Is BIG Necessarily Better?

A guest lecture for CS5344, 21/3/2014

Wong Limsoon

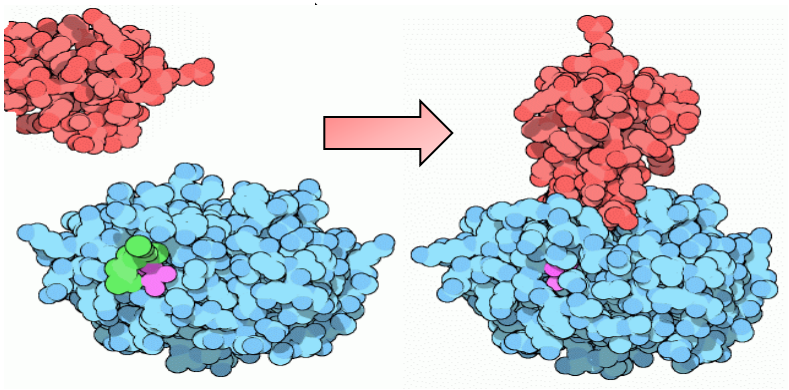
NUS National University of Singapore

More data may not be better…

A few stories

- **Discovering protein complexes from PPIN**
- **Identifying causal genes, Part 1**
- **Identifying causal genes, Part 2**
- **Finding interesting patterns**

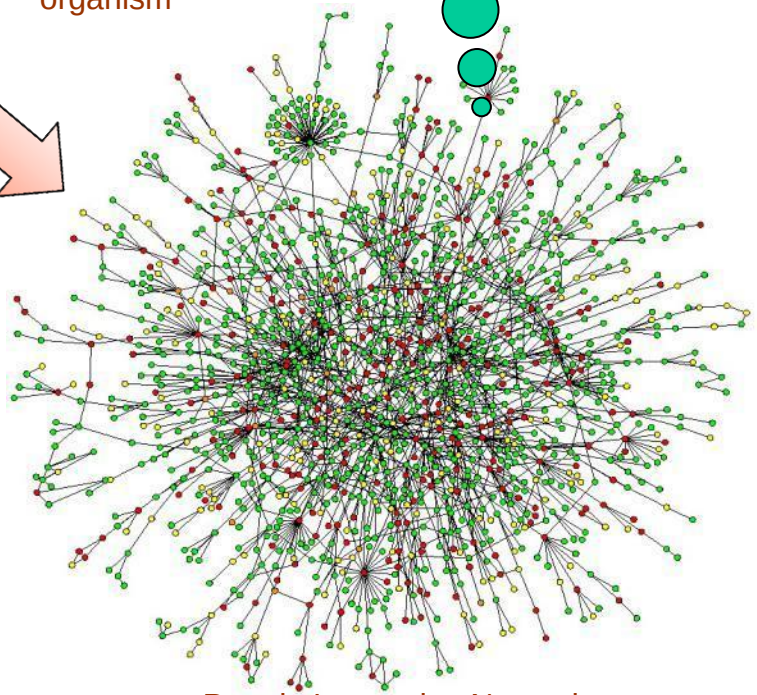
Protein-protein interaction networks



Individual proteins come together and interact

PPIN
Valuable source of knowledge

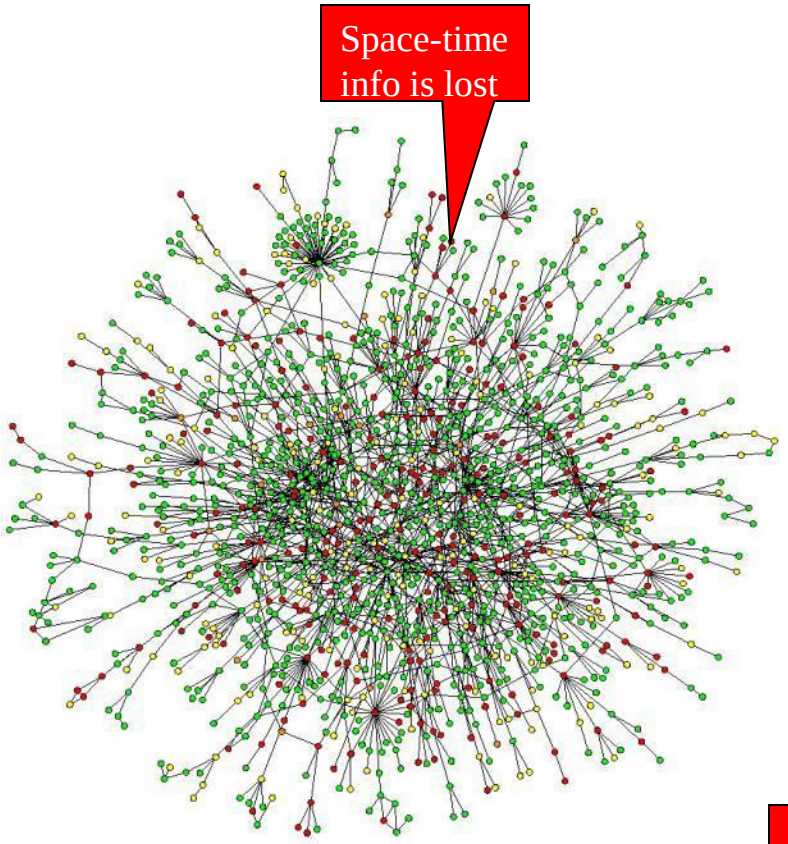
Collection of such interactions in an organism



Protein Interaction Network

- **Proteins come together & interact**
- **The collection of these interactions form a Protein Interaction Network or PPIN**

Detection & analysis of protein complexes in PPIN



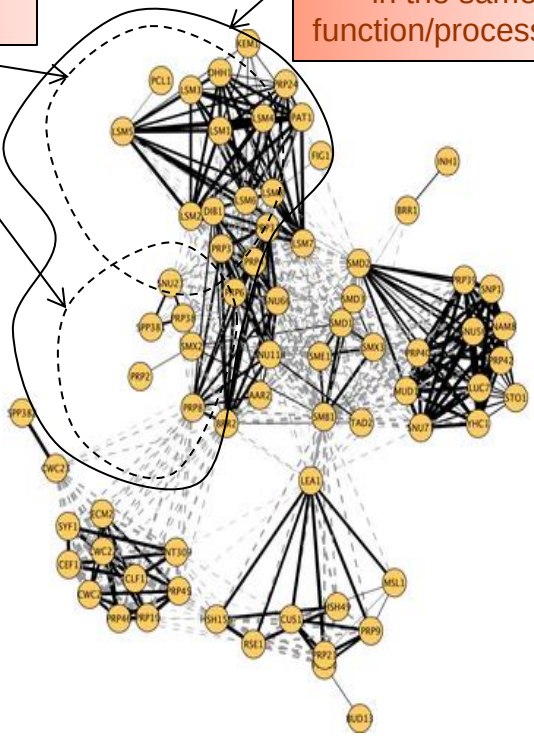
PPIN derived from several high-throughput expt

Identifying embedded complexes

Space-time info is "recovered"

Individual complexes (Some might share proteins)

Entire module might be involved in the same function/process



Embedded **complexes** identified from PPIN

NUS National University of Singapore

Difficulties

- **Protein complexes are discovered from PPIN by, e.g., clustering approaches**

- **But success has been limited**
  - Noise in PPI data
    - **Spuriously-detected interactions (false positives), and missing interactions (false negatives)**
  - Transient interactions
    - **Many proteins that actually interact are not from the same complex, they bind temporarily to perform a function**
  - Also, not all proteins in the same complex may actually interact with each other

