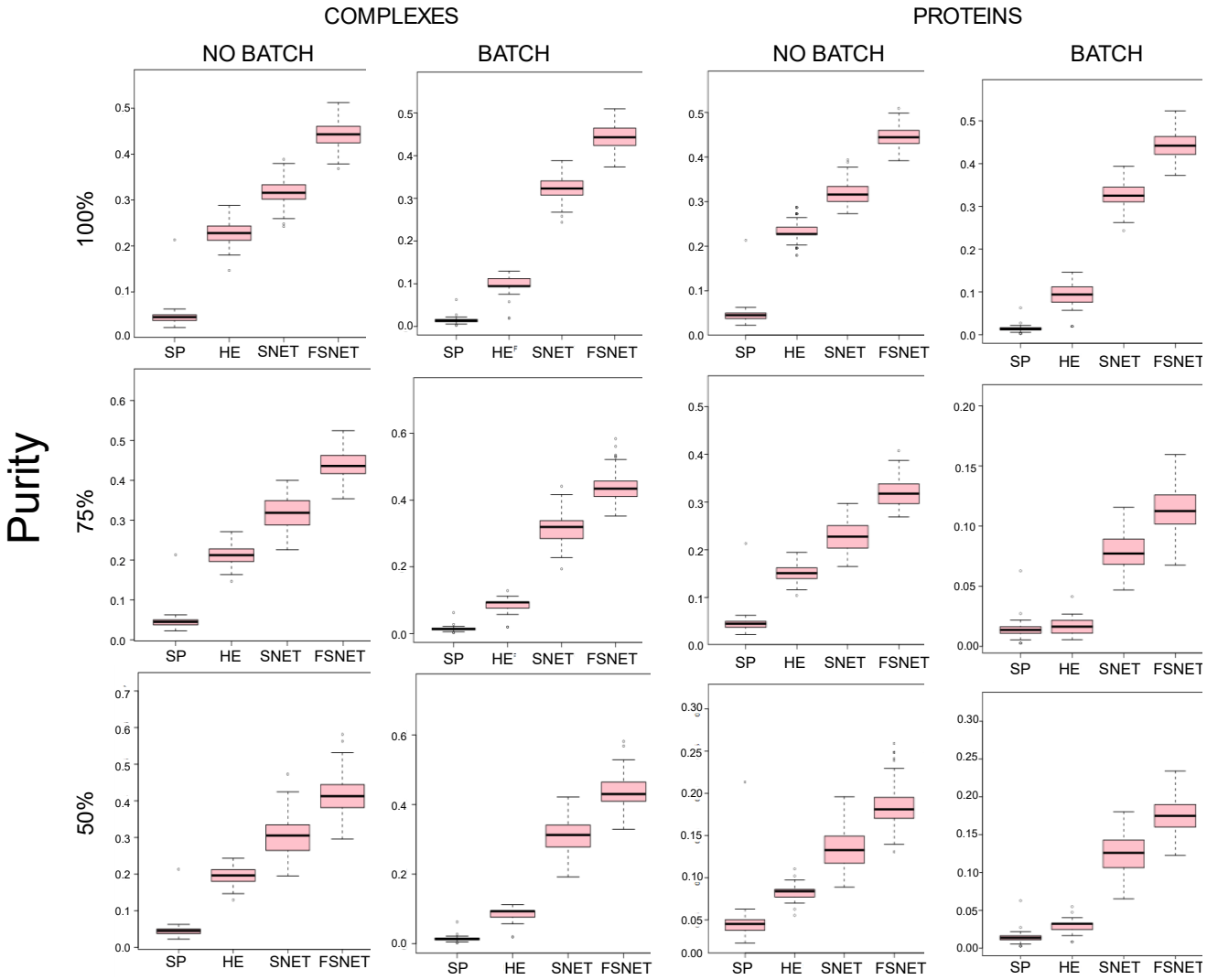
- **SP is the protein-based two-sample t-test**
- **HE is a two-step procedure deploying SP first, followed by the Fisher's exact test on networks**
- **Significant artificial complexes are constructed with various level of purity (i.e. proportion of significant proteins in the complex). Equal # of non-significant 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)

