CS4220: Knowledge Discovery Methods for Bioinformatics

Unit 3: Gene Expression Profile Analysis

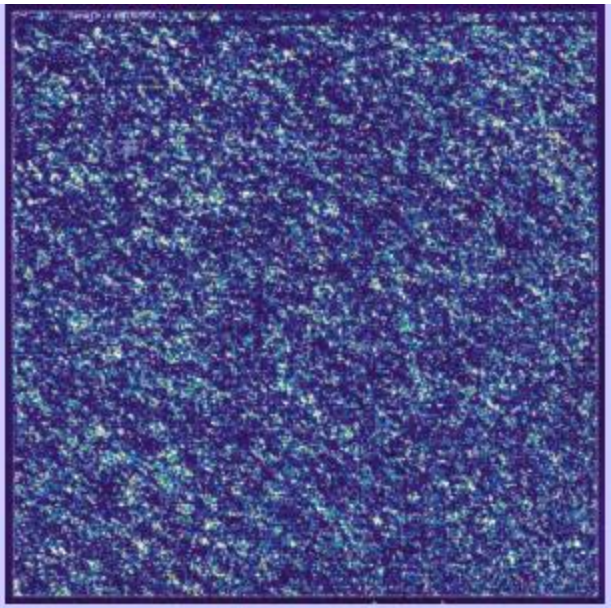
Wong Limsoon

Plan

- **Basic gene expression profile analysis**
- **Some issues in gene expression analysis**
- **Batch effect & normalization**
- **Improving reproducibility**
- **More advanced analysis**
  - A Novel Principle for Childhood ALL Relapse Prediction

Basic Gene Expression Profile Analysis

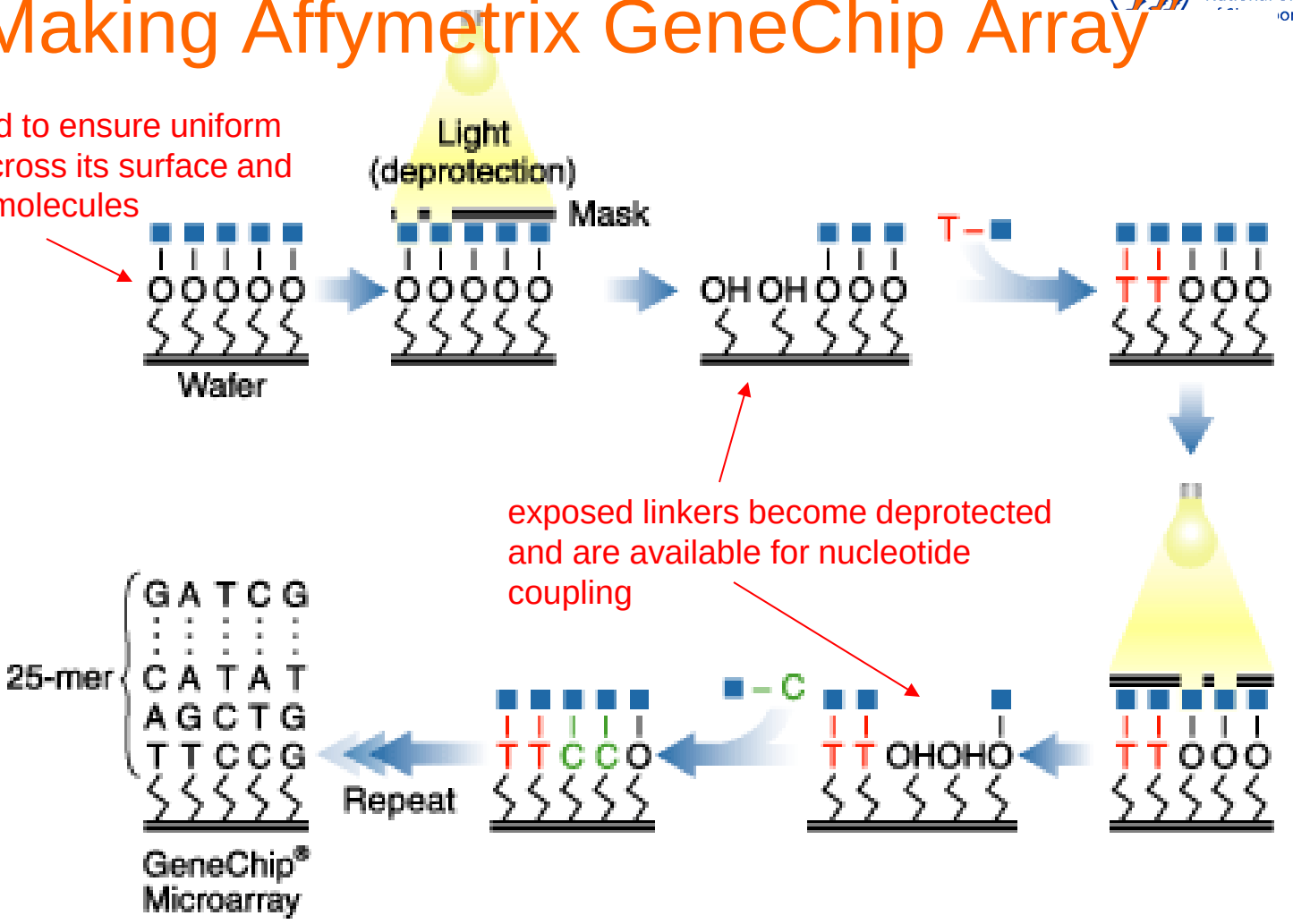
Affymetrix GeneChip Array



Source: Affymetrix

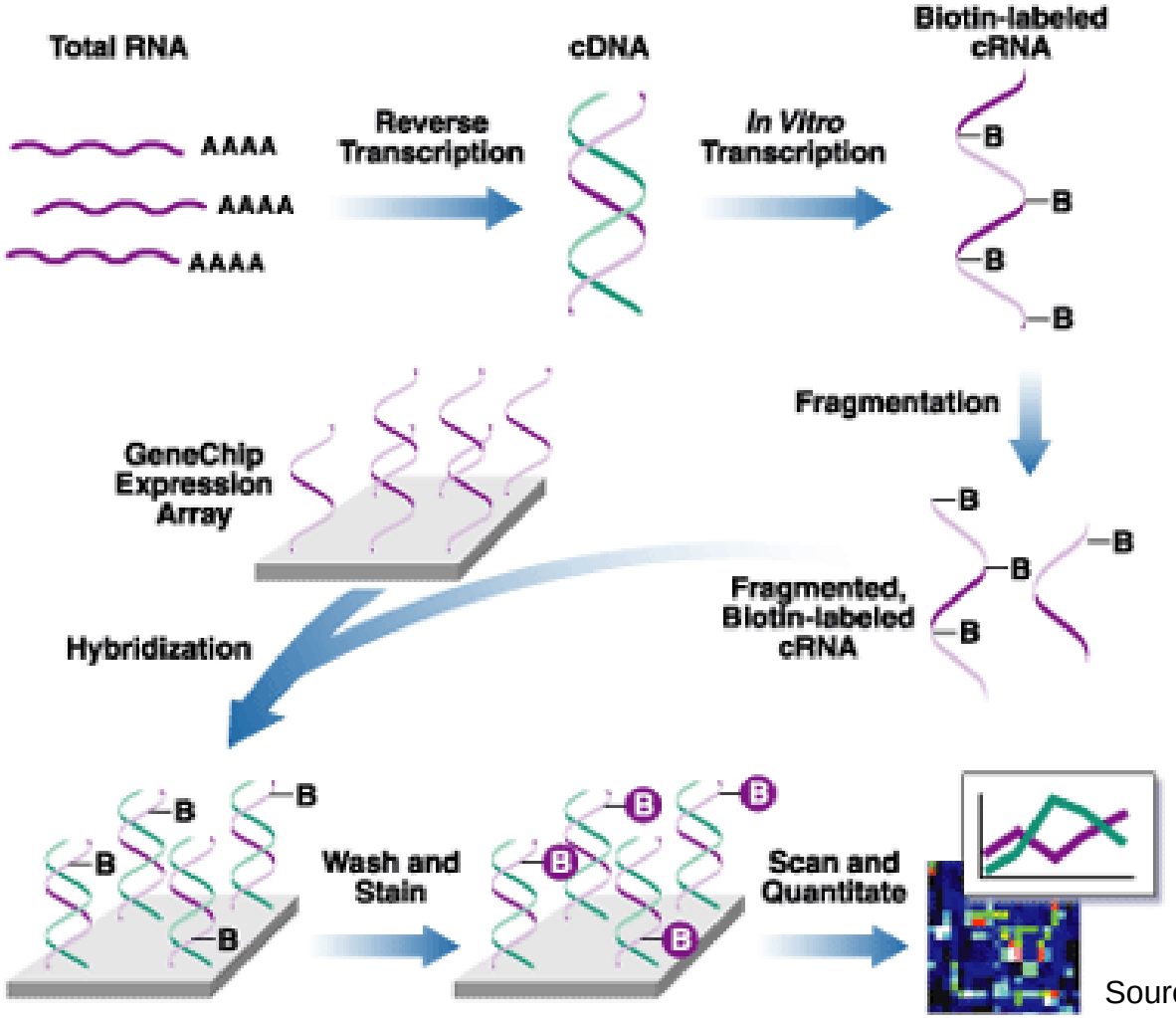
Making Affymetrix GeneChip Array

quartz is washed to ensure uniform hydroxylation across its surface and to attach linker molecules



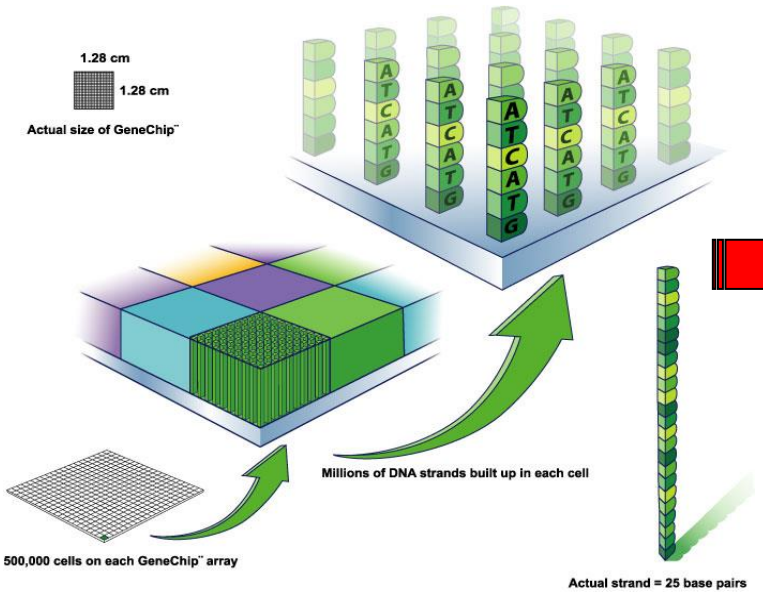
Source: Affymetrix

Gene Expression Measurement by Affymetrix GeneChip Array



Source: Affymetrix

Diagnosis Using Microarray



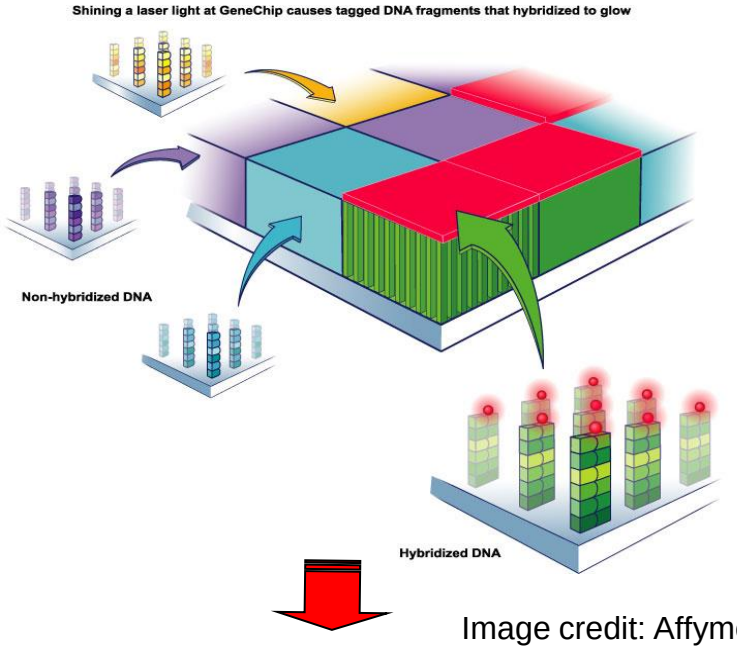
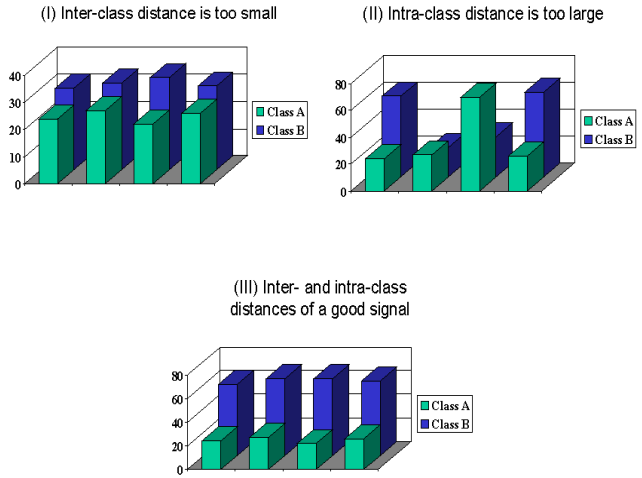
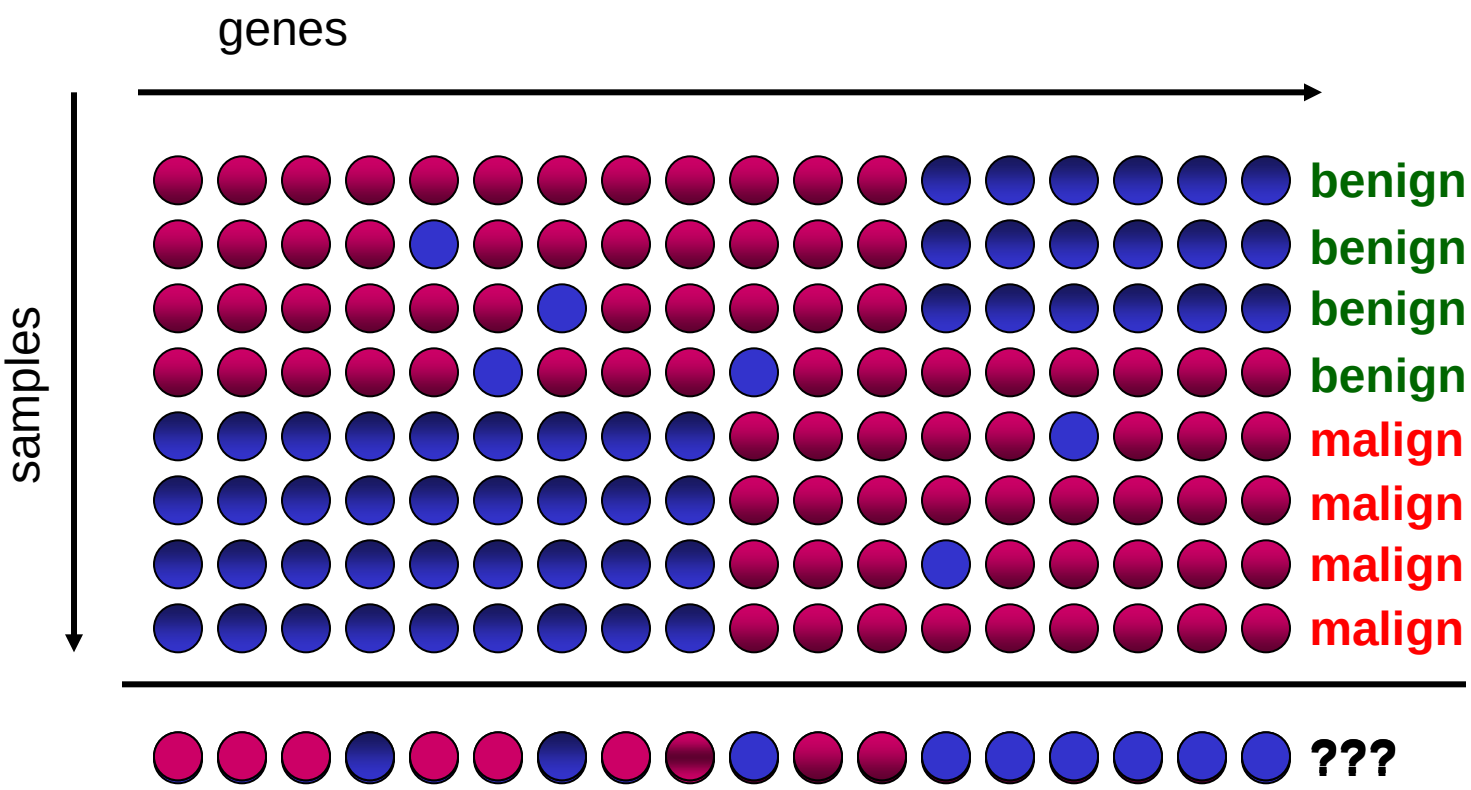


Image credit: Affymetrix

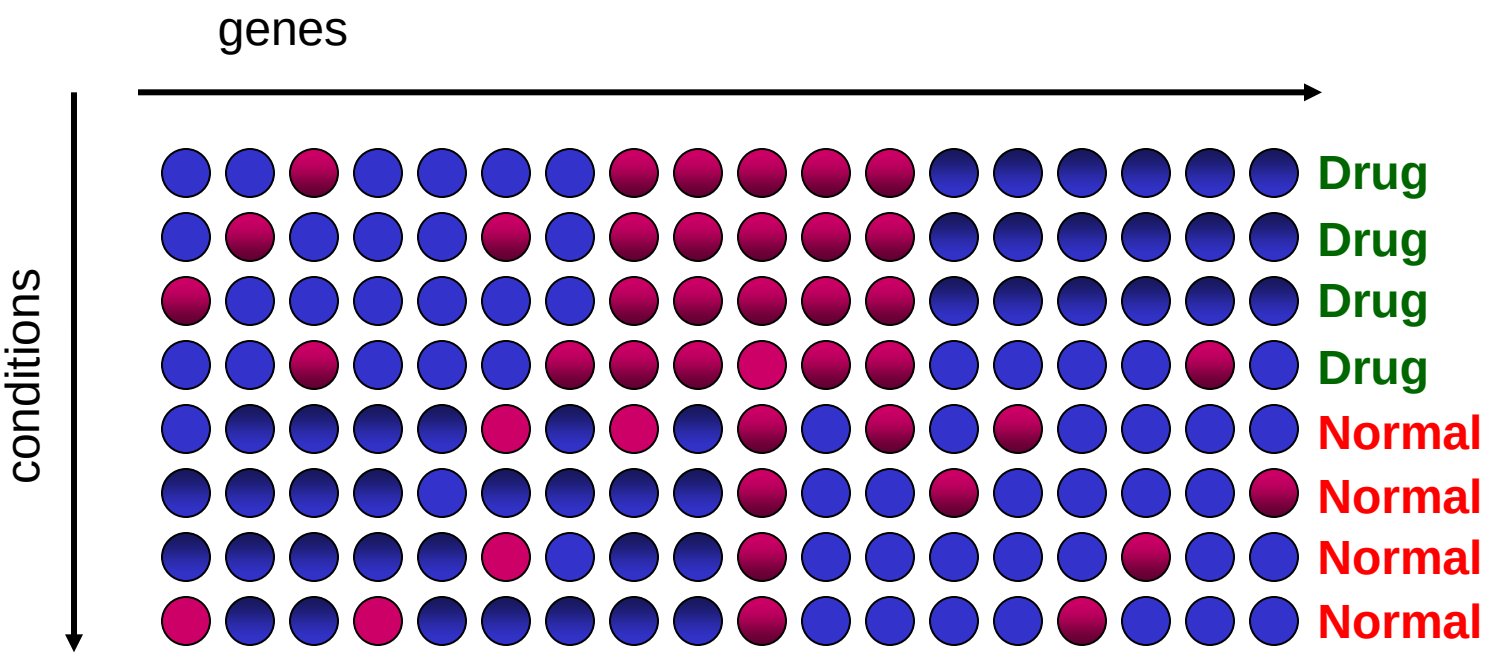


	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAv	Avg Diff	Abs Call	
AFFX-MurI	5	2	19	297.5	A	M16762 Mouse int
AFFX-MurI	3	2	19	554.2	A	M37897 Mouse int
AFFX-MurI	4	2	19	308.6	A	M25892 Mus musc
AFFX-MurF	1	3	19	141	A	M83649 Mus musc
AFFX-BioE	13	1	19	9340.6	P	J04423 E coli bioB
AFFX-BioE	15	0	19	12862.4	P	J04423 E coli bioB
AFFX-BioE	12	0	19	8716.5	P	J04423 E coli bioB
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacterioph
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacterioph
AFFX-BioE	7	5	18	-483	A	J04423 E coli bioB
AFFX-BioE	5	4	18	313.7	A	J04423 E coli bioB
AFFX-BioE	7	6	20	-1016.2	A	J04423 E coli bioB

Application: Disease Subtype Diagnosis



Application: Drug Action Detection



Which group of genes are the drug affecting on?

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**

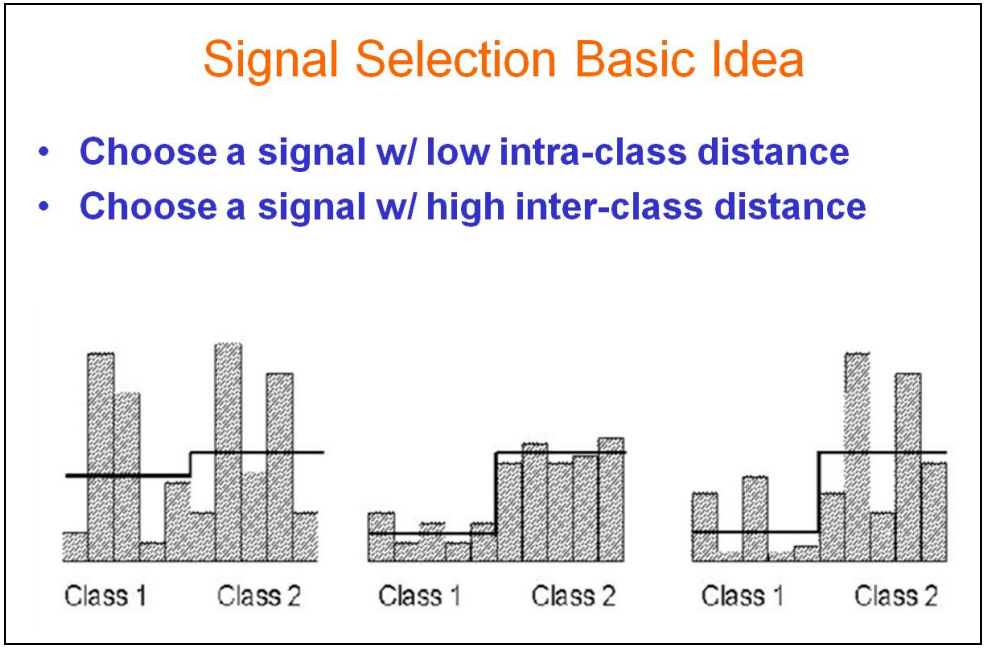


Image credit: Golub et al., *Science*, 286:531–537, 1999

If you don't remember this, you should go back to revise your CS2220 notes ☺

Terminology: DE gene = differentially expressed gene

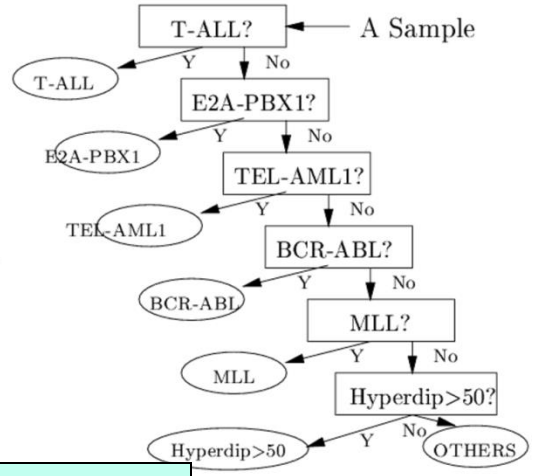
You can build a gene expression profile classifier in a simple or in a more complex way

- **Parallel-multiclass classification scheme**

- **Tree-structured classification scheme**

22

Childhood ALL Subtype Diagnosis Workflow



A tree-structured diagnostic workflow was recommended by our doctor collaborator

Next, we take an example to demonstrate the scores used by PCL. A BCR-ABL test sample contained almost all of the top-20 BCR-ABL discriminators. So, a score of 19.6 was assigned to it. Several top-20 OTHERS discriminators together with some beyond the top-20 list were also contained in this test sample. So, another score of 6.97 was assigned. This test sample did not contain any discriminators of E2A-PBX1, Hyperdip>50, or T-ALL. So, the scores are as follows:

subtype	BCR-ABL	E2A-PBX1	Hyperdip>50	T-ALL	MLL	TEL-AML1	OTHERS
score	19.63	0.00	0.00	0.00	0.71	2.96	6.97