Cytochrome BC1 Complex

- **Involved in electron-transport chain in mitochondrial inner membrane**

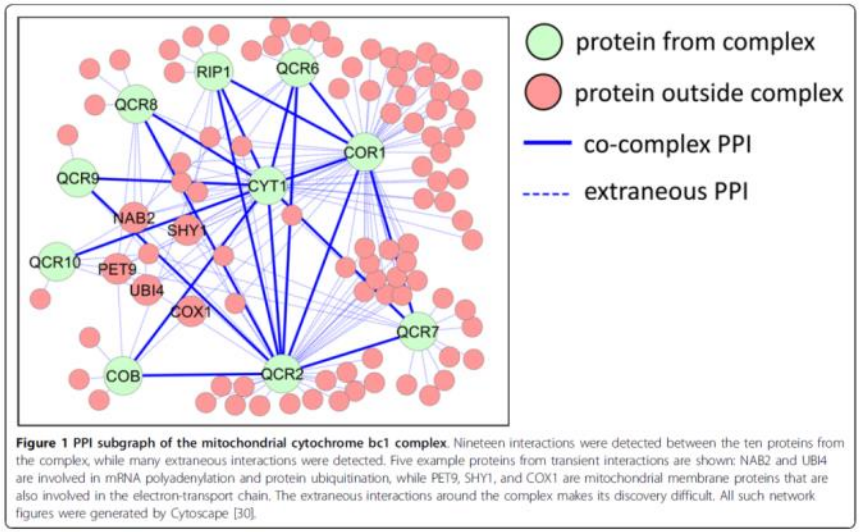


Figure 1 PPI subgraph of the mitochondrial cytochrome bc1 complex. Nineteen interactions were detected between the ten proteins from the complex, while many extraneous interactions were detected. Five example proteins from transient interactions are shown: NAB2 and UBI4 are involved in mRNA polyadenylation and protein ubiquitination, while PET9, SHY1, and COX1 are mitochondrial membrane proteins that are also involved in the electron-transport chain. The extraneous interactions around the complex makes its discovery difficult. All such network figures were generated by Cytoscape [30].

- **Discovery of BC1 from PPI data is difficult**
  - Sparseness of its PPI subnetwork
    - **Only 19 out of 45 possible interactions were detected between the complex's proteins**
  - Extraneous interactions with other proteins outside the complex
    - **E.g., UBI4 is involved in protein ubiquitination, and binds to many proteins to perform its function**

Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

NUS National University of Singapore

Perhaps "big data" can help?

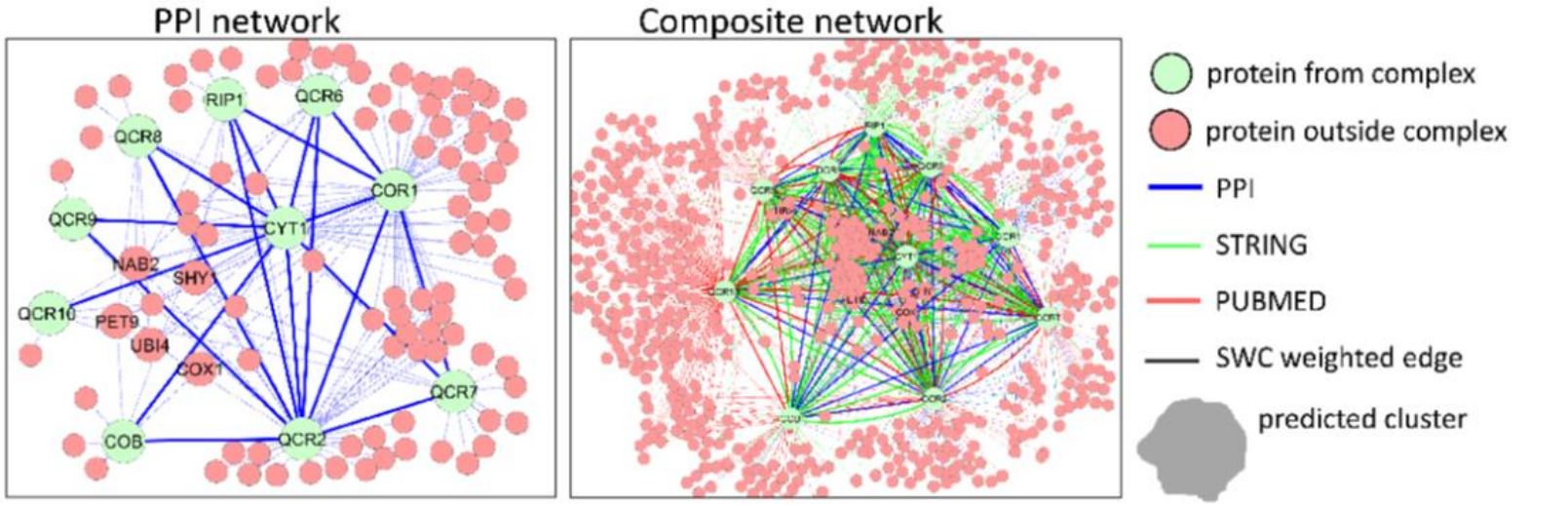
- **Composite network**
  - Vertices represent proteins, edges represent relationships between proteins. Put an edge betw proteins u, v, iff u and v are related according to any of the data sources

Data source	Database	Scoring method
PPI	BioGRID, IntACT, MINT	Iterative AdjustCD.
L2-PPI (indirect PPI)	BioGRID, IntACT, MINT	Iterative AdjustCD
Functional association	STRING	STRING
Literature co-occurrence	PubMed	Jaccard coefficient

	Yeast			Human		
	# Pairs	% co-complex	coverage	# Pairs	% co-complex	coverage
PPI	106328	**5.8%**	**55%**	48098	10%	14%
L2-PPI	181175	1.1%	18%	131705	5.5%	20%
STRING	175712	5.7%	89%	311435	3.1%	27%
PubMed	161213	4.9%	70%	91751	4.3%	11%
All	531800	**2.1%**	**98%**	522668	3.4%	49%

Yong, et al. **Supervised maximum-likelihood weighting of composite protein networks for complex prediction**. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

More is not always better, unless..

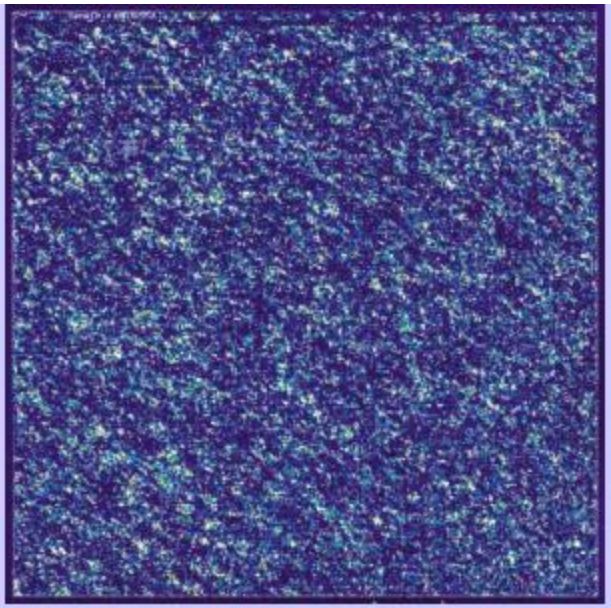


While proteins in BC1 become fully connected in the composite network, there is a blow-up in extraneous proteins. So clustering won't discover the complex, unless you know how to remove the extraneous proteins