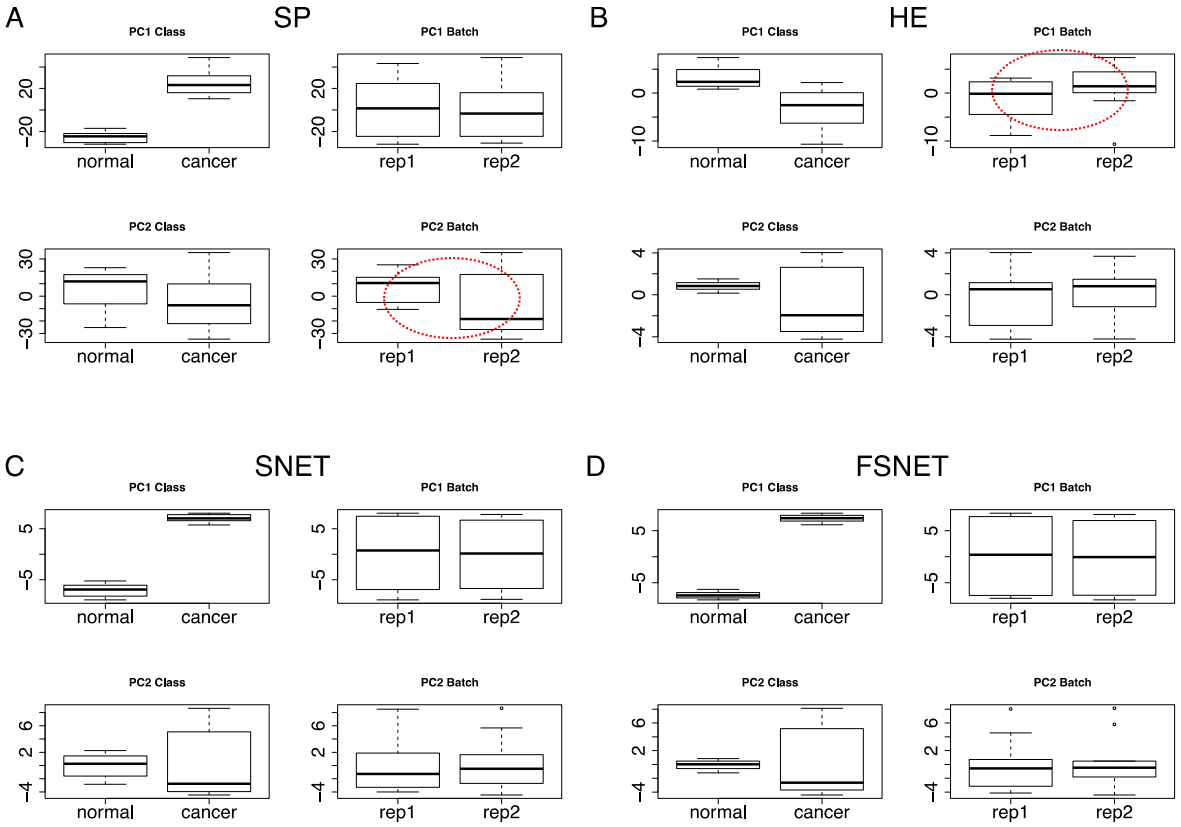



F-score distributions SNET and FSNET is robust against batch effects relative to traditional methods e.g. SP and HE

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)