Source: Li et al. *Bioinformatics*, 19:71--78, 2003

Hierarchical Clustering

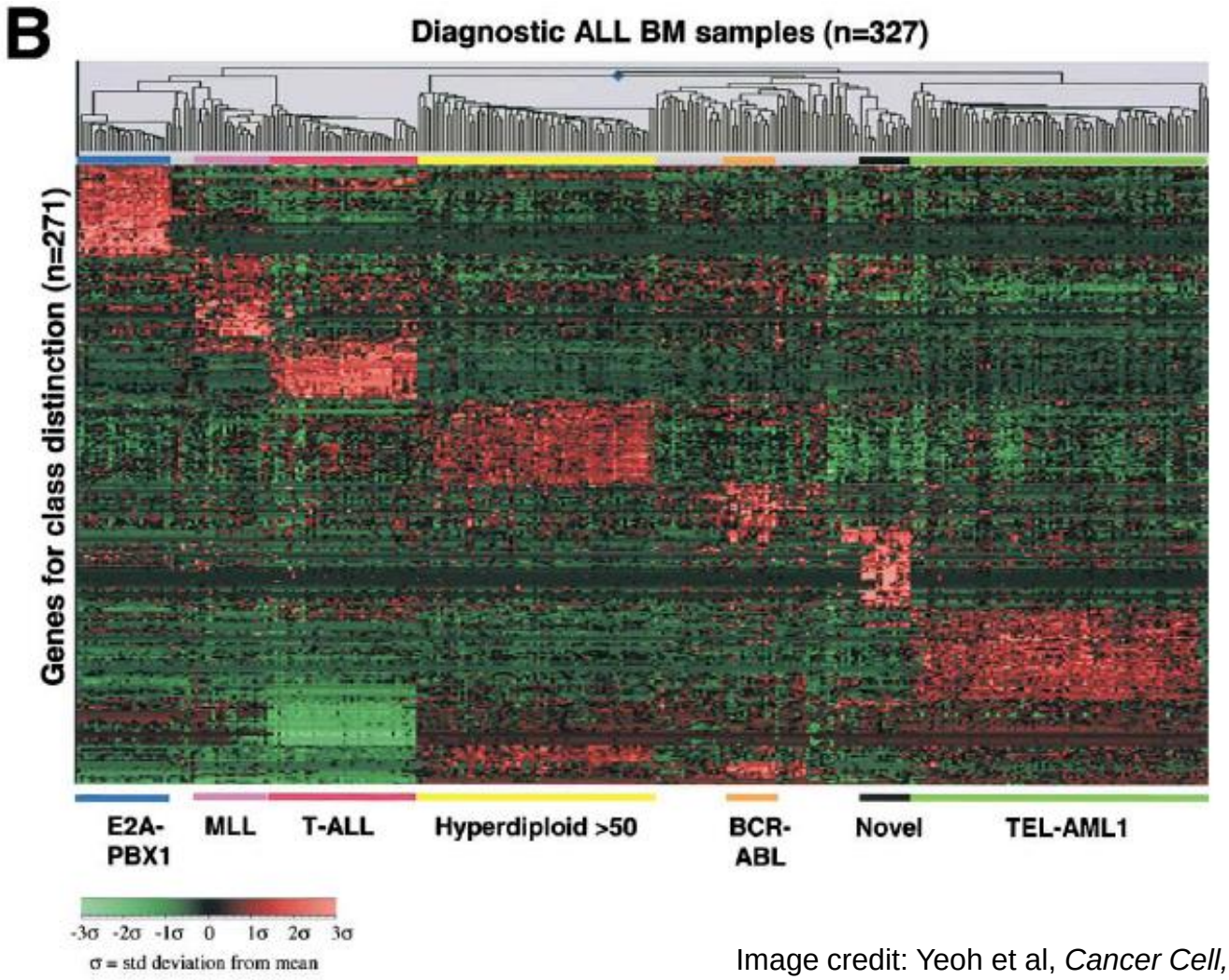
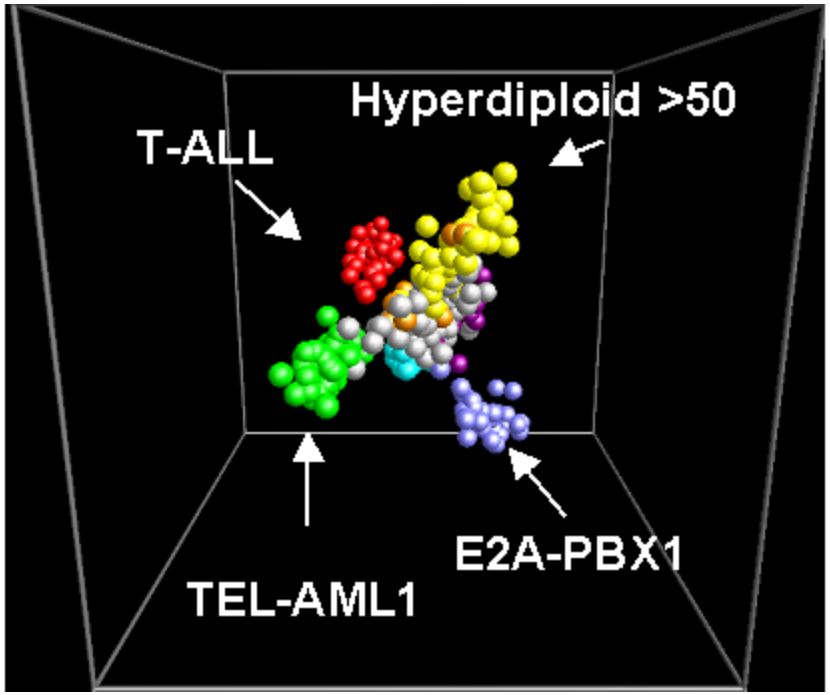


Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

PCA Plots



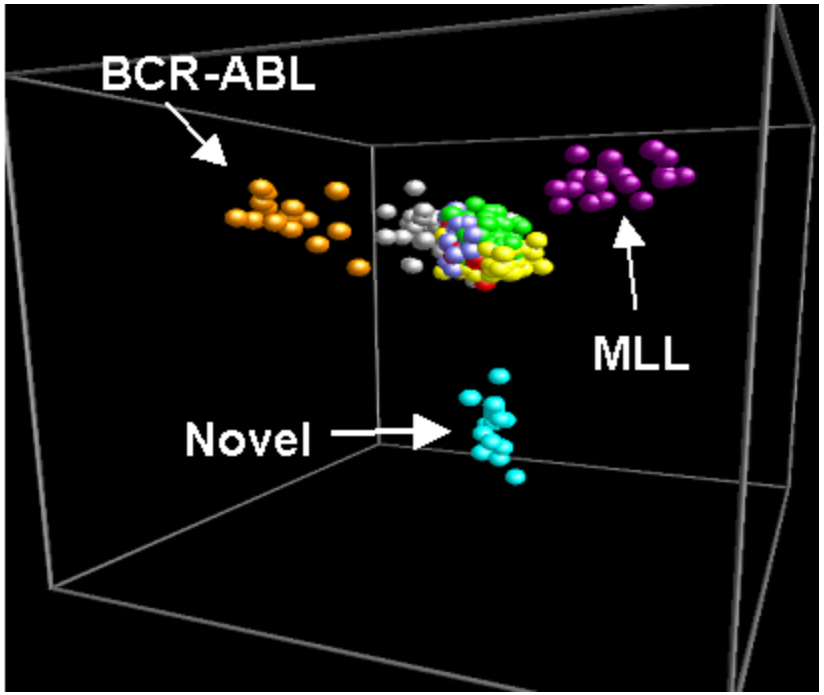


Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

Some Issues in Gene Expression Analysis

Some Headaches

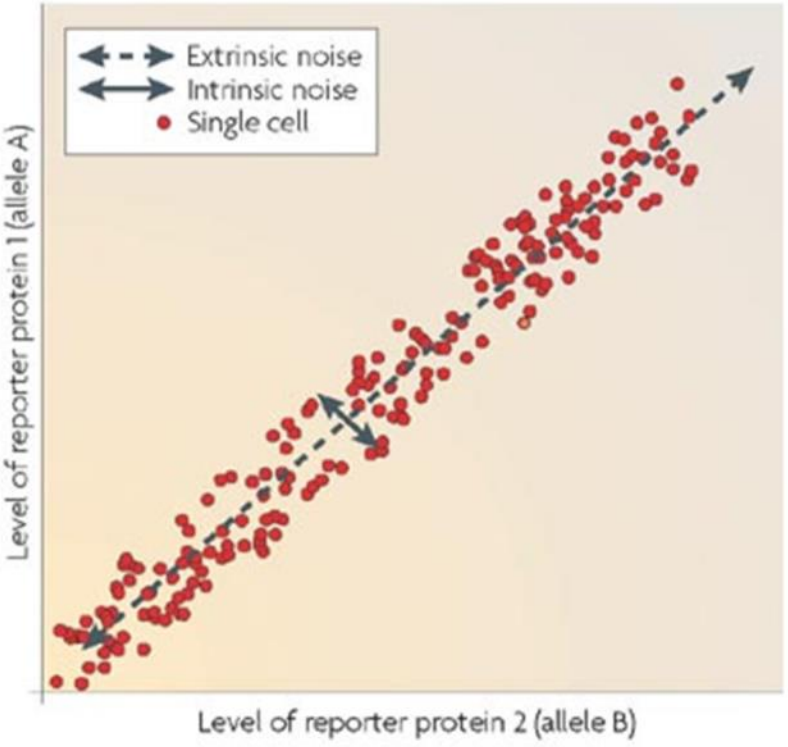
- **Natural fluctuations of gene expression in a person**

- **Noise in experimental protocols**
  - Numbers mean diff things in diff batches
  - Numbers mean diff things in data obtained from diff platforms

⇒ **Selected genes may not be meaningful**
  - Diff genes get selected in diff expts

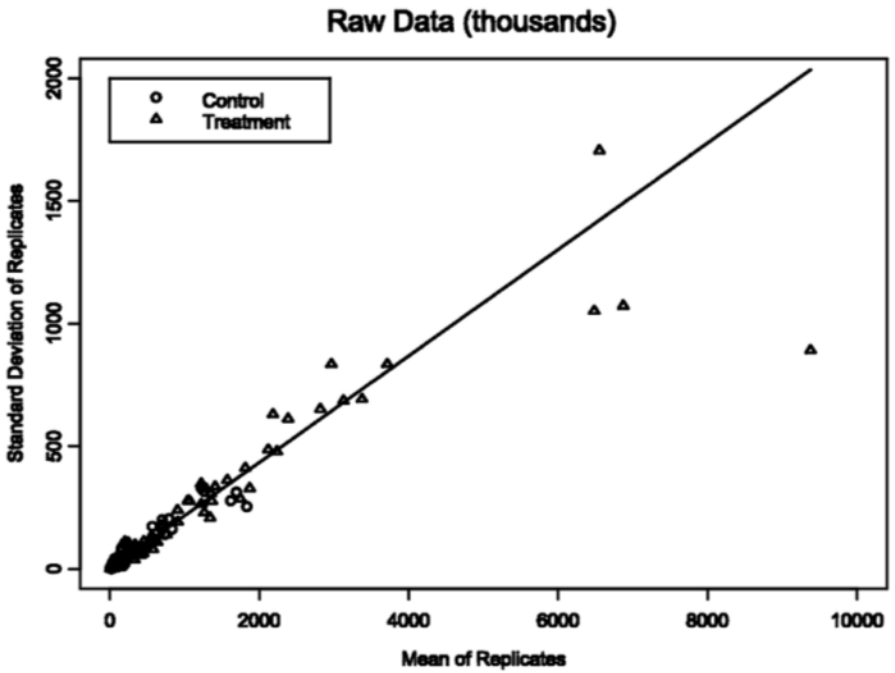
Natural Fluctuations & Expt Noise

Intrinsic & extrinsic noise



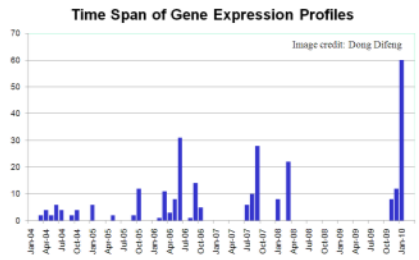
Nat Rev Genet, 9:583-593, 2008

Measurement errors

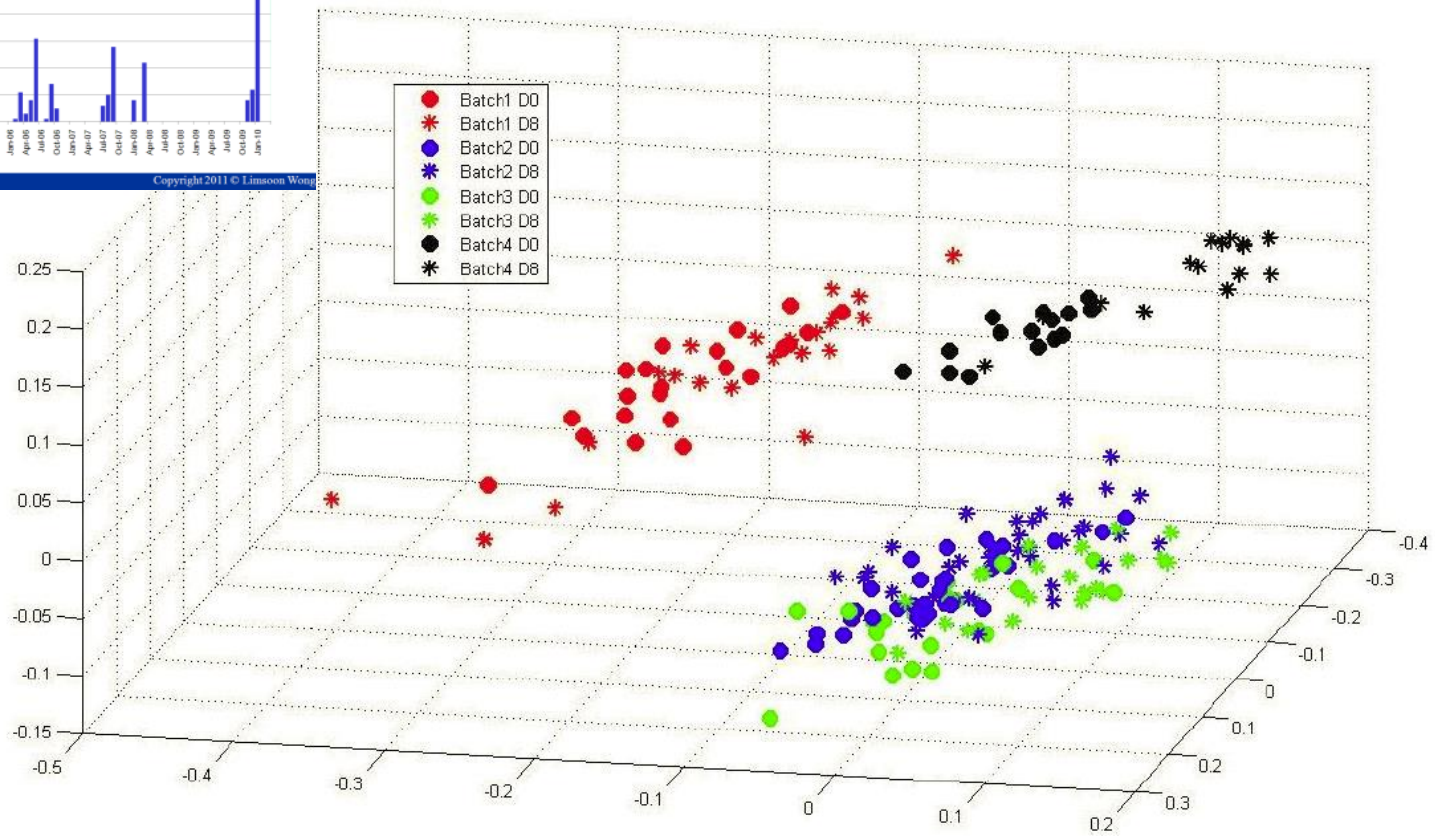


J Comput Biol, 8(6):557-569, 2001

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



Batch Effects



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

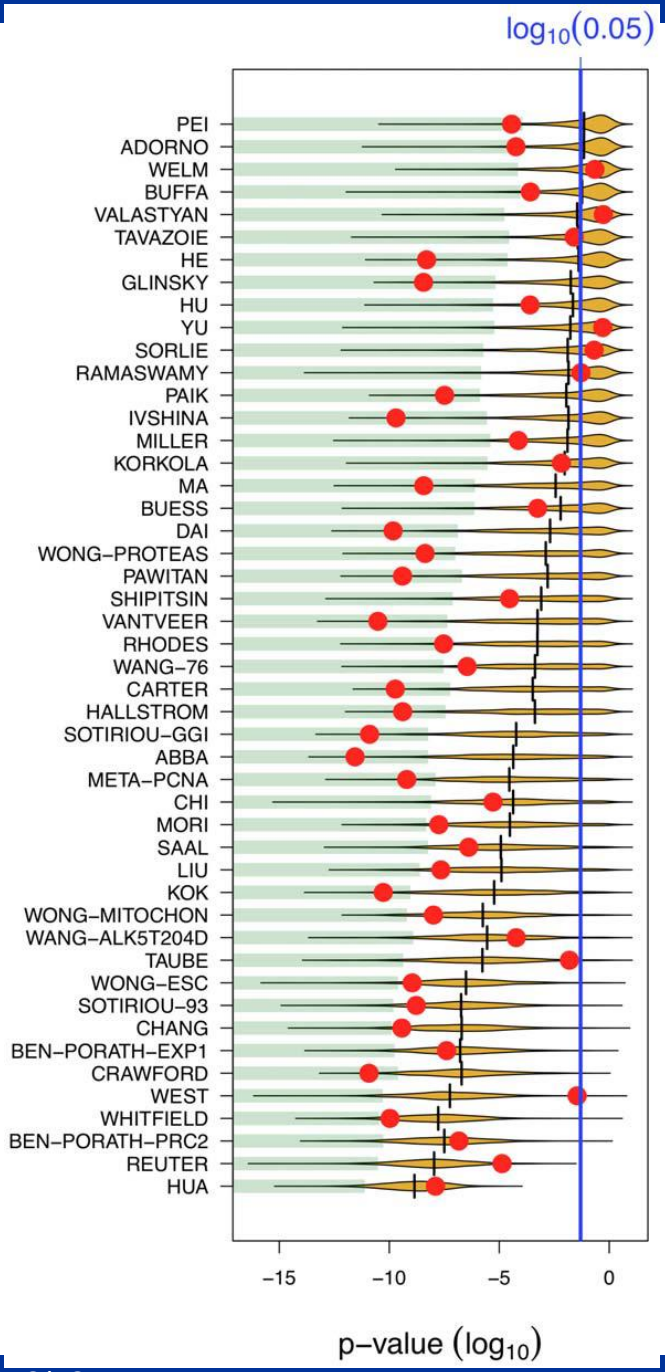
Image credit: Difeng Dong's PhD dissertation, 2011

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.

Batch Effect & Normalization

Approaches to Normalization

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

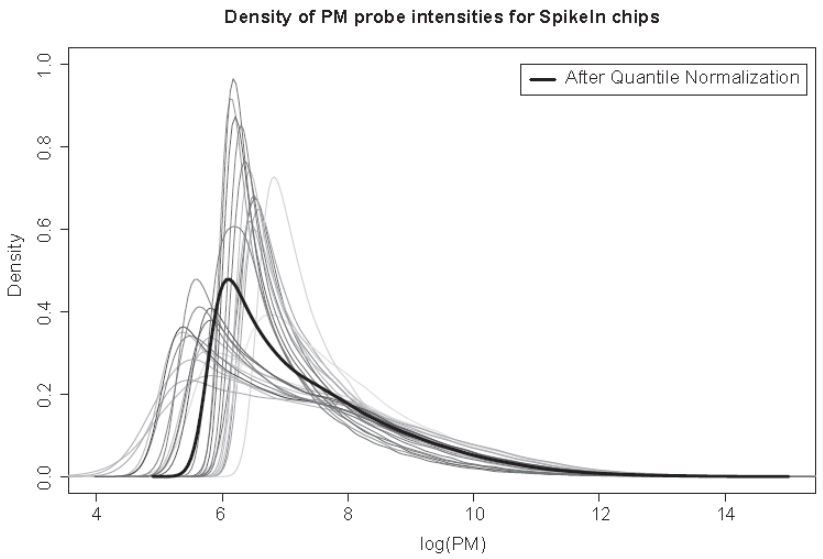
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

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

- **Quantile normalization**

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

After quantile normalization

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

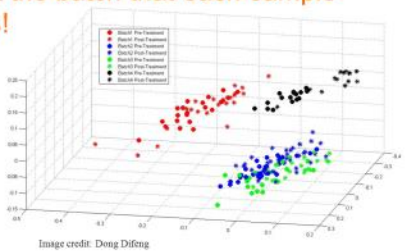


Image credit: Dong Difeng

⇒ Need normalization to correct for batch effect

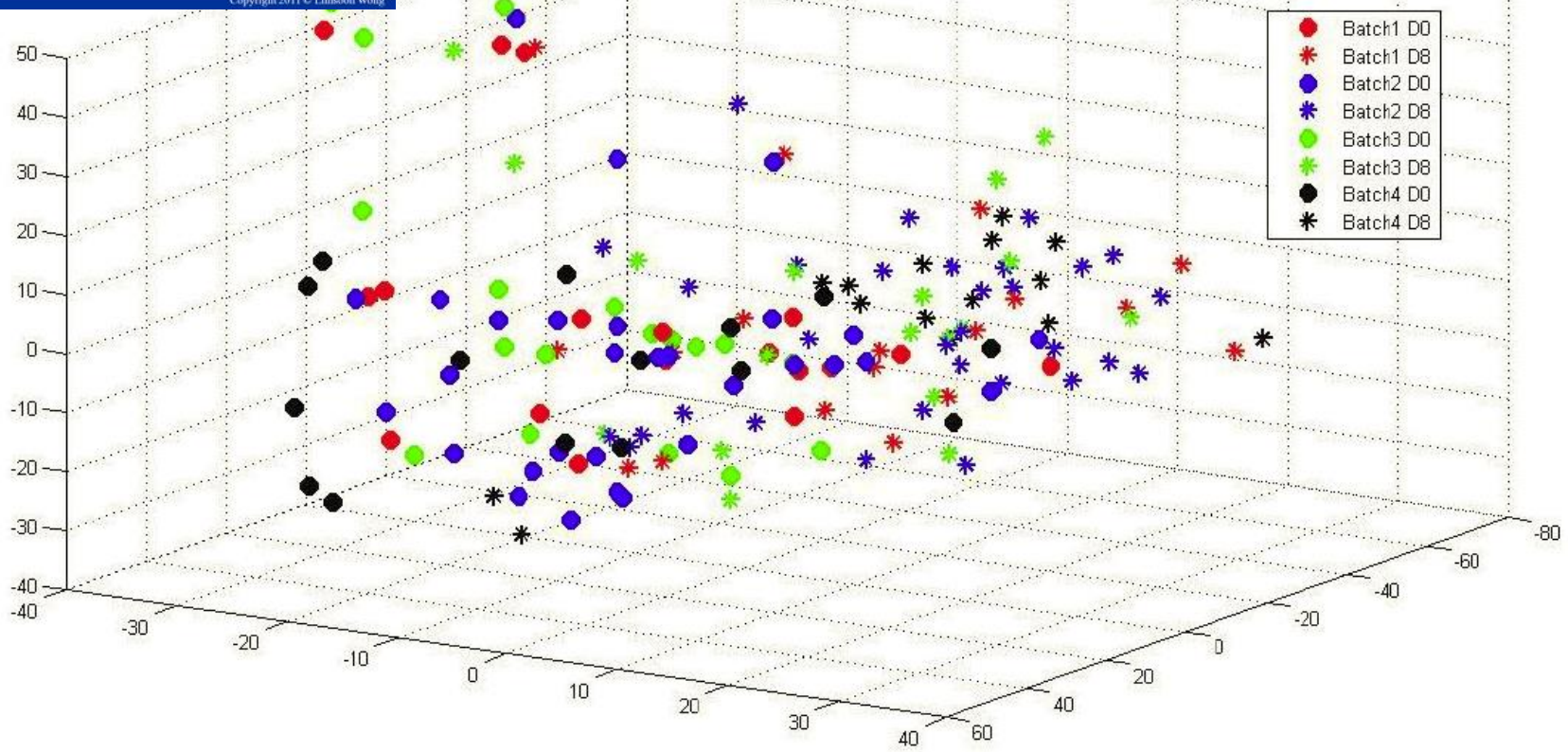
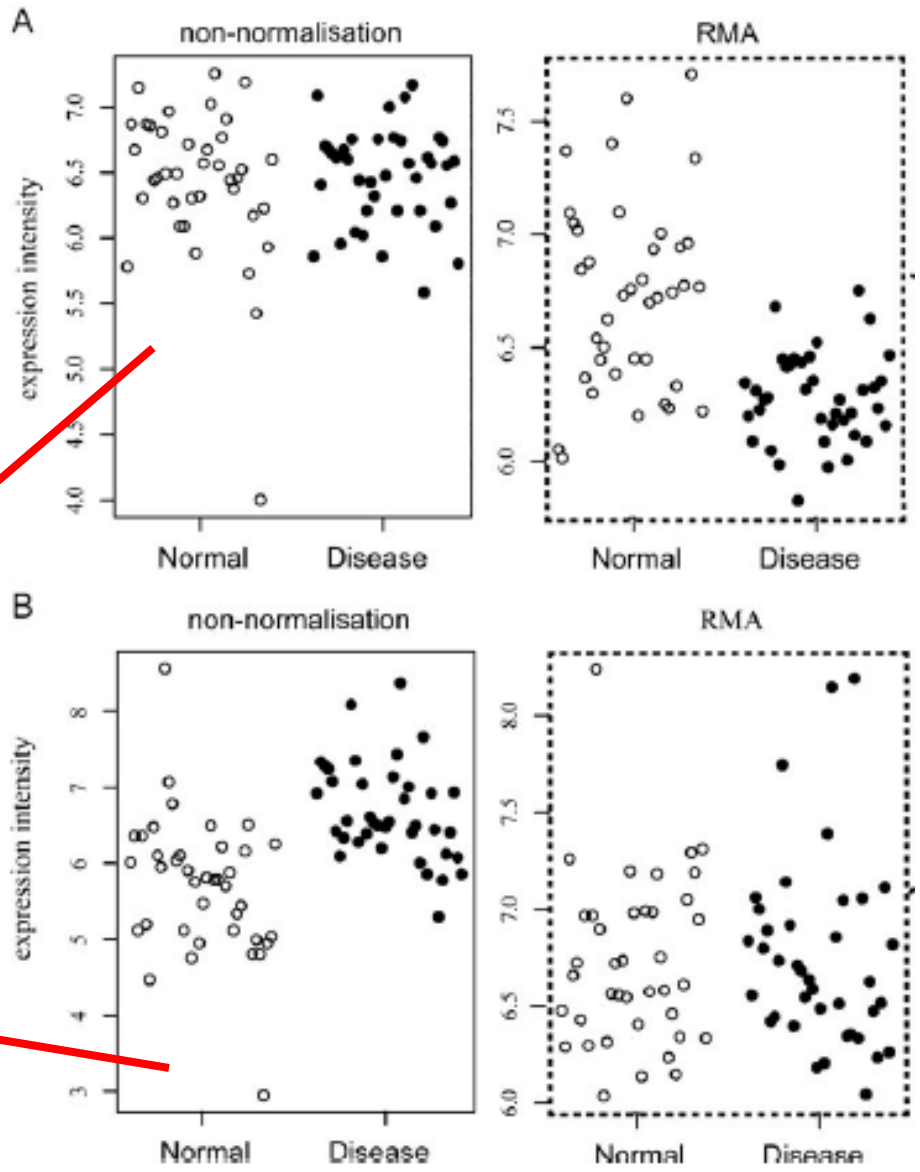


Image credit: Difeng Dong's PhD dissertation, 2011

Caution: "Over normalize" 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 higher than 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 nornmalized data



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

Embracing Noise to Improve Cross-Batch Prediction Accuracy

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

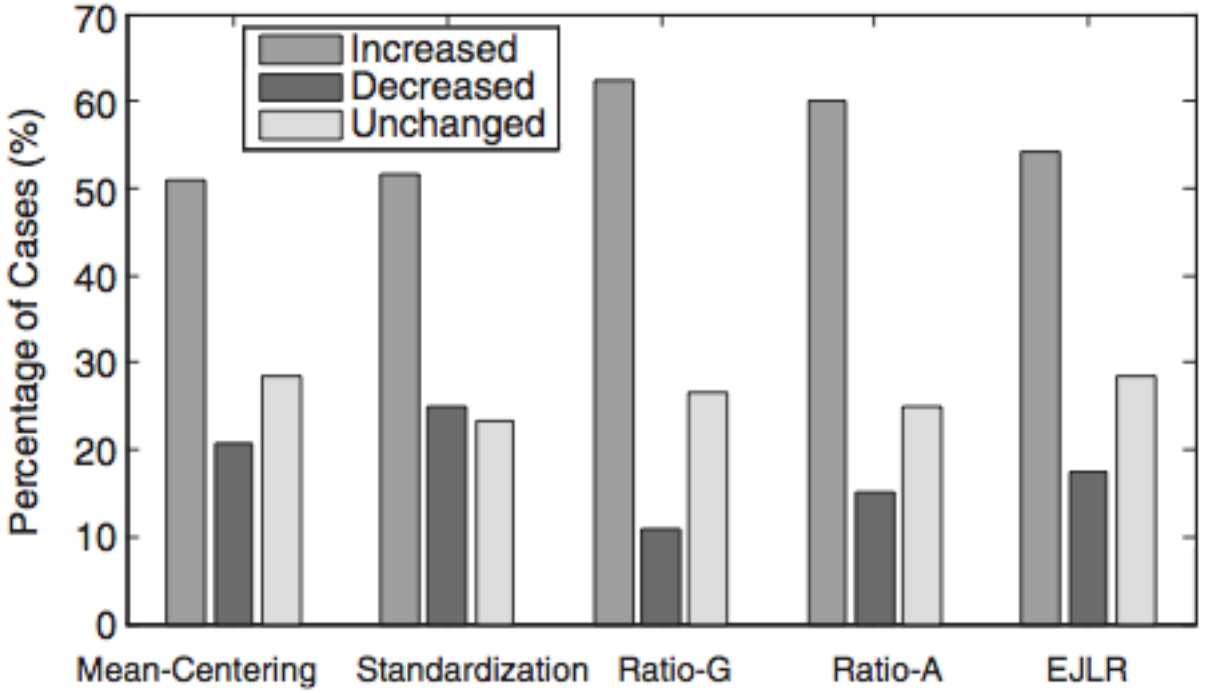


Figure 10 Percentages of increased, decreased and unchanged cases in prediction performance after applying different batch effect removal methods. The total number of cases explored is 120.