A few stories

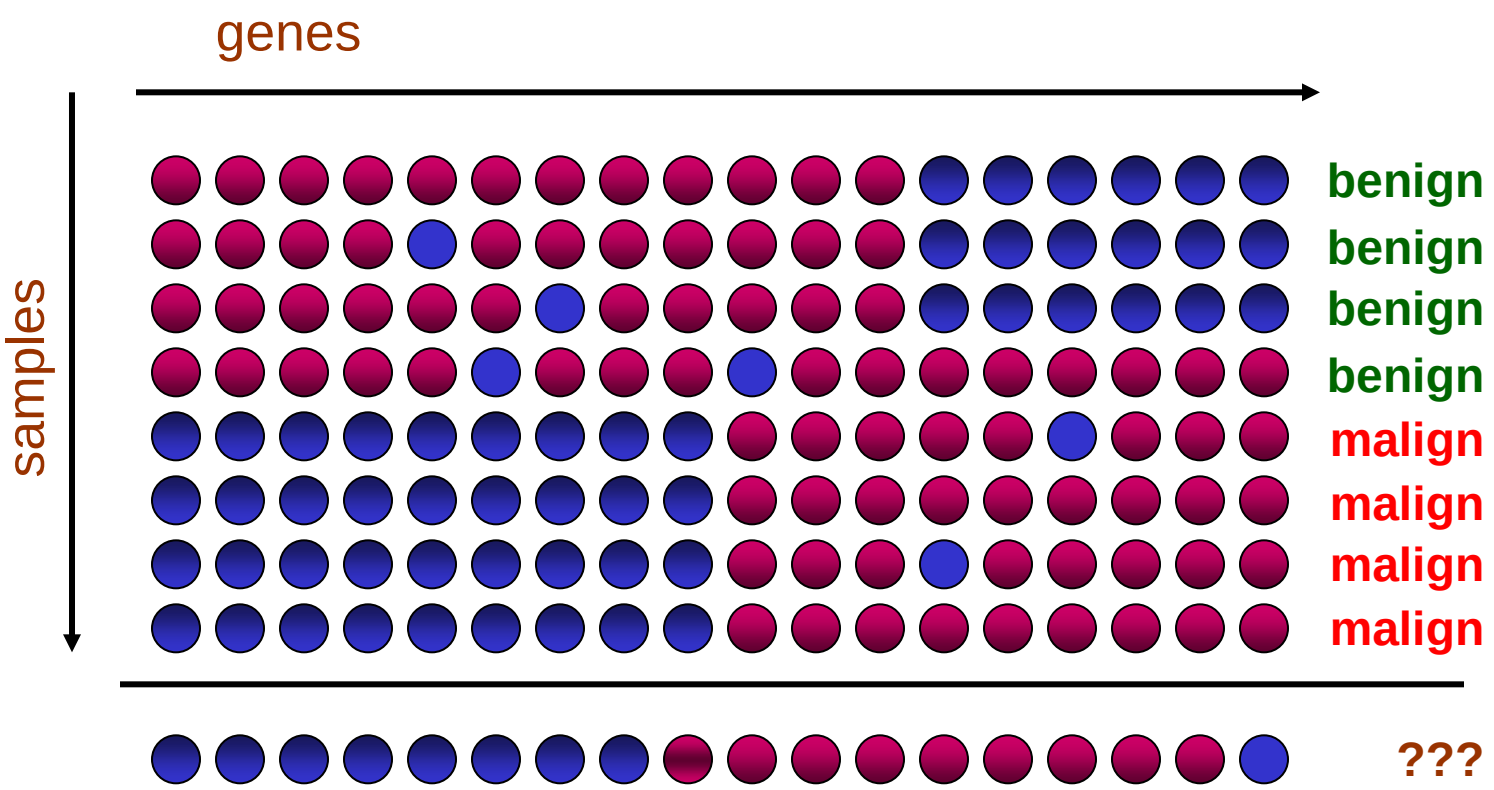
- Discovering protein complexes from PPIN
- **Identifying causal genes, Part 1**
- Identifying causal genes, Part 2
- Finding interesting patterns