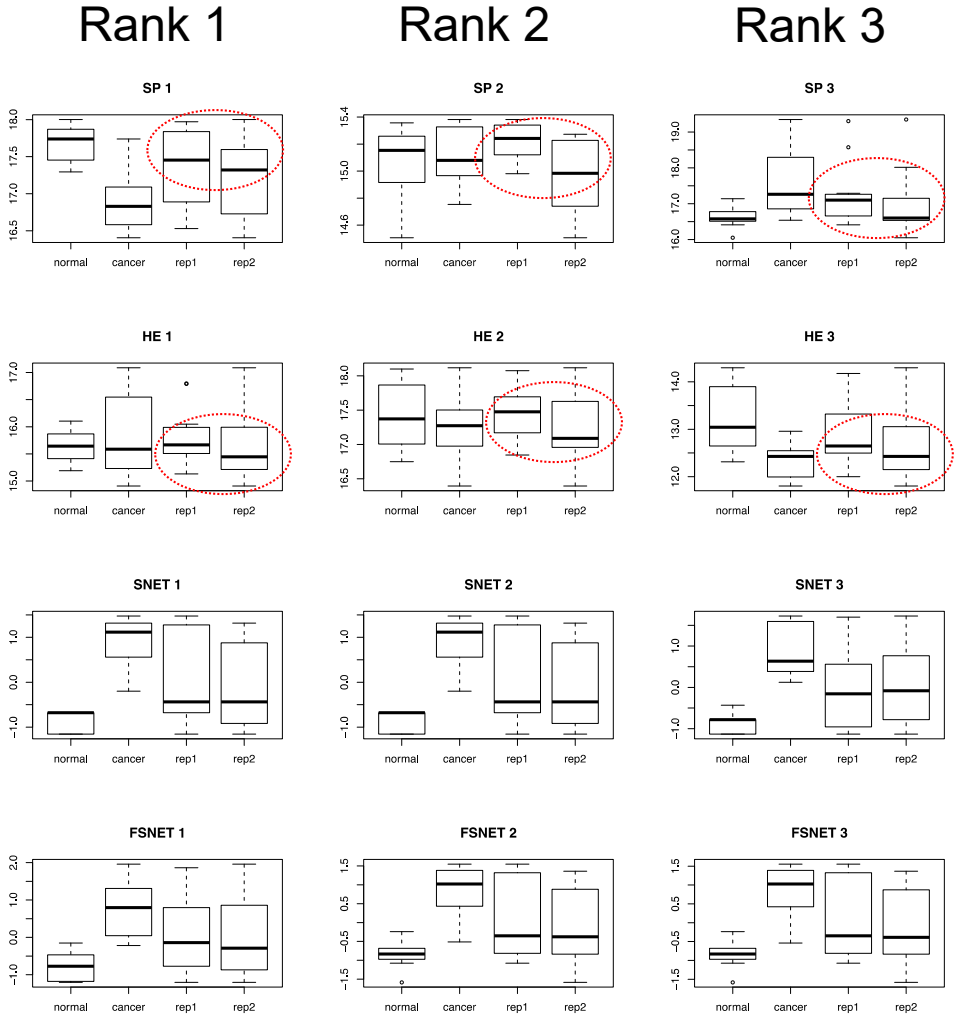



Protein complexes used as reference

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

NUS National University of Singapore

# SUBPOPULATION-SENSITIVE FEATURE SELECTION

Belorkar, Vadigepalli, Wong, “SPSNet: Subpopulation-sensitive network-based analysis of heterogeneous gene expression data”, manuscript, 2017

- **While SNET, FSNET, PFSNET, etc. are batch-effect resistant, they are design for feature-selection from homogeneous phenotypes**

- **They loses sensitivity when the phenotypes are heterogeneous subpopulations**