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.

Batch Effect 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"

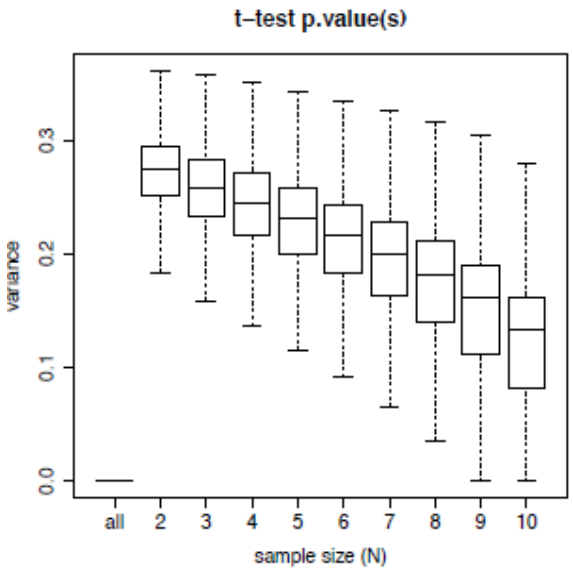
The "Embracing Noise" Approach

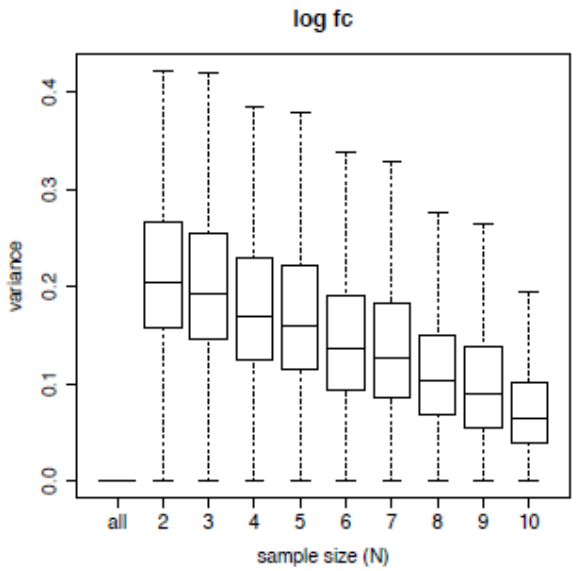
- **Ranking values (c.f. quantile normalization)**
  - Instead of absolute values
  - Inspired by MAQC project
    - **"Absolute values may be different (among batches) but relative values are conserved between different platforms"**

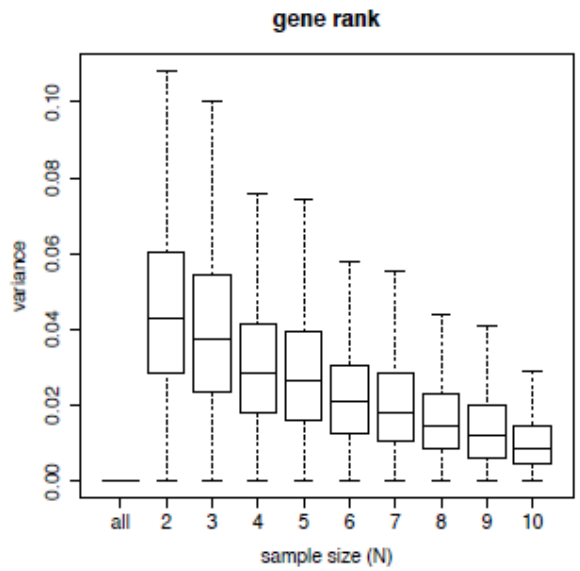
- **Stochastic sampling with replacement**
  - Bootstrapping suppresses noise
    - **Training clones produced are likely to be enriched with more "clean" samples**

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

Koh & Wong, *BMC Systems Biology*, 6(Suppl 2):S3, 2012

Bootstrapping suppresses noise

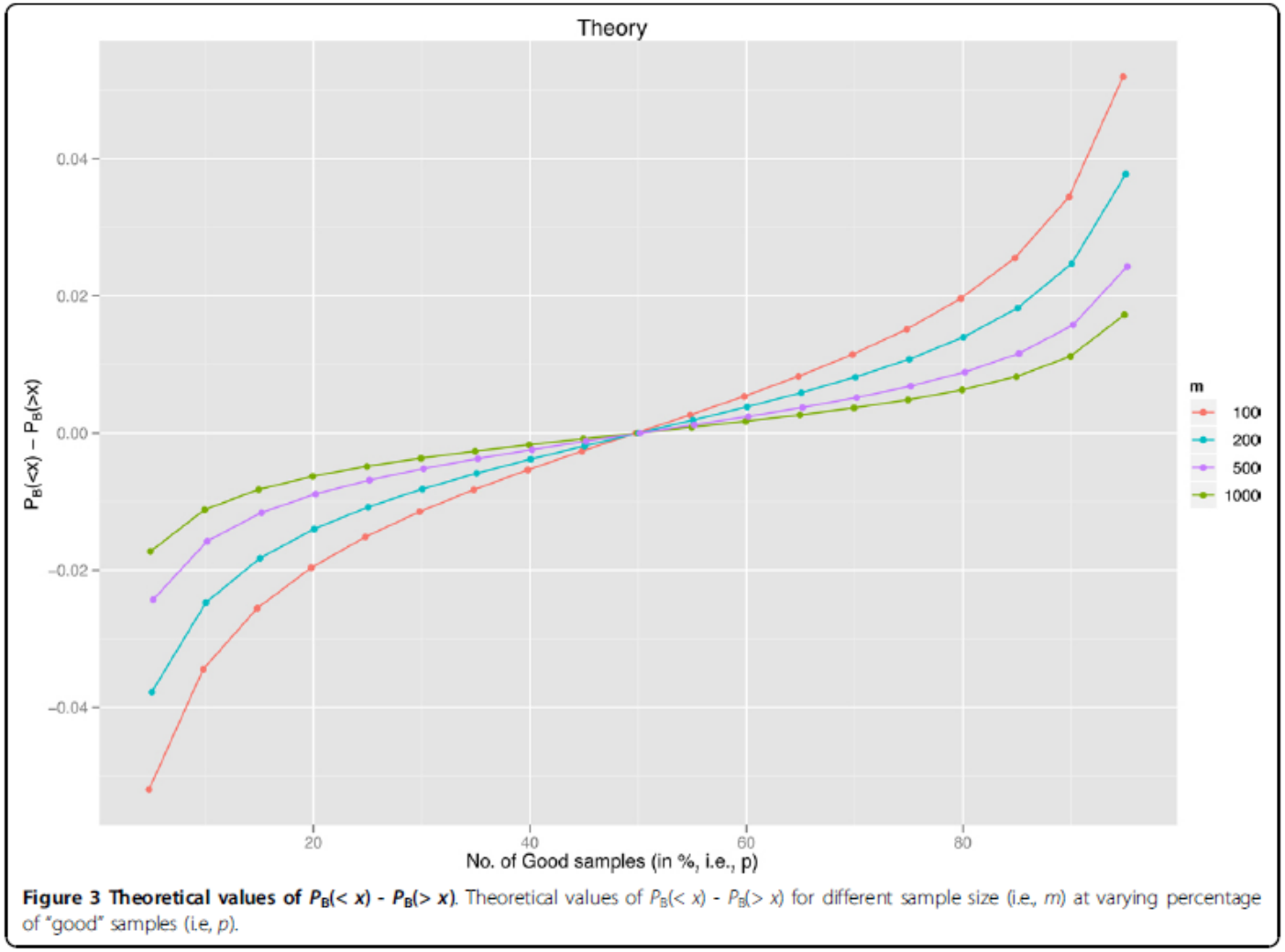


Figure 3 Theoretical values of $P_B(< x)$ - $P_B(> x)$. Theoretical values of $P_B(< x)$ - $P_B(> x)$ for different sample size (i.e., m) at varying percentage of "good" samples (i.e, p).

Suppose a set S of m samples is given. Suppose x of the samples are "bad" (i.e., incorrect or very noisy) and y of the samples are "good" (i.e., correct or little noise). Let $p = y/m$ and $q = x/m = (1-p)$. Let B be a bag of m samples randomly drawn with repetitions from S. Therefore, according to the principle of Bernouli trials, the chances of B having k "bad" samples is given by $P_B(k) = ({}^mC_k)(p^{m-k})(q^k)$, where ${}^mC_k$ means "m choose k". Then the chances of B having fewer "bad" samples than S is given by $P_B(<x) = \sum_{k<x} P_B(k)$, while the chances of B having more "bad" samples than S is given by $P_B(>x) = \sum_{k>x} P_B(k)$, and the chances of B having the same number of "good" and "bad" samples as S is given by $P_B(=x) = P_B(x)$.

Now, we expand $P_B(<x)$ and get $P_B(<x) = \sum_{k<x} P_B(k) = \sum_{k<x-1} P_B(k) + P_B(x-1) = \sum_{k<x-1} P_B(k) + ({}^mC_{x-1})(p^{m-(x-1)})(q^{x-1}) > ({}^mC_{x-1})(p^{m-(x-1)})(q^{x-1}) = ({}^mC_{x-1})(p/q)(p^{m-x})(q^x) = ({}^mC_{x-1})((m-x)/x)(p^{m-x})(q^x) = ({}^mC_x)(p^{m-x})(q^x) = P_B(x)$. That is, $P_B(<x) > P_B(=x)$.

Next, we expand $P_B(>x)$ and get $P_B(>x) = \sum_{k>x} P_B(k) = \sum_{2x-1\geq k>x} P_B(k) + \sum_{k>2x-1} P_B(k)$. By the Chernoff-Hoeffding bound, $\sum_{k>2x-1} P_B(>k) \leq exp(-((x-1)/x)^2 x/3) = exp(-(x-1)^2/3x)$, which rapidly decays to essentially 0 for x as small as 10 (i.e., when there are 10 "bad" samples). Meanwhile, if $p > q$, by a demonstration similar to what we have done for the case of $P_B(<x) > P_B(x)$, we can show that $\sum_{2x-1\geq k>x} P_B(k) \leq P_B(<x)$. Putting the two observations together, we get $P_B(<x) \geq P_B(>x)$ when $p > q$ and x as small as 10. This completes the proof that bootstrapping generates replicates that are more likely to be enriched with "good" samples than the original training set.

Formal proof of bagging suppresses noise

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

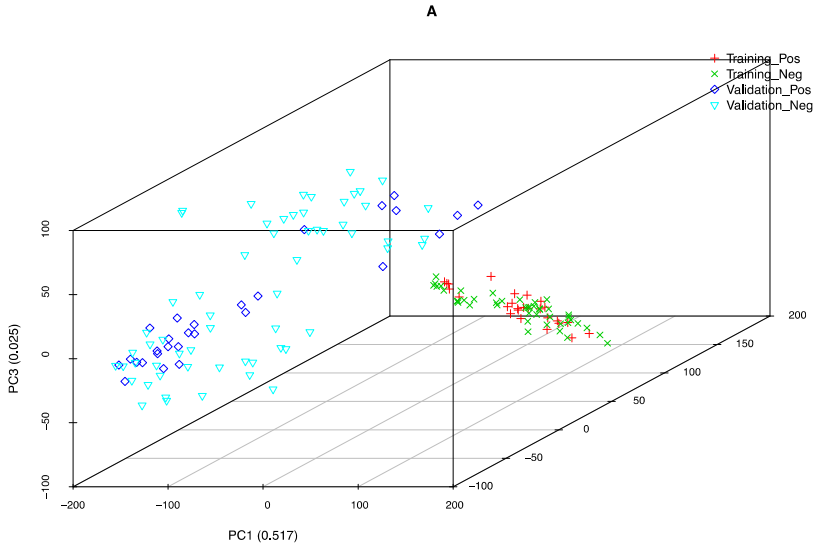
Koh, et al. **Improved statistical model checking methods for pathways analysis**. *BMC Bioinformatics*, 13(Suppl 17):S15, 2012

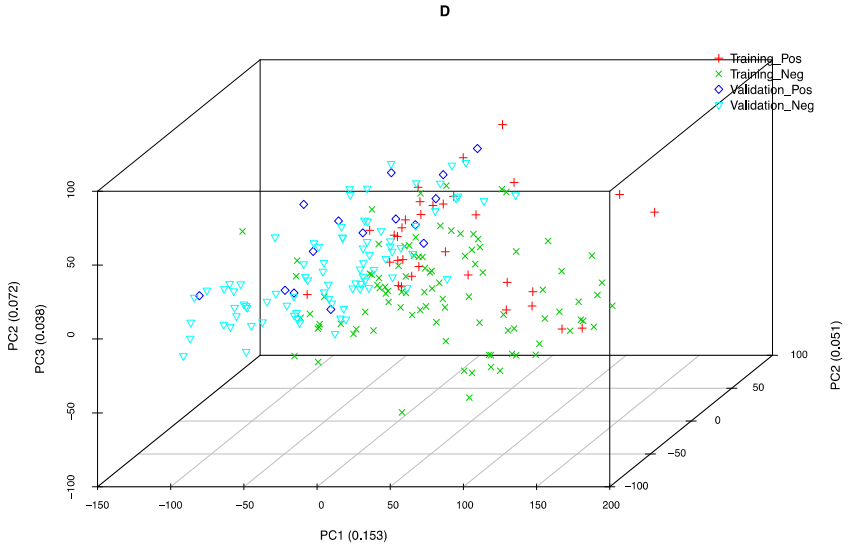
Datasets

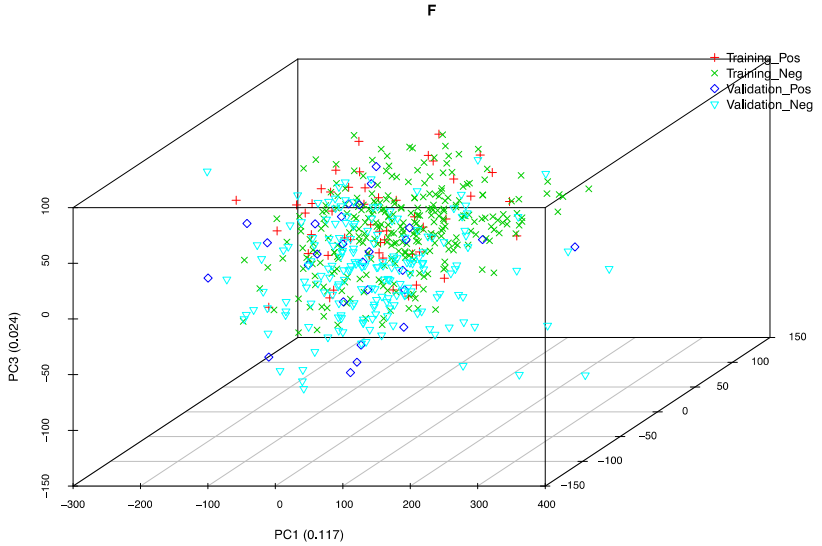
Data set code	Data set description	Training set			Validation set		
		Number of samples	Positives	Negatives	Number of Samples	Positives	Negatives
A	Lung tumorigen vs. non-tumorigen (Mouse)	70	26	44	88	28	60
D	Breast cancer pre-operative treatment response (pathologic complete response)	130	33	97	100	15	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

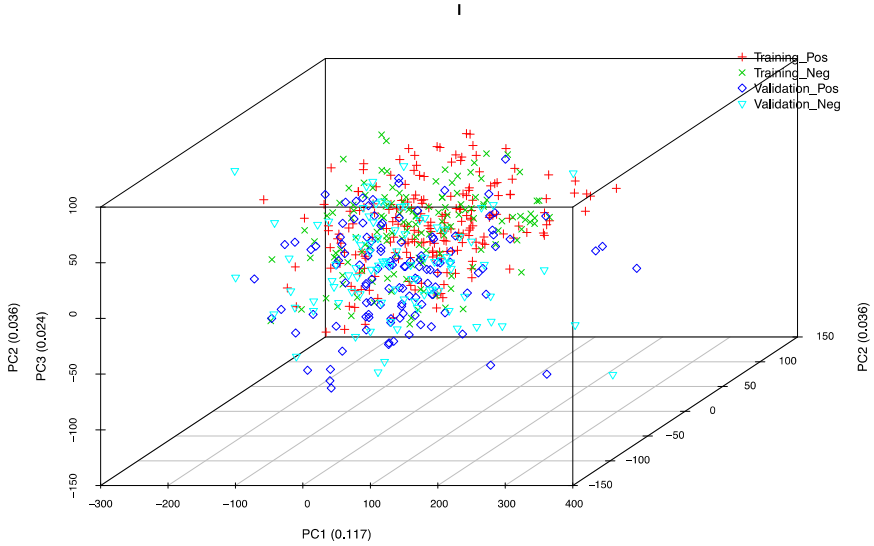
- Two additional data sets of size 25% or 50% of the above (with same class ratio)
- Total of 12 training sets and 12 validation sets

PCA of Datasets



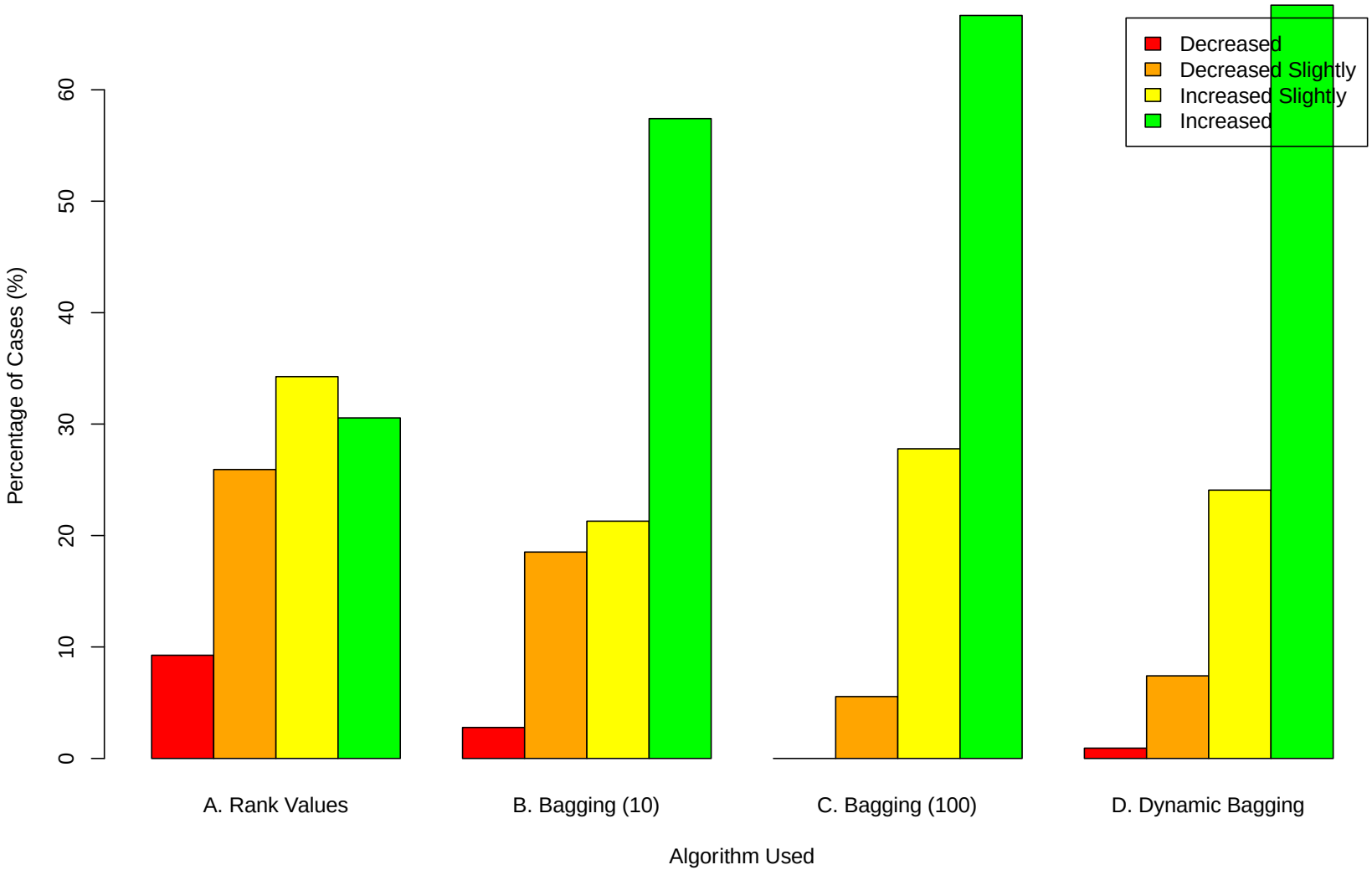


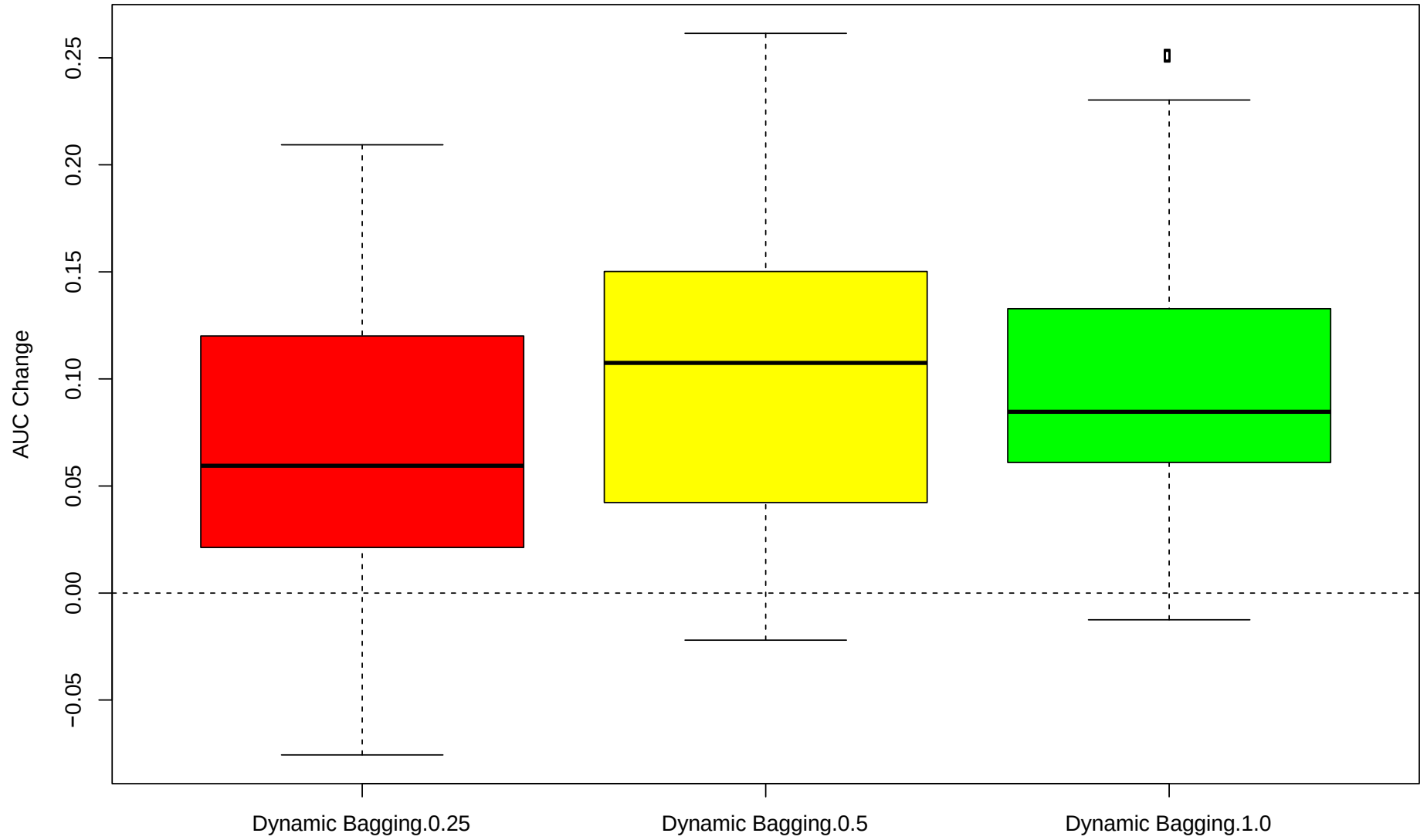




Experiments

- **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

Improving Reproducibility of Gene Expression Profile Analysis

Individual Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples

- **Prob(a gene is correlated) = $1/2^6$**
- **# of genes on array = 100,000**

$\Rightarrow$ **E(# of correlated genes) = 1,562**

- **How many genes on a microarray are expected to perfectly correlate to these samples?**