Microarray

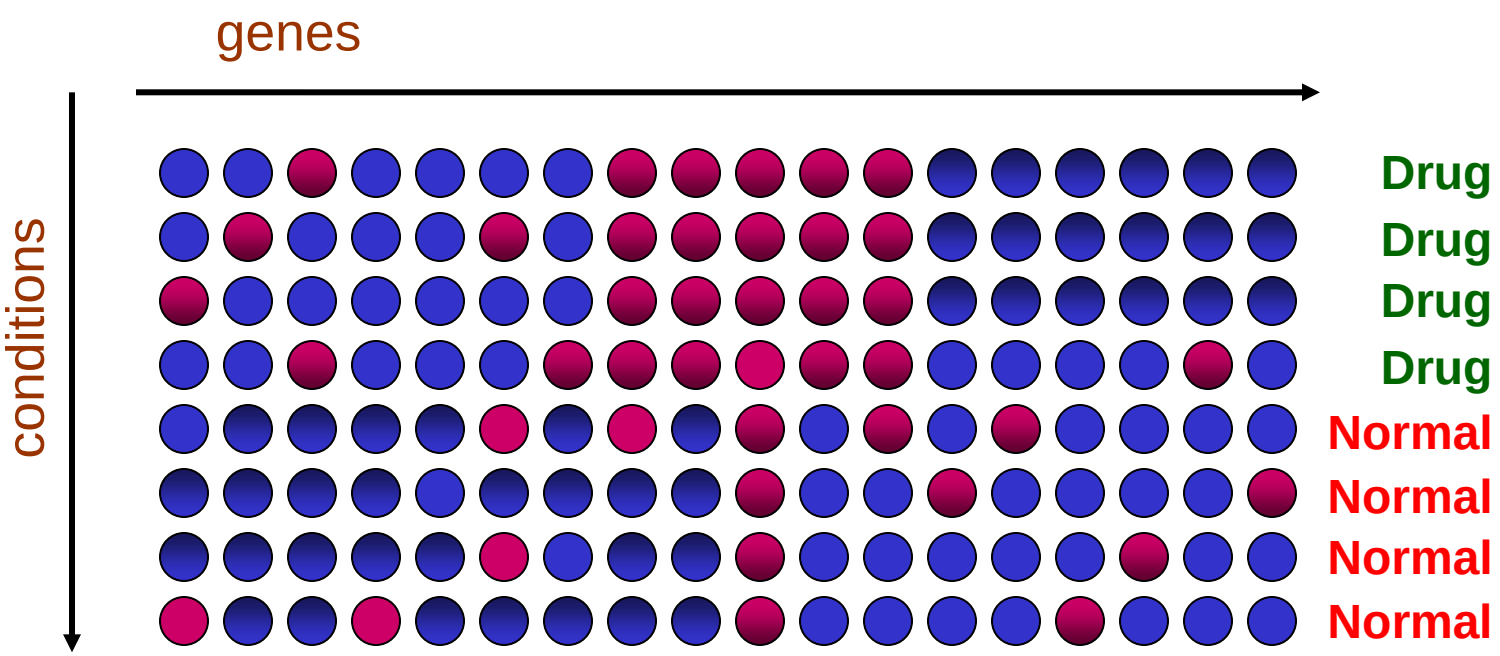


Source: Affymetrix

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**

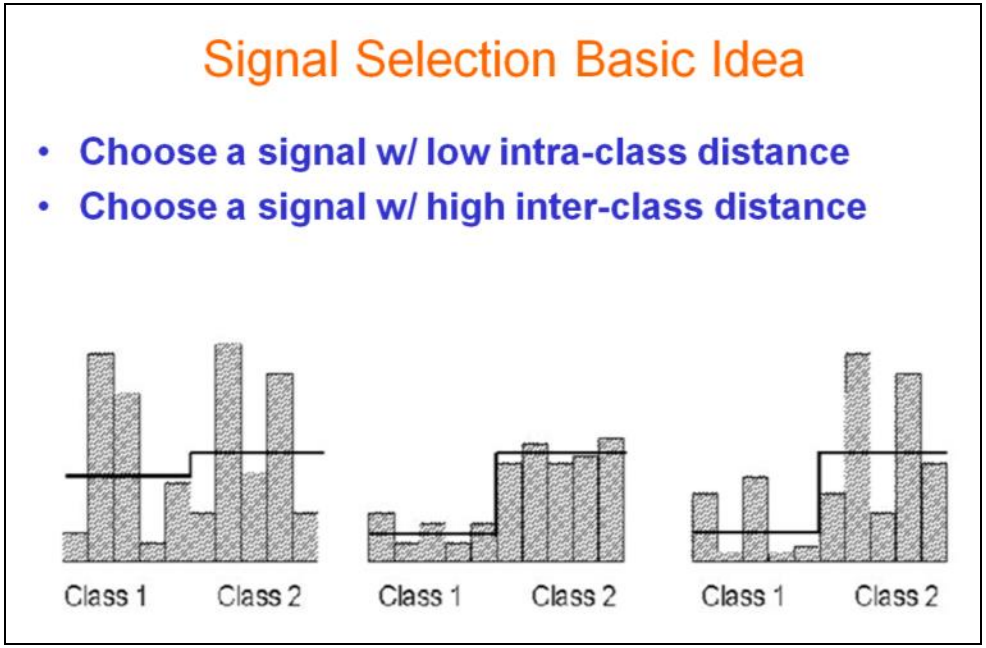


Image credit: Golub et al., *Science*, 286:531–537, 1999

Terminology: DE gene = differentially expressed gene

Hierarchical clustering

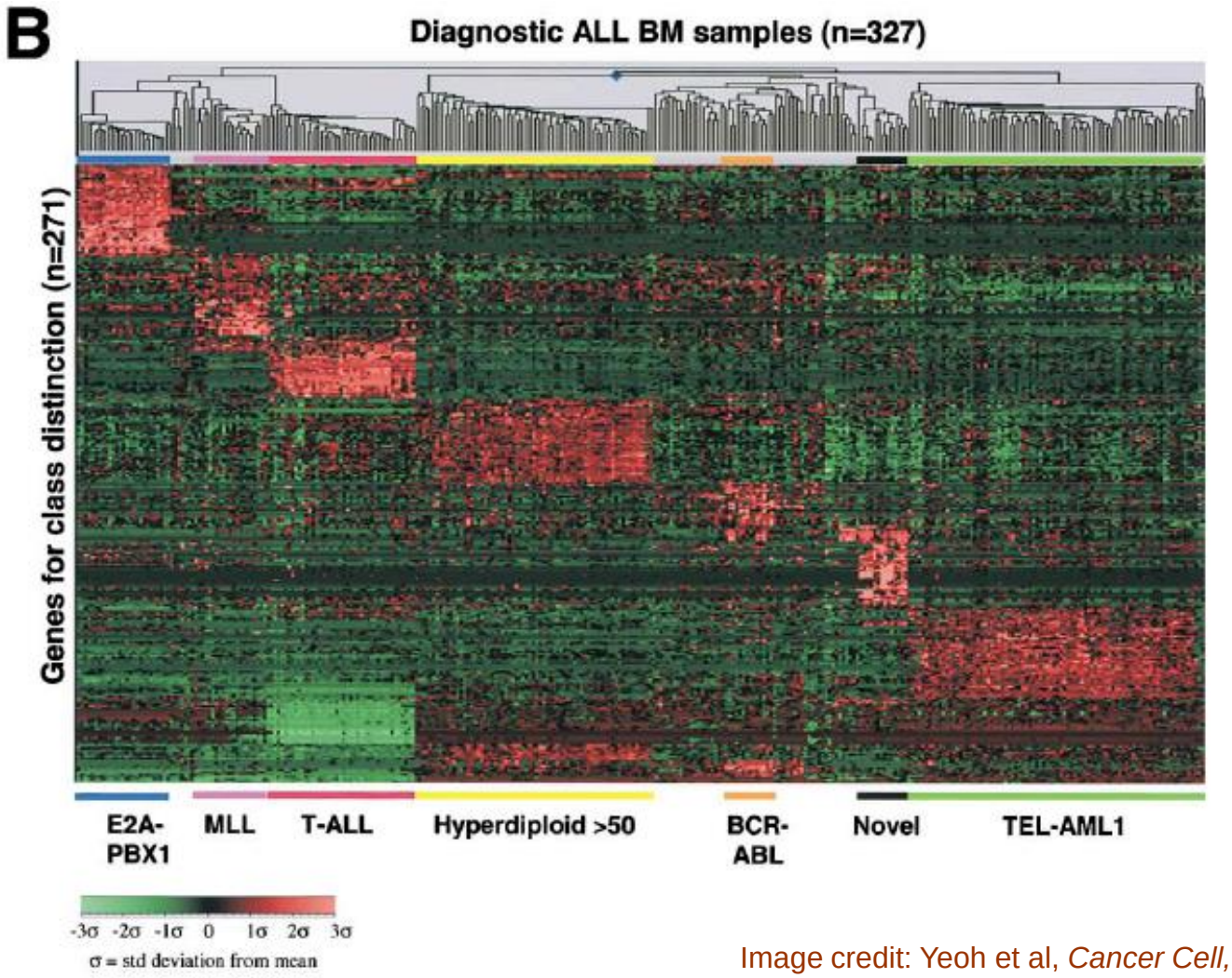


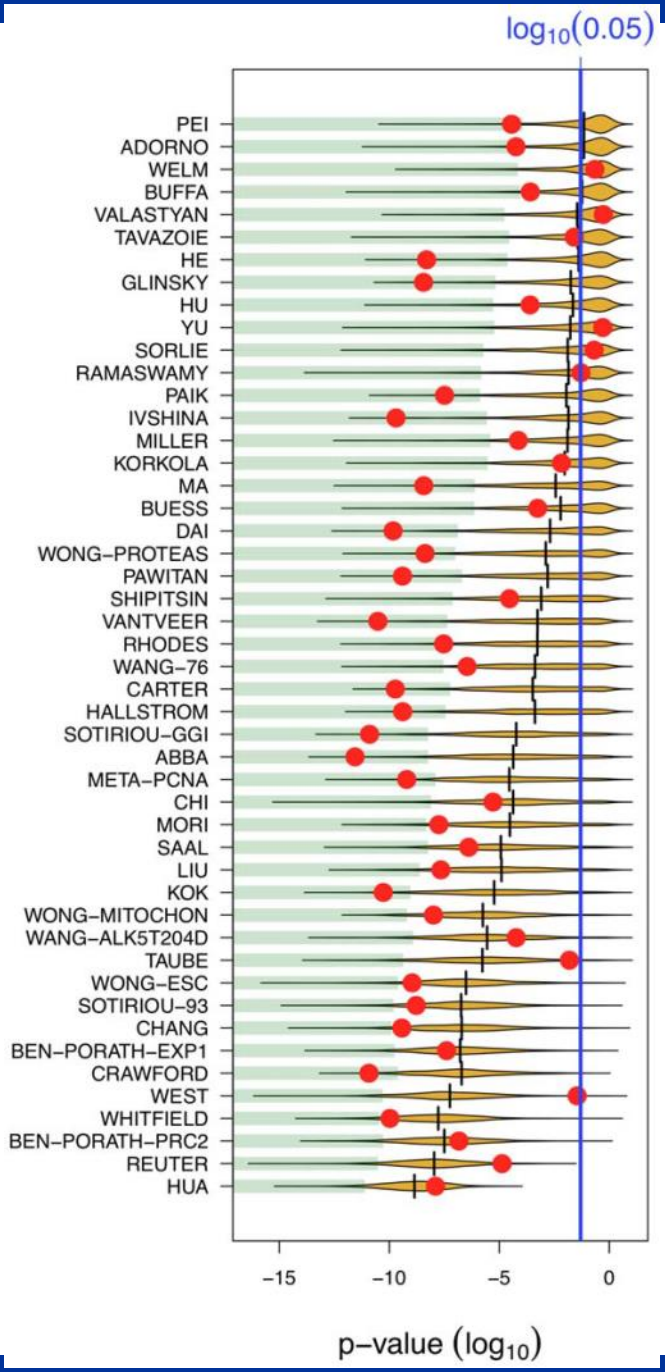
Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