NUS National University of Singapore

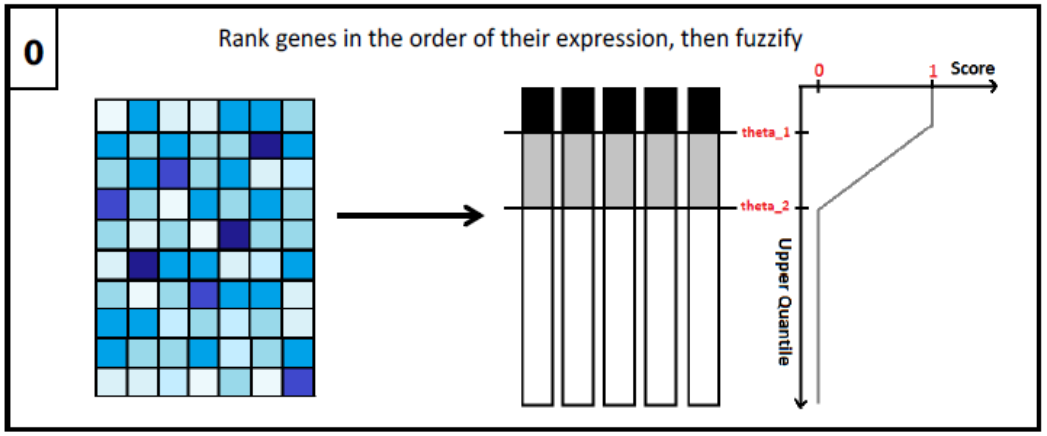
## Hypothesis

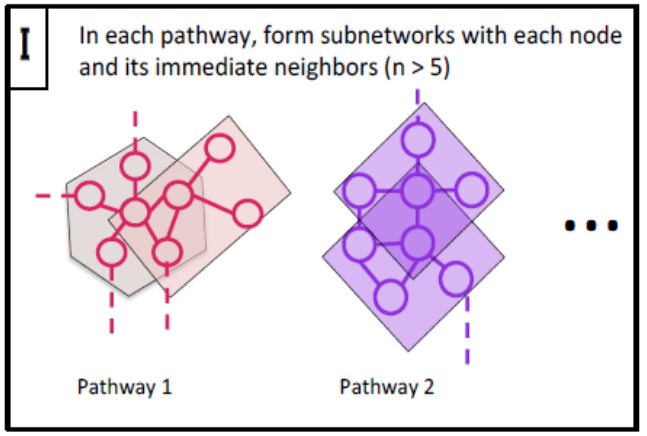
- **Each subpopulation in a heterogeneous dataset should "uniquely dominate" a few subnetworks**
  - Subpopulation X dominates subnetwork Y if genes from Y are highly expressed in subjects in X
  - Subpopulation X uniquely dominates subnetwork Y if X dominates Y, and no other subpopulation dominates Y

## Idea

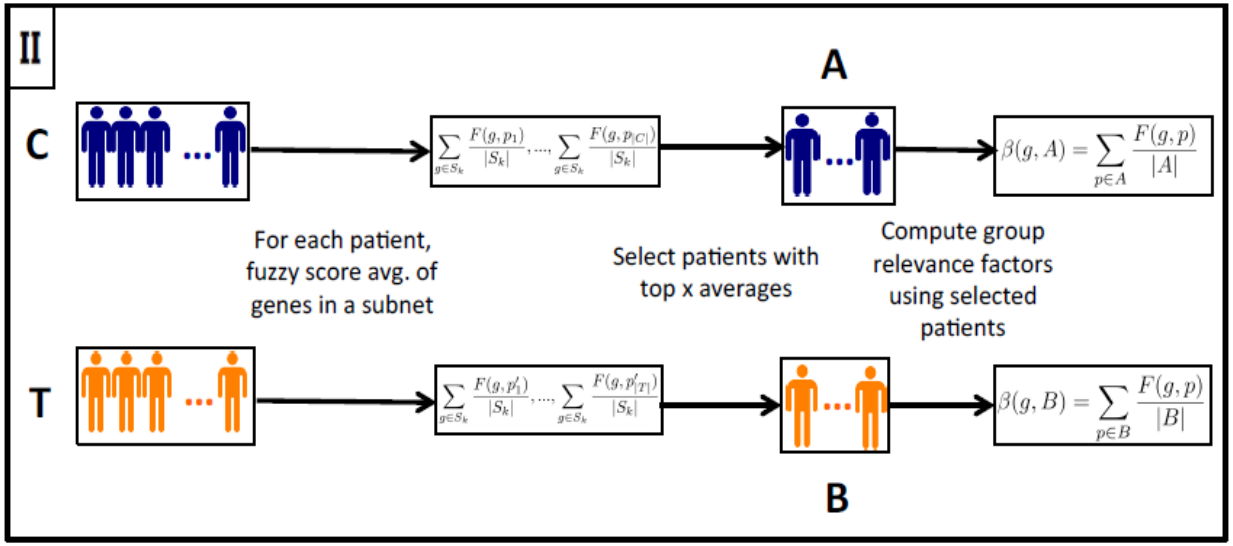
- **For a subnetwork, use the top n subjects as a reference for an undeclared subtype. Then run PFSNET on this subnetwork using this reference**