$\Rightarrow$ **Many false positives**

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

Group of Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **What is the chance of a group of 5 genes being perfectly correlated to these samples?**

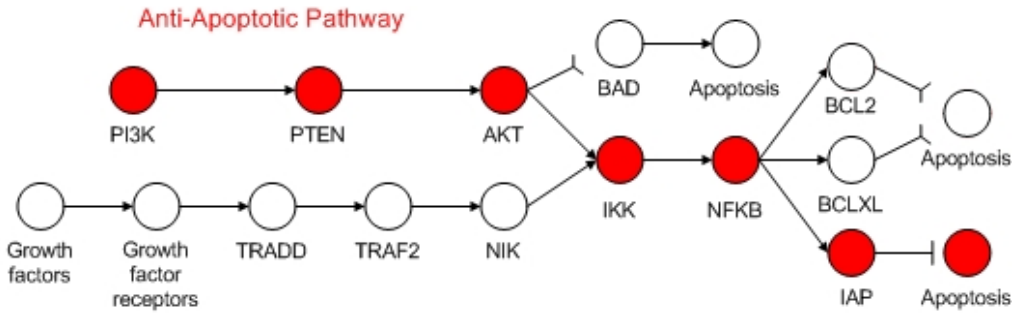
- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- **# of groups = $^{100000}C_5$**

$\Rightarrow$ **E(# of groups of genes correlated) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$**

$\Rightarrow$ **Even more false positives?**

- **Perhaps no need to consider every group**

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**

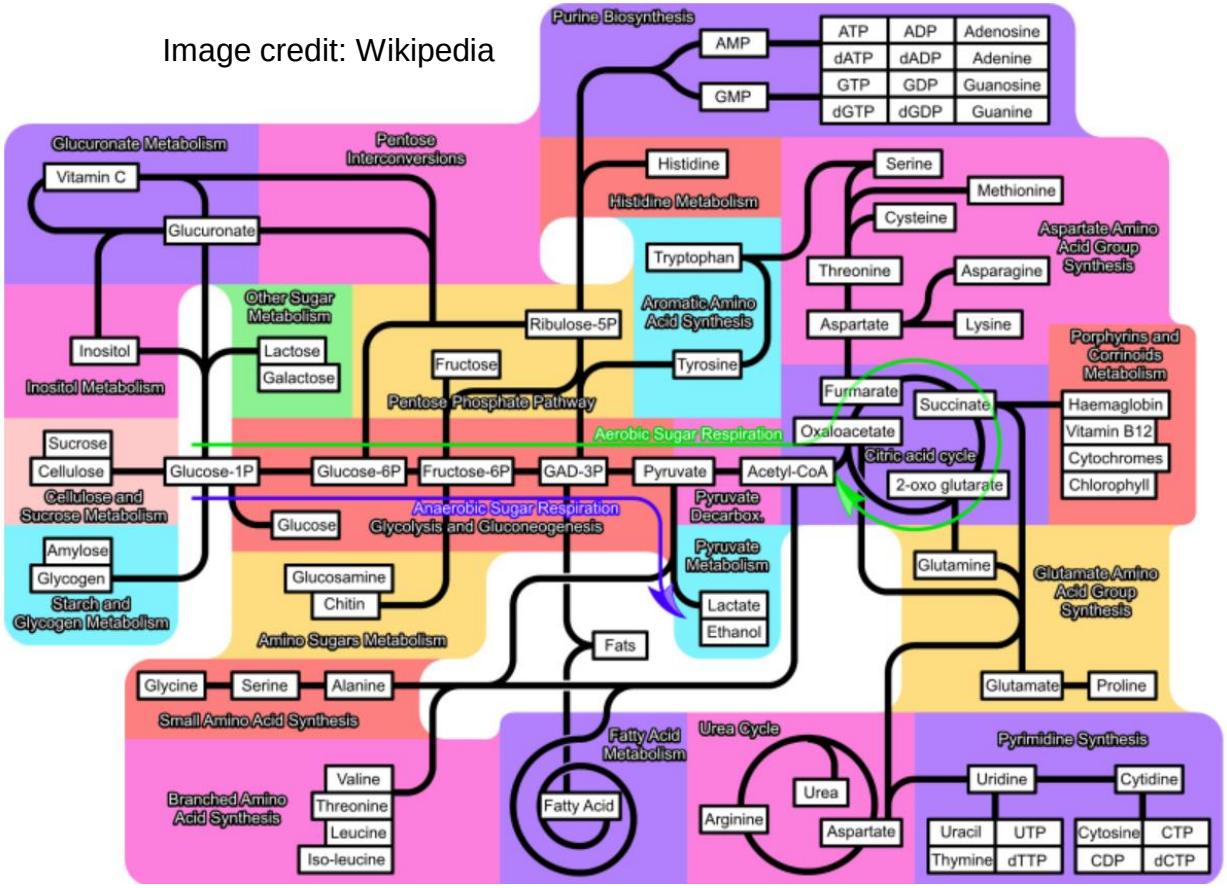
Types of Biological Networks

- **Natural biological pathways**
  - Metabolic pathway
  - Gene regulation network
  - Cell signaling network

- **Protein-protein interaction networks**

Metabolic Pathway

Image credit: Wikipedia



- **A series of biochem reactions in a cell**
  - Catalyzed by enzymes
  - Step-by-step modification of an initial molecule to form another product that can
    - **be used /store in the cell**
    - **initiate another metabolic pathway**

Gene Regulation Network

- **Gene regulation is the process that turns info from genes into gene products**
- **Gives a cell control over its structure & function**
  - Cell differentiation
  - Morphogenesis
  - Adaptability, …

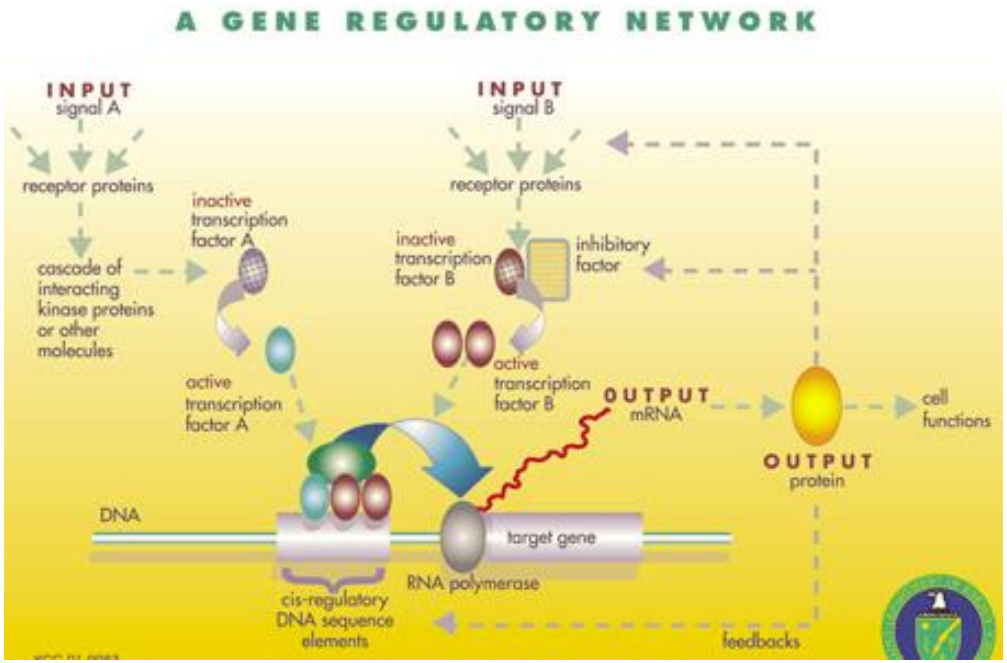


Image credit: Genome to Life

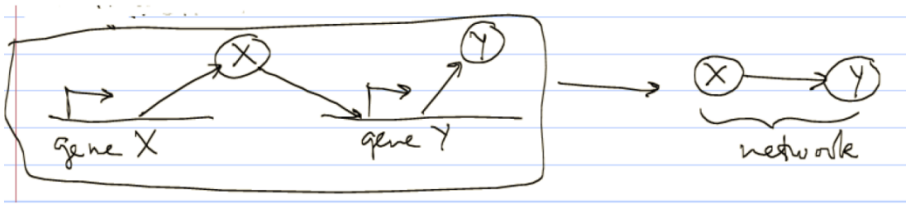


Image credit: Natasa Przulj

Cell Signaling Network

- **It is the entire set of changes induced by receptor activation**
  - Governs basic cellular activities and coordinates cell actions
- **Cells communicate with each other**
  - Direct contact (juxtacrine signaling)
  - Short distances (paracrine signaling)
  - Large distances (endocrine signaling)
- **Errors result in cancer, diabetes, ...**

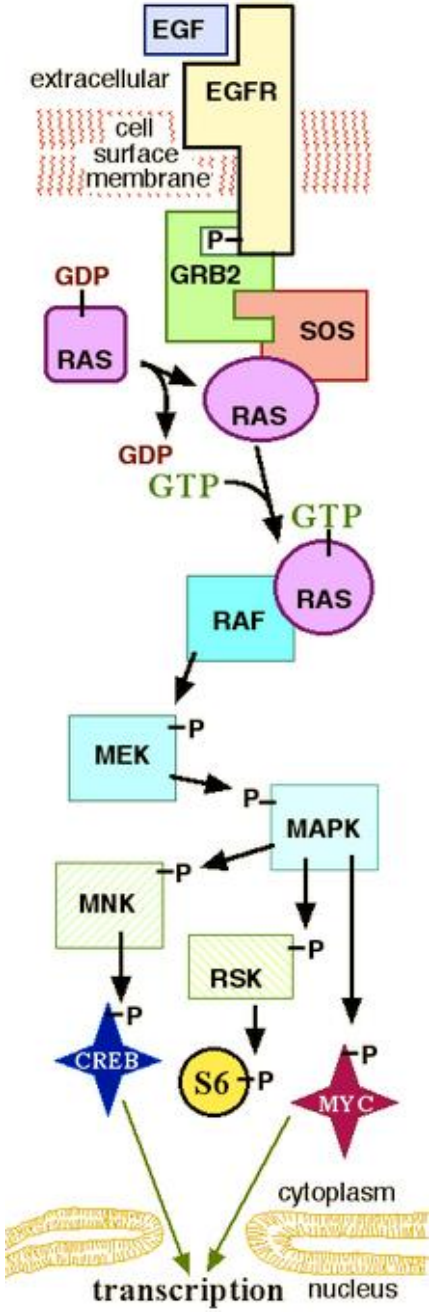
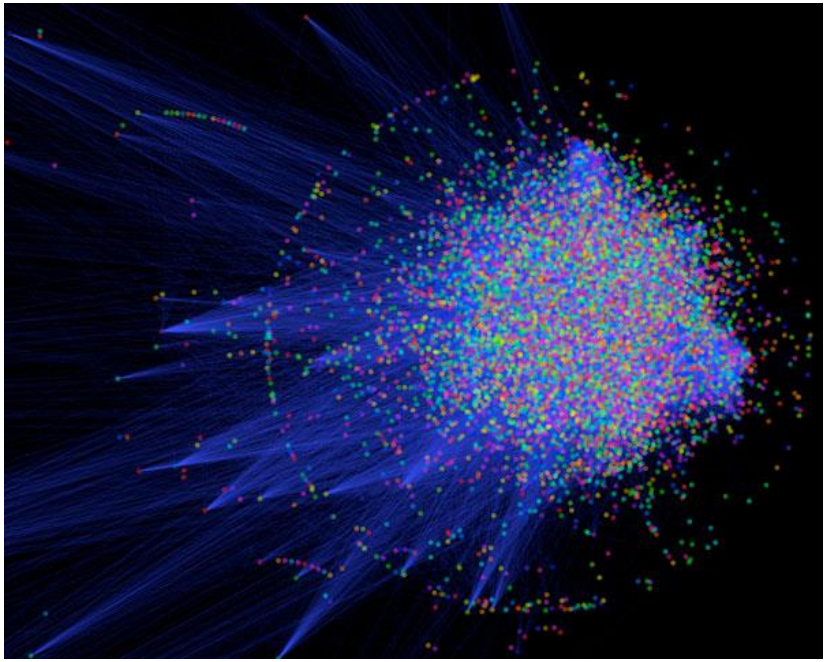


Image credit: Wikipedia

Protein Interaction Network (PPIN)

- **PPI usual refers to physical binding between proteins**
  - Stable interaction
    - **Protein complex**
    - **~70% of PPIs**
  - Transient interaction, modifying a protein for further actions
    - **Phosphorylation**
    - **Transportation**
    - **~30% of PPIs**



Visualization of the human interactome.
Image credit: Wikipedia

- **PPIN is usually a set of PPIs; it is not put into biological context**

NUS National University of Singapore

Sources of Biological Pathways

Database	Remarks
KEGG	KEGG (`http://www.genome.jp/kegg`) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.
WikiPathways	WikiPathways (`http://www.wikipathways.org`) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.
Reactome	Reactome (`http::/www.reactome.org`) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.
Pathway Commons	Pathway Commons (`http://www.pathwaycommons.com`) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.
PathwayAPI	PathwayAPI (`http://www.pathwayapi.com`) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh *et al.*, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.

Goh et al. "How advancement in biological network analysis methods empowers proteomics". *Proteomics*, 12(4-5):550-563, 2012

Sources of Protein Interactions

Database	# nodes, # edges	URL	Build Focus	Reference
BioGRID	10k, 40k	http://thebiogrid.org	Literature	(Stark *et al.*, 2006)
DIP	2.6k, 3.3k	http://dip.doe-mbi.ucla.edu	Literature	(Xenarios *et al.*, 2002)
HPRD	30k, 40k	http://www.hprd.org	Literature	(Prasad *et al.*, 2009)
IntAct	56k, 267k	http://www.ebi.ac.uk/intact	Literature	(Aranda *et al.*, 2010)
MINT	30k, 90k	http://mint.bio.uniroma2.it/mint	Literature	(Chatr-aryamontri *et al.*, 2007)
STRING	5200k, ?	http://string-db.org	Literature, Prediction	(Szklarczyk *et al.*, 2011)

Goh et al. "How advancement in biological network analysis methods empowers proteomics". *Proteomics*, 12(4-5):550-563, 2012

and Protein Complexes

- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, *NAR*, 2010

Taming false positives by considering pathways instead of all possible groups

NUS National University of Singapore

Group of Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **What is the chance of a group of 5 genes being perfectly correlated to these samples?**

- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- ~~**# of groups = $^{100000}C_5$**~~
- ~~**E(# of groups of genes correlated) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$**~~

⇒ **Even more false positives?**

- **Perhaps no need to consider every group**

# of pathways = 1000

E(# of pathways correlated) = $1000 * (1/2^6)^5 = 9.3*10^{-7}$

Towards More Meaningful Genes

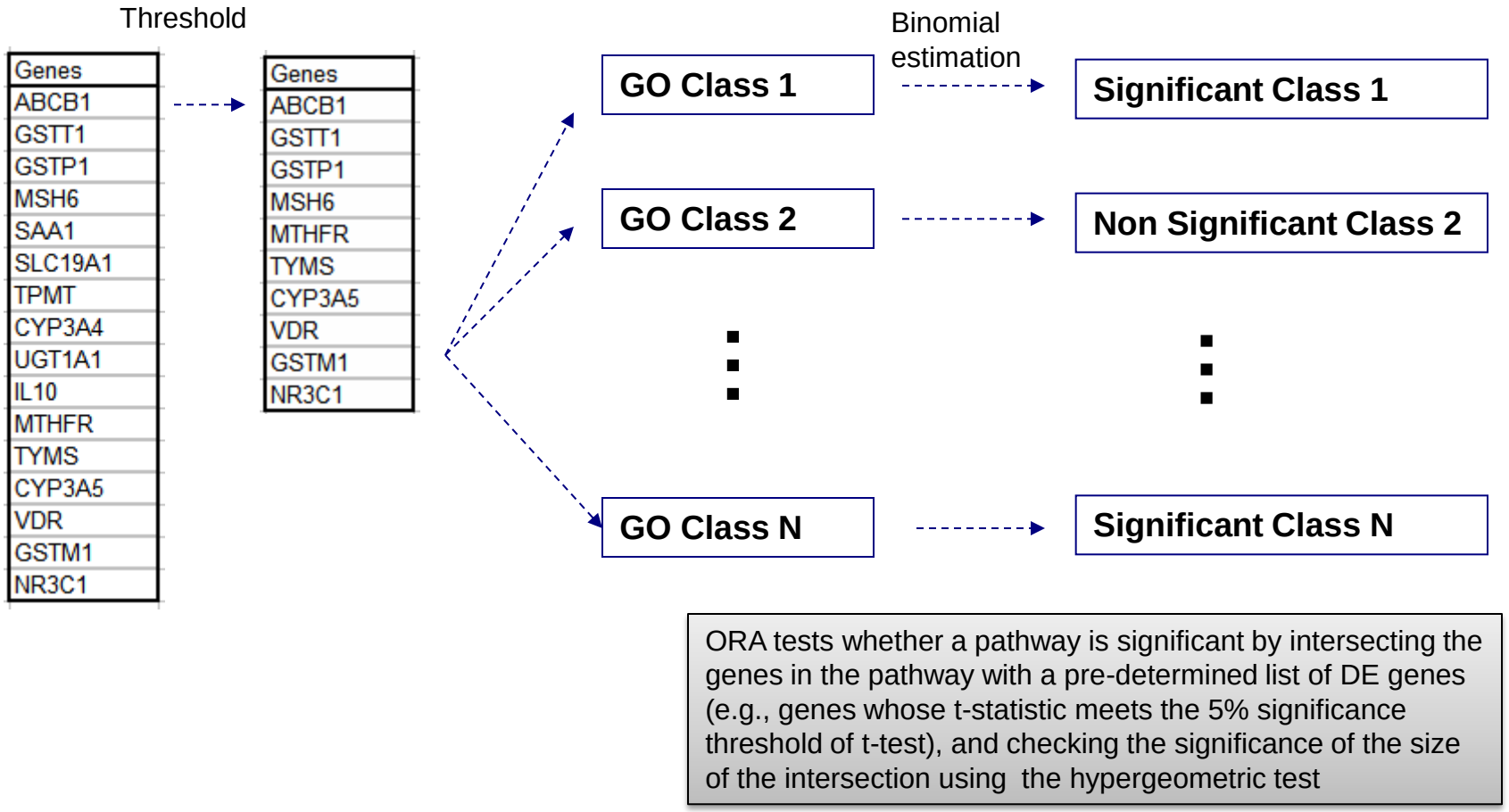
- **ORA**
  - Khatri et al
  - *Genomics*, 2002
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - *PNAS*, 2005
- **Snet, PFSNet**
  - Lim & Wong
  - *Bioinformatics*, 2014

Overlap Analysis

Direct-Group Analysis

Network-Based Analysis

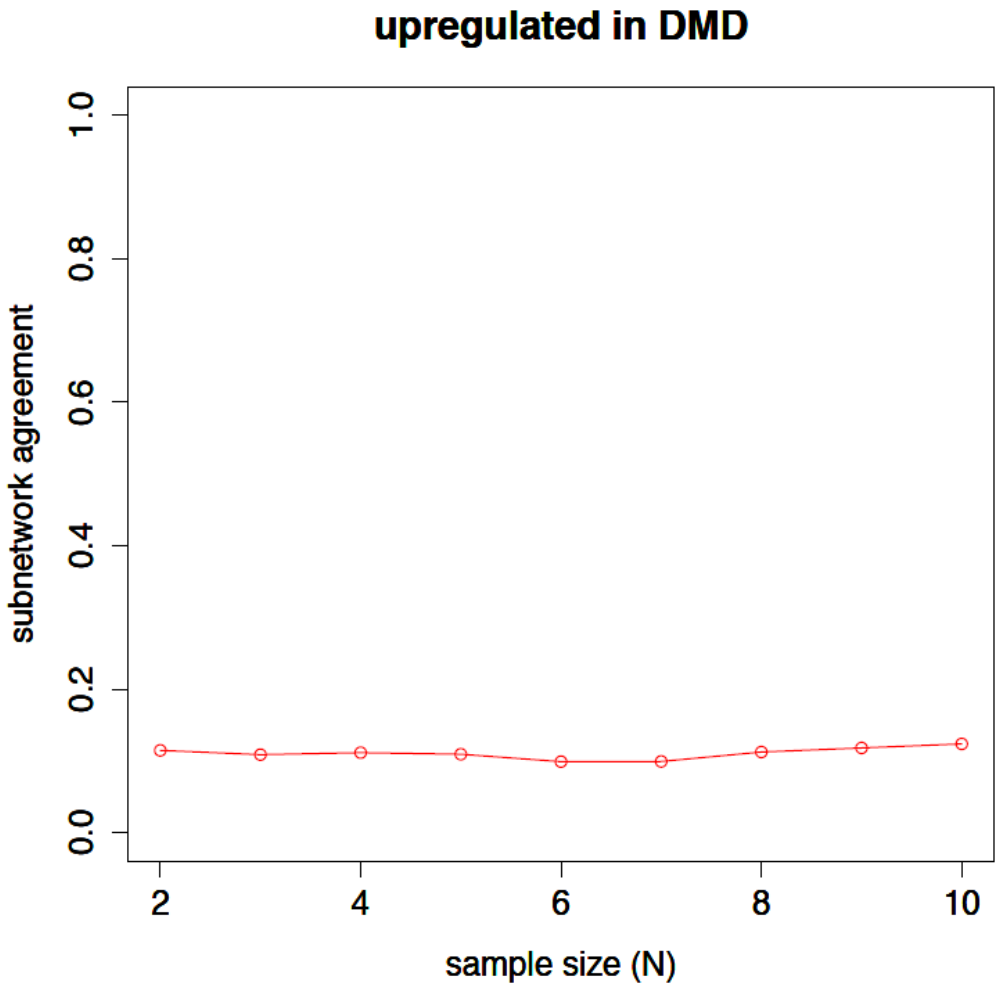
Overlap Analysis: ORA



ORA tests whether a pathway is significant by intersecting the genes in the pathway with a pre-determined list of DE genes (e.g., genes whose t-statistic meets the 5% significance threshold of t-test), 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.

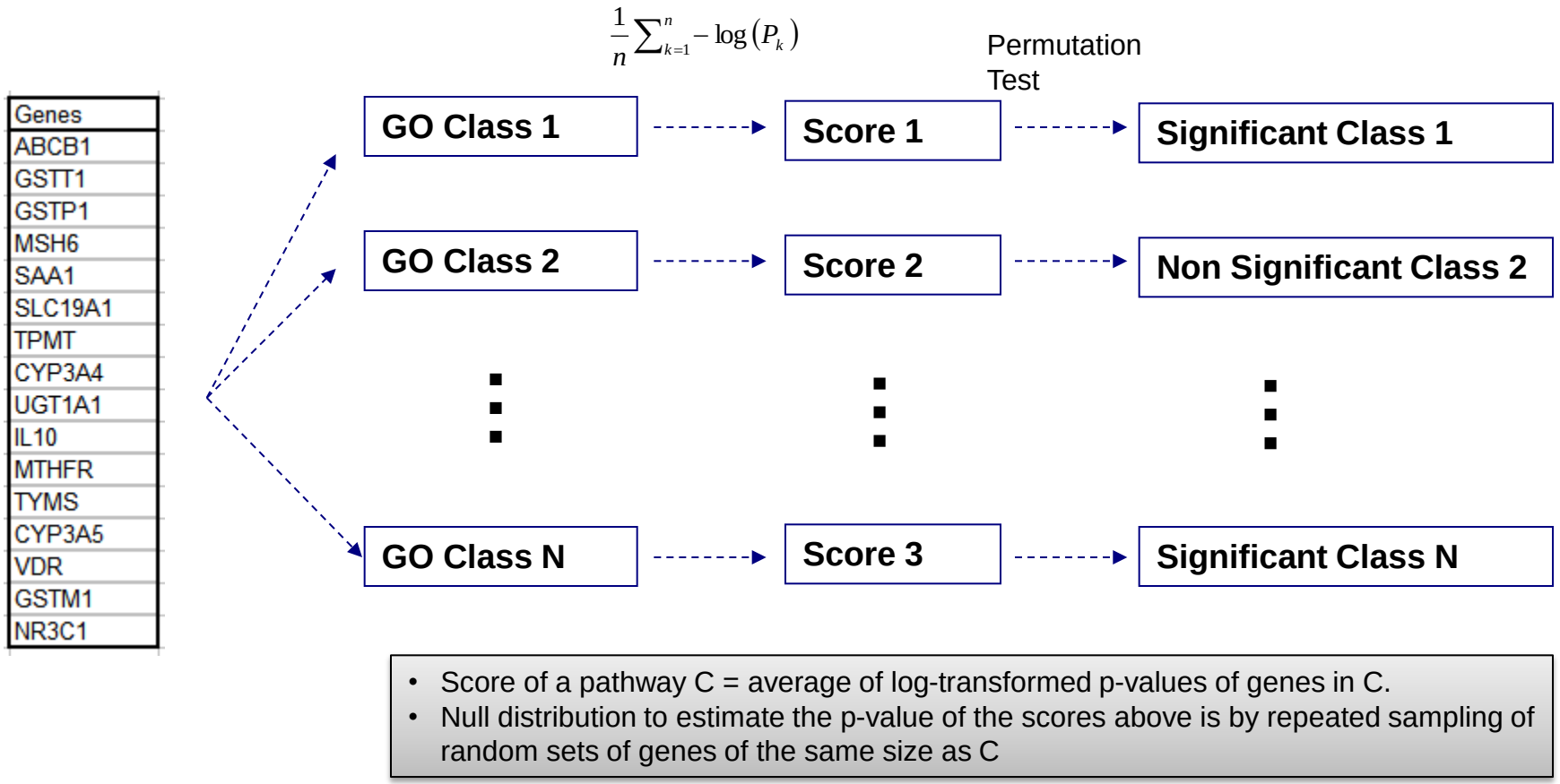
Disappointing Performance. Why?