Percentage of overlapping genes

- **Low % of overlapping genes from diff expt in general**
  - Prostate cancer
    - **Lapointe et al, 2004**
    - **Singh et al, 2002**
  - Lung cancer
    - **Garber et al, 2001**
    - **Bhattacharjee et al, 2001**
  - DMD
    - **Haslett et al, 2002**
    - **Pescatori et al, 2007**

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

Zhang et al, *Bioinformatics*, 2009



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

Venet et al. "Most random gene expression signatures are significantly associated with breast cancer outcome". *PLoS Computational Biology*, 7(10):e1002240, 2011.

Too many genes is bad, Need to eliminate irrelevant ones

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **Prob(a gene is correlated) = $1/2^6$**
- **# genes on array =30k**
- **E(# of correlated genes) = 469**

- **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!**

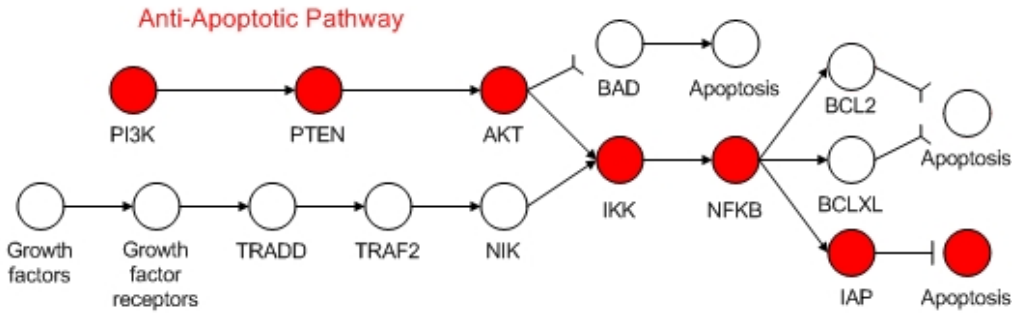
The situation is worse for people looking for genetic mutations that cause a disease

- **10,000,000 SNPs, with 5% MAF**
- **3 control vs 3 disease samples**
- **Prob(a SNP is correlated) = 0.0001**
- **E(# of correlated SNP) = 1,071**

A few stories

- Discovering protein complexes from PPIN
- Identifying causal genes, Part 1
- **Identifying causal genes, Part 2**
- Finding interesting patterns

Biology to the rescue: 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**

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. *Proteomics*, 12(4-5):550-563, 2012.

Big data of biological pathways

Human apoptosis pathway

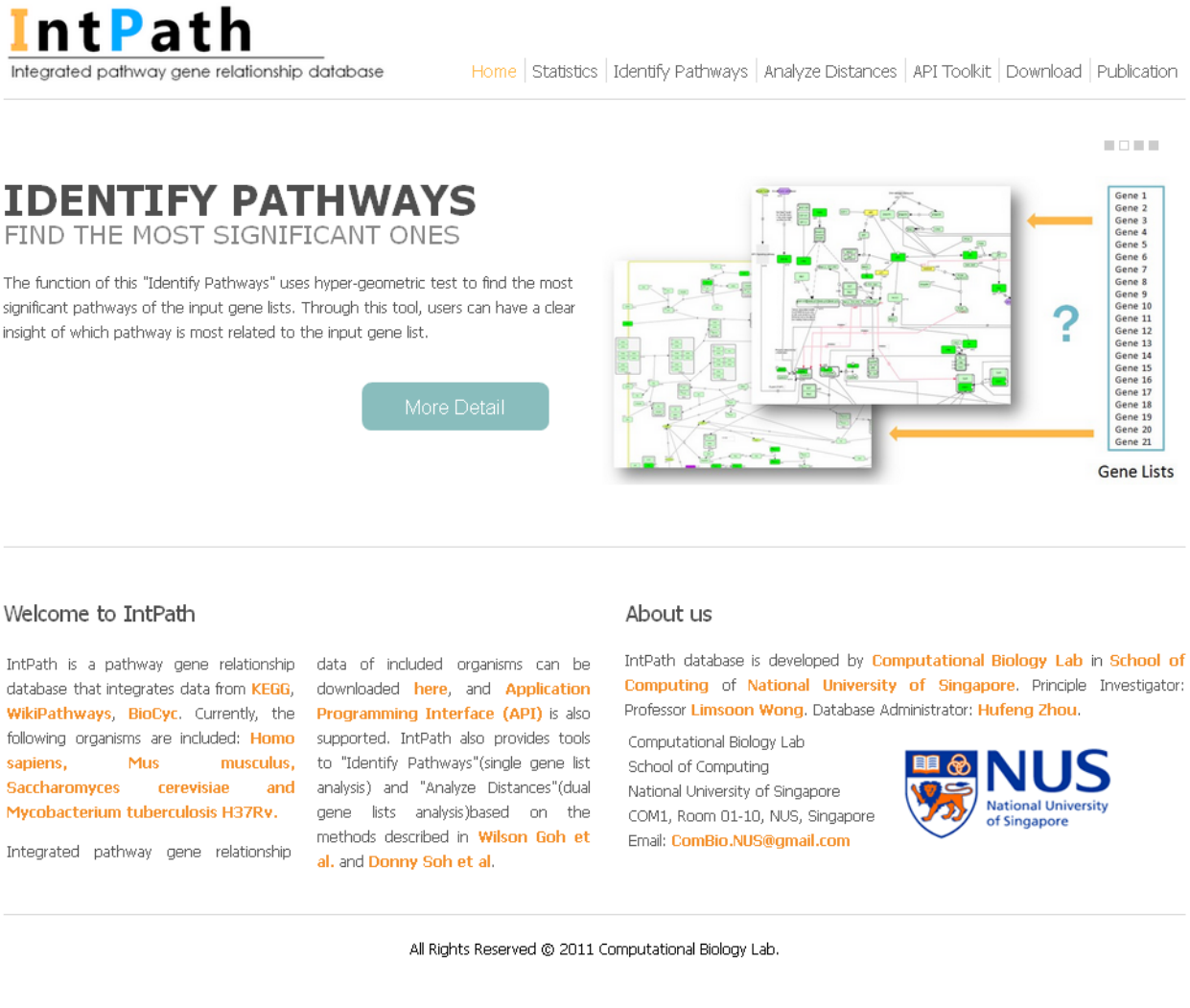
Apoptosis Pathway			
	Wiki x KEGG	Wiki x Ingenuity	KEGG x Ingenuity
Gene Pair Count:	144 vs 172	144 vs 3557	172 vs 3557
Gene Count:	85 vs 80	85 vs 176	80 vs 176
Gene Overlap:	38	28	30
Gene % Overlap:	48%	33%	38%
Gene Pair Overlap:	23	14	24
Gene Pair % Overlap:	16%	10%	14%

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

- **The various data sources have low overlap**

$\Rightarrow$ **Good to unify them to get more complete pathways, right?**

A unified database of biological pathways



Zhou, et al. *BMC Systems Biology*, 6(Suppl 2):S2, 2012.