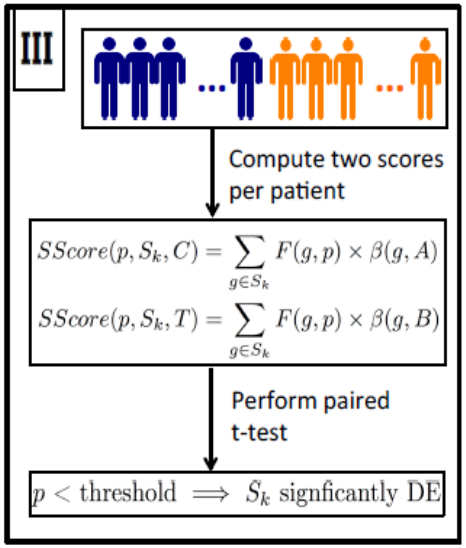
# SPSNet



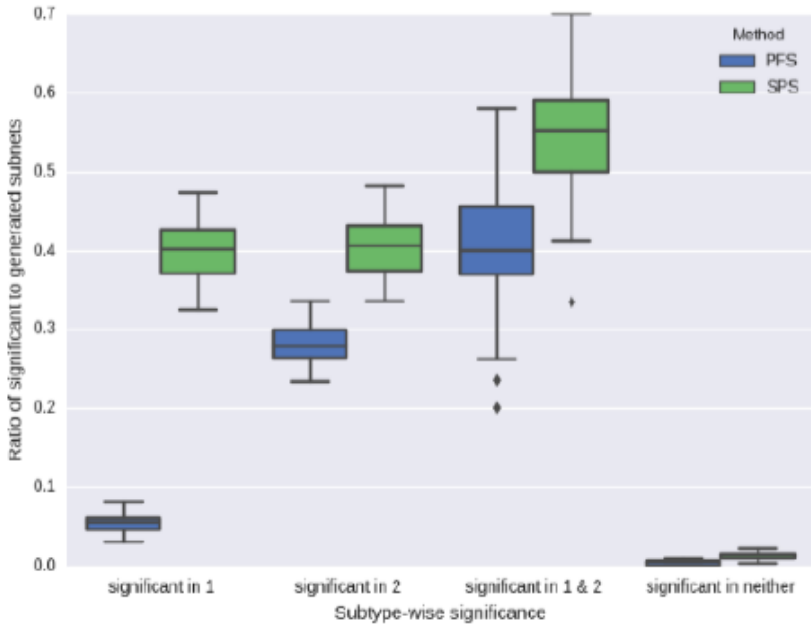

For each subnetwork $S_k$



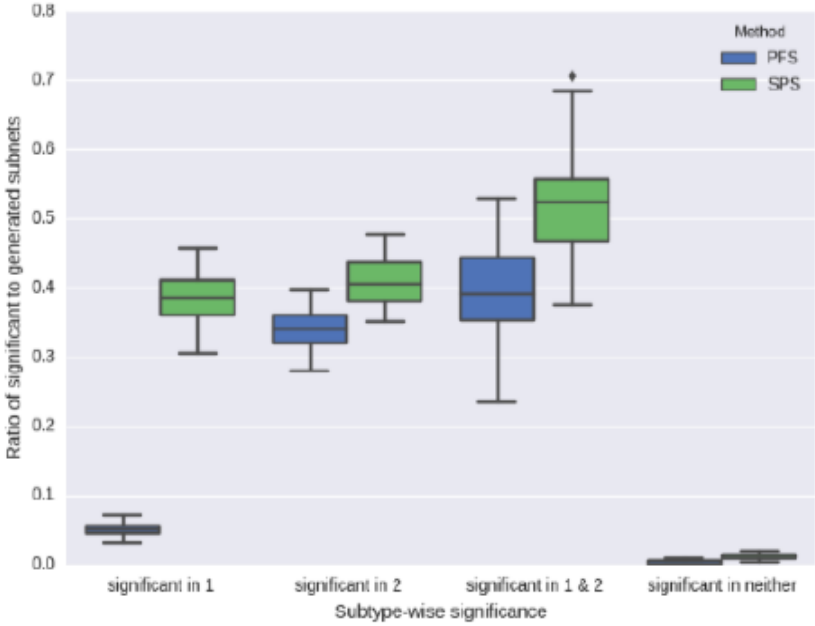



# Recall & precision on simulated datasets

- **SPSNET recalls more planted significant subnetworks than PFSNET, while keeping false positives in check**




(c) Dataset 2: 40% subtype 1, 60% subtype 2




(d) Dataset 2: 20% subtype 1, 80% subtype 2

# Isolating subpopulations

- **Mix T-ALL + TEL-AML1 vs. normal**

| | 30 TEL-AML1 + 29 T-ALL | 30 TEL-AML1 + 20 T-ALL | 30 TEL-AML1 + 10 T-ALL |
|---|---|---|---|
| PFSNet | 0.116 | 0.12 | 0.079 |
| SPSNet | **0.323** | **0.342** | **0.288** |

Silhouette scores based on PC1-3 of feature matrices built using scores of significant subnetworks in PFSNET and SNET