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

Pathway data
- PathwayAPI, Soh et al., 2010

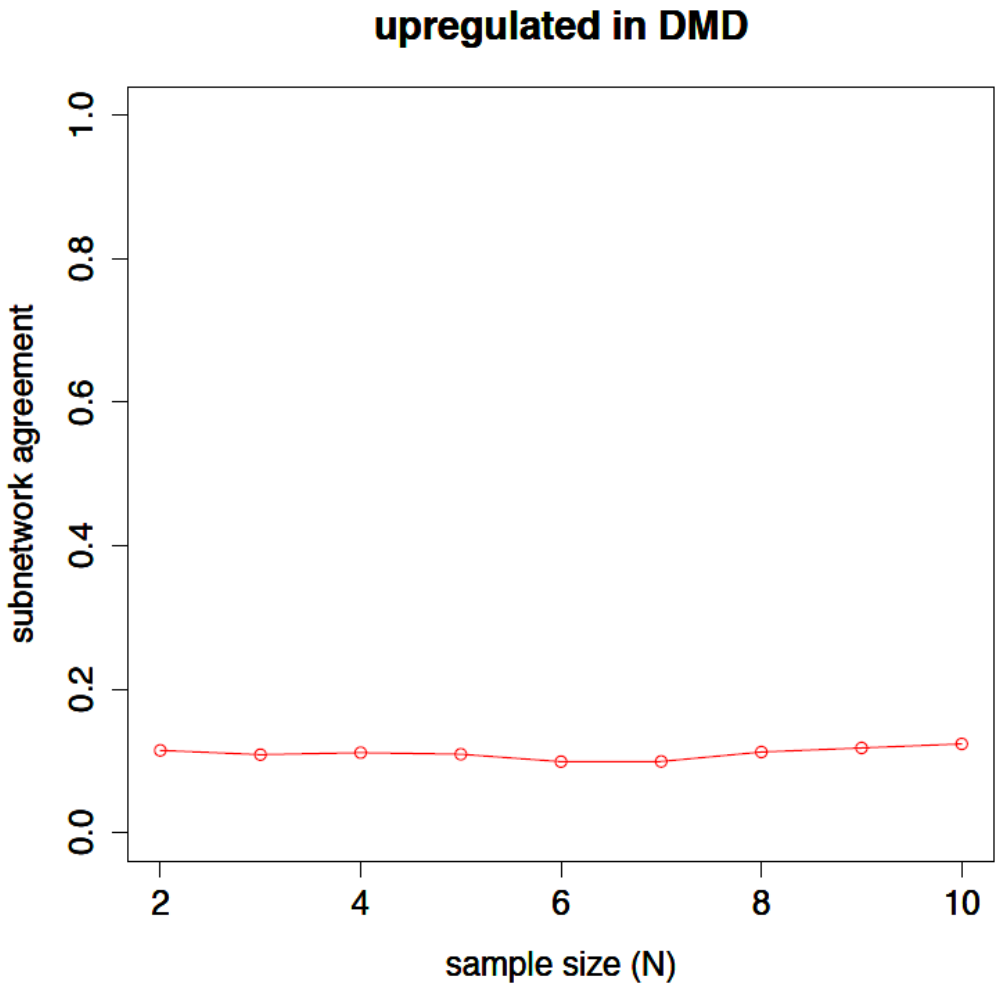
Direct-Group Analysis: FCS



- Score of a pathway C = average of log-transformed p-values of genes in C.
- Null distribution to estimate the p-value of the scores above is by repeated sampling of random sets of genes of the same size as C

Pavlidis et al. “Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex”. *Neurochem Res*., 29(6):1213-1222, 2004.

Where will FCS be in comparison to ORA below? Why?



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

Pathway data
- PathwayAPI, Soh et al., 2010

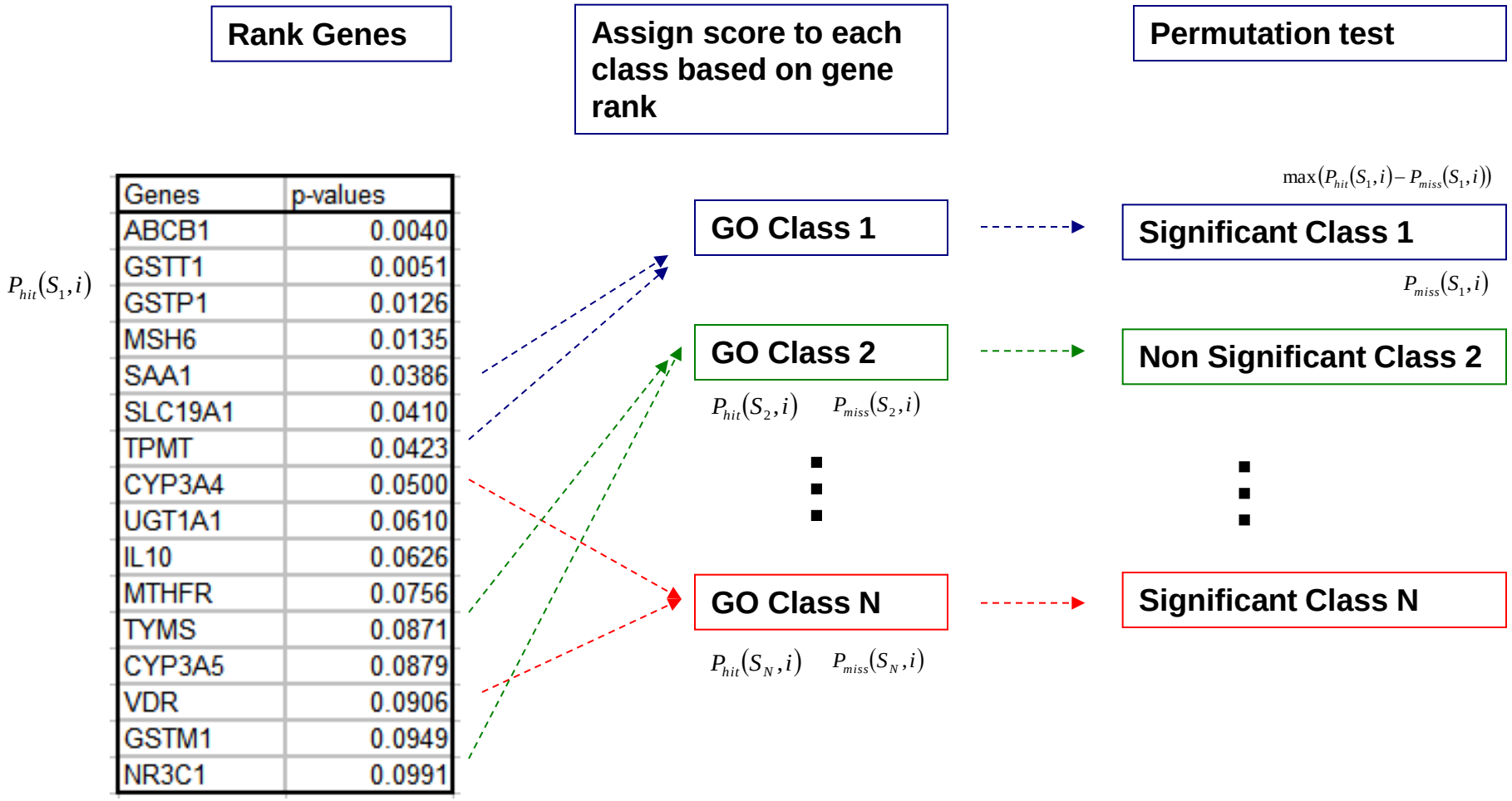
A problem w/ FCS

- **Its null hypothesis:**
  - "genes in a pathway C are independently expressed & not diff from other genes
- **But …**
  - Genes in a pathway are not independent

  $\Rightarrow$ Becomes over sensitive
- **Solution: generate null distribution by randomizing patient class labels**

Goeman & Buhlmann. "Analyzing gene expression data in terms of gene sets: Methodological issues". *Bioinformatics*, 23(8):980-987, 2007

Direct-Group Analysis: GSEA



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

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

GSEA: Key Points

- **"Enrichment score"**
  - The degree that the genes in pathway C are enriched in the extremes of ranked list of all genes
  - Measured by Komogorov-Smirnov statistic

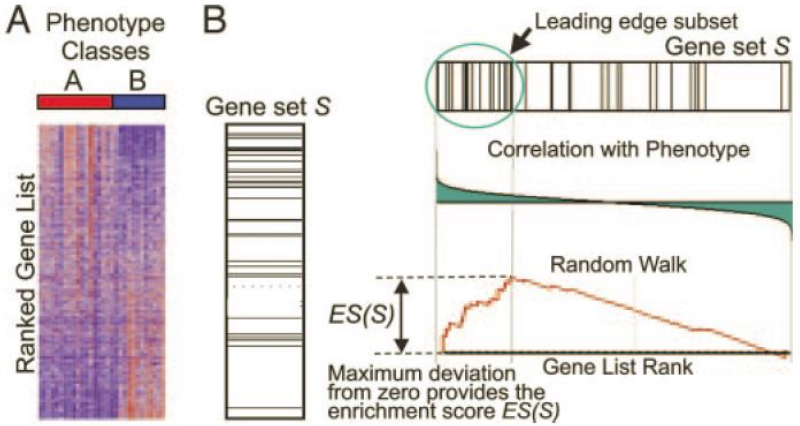
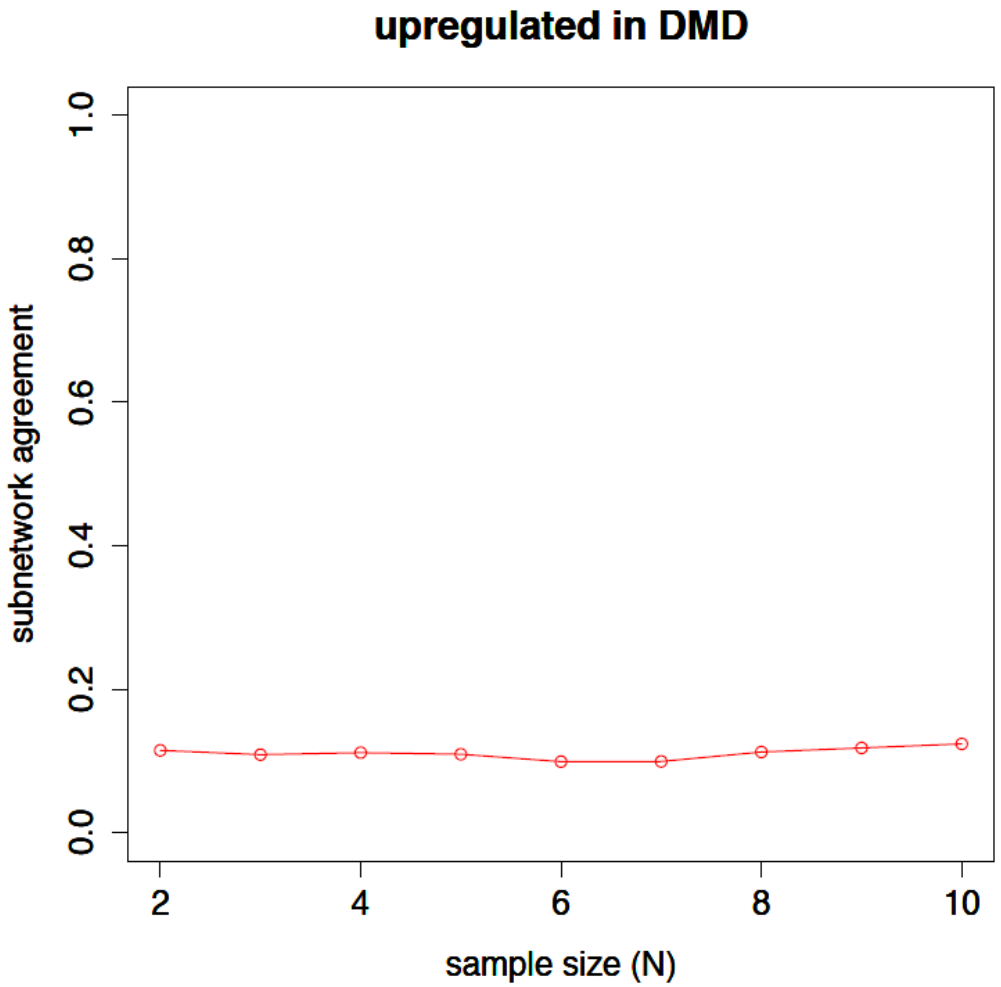


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Subramanian et al., *PNAS*, 102(43):15545-15550, 2005

- **Null distribution to estimate the p-value of the scores above is by randomizing patient class labels**

Where will GSEA be in comparison to ORA and FCS? Why?

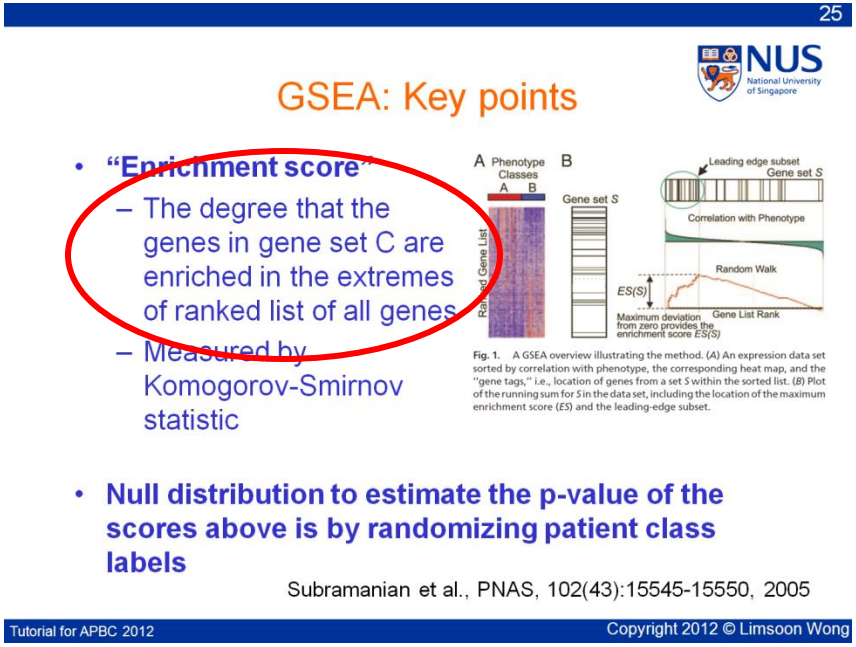


DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010

Wong. "Using Biological Networks in Protein Function Prediction and Gene Expression Analysis". *Internet Mathematics*, 7(4):274--298, 2011.

A problem w/ GSEA



- **Its enrichment score considers all genes in C**
- **But …**
  - Not all branches of a large pathway have to "go wrong"

  ⇒ Cannot detect if only a small part of a pathway malfunctions
- **Solution: Break pathways into subnetworks**

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

Network-Based Analysis: SNet

- **Group samples into type D and ¬D**
- **Extract & score subnetworks for type D**
  - Get list of genes highly expressed in most D samples
    - **These genes need not be differentially expressed!**
  - Put these genes into pathways
  - Locate connected components (ie., candidate subnetworks) from these pathway graphs
  - Score subnetworks on D samples and on ¬D samples
- **For each subnetwork, compute t-statistic on the two sets of scores**
- **Determine significant subnetworks by permutations**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

NUS National University of Singapore

SNet: Score Subnetworks

Step 2: Subnetwork Scoring We assign a score vector $SN_{sn,d}^{u\_score}$ with respect to phenotype d to each subnetwork sn within SN^{List} according to Equation 1.

$$SN_{sn,d}^{u\_score} = \langle SN_{sn,1,d}^{i\_score}, SN_{sn,2,d}^{i\_score}, ..., SN_{sn,n,d}^{i\_score} \rangle \quad (1)$$

Where n is the number of patients in phenotype d. The formula $SN_{sn,i,d}^{i\_score}$ for the i^{th} patient (also the i^{th} element of this vector) is given by:

$$SN_{sn,i,d}^{i\_score} = \sum_{j=1}^{g} G_{sn,j,d}^{score} \quad (2)$$

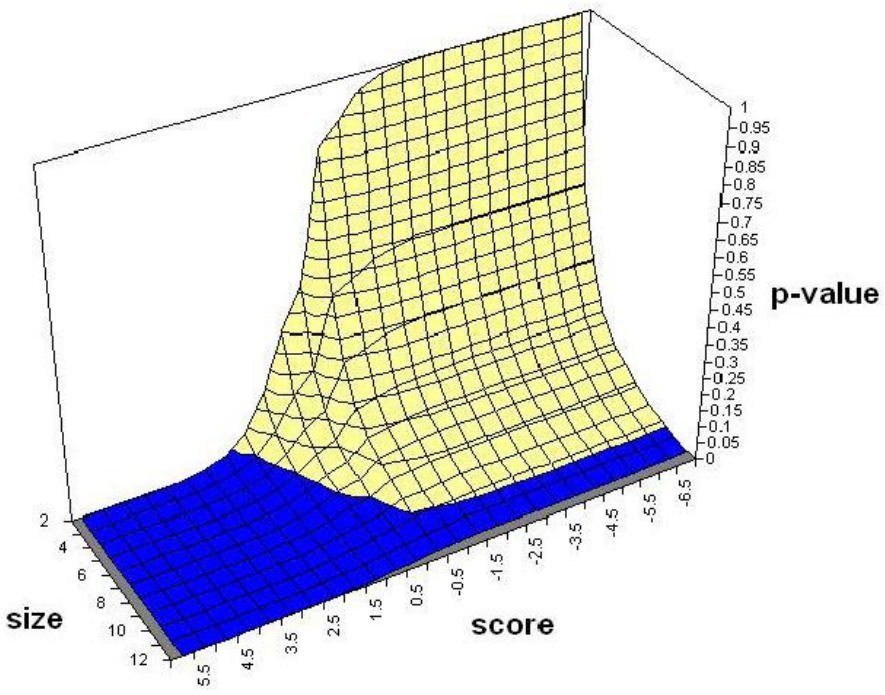
$G_{sn,j,d}^{score}$ refers to the score of the j^{th} gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G_{sn,j,d}^{score}$ is given by Equation 3) and is simply given by:

$$G_{sn,j,d}^{score} = k/n \quad (3)$$

Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d. The entire Step 2 is repeated for the other disease phenotype $\neg d$, giving us the score vectors, $SN_{sn,d}^{u\_score}$ and $SN_{sn,\neg d}^{u\_score}$ for the same set of connected components. The t-test is finally calculated between these two vectors, creating a final t-score for each subnetwork sn within SN_{List}.

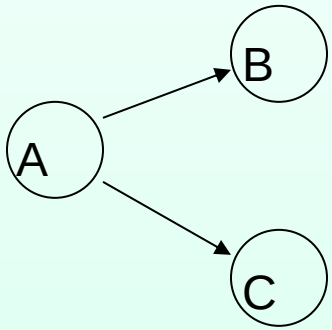
SNet: Significant Subnetworks

- **Randomize patient samples many times**
- **Get t-score for subnetworks from the randomizations**
- **Use these t-scores to establish null distribution**
- **Filter for significant subnetworks from real samples**



Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Key Insight # 1



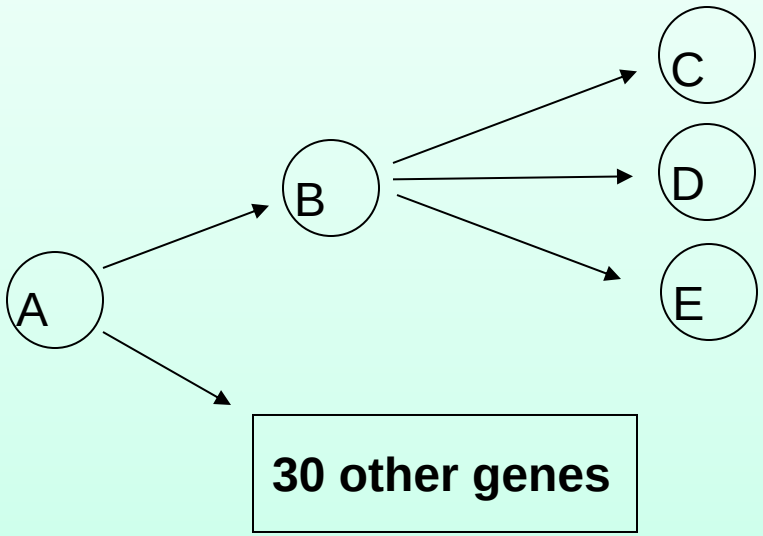
Genes A, B, C are high in phenotype *D*

A is high in phenotype ~*D* but B and C are not

Conventional techniques: Gene B and Gene C are selected. Possible incorrect postulation of mutations in gene B and C

- **SNet does not require all the genes in subnet to be diff expressed**
- **It only requires the subnet as a whole to be diff expressed**
- **Able to capture entire relationship, postulating a mutation in gene A**

Key Insight # 2



A branch within pathway consisting of genes A, B, C, D and E are high in phenotype *D*

Genes C, D and E not high in phenotype *~D*

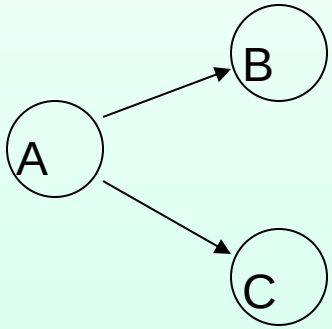
30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

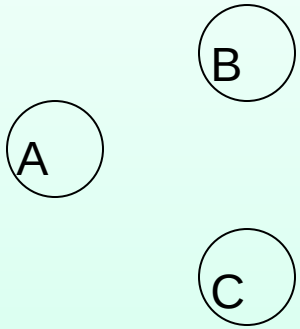
- **SNet: Able to capture the subnetwork branch within the pathway**

Key Insight # 3

Pathway 1



Pathway 2



Genes A, B and C are present in two separate pathways

A, B and C are high in phenotype *D*, but not high in phenotype *~D*

Conventional techniques:

Both pathways are scored equally. So both got selected, resulting in pathway 2 being a false positive

- **SNet: Able to select only pathway 1, which has the relevant relationship**

Let's see whether SNet gives us subnetworks that are

(i) more consistent between datasets of the same types of disease samples

(ii) larger and more meaningful

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Better Subnetwork Overlap

Table 1. Table showing the percentage overlap significant subnetworks between the datasets. Each row refers to a separate disease (as indicated in the first column). Each disease is tested against two datasets depicted in the second and third column. The overlap percentages refer to the pathway overlaps obtained from running SNet (column 4) and GSEA (column 5) The actual number of overlaps are parenthesized in the same columns.

Disease	Dataset 1	Dataset 2	SNet	GSEA
Leuk	Golub	Armstrong	83.3% (20)	0.0% (0)
Subtype	Ross	Yeoh	47.6% (10)	23.1% (6)
DMD	Haslett	Pescatori	58.3% (7)	55.6% (10)
Lung	Bhatt	Garber	90.9% (9)	0.0% (0)

- **For each disease, take significant subnetworks from one dataset and see if it is also significant in the other dataset**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Better Gene Overlaps

Table 2. Table showing the number and percentage of significant overlapping genes. γ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding algorithms.

Disease	γ	SNet	GSEA	SAM	t-test
Leuk	84	91.3%	2.4%	22.6%	14.3%
Subtype	75	93.0%	4.0%	49.3%	57.3%
DMD	45	69.2%	28.9%	42.2%	20.0%
Lung	65	51.2%	4.0%	24.6%	26.2%

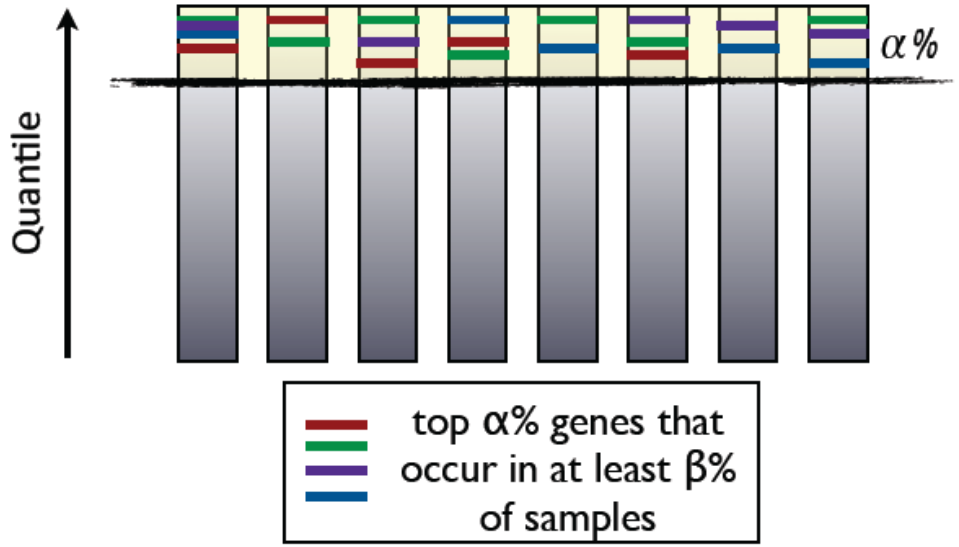
- **For each disease, take significant subnetworks extracted independently from both datasets and see how much their genes overlap**

Soh et al. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Larger Subnetworks

Table 3. Table comparing the size of the subnetworks obtained from the t-test and from SNet. The first column shows the disease and the second column shows the number of genes which comprised of the subnetworks. The third and fourth column depicts the number of genes present within each subnetwork for the t-test and SNet respectively. So for instance in the leukemia dataset, we have 8 subnetworks with size 2 genes, 1 subnetwork with size 3 genes for the t-test. For SNet, we have 2 subnetworks with size 5 genes, 3 subnetworks with size 6 genes, 2 subnetworks with size 7 genes and 1 subnetwork with a size of $\geq$ 8 genes

Disease	γ	Num Genes (t-test)				Num Genes (SNet)			
		2	3	4	5	5	6	7	$\geq$ 8
Leuk	84	8	1	0	0	2	3	2	1
Subtype	75	5	1	1	1	1	0	1	6
DMD	45	3	1	0	0	1	0	0	5
Lung	65	3	2	1	0	5	3	0	1



Issue #1 with SNet

Fig. 2. In SNet, the top $\alpha\%$ of genes of each sample in phenotype D is highlighted in yellow. A subset of these genes that are represented in color bands are in at least $\beta\%$ of the samples are then taken to induce subnetworks.