Using biology background: GSEA

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

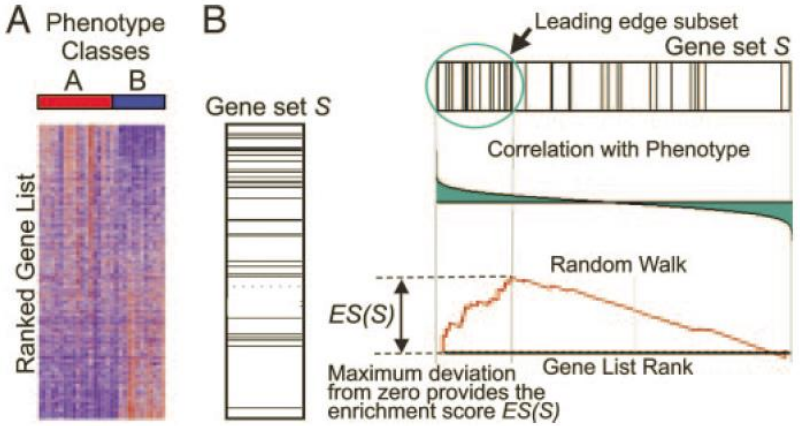


Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location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

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

Unfortunately, it doesn't always work

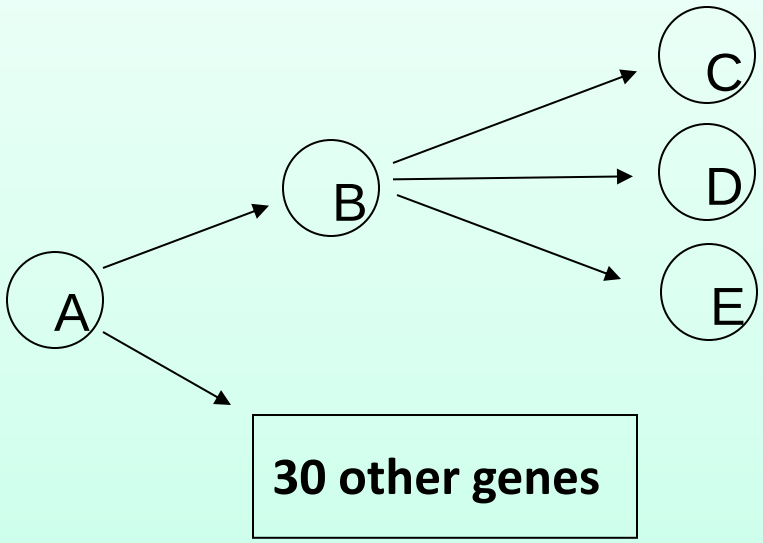
Table 2. Table showing the number and percentage of significant overlapping genes. γ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding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

- **Surprisingly, GSEA fails on large unified pathways!**

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

More is not always better, unless …



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

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

30 other genes not diff expressed

GSEA: Entire network is likely to be missed

- **Need to know how to capture the subnetwork branch within the pathway**

A few stories

- Discovering protein complexes from PPIN
- Identifying causal genes, Part 1
- Identifying causal genes, Part 2
- **Finding interesting patterns**

Different rules may be produced, Different drugs may be considered better

- **Mining only men's data**

Men

	A	B
lived	20	50
died	80	160

- **Mining combined data**

Women

	A	B
lived	40	15
died	20	5

- **You get rules like**
  - "drug A → lived"
    - **Supp = 20/310 = 6%**
    - **Conf = 20/100 = 20%**
  - "drug B → live"
    - **Supp = 50/310 = 16%**
    - **Conf = 50/160 = 31%**

- **You get rules like**
  - "drug A → lived"
    - **Supp = 60/390 = 2%**
    - **Conf = 60/160 = 38%**
  - "drug B → lived"
    - **Supp = 65/390 = 17%**
    - **Conf = 65/180 = 36%**

Statistics lies, unless …

Overall

	A	B
lived	60	65
died	100	165

Looks like treatment A is better

Women

	A	B
lived	40	15
died	20	5

Men

	A	B
lived	20	50
died	80	160

Looks like treatment B is better

History of heart disease

	A	B
lived	10	5
died	70	50

No history of heart disease

	A	B
lived	10	45
died	10	110

Looks like treatment A is better

Challenge : Separating causal factors from confounding factors; making good inferences.

Rules are just rules, unless …

Data mining sensor & telemetry data in a factory may give you rules like …

Fuse blow → Robot stop

··· a thousand other rules ···

Circuit overload → Fuse blow

··· a thousand other rules ···

Insufficient lubrication → Circuit overload

··· a thousand other rules ···

Oil pump clogged → Insufficient lubrication

··· a thousand other rules ···

Metal shavings → Oil pump clogged

Challenge : Asking "why" 5 levels deep, and getting to the root cause.

NUS National University of Singapore

What have we learned?

- **More data can offer a more complete picture, fill in gaps, etc.**
- **More data can also introduce noise into an analysis**
- **Unless you know how to tame this noise, more data may not lead to a better analysis**

How we can get more out of big data

A few suggestions

- **Look deeper into your $\chi 2$ test statistic**
- **Explore more stratifications of your data**
- **Know when to discard the 1st PC in PCA**
- **Make sure you get the null hypothesis right, and exploit domain knowledge properly**

Comparison betw proportions

Treatment	Improvement	No improvement	Total
Arthritic drug	18	6	24
placebo	9	11	20
Total	27	17	44

- **Proportion improved in drug group = 18/24 = 75%**
- **Proportion improved in placebo group = 9/20 = 45.0%**

- **Question: What is the probability that the observed difference of 30% is purely due to sampling error, i.e. chance in sampling?**
- **Use $\chi 2$ –test**

$\chi 2$ test for statistical association

treatment	Improvement	No improvement	Total
Arthritic drug	18 (a)	6 (b)	24
placebo	9 (c)	11 (d)	20
Total	27	17	44

- **Prob of selecting a person in drug group = 24/44**
- **Prob of selecting a person with improvement = 27/44**
- **Prob of selecting a person from drug group who had shown improvement= (24/44)*(27/44) = 0.3347 (assuming two independent events)**
- **Expected value for cell (a) =0.3347*44 = 14.73**

$\chi 2$ test for statistical association

treatment	Improvement	No improvement	Total
Arthritic drug	18 (14.73)	6 (9.27)	24
placebo	9 (12.27)	11 (7.73)	20
Total	27	17	44

- **General formula for $\chi 2$**

$$\chi^2 = \sum \frac{(obs - \exp)^2}{\exp}$$

- **Note: $\chi 2$ test is always performed on categorical variables using absolute frequencies, never percentage or proportion**

$\chi 2$ test for statistical association

- **For the given problem:**

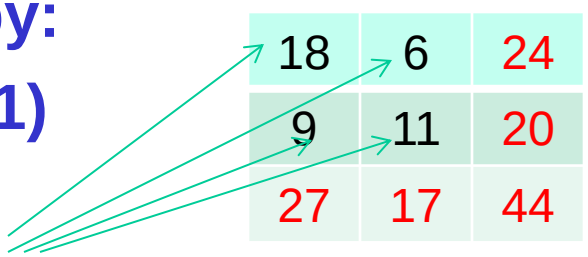
$$\sum \frac{(obs-\exp)^2}{\exp} = \frac{(18-14.73)^2}{14.73} + \frac{(6-9.27)^2}{9.27} + \frac{(9-12.27)^2}{12.27} + \frac{(11-7.73)^2}{7.73}$$

$$= 4.14 \text{ with } 1 \text{ degree of freedom}$$

- **$\chi 2$ degree of freedom is given by:**

  (no. of rows-1)*(no. of cols-1)

  = (2-1)*(2-1) = 1



How many of these 4 cells are free to vary if we keep the row and column totals constant?