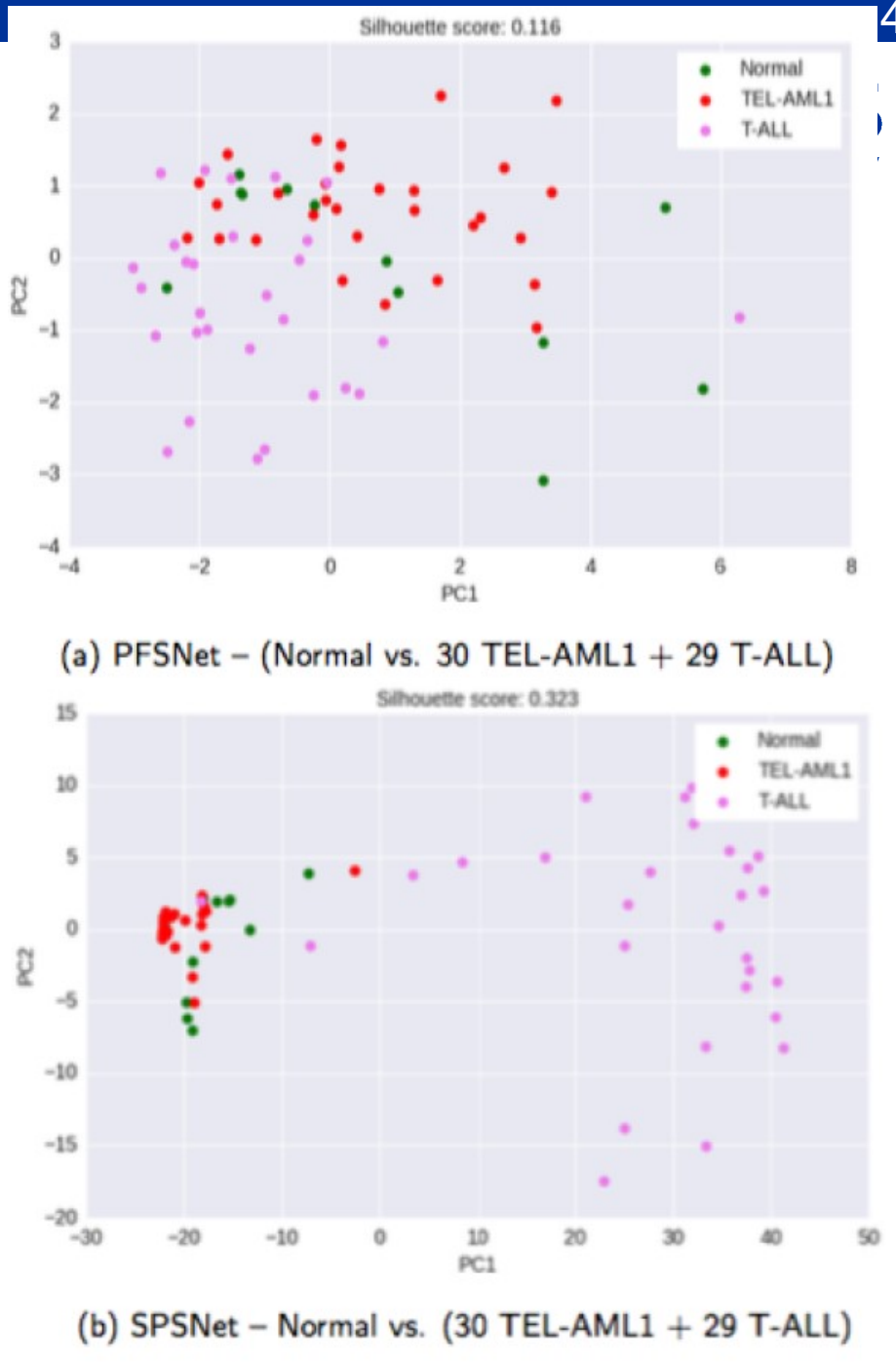
- **Mix two batches of HCC tumour vs non-tumour**

| | Normal vs HCC (first 3 PCs, with batch labels) | Normal vs HCC ($2^{nd}$, $3^{rd}$ PC, without batch labels) |
|---|---|---|
| PFSNet | 0.145 | 0.117 |
| SPSNet | **0.268** | **0.298** |