- What if the real important genes are close to, but not in, the top $\alpha\%$ most highly expressed genes?
- Blindly increasing α does not help, as this will bring in lots of false-positive genes

Issue #2 with SNet

$$SN^{i\_score}_{sn,i,d} = \sum_{j=1}^{g} G^{score}_{sn,j,d} \quad (2)$$

$G^{score}_{sn,j,d}$ refers to the score of the j^{th} gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G^{score}_{sn,j,d}$ is given by Equation 3) and is simply given by:

$$G^{score}_{sn,j,d} = k/n \quad (3)$$

Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d.

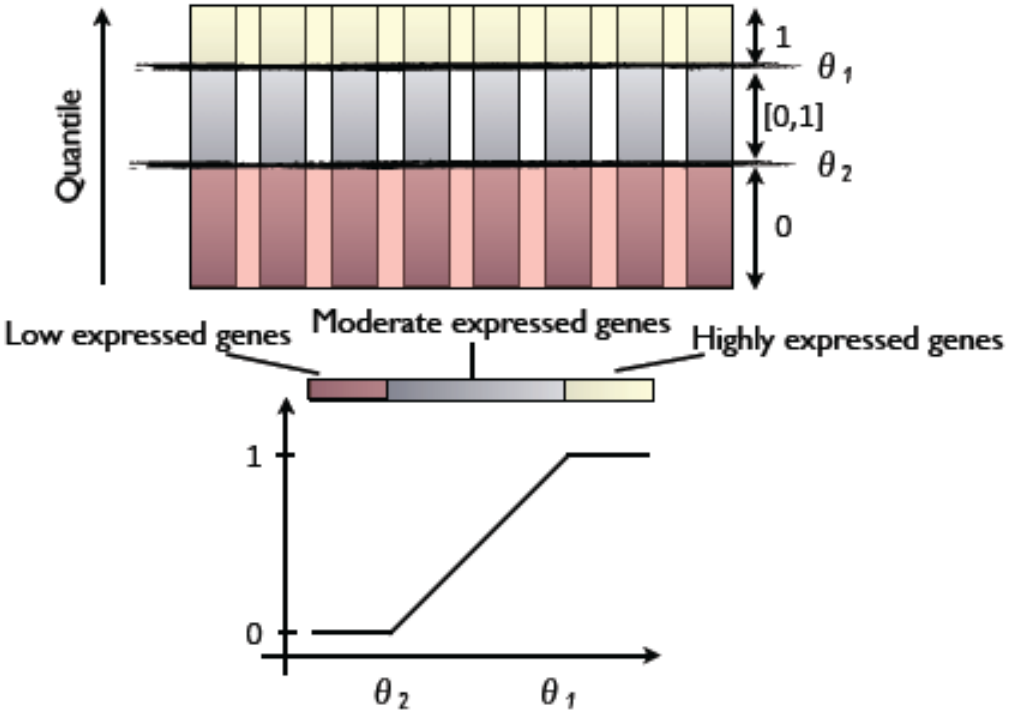
- **SNet weighs genes & scores subnetworks only on the basis of phenotype D**
- **Why not consider phenotype ~D as well?**

PFSNet

- **Deal with issue #1 of SNet using "fuzzification"**
- **Deal with issue #2 of SNet using paired t-test**

$\Rightarrow$ **PFSNet – Paired Fuzzy SNet**

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



Fuzzification

Our goal in this step is to compute a gene list, which segregates the pathways into smaller components. The voting criteria that determines whether the gene g_i is accepted into this gene list is given below:

$$\sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|} > \beta \qquad (1)$$

where D is the phenotype for which the subnetwork is generated, p_j ranges over the patients of phenotype D and fs is the fuzzy function which converts the gene expression value e_{g_i,p_j} to a value between 0 and 1.

In PFSNet, instead of computing the gene scores with respect to phenotype D, we also compute the gene scores with respect to phenotype $\neg D$. Hence, each node is given scores which we denote as $\beta_1^*(g_i)$ and $\beta_2^*(g_i)$, computed as follows:

$$\beta_1^*(g_i) = \sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|}, \quad \beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{fs(e_{g_i,p_j})}{|\neg D|} \qquad (4)$$

Accordingly, for every subnetwork S, each patient of phenotype D can be scored under β_1^* and β_2^*, as follows:

$$Score_1^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_1^*(g_i), \qquad (5)$$

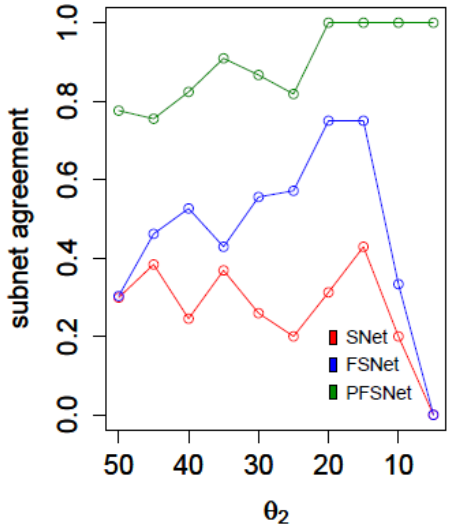
$$Score_2^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_2^*(g_i) \qquad (6)$$

Paired T-Test

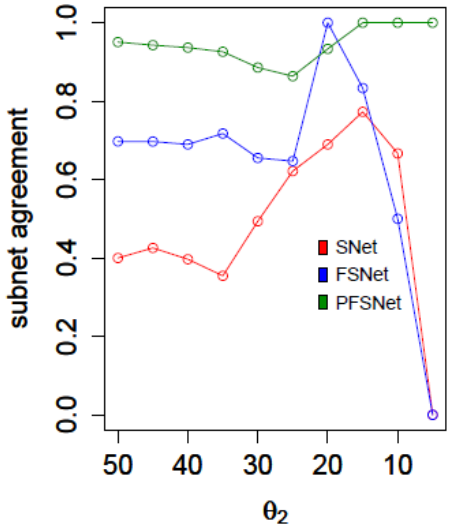
- **$Score^{Pk}_1(S)$ and $Score^{Pk}_2(S)$ are computed for the same sample Pk and subnetwork S**

$\Rightarrow$ **Can do paired t-test**

  - Null hypothesis: If S is irrelevant to D vs ~D, we expect $Score^{Pk}_1(S) - Score^{Pk}_2(S)$ to be around 0

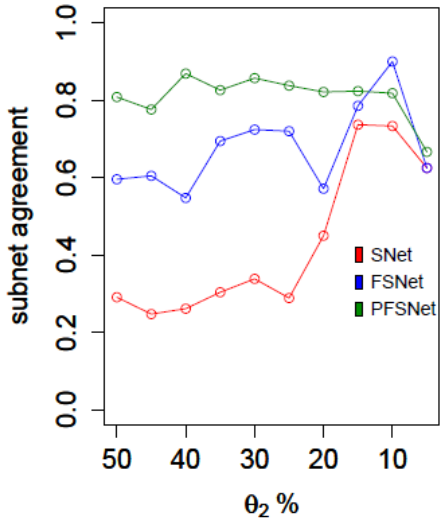


upregulated in ALL

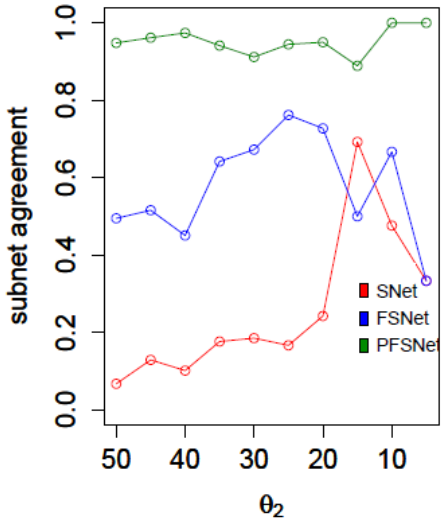


upregulated in AML

Fig. 4: Consistency of subnetworks in Leukemia dataset



upregulated in DMD

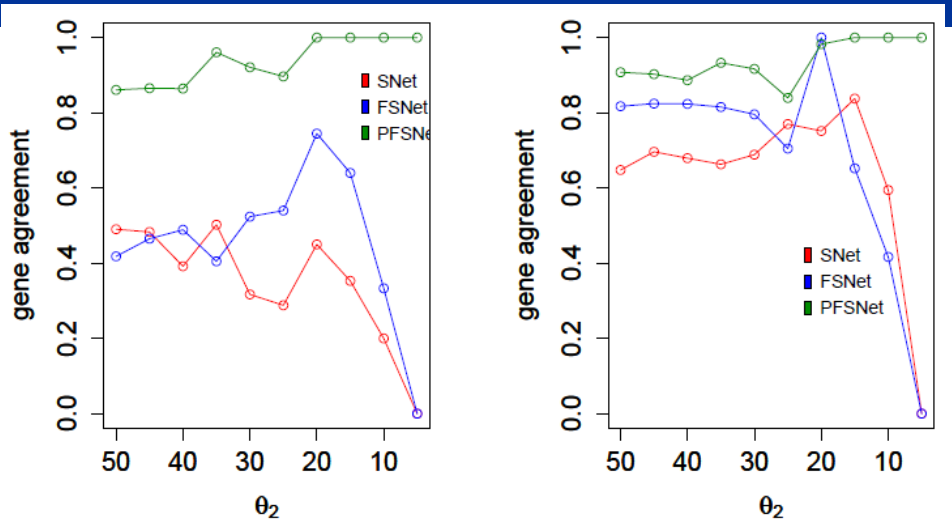


upregulated in NORM

Fig. 6: Consistency of subnetworks in DMD dataset

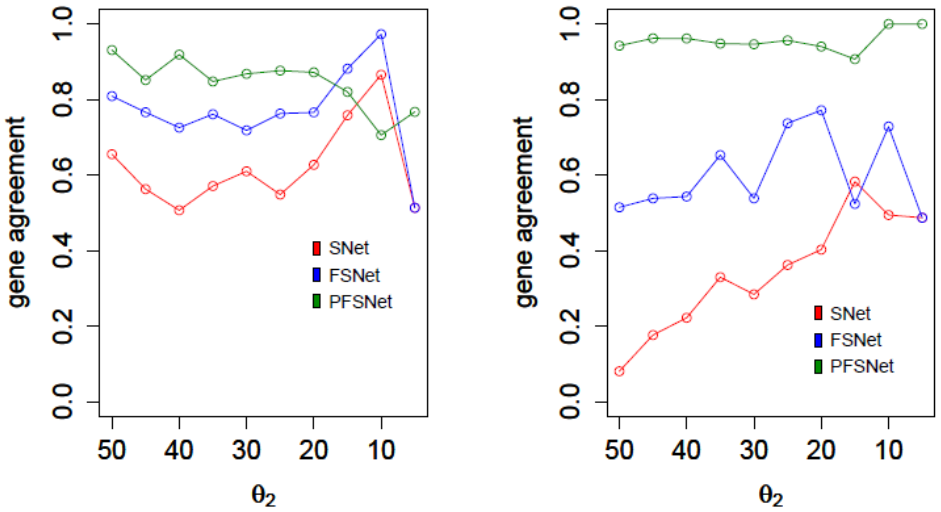
PSFNet vs SNet: Subnet Agreement

Overlap = |A∩B| / |A∪B|



upregulated in ALL upregulated in AML

Fig. 7: Consistency of genes in Leukemia dataset



upregulated in DMD upregulated in NORM

Fig. 9: Consistency of genes in DMD dataset

PSFNet vs SNet: Gene Agreement

Overlap = $|A \cap B| / |A \cup B|$

PFSNet vs GSEA & GGEA: Pathway Agreement

Dataset	PFSNet	FSNet	GSEA	GGEA
Leukemia	1.00	0.75	0.12	0.18
ALL (subtype)	0.56	0.38	0.34	0.37
DMD	0.82	0.79	0.57	0.51

For PFSNet and FSNet, threshold values of $\theta_1 = 0.95, \theta_2 = 0.85$ are used.

Overlap = $|A \cap B| / |A \cup B|$

PFSNet vs T-Test: Gene Agreement

Table 2. Comparing gene-level agreement of PFSNet, FSNet, SNet, GSEA, SAM, t-test.

Dataset	PFSNet		FSNet		SNet		GSEA		SAM(5% sig)		SAM(top 100)		t-test(5% sig)		t-test(top 100)	
	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D
Leukemia	1.00	0.81	0.64	0.42	0.35	0.58	0.12	0.20	0.50	0.47	0.01	0.01	0.35	0.29	0.19	0.07
ALL (subtype)	0.54	0.70	0.38	0.41	0.29	0.57	0.04	0.04	0.19	0.27	0.12	0.21	0.08	0.10	0.01	0.00
DMD	0.82	0.72	0.88	0.75	0.76	0.54	0.44	0.20	0.34	0.08	0.27	0.19	0.41	0.19	0.11	0.25

For PFSNet and FSNet, threshold values of $\theta_1 = 5\%, \theta_2 = 15\%$ are used. D represents subnetworks enriched in phenotype D and $\neg D$ represents subnetworks enriched in phenotype $\neg D$. For GSEA, the "leading edge genes" were used. For SAM and t-test, we took genes at 5% significance level and also the top n genes indicated in brackets.

Overlap = $|A \cap B| / |A \cup B|$

PFSNet vs GSEA & GGEA: Pathway Agreement

Dataset	PFSNet	FSNet	GSEA	GGEA
Leukemia	1.00	0.75	0.12	0.18
ALL (subtype)	0.56	0.38	0.34	0.37
DMD	0.82	0.79	0.57	0.51

Testing subnets from PFSNet using GSEA & GGEA

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

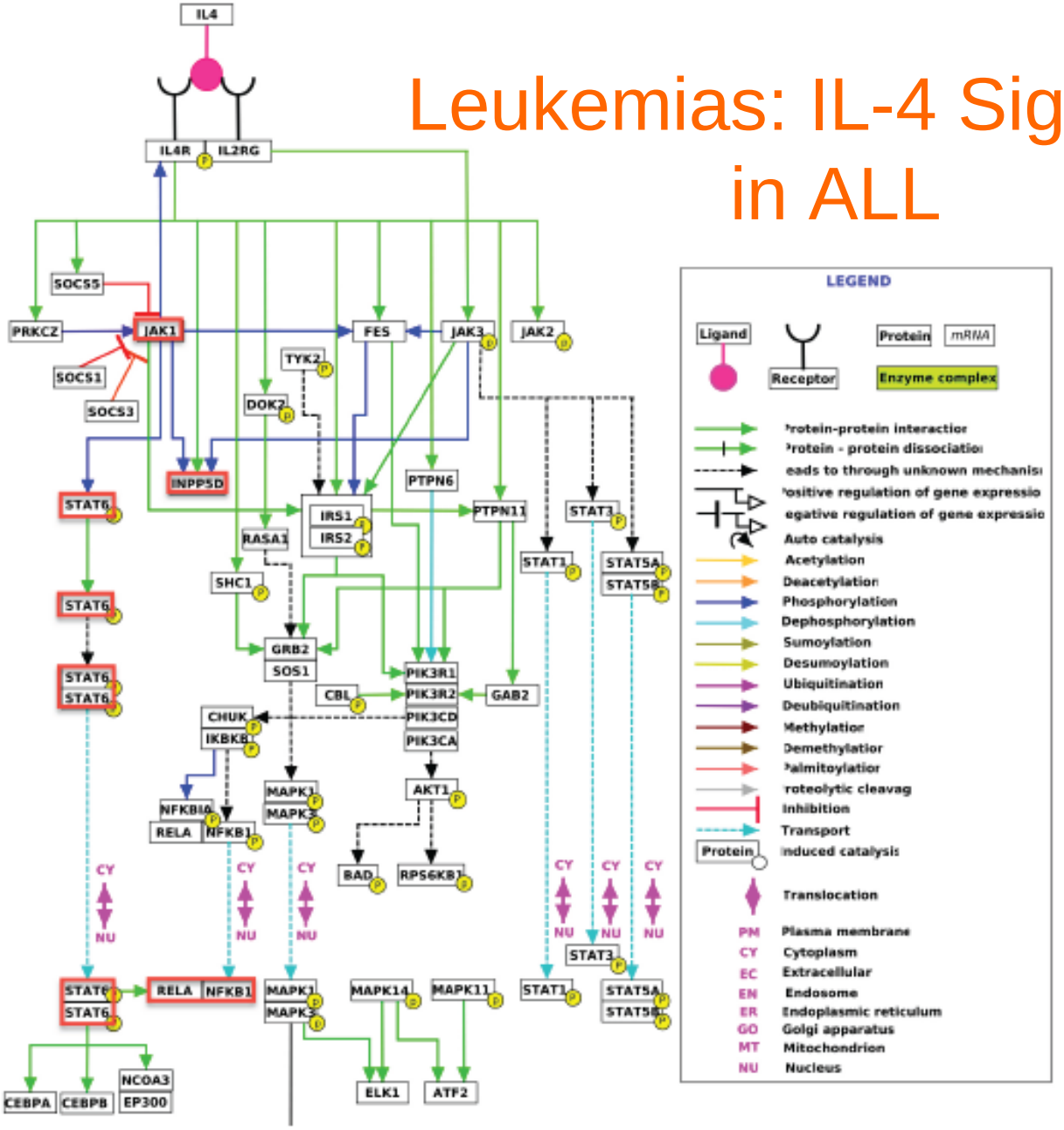
Top 5 Subnets

Table 4. Top 5 subnetworks that have biological significance.

Leukemia	ALL subtype	DMD
Proteasome Degradation	Wnt Signaling*#	Striated Muscle Contraction*#
IL-4 Signaling*#	Antigen Processing	Integrin Signaling
Antigen Processing*	Jak-STAT Signaling*#	VEGF Signaling*
B-Cell Receptor Signaling#	T-Cell Receptor Signaling	Tight Junction
Wnt Signaling*#	Adherens Junction*#	Actin Cytoskeleton Signaling

* indicates subnetworks that were not found in SNet and # indicates pathways that were missed by GSEA

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.

What have we learned?

- **Common headaches in gene expression analysis**
  - Natural fluctuation, protocol noise, batch effect
- **Use of biological background info to tame false positives**
- **Overlap analysis → direct-group analysis → network-based analysis**
- **Subnetwork-based methods yield more consistent and larger disease subnetworks**

Still a major challenge

- **Suppose there are very few samples, so few that you cannot estimate the p-value by permuting class labels**

- **What do you do?**

From pathways to models, From static to dynamic:

A couple of very recent papers that are worth your leisure reading…

- Geistlinger et al. **From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems**. *Bioinformatics*, 27(13):i366—i373, 2011
- Zampieri et al. **A system-level approach for deciphering the transcriptional response to prion infection**. *Bioinformatics*, 27(24): 3407--3414, 2011

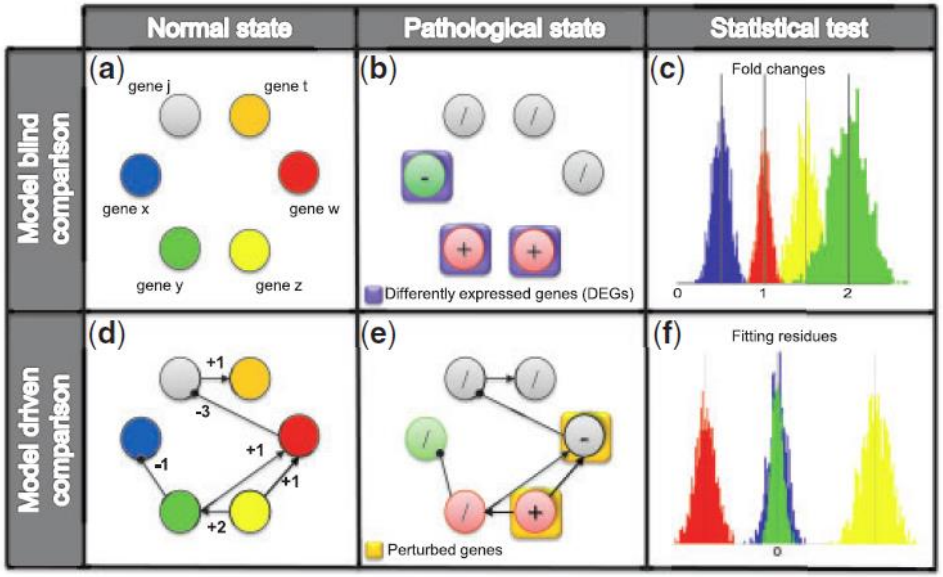


Fig. 1. System response inference: a toy genetic network consisting of six genes exemplifies the advantages of using a system-level data comparison (**a**). Standard statistical tests (i.e. *t*-test) unveil significant fold change in gene expression variations for each transcript individually (**b**), neglecting the underlying regulatory network. Such statistical test can identify whether the expression level of a transcript is significantly changed with respect to a reference. Putative gene expression changes are reported in panel (**c**). In this specific example, two genes are identified to be overexpressed [red/+ nodes] and one downregulated (green/- node), while the remaining three do not show any changes (grey nodes). By knowing the corresponding genetic regulatory network (**d**), we can discriminate the coherent variations from the unexpected ones. As shown in the example, two of the genes that showed a significant expression variations are consistent with model predictions i.e. the expression changes of genes *x* and *y* can be explained by the variation of gene *z*. This is reflected by a skew distribution of discrepancies (i.e. residues), between model predictions and observed data, centered around 0 (**f**). At the same time, one transcript, *w*, is not responding coherently to the initial model. The fact that its expression is unchanged, when it should have been increased, might relate to an anomalous direct effect of the pathology, preventing a synergistic response between all the genes in the system. Hence, the list of 'perturbed genes' can be sensibly different from the standard DEGs identified from individual fold change analysis (**b**/**e**).

Must Read

- [NEA] Sivachenko et al. **Molecular networks in microarray analysis**. *JBCB*, 5(2b):429-546, 2007

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

- Wong. **Using Biological Networks in Protein Function Prediction and Gene Expression Analysis.** *Internet Mathematics*, 7(4):274--298, 2011

Good to Read

- Zhang et al. **Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes**. *Bioinformatics*, 25(13):1662-1668, 2009
- [ORA] Khatri & Draghici**. Ontological analysis of gene expression data: Current tools, limitations, and open problems**. *Bioinformatics*, 21(18):3587-3595, 2005
- [FCS] Pavlidis et al. **Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex.** *Neurochem Res*., 29(6):1213-1222, 2004
- [FCS] Goeman et al. **A global test for groups of genes: Testing association with a clinical outcome**. *Bioinformatics*, 20(1):93-99, 2004
- [GSEA] Subramanian et al. **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *PNAS*, 102(43):15545-15550, 2005
- [SNet] Soh et al. **Finding consistent disease subnetworks across microarray datasets**. *BMC Genomics*, 12(Suppl. 13):S15, 2011

A Novel Principle for Childhood ALL Relapse Prediction

This part of the lecture is show you an example of more advanced forms of gene expression analysis

Childhood Acute Lymphoblastic Leukemia

- **The most common cancer in children**
  - 3,000 new cases in US
  - 2,000 new cases in ASEAN countries

- **80% achieve long-term relapse-free survival, but**
  - 20% relapse and eventually die
  - Large fraction of them suffer severe side effects

⇒ **Predict relapse early and treat more aggressively**

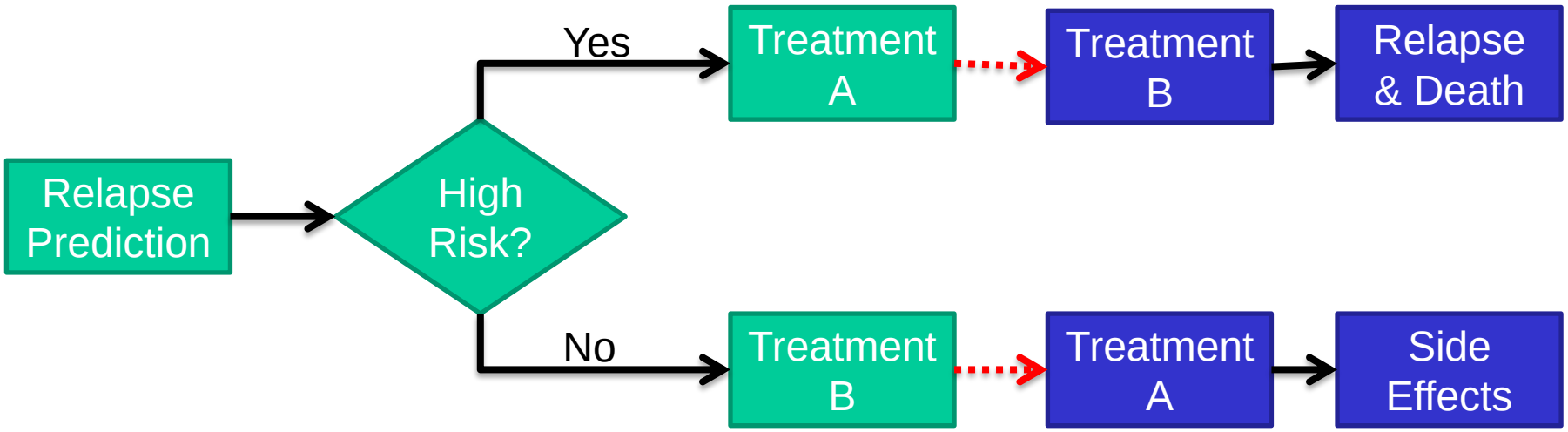
Outline

- **Background**
- **Hypotheses**
- **Framework**
- **Methodologies**
  - Data Preparation
  - Model Construction
  - Relapse Prediction
- **Validation**
- **Conclusion**

Main reference for this work

- **Difeng Dong, "Relapse Prediction in Childhood Acute Lymphoblastic Leukemia by Time-Series Gene Expression Profiling", PhD thesis, November 2011, National University of Singapore**