χ^2 table

Critical values in the distributions of chi-squared for different degrees of freedom

df	Probability .05	.02	.01	.001
1	3.841	5.412	6.635	10.827
2	5.991	7.824	9.210	13.815
3	7.815	9.837	11.345	16.266
4	9.488	11.668	13.277	18.467
5	11.070	13.388	15.086	20.515
6	12.592	15.033	16.812	22.457
7	14.067	16.622	18.475	24.322
8	15.507	18.168	20.090	26.125
9	16.919	19.679	21.666	27.877
10	18.307	21.161	23.209	29.588
11	19.675	22.618	24.725	31.264
12	21.026	24.054	26.217	32.909
13	22.362	25.372	27.688	34.528
14	23.585	26.873	29.141	36.123
15	24.996	28.259	30.578	37.697
16	26.296	29.633	32.000	39.252
17	27.587	30.995	33.409	40.790
18	28.869	32.346	34.805	42.312
19	30.144	33.687	36.191	43.820
20	31.410	35.020	37.566	35.315
21	32.671	36.343	38.932	46.797
22	33.924	37.659	40.289	48.268
23	35.172	38.968	41.638	49.728
24	36.415	40.270	42.980	51.179
25	37.652	41.566	44.314	52.620
26	38.885	42.856	45.642	54.052
27	40.113	44.140	46.963	55.476
28	41.337	45.419	48.278	56.893
29	42.557	46.693	49.588	58.302
30	43.773	47.962	50.892	59.703

observed χ^2 value of 4.14 exceeds critical value of 3.841 for P=0.05 but not critical value of 5.412 for P=0.02 at 1 d.f.

i.e. $0.05 > P > 0.02$

$\chi 2$ test for statistical association

- **Probability of getting an observed difference of 30% in improvement rates if the Null hypothesis of no association is correct is betw 2% and 5%**

- **Hence, there is some statistical evidence from this study to suggest that treatment of arthritic patient with the drug can significantly improve grip strength**

Extending to RxC tables

Type of vaccines	Had flu	Avoided flu	total
I	43	237	280
II	52	198	250
III	25	245	270
IV	48	212	260
V	57	233	290
Total	225	1125	1350

- **Null hypothesis assumes all vaccines tested had equal efficacy**

Computation of the $\chi 2$

Type of vaccines	Had flu	$(O-E)^2/E$	Avoided flu	$(O-E)^2/E$
I	43 (46.7)	0.293	237 (233.3)	0.059
II	52 (41.7)	2.544	198 (208.3)	0.509
III	25 (45.0)	8.889	245 (225.0)	1.778
IV	48 (43.3)	0.510	212 (216.7)	0.102
V	57 (48.3)	1.567	233 (241.7)	0.313
Total	225	13.803	1125	2.761

- **$\chi 2$ =13.803+2.761 = 16.564 with 4 d.f.**

χ^2 table

Critical values in the distributions of chi-squared for different degrees of freedom

df	Probability .05	.02	.01	.001
1	3.841	5.412	6.635	10.827
2	5.991	7.824	9.210	13.815
3	7.815	9.837	11.345	16.266
4	9.488	11.668	13.277	18.467
5	11.070	13.388	15.086	20.515
6	12.592	15.033	16.812	22.457
7	14.067	16.622	18.475	24.322
8	15.507	18.168	20.090	26.125
9	16.919	19.679	21.666	27.877
10	18.307	21.161	23.209	29.588
11	19.675	22.618	24.725	31.264
12	21.026	24.054	26.217	32.909
13	22.362	25.372	27.688	34.528
14	23.585	26.873	29.141	36.123
15	24.996	28.259	30.578	37.697
16	26.296	29.633	32.000	39.252
17	27.587	30.995	33.409	40.790
18	28.869	32.346	34.805	42.312
19	30.144	33.687	36.191	43.820
20	31.410	35.020	37.566	35.315
21	32.671	36.343	38.932	46.797
22	33.924	37.659	40.289	48.268
23	35.172	38.968	41.638	49.728
24	36.415	40.270	42.980	51.179
25	37.652	41.566	44.314	52.620
26	38.885	42.856	45.642	54.052
27	40.113	44.140	46.963	55.476
28	41.337	45.419	48.278	56.893
29	42.557	46.693	49.588	58.302
30	43.773	47.962	50.892	59.703

observed χ^2 value of 16.564 with 4 d.f. exceeds critical value of 13.277 for P=0.01 but not critical value of 18.467 for P=0.001.

i.e. $0.01 > P > 0.001$

Digging deeper

Type of vaccines	Had flu	$(O-E)^2/E$	Avoided flu	$(O-E)^2/E$
I	43 (46.7)	0.293	237 (233.3)	0.059
II	52 (41.7)	2.544	198 (208.3)	0.509
III	**25 (45.0)**	**8.889**	**245 (225.0)**	**1.778**
IV	48 (43.3)	0.510	212 (216.7)	0.102
V	57 (48.3)	1.567	233 (241.7)	0.313
Total	225	13.803	1125	2.761

- **Vaccine III contributes to the overall $\chi 2$= (8.889+1.778)/16.564 = 64.4%**

$\chi 2$ with Vaccine III removed

Type of vaccines	Had flu	Avoided flu	total
I	43	237	280
II	52	198	250
IV	48	212	260
V	57	233	290

- **$\chi 2$ =2.983 with 3 d.f.**
- **0.1<p<0.5, not statistically significant**

i.e., vaccine III is necessary for significance

Vaccine III vs. rest

Type of vaccines	Had flu	Avoided flu	total
III	25	245	270
I, II, IV, V	200	880	1080
Total	225	1125	1350

- $\chi 2 = 12.7$ with 1 d.f.
- $P<0.001$
- There appear to be strong statistical evidence that the protective effect of vaccine III is significantly better than the other vaccines

i.e., vaccine III is sufficient for significance

A few suggestions

- Look deeper into your $\chi 2$ test statistic
- **Explore more stratifications of your data**
- Know when to discard the 1st PC in PCA
- Make sure you get the null hypothesis right, and exploit domain knowledge properly

Hypothesis testing

- **A hypothesis compares two or more groups**
  - Do smokers have higher cancer rates than non-smokers?
  - Are children more vulnerable to H1N1 flu than adults?

- **Statistical hypothesis testing**
  - Test whether a hypothesis is supported by data using statistical methods

Conventional hypothesis generation

- **Postulate a hypothesis**
  - Is drug A more effective than drug B?