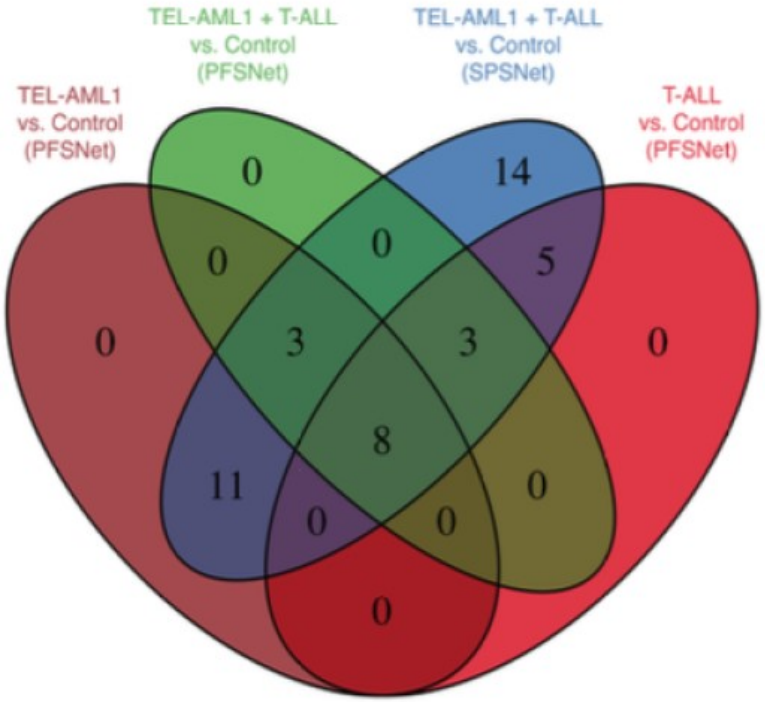
Silhouette scores based on PCA of feature matrices built using scores of significant subnetworks in PFSNET and SNET

- **SPSNET is much better than PFSNET at separating hidden subpopulations/batch effects**