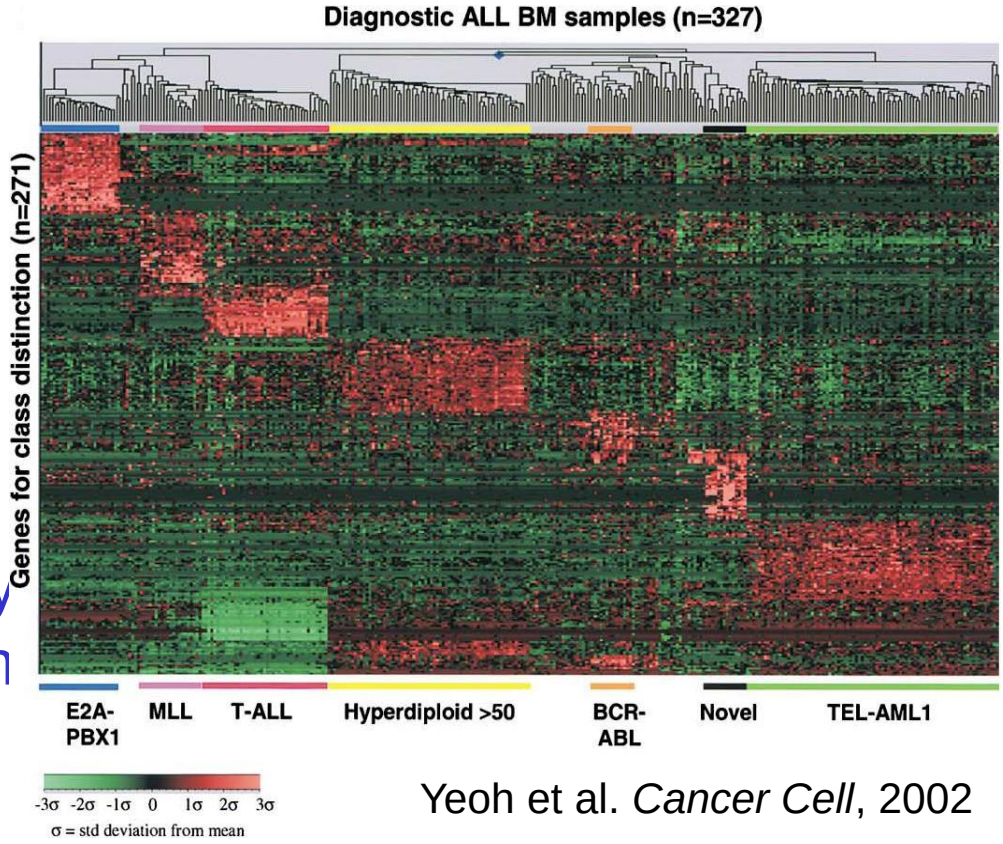
Contemporary ALL Treatment Framework



- **Treatment A: Intensive Treatment**
- **Treatment B: Moderate Treatment**

Previous Work

- **Correlate GEP to childhood ALL subtypes**
  - Identified subtype-based genetic signatures
  - Diagnostic accuracy is >95%, better than routine diagnostic methods

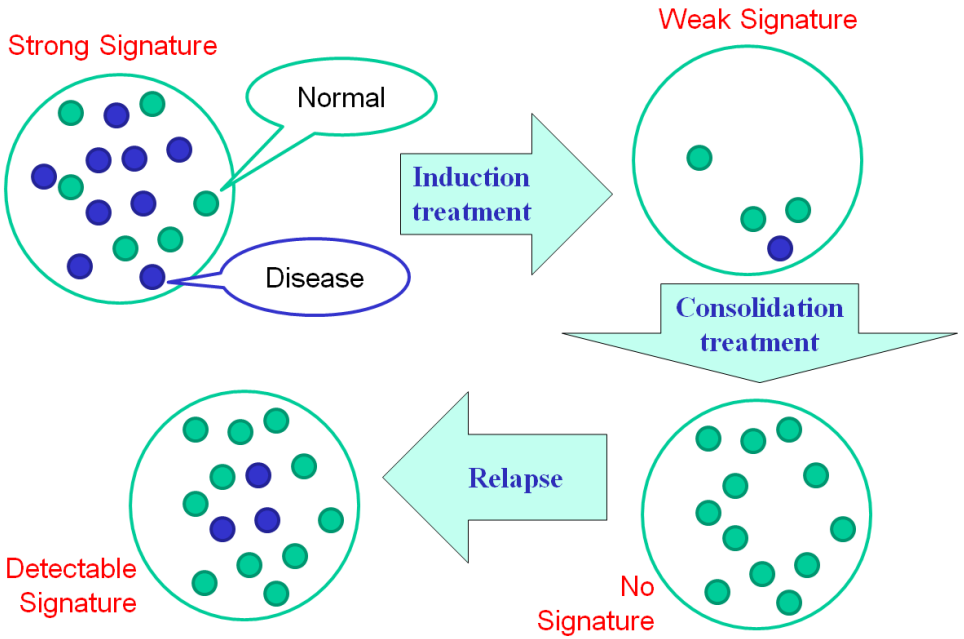


Yeoh et al. *Cancer Cell*, 2002

What does intensity of genetic signature means?

Hypotheses

- **Treatment gradually removes leukemic cells in patient**
- **Diagnostic GEP captures leukemic subtype signature**



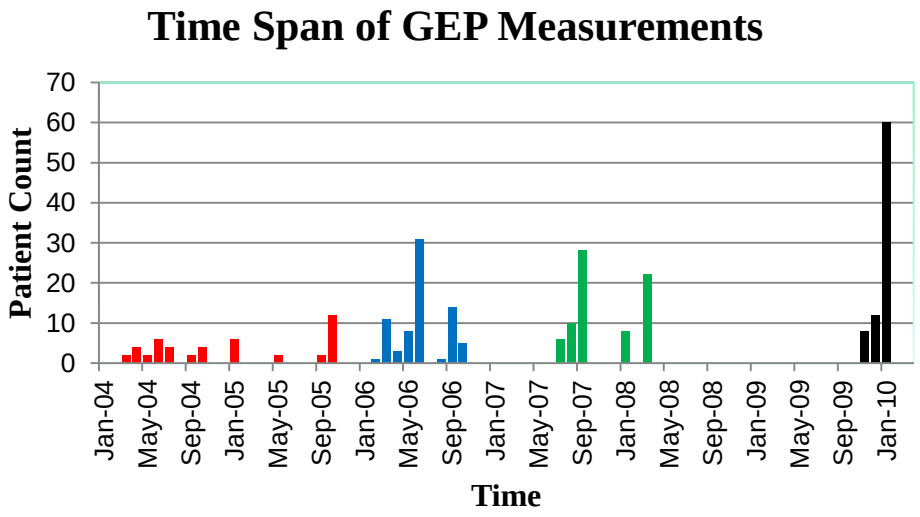
- **Hypothesis 1: Time-series GEP captures reduction of leukemic cells during treatment**
- **Hypothesis 2: Poor genetic response suggests high risk of relapse**

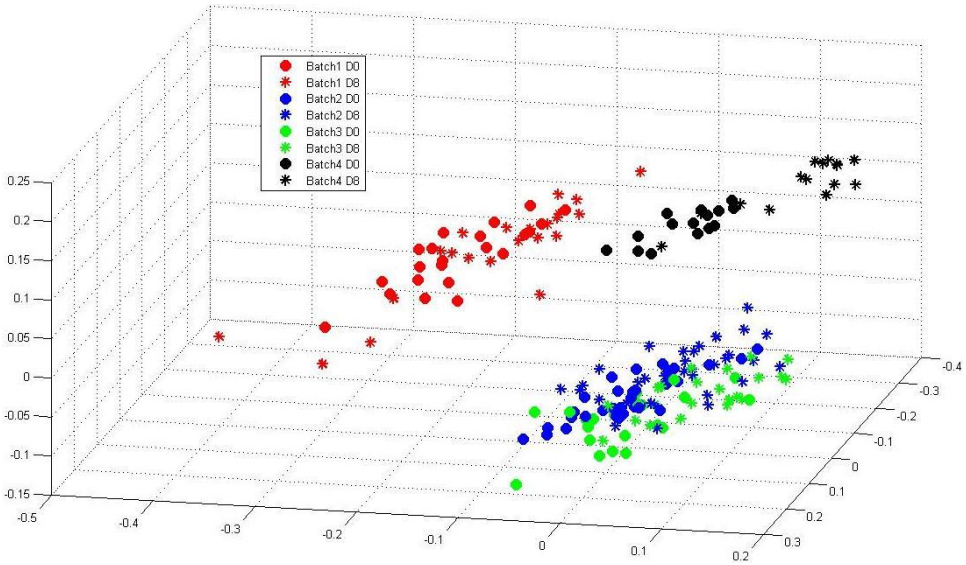
Framework

- **Time-series GEP data preparation (normalization)**
- **H1: Time-series GEP captures reduction of leukemic cells during treatment**
  - Unsupervised hierarchical clustering
  - Signature dissolution analysis
  - Genetic status shifting (GSS) model
- **H2: Poor genetic response ➔ high risk of relapse**
  - Prediction based on GSS distance
- **Validation in independent datasets**

GEP Data Preparation

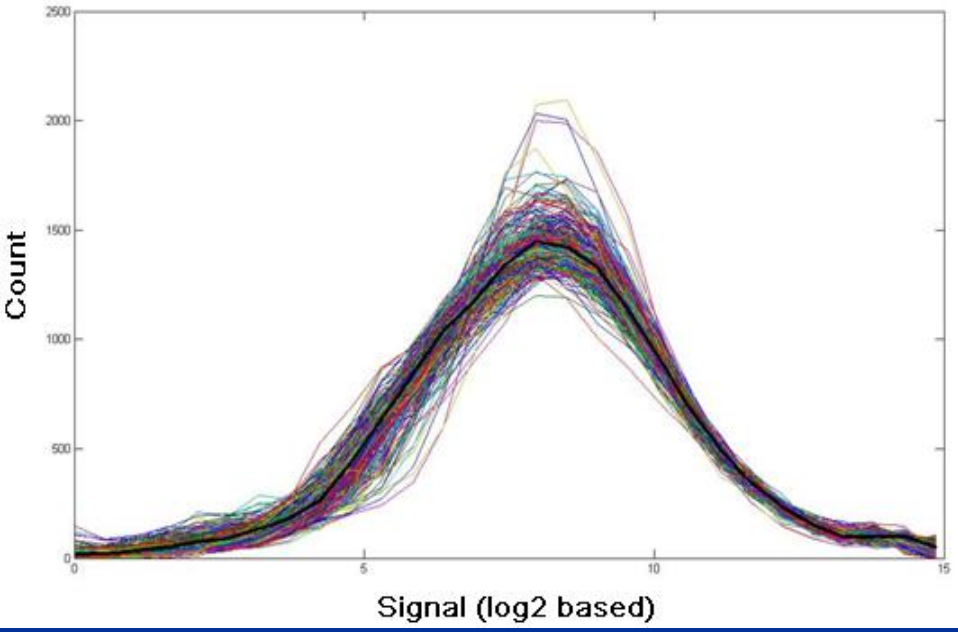
- **96 patients, 10 relapses vs 86 remissions**
- **GEP collected on 4 time points, D0, D8, D15, D33, a matrix of >30,000 genes * >300 samples**
- **Data generated by MAS5.0**

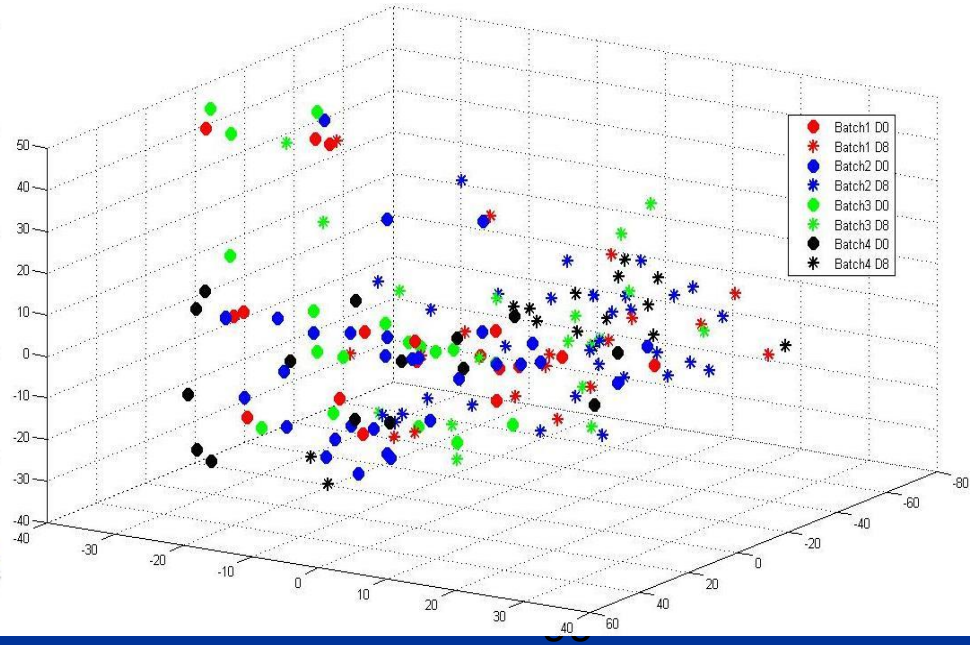




GEP Data Normalization

- **Scaling factor >20 ➔ Remove**
- **Noise mainly in low-expression genes**
- **Genes with > 70% absent calls ➔ Remove**
- **4,736 genes remain**
- **Perform quantile normalization**





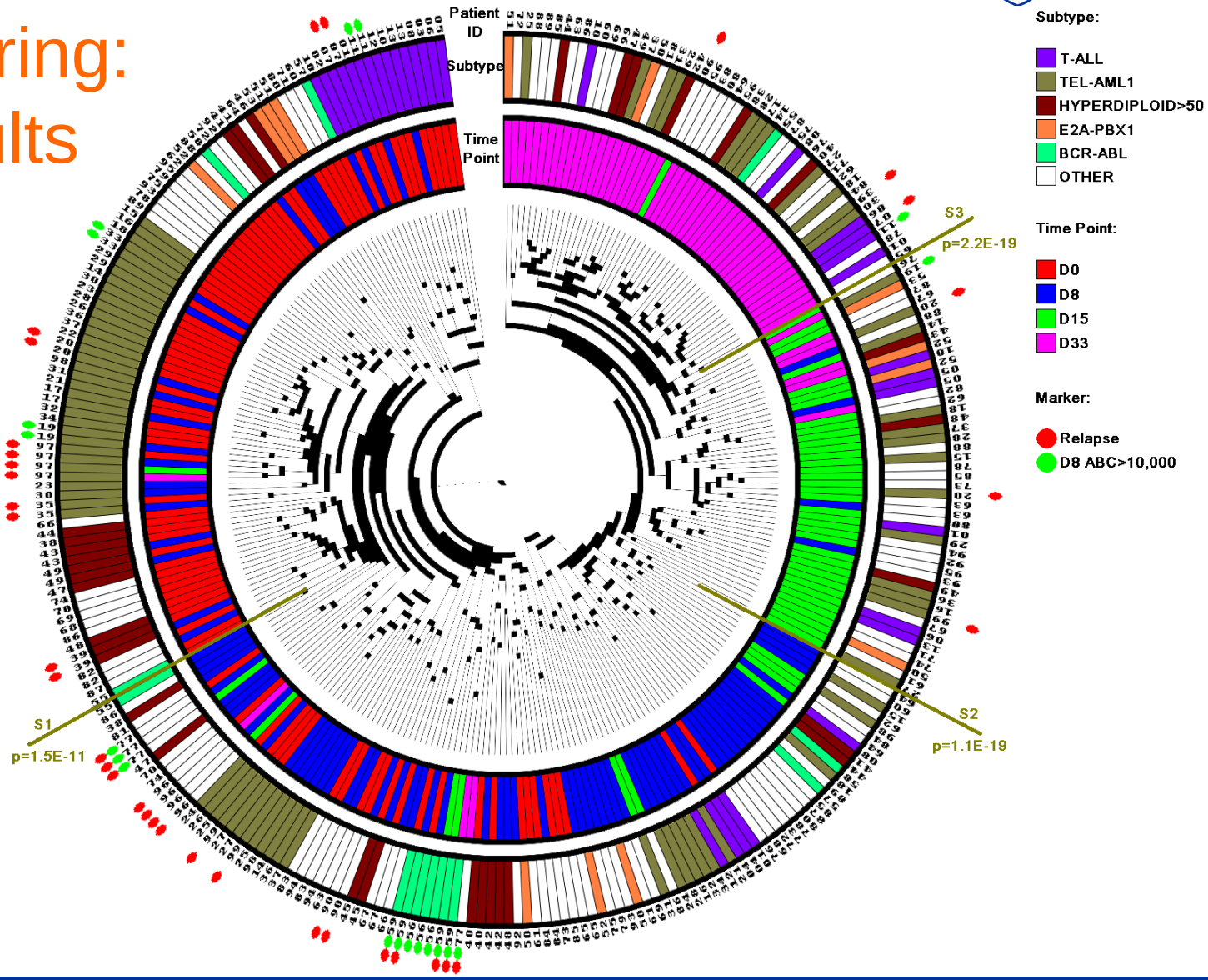
Framework

- **Time-series GEP data preparation (normalization)**
- **H1: Time-series GEP captures reduction of leukemic cells during treatment**
  - Unsupervised hierarchical clustering
  - Signature dissolution analysis
  - Genetic status shifting (GSS) model
- **H2: Poor genetic response ➔ high risk of relapse**
  - Prediction based on GSS distance
- **Validation in independent datasets**

Unsupervised Clustering

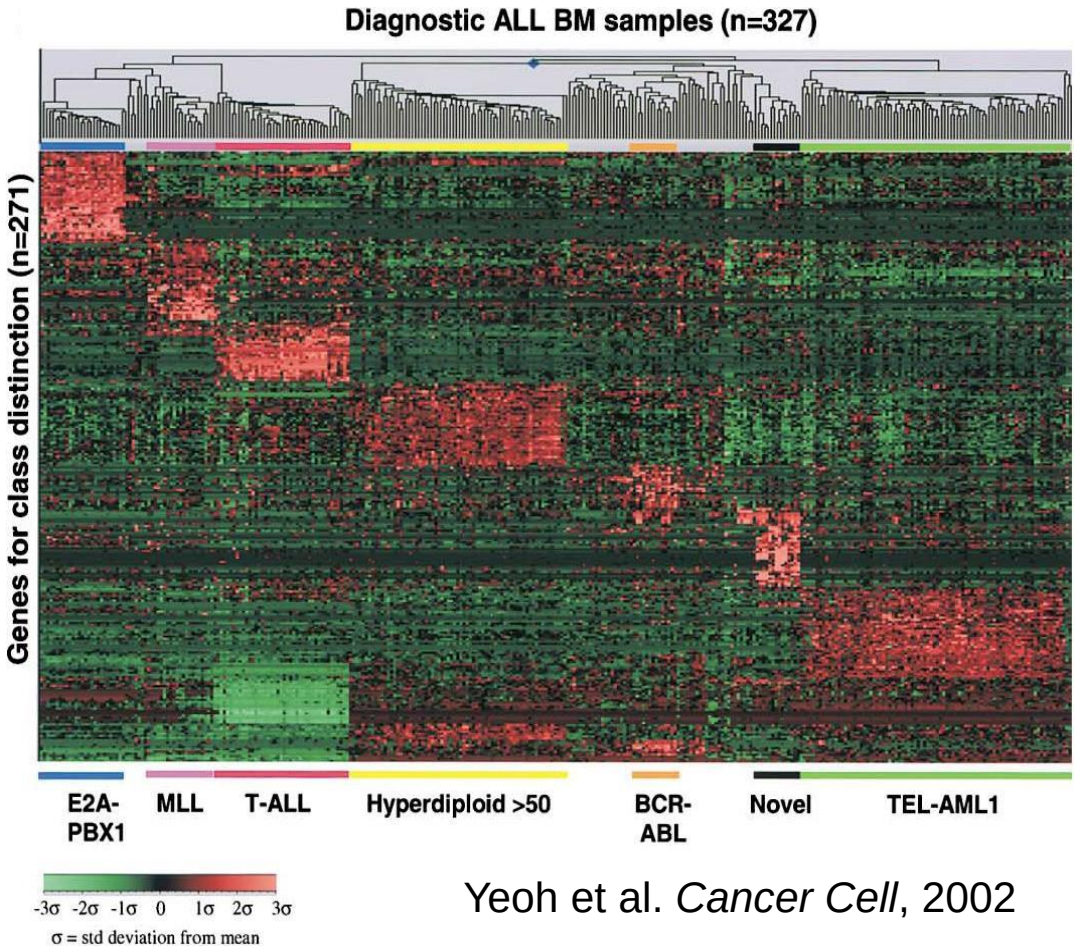
- **Top 10% of genes with largest variance across whole dataset**
  - 1,474 genes
  - Noise mainly in low-expression genes
- **Unsupervised hierarchical clustering on patients**
  - Pearson's correlation
  - Completed linkage

Unsupervised Clustering: Results



Signature Dissolution Analysis

- **How do intensity of genetic signatures change during treatment?**



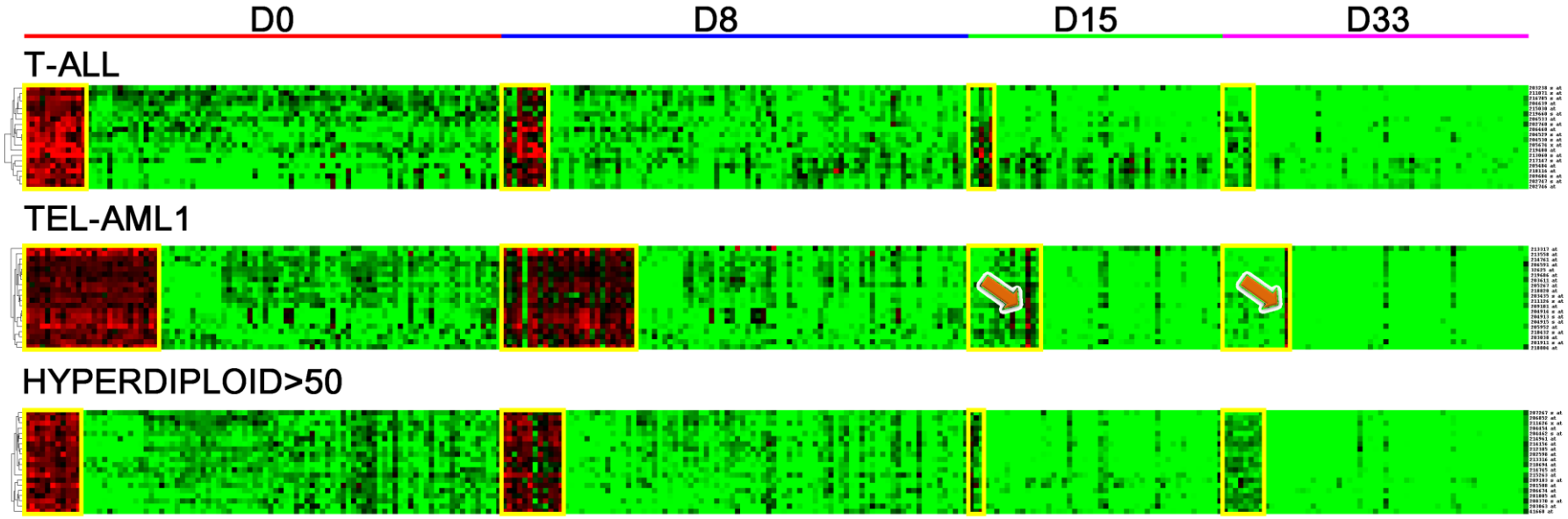
Yeoh et al. *Cancer Cell*, 2002

Signature Dissolution Analysis

- **Consider the 3 largest subtypes**
  - TEL-AML1, n = 26
  - T-ALL, n = 12
  - Hyperdiploid>50, n = 12

- **Select genetic signature genes for them**
  - Organize diagnostic samples into Subtype A vs ~A
  - Only consider genes highly expressed in A
  - Pick 20 most diff expressed genes by t-test
    - **The selected signature genes was validated by testing their prediction accuracy on public data**
    - **Accuracy achieved: 95%**

Signature Dissolution Analysis



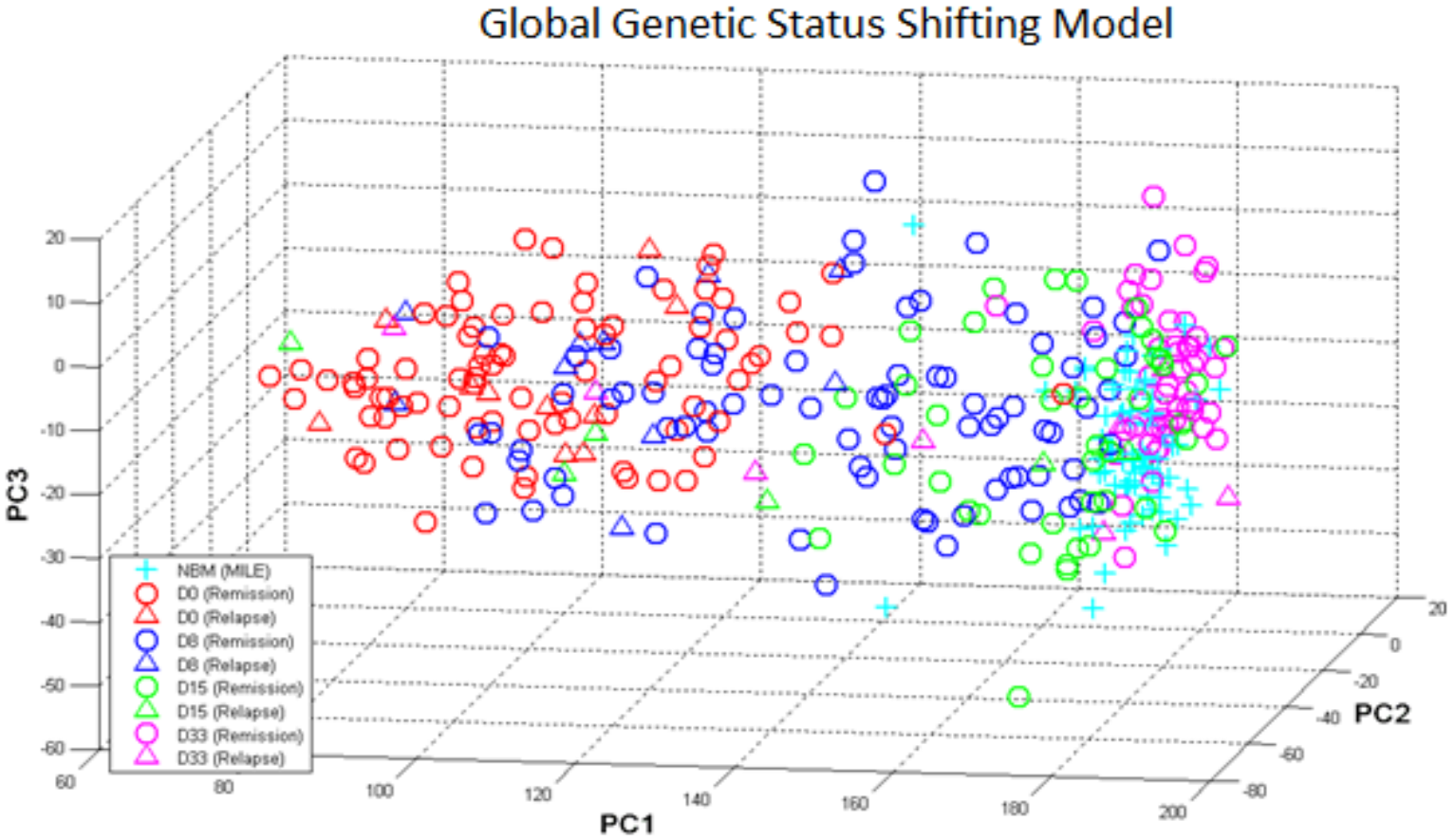
- **Red is highly expressed**
- **Green is lowly expressed**

Global Genetic Status Shifting (GSS) Model

- **Select drug responsive genes**
  - Diff expressed genes betw D0 & D8 by t-test ($q<0.0001$)
  - >2 fold change betw D0 and D8

  $\Rightarrow$461 up- and 99 down-regulated genes

- **Apply principal component analysis**
  - Genes are considered as features
  - PCA
  - Each point is a sample

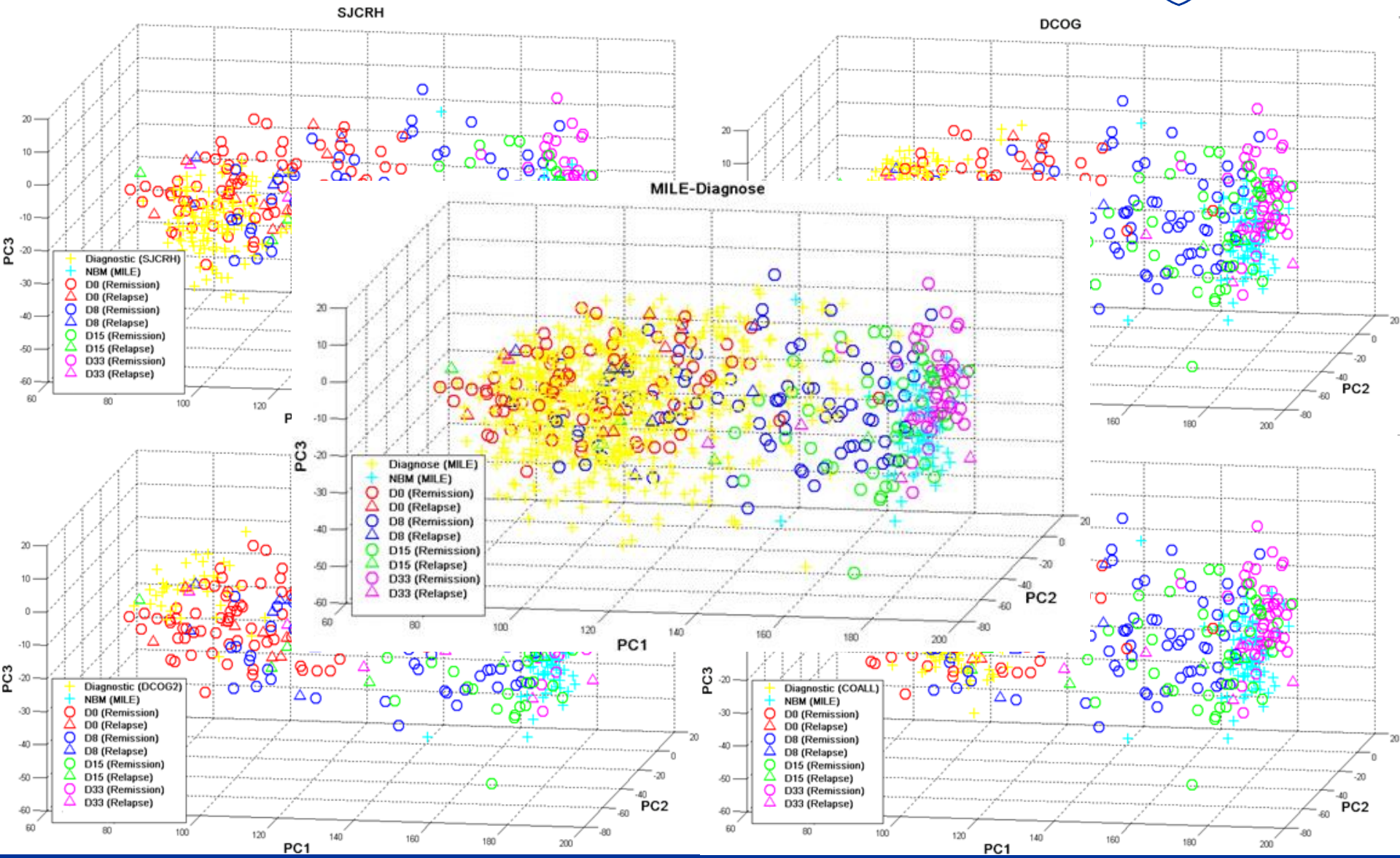


(a)

PC	1	2	3	4	5	6	7	8	Total
Variance	49.08%	7.56%	5.19%	3.64%	2.10%	1.81%	1.61%	1.35%	72.34%

(b)

Verifying the Global GSS Model



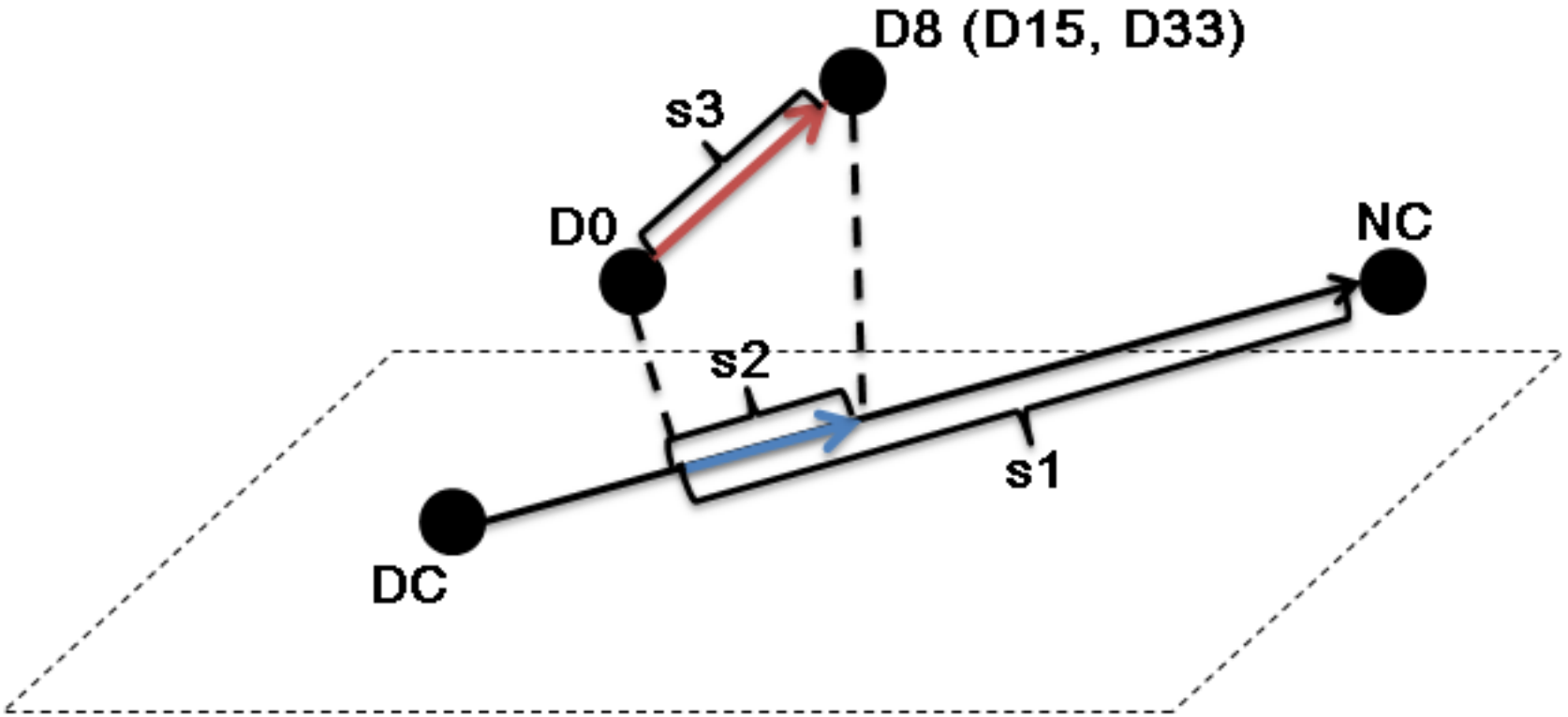
GO and Pathway Ingenuity Analysis on Drug-Responsive Genes in the Global GSS Model

- **UP: Reconstruction of immune system and restoration of normal hematogenesis**
- **DOWN: Cell development and DNA synthesis**
- **DOWN: Negative regulation of apoptosis**

Framework

- **Time-series GEP data preparation (normalization)**
- **H1: Time-series GEP captures reduction of leukemic cells during treatment**
  - Unsupervised hierarchical clustering
  - Signature dissolution analysis
  - Genetic status shifting (GSS) model
- **H2: Poor genetic response ➔ high risk of relapse**
  - Prediction based on GSS distance
- **Validation in independent datasets**

Genetic Status Shifting Distance



DC: Disease Centroid

NC: NBM Centroid

ASD = s3 (Absolute shiftig distance)

ESD = s2 (Effective shiftig distance)

ESR = s2/s1 (Effective shiftig ratio)

Relapse Prediction by ESD

RANK	SAMPLE	ESD-D8	RANK	SAMPLE	ESD-D8	RANK	SAMPLE	ESD-D8
1	56_KL464	-2.21	30	29_KL439	26.20	59	05_KL354	52.56
2	77_KL401	-1.51	31	49_R432	26.77	60	80_KL423	53.33
3	33_R247	-0.46	32	83_KL457	29.13	61	75_KL385	54.31
4	19_KL205	-0.21	33	04_KL322	30.64	62	69_KL313	54.40
5	45_R194	2.13	34	74_KL383	31.52	63	64_KKH29	55.12
6	59_R281	2.50	35	60_KKH30	32.88	64	93_R337	55.35
7	97_R208	2.94	36	38_KL218	33.00	65	26_KL369	56.83
8	67_KL287	4.63	37	10_R257	33.39	66	14_KKH19	57.79
9	96_R202	5.14	38	40_KL430	33.55	67	28_KL375	58.21
10	11_R280	5.90	39	86_KL509	34.48	68	62_KKH22	60.55
11	27_KL374	6.31	40	51_KL461	34.83	69	21_KL300	60.80
12	70_KL320	6.64	41	79_KL421	35.43	70	01_KKH18	62.07
13	41_KL441	7.86	42	94_R354	35.66	71	32_R233	67.35
14	20_KL274	8.91	43	24_KL328	35.78	72	34_R256	67.95
15	82_KL454	9.53	44	47_R334	37.35	73	68_KL304	70.06
16	13_R410	11.45	45	30_KL444	38.26	74	18_KKH28	70.41
17	39_KL395	11.49	46	57_KL535	39.40	75	36_R343	71.95
18	17_KKH27	12.64	47	73_KL381	39.51	76	58_KL543	71.96
19	35_R313	13.76	48	66_KL247	39.69	77	08_KL456	72.39
20	55_KL419	18.60	49	16_KKH21	40.05	78	12_R297	74.21
21	99_KL416	19.83	50	84_KL458	40.12	79	88_KL544	75.25
22	23_KL321	20.37	51	37_R355	41.68	80	72_KL377	76.58
23	50_KL360	20.80	52	25_KL357	42.14	81	44_KL541	79.45
24	90_R253	21.40	53	15_KKH20	43.35	82	87_KL522	79.91
25	61_KKH13	21.67	54	52_R252	45.31	83	89_R245	85.37
26	43_KL536	23.00	55	85_KL485	45.82	84	63_KKH25	85.67
27	65_KL224	23.30	56	92_R332	49.15	85	78_KL412	92.38
28	07_KL417	24.00	57	98_KL387	49.17	86	95_R431	104.57
29	42_KL507	24.00	58	48_R339	49.26			

Comparison with Other Clinical & Computational Protocols

Method	Prognostic Feature	Sensitivity	Specificity	Accuracy
Holleman-DT	D0 GEP	60.00%	69.51%	64.76%
Holleman-NB	D0 GEP	60.00%	69.51%	64.76%
Holleman-SVM	D0 GEP	60.00%	69.51%	64.76%
Bhojwani	D0 GEP	20.00%	79.27%	49.63%
NCI	D0 GEP	80.00%	58.14%	69.07%
Cytogenetics	Diagnostic Cytogenetics	30.00%	**94.19%**	62.09%
MRD-D33	D33 MRD	77.78%	54.12%	65.95%
D8 Response	D8 Blast Count	30.00%	85.53%	57.76%
ASD-D8	D0 and D8 GEP	90.00%	73.68%	81.84%
ESD-D8	D0 and D8 GEP	**100.00%**	75.00%	**87.50%**
ESR-D8	D0 and D8 GEP	90.00%	73.68%	81.84%

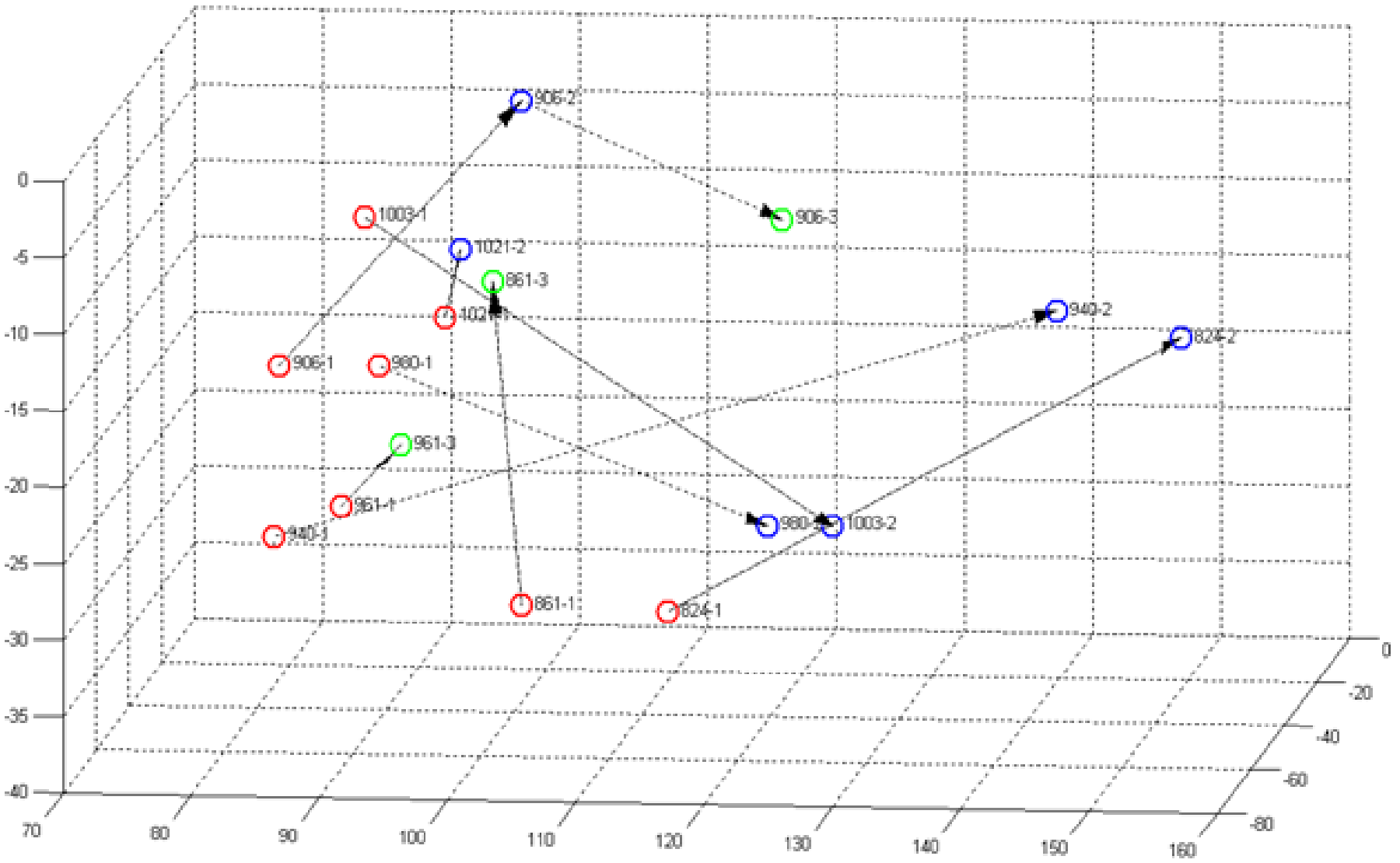
Framework

- **Time-series GEP data preparation (normalization)**
- **H1: Time-series GEP captures reduction of leukemic cells during treatment**
  - Unsupervised hierarchical clustering
  - Signature dissolution analysis
  - Genetic status shifting (GSS) model
- **H2: Poor genetic response ➔ high risk of relapse**
  - Prediction based on GSS distance
- **Validation in independent datasets**

Validation on new Childhood ALL Dataset

- **8 childhood ALL patients from Europe**
- **GEP on D0, D8, and D15**
- **Standard data preprocessing**
- **Use the same drug-responsive genes as the global GSS model**
- **Apply PCA and use ESD to make prediction**

Genetic Status Shifting Model -- Validation



Recall… so threshold is ESD = 24

RANK	SAMPLE	ESD-D8	RANK	SAMPLE	ESD-D8	RANK	SAMPLE	ESD-D8
1	56_KL464	-2.21	30	29_KL439	26.20	59	05_KL354	52.56
2	77_KL401	-1.51	31	49_R432	26.77	60	80_KL423	53.33
3	33_R247	-0.46	32	83_KL457	29.13	61	75_KL385	54.31
4	19_KL205	-0.21	33	04_KL322	30.64	62	69_KL313	54.40
5	45_R194	2.13	34	74_KL383	31.52	63	64_KKH29	55.12
6	59_R281	2.50	35	60_KKH30	32.88	64	93_R337	55.35
7	97_R208	2.94	36	38_KL218	33.00	65	26_KL369	56.83
8	67_KL287	4.63	37	10_R257	33.39	66	14_KKH19	57.79
9	96_R202	5.14	38	40_KL430	33.55	67	28_KL375	58.21
10	11_R280	5.90	39	86_KL509	34.48	68	62_KKH22	60.55
11	27_KL374	6.31	40	51_KL461	34.83	69	21_KL300	60.80
12	70_KL320	6.64	41	79_KL421	35.43	70	01_KKH18	62.07
13	41_KL441	7.86	42	94_R354	35.66	71	32_R233	67.35
14	20_KL274	8.91	43	24_KL328	35.78	72	34_R256	67.95
15	82_KL454	9.53	44	47_R334	37.35	73	68_KL304	70.06
16	13_R410	11.45	45	30_KL444	38.26	74	18_KKH28	70.41
17	39_KL395	11.49	46	57_KL535	39.40	75	36_R343	71.95
18	17_KKH27	12.64	47	73_KL381	39.51	76	58_KL543	71.96
19	35_R313	13.76	48	66_KL247	39.69	77	08_KL456	72.39
20	55_KL419	18.60	49	16_KKH21	40.05	78	12_R297	74.21
21	99_KL416	19.83	50	84_KL458	40.12	79	88_KL544	75.25
22	23_KL321	20.37	51	37_R355	41.68	80	72_KL377	76.58
23	50_KL360	20.80	52	25_KL357	42.14	81	44_KL541	79.45
24	90_R253	21.40	53	15_KKH20	43.35	82	87_KL522	79.91
25	61_KKH13	21.67	54	52_R252	45.31	83	89_R245	85.37
26	43_KL536	23.00	55	85_KL485	45.82	84	63_KKH25	85.67
27	65_KL224	23.30	56	92_R332	49.15	85	78_KL412	92.38
28	07_KL417	24.00	57	98_KL387	49.17	86	95_R431	104.57
29	42_KL507	24.00	58	48_R339	49.26			

Result on new Childhood ALL Dataset

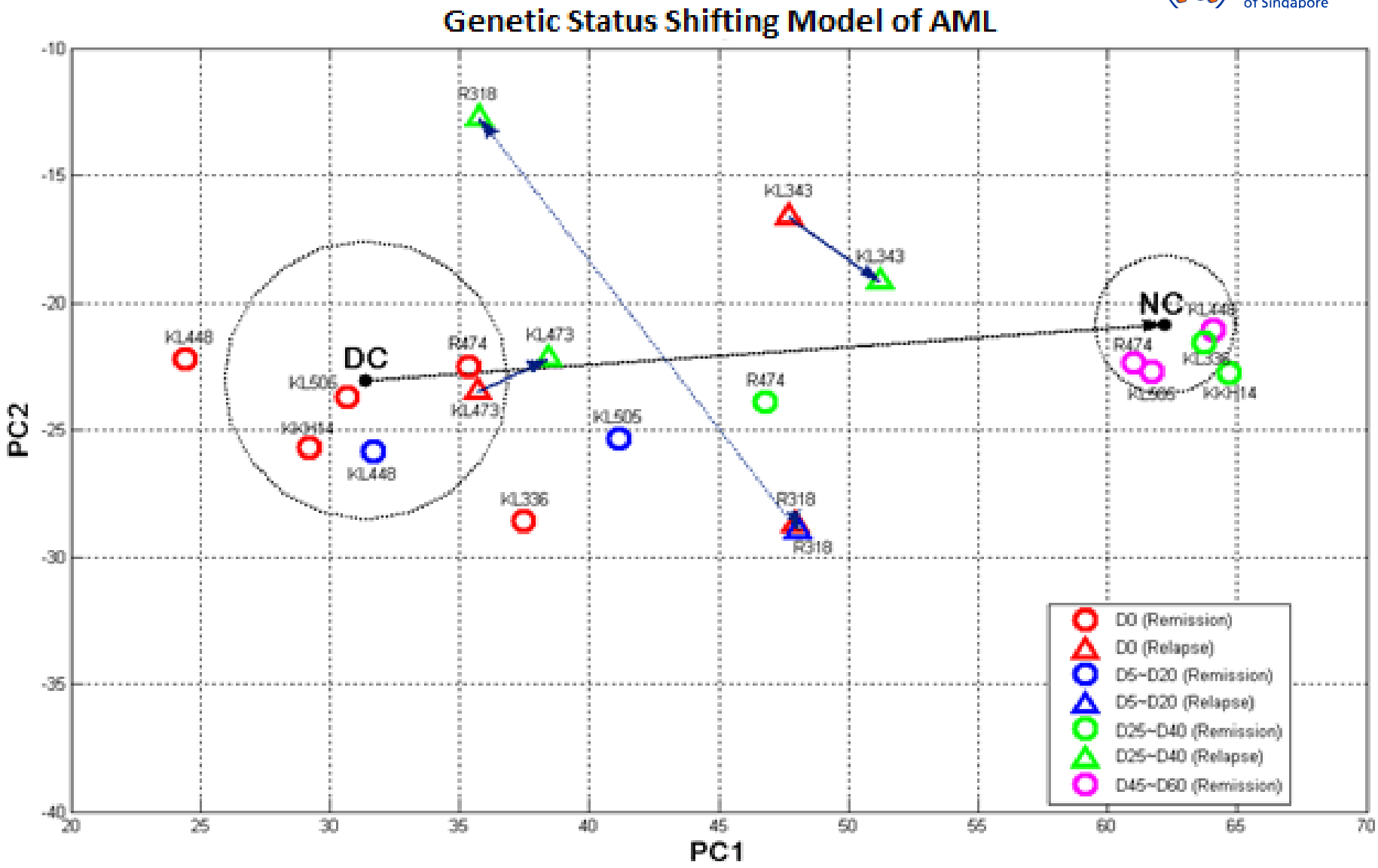
ID	Type	Subtype	ESD	Chance of adverse event	Risk
861-M	ALL	BCR-ABL	-4.67	Very high risk	100%
1021-M	ALL	BCR-ABL	1.86	Very high risk	90%
961-M	ALL	HYPERDIP	6.26	High risk	60%
906-M	ALL	OTHERS	17.16	High risk	30%
980-M	ALL	OTHERS	30.84	Intermediate to low risk	0%
824-M	ALL	TEL-AML1	39.16	Intermediate to low risk	0%
1003-M	ALL	HYPERDIP	41.48	Intermediate to low risk	0%
940-M	ALL	TEL-AML1	59.07	Intermediate to low risk	0%

Validation on AML Datasets: How general is the principle of the GSS Model

- **8 acute myeloid leukemia (AML) patients**
- **Similar treatment philosophy but much lower long-term event-free survival rate (40%)**

- **Unsynchronized GEPs betw D0 and D60**
- **Standard data preprocessing**

- **Select drug-responsive genes by MILE-AML vs MILE-NBM**
- **PCA and use ASD and ESD to make prediction**



NUS National University of Singapore

Results on AML Dataset

Rank	SAMPLE	ASD	Outcome		Rank	Sample	ESD	Outcome
1	**R318-D5**	**0.28**	**R**		**1**	**R318-D33**	**-11.03**	**R**
2	**KL473-D32**	**3.04**	**R**		**2**	**R318-D5**	**0.04**	**R**
3	**KL343-D36**	**4.33**	**R**		**3**	**KL473-D32**	**2.83**	**R**
4	KL448-D17	8.11			**4**	**KL343-D36**	**3.34**	**R**
5	KL505-D14	10.61			5	KL448-D17	6.99	
6	R474-D33	11.52			6	KL505-D14	10.33	
7	**R318-D33**	**20.10**	**R**		7	R474-D33	11.31	
8	R474-D60	25.67			8	R474-D60	25.62	
9	KL336-D31	27.14			9	KL336-D31	26.65	
10	KL505-D45	31.07			10	KL505-D45	31.04	
11	KKH14-D36	35.61			11	KKH14-D36	35.61	
12	KL448-D51	39.71			12	KL448-D51	39.67	

Conclusions

- **A novel GSS model to predict relapse of childhood ALL with significant improvement**
  - Accuracy 20% better than current methods
  - Validated in an independent cohort
- **Same principle applicable to predict AML relapse, a disease with only 40% of patients survived in 5 years**
- **This is the first time-series GEPs study in leukemias**

Acknowledgements


Ah Fu


Donny Soh


Dong Difeng


Kevin Lim

- **Much of this lecture is based on the works of my past/current students**
  - Koh Chuan Hock (Ah Fu)
  - Donny Soh
  - Dong Difeng
  - Kevin Lim