- **How?**
  - Collect data and eye ball a pattern!

PID	Race	Sex	Age	Smoke	Stage	Drug	Response
1	Caucasian	M	45	Yes	1	A	positive
2	Chinese	M	40	No	2	A	positive
3	African	F	50	Yes	2	B	negative
…	…	…	…	…	…	…	…
N	Caucasian	M	60	No	2	B	negative

P-value

- **Use statistical methods to decide whether a hypothesis "Is drug A more effective than drug B? " is supported by data**
  - E.g., χ2-test

	Response= positive	Response= Negative	Proportion of positive responses
Drug=A	890	110	89%
Drug=B	830	170	83%

- **p-value = 0.0001**
  - Prob of observed diff betw the two drugs given assumption that the they have same effect

Limitations of conventional approach

- **Hypothesis-driven**
  - Scientist has to think of a hypothesis first
  - Allow just a few hypotheses to be tested at a time

- **So much data have been collected …**
  - No clue on what to look for
  - Know something; but do not know all
  - Impossible to inspect so much data manually

⇒ **Exploratory hypothesis testing in a data-driven manner**

Exploratory hypothesis testing

- **Data-driven hypothesis testing**
  - Have a dataset but dunno what hypotheses to test
  - Use computational methods to automatically formulate and test hypotheses from data

- **Problems to be solved:**
  - How to formulate hypotheses?
  - How to automatically generate & test hypotheses?

Formulation of a hypothesis

- **"For Chinese, is drug A better than drug B?"**
- **Three components of a hypothesis:**
  - Context (under which the hypothesis is tested)
    - **Race: Chinese**
  - Comparing attribute
    - **Drug: A or B**
  - Target attribute/target value
    - **Response: positive**
- **⟨{Race=Chinese}, Drug=A|B, Response=positive⟩**

Testing a hypothesis

- **⟨{Race=Chinese}, Drug=A|B, Response=positive⟩**

context	Comparing attribute	response= positive	response= negative
{Race=Chinese}	Drug=A	N^A_{pos}	$N^A - N^A_{pos}$
	Drug=B	N^B_{pos}	$N^B - N^B_{pos}$

- **To test this hypothesis we need info:**
  - N^A =support({Race=Chinese, Drug=A})
  - N^A_{pos} =support({Race=Chinese, Drug=A, Res=positive})
  - N^B =support({Race=Chinese, Drug=B})
  - N^B_{pos} =support({Race=Chinese, Drug=B , Res=positive})

⇒ **Frequent pattern mining**

Significance of observed diff

- **When a single hypothesis is tested, a p-value of 0.05 is recognized as low enough**
  - If we test 1000 hypotheses, ~50 hypotheses will pass the 0.05 threshold by random chance
- **Control false positives**
  - Bonferroni's correction
    - **Family-Wise Error Rate: Prob of making one or more false discoveries**
  - Benjamini and Hochberg's method
    - **False Discovery Rate: Proportion of false discoveries**
  - Permutation method

Need for hypothesis analysis

- **Exploration is not guided by domain knowledge**

  $\Rightarrow$Spurious hypotheses has to be eliminated

- **Reasons behind significant hypotheses**
  - Find attribute-value pairs that change the diff a lot
    - **DiffLift: How much diff betw the two groups is lifted**
    - **Contribution: Freq of attribute-value pairs**

DEFINITION 3 (*DiffLift*($A{=}v|H$)). *Let* $H = \langle P, A_{diff} = \{v_1, v_2\}, A_{target}, v_{target}\rangle$ *be a hypothesis,* A_{target} *be categorical,* $P_1 = P \cup \{A_{diff} = v_1\}$ *and* $P_2 = P \cup \{A_{diff} = v_2\}$ *be the two sub-populations of* H, $A = v$ *be an item not in* H, *that is,* $A \neq A_{diff}$, $A \neq A_{target}$ *and* $A = v \notin P$. *After adding item* $A = v$ *to* H, *we get two new sub-populations:* $P_1' = P_1 \cup \{A = v\}$ *and* $P_2' = P_2 \cup \{A = v\}$. *The lift of difference after adding* $A = v$ *to* H *is defined as* $DiffLift(A{=}v|H) = \frac{p_1' - p_2'}{p_1 - p_2}$, *where* p_i *is the proportion of* v_{target} *in sub-population* P_i, *and* p_i' *is the proportion of* v_{target} *in sub-population* P_i', $i{=}1, 2$.

DEFINITION 6 (*Contribution*($A = v|H$)). *Let* H *be a hypothesis,* $A = v$ *be an attribute value not in* H, P_1 *and* P_2 *be the two sub-populations of* H, P_1' *and* P_2' *be the two sub-populations after adding* $A = v$ *to* H. *The contribution of* $A = v$ *to* H *is defined as* $Contribution(A = v|H) = \frac{\frac{n_1'}{n_1}(p_1' - p_1) - \frac{n_2'}{n_2}(p_2' - p_2)}{p_1 - p_2}$, *where* p_i *is the proportion of* v_{target} *in sub-population* P_i, *and* p_i' *is the proportion of* v_{target} *in sub-population* P_i', $i = 1, 2$.

Spurious hypotheses

	response= positive	response= negative	proportion of positive response
Drug=A	890	110	89.0%
Drug=B	830	170	83.0%
Drug=A, Stage=1	800	80	90.9%
Drug=B, Stage=1	190	10	95%
Drug=A, Stage=2	90	30	75%
Drug=B, Stage=2	640	160	80%

- **Simpson's Paradox**
  - "Stage" has assoc w/ both "drug" & "response":
    - **Doc's tend to give drug A to patients at stage 1, & drug B to patients at stage 2**
    - **Patients at stage 1 are easier to cure than patients at stage 2**
  - Attribute "stage" is called a confounding factor

Reasons behind significant hypotheses

	Failure rates
Product A	4%
Product B	2%
Product A, time-of-failure=loading	6.0%
Product B, time-of-failure=loading	1.9%
Product A, time-of-failure=in-operation	2.1%
Product B, time-of-failure=in-operation	2.1%
Product A, time-of-failure=output	2.0%
Product B, time-of-failure=output	1.9%

- **Problem is narrowed down**
  - Product A has exceptionally higher drop rate than product B only at the loading phase

Problem statement: Exploratory hypothesis testing

- **Given**
  - Dataset D, min_sup, max_pvalue, min_diff
  - A_{target}= v_{target}
  - $\mathcal{A}_{grouping}$: context/comparing attributes

- **Find all H = ⟨P,A_{diff} = v_1|v_2, A_{target}= v_{target} ⟩**
  - $A_{diff} \in \mathcal{A}_{grouping}$ & $\forall(A{=}v)$ in P, $A \in \mathcal{A}_{grouping}$
  - $sup(P_i) \geq$ min_sup, where $P_i = P \cup \{A_{diff} = v_i\}$, i=1, 2
  - p-value(H) ≤ max_pvalue
  - $|p_1 - p_2| \geq$ min_diff, where p_i is proportion of v_{target} in sub-population P_i, i=1, 2

NUS National University of Singapore

Problem statement:
Hypothesis analysis

- **Given a significant hypothesis H, generate the following info for further analysis**
  - Simpson's Paradoxes formed by H with attributes not in H
  - List of attribute-value pairs not in H ranked in descending order of DiffLift(A=v|H) and Contribution(A=v|H)
  - List of attributes not in H ranked in descending order of DiffLift(A|H) and Contribution(A|H)

Algo for exploratory hypothesis testing

- **A hypothesis is a comparison betw two or more sub-populations, and each sub-population is defined by