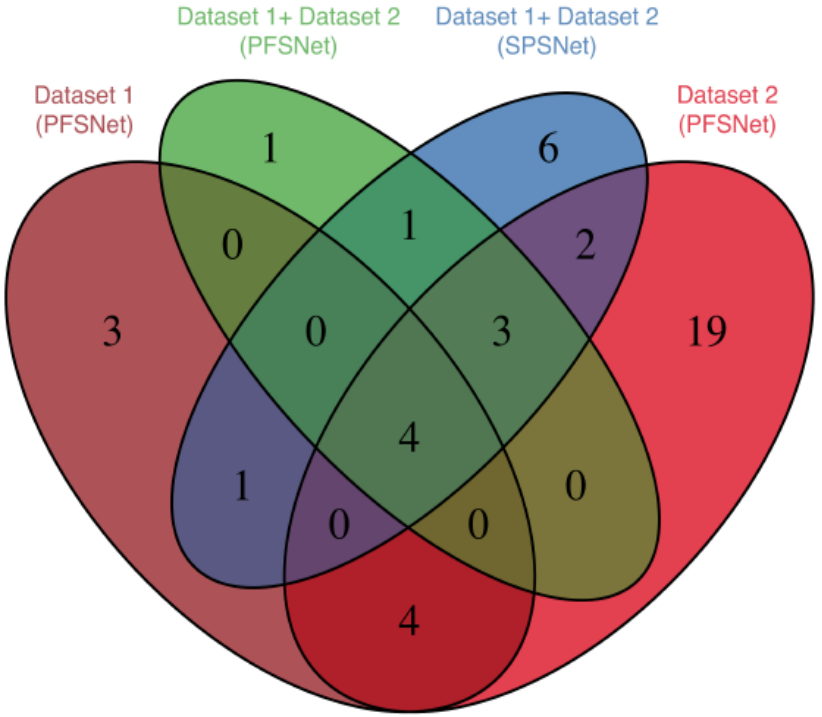
Mix T-ALL + TEL-AML1 vs normal



# Heterogeneous vs homogeneous



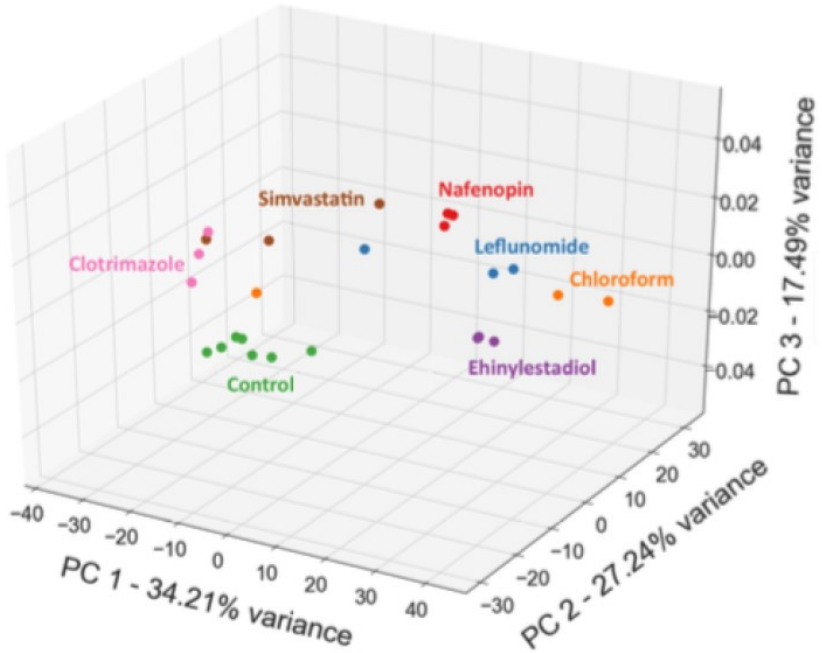

30 TEL-AML1 + 29 T-ALL vs normal



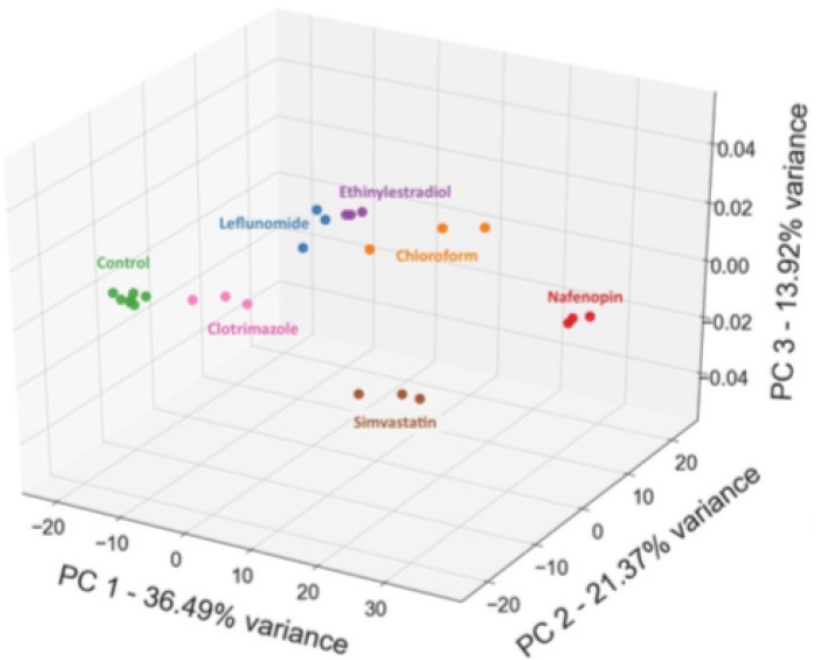

HCC two batches, tumour vs non-tumour

- **SPSNET finds more subpopulation-specific subnetworks than PFSNET**

# SPSNET works when there are >2 subpopulations too




(a) SPSNet: across 6 modes of action




(b) ANOVA: across 6 modes of action

- **Rat toxicogenomics RNA-seq : 1 control vs 5 drugs**
- **SPSNET (no drug info) works as well as ANOVA (w/ drug info)**

NUS National University of Singapore

# SUMMARY

# What have we learned?

- **Common normalization methods have problems**
  - Fail to remove batch effects
  - Remove subpopulation effects along with batch
  - Introduce false effects
- **GFS is a better normalization method**
- **SNET/FSNET/PFSNET are batch effect-resistant**
- **SPSNET is subpopulation-sensitive, works well for datasets with undeclared heterogeneity**
- **These methods work well on microarray, RNA-seq, and SWATH MS proteomics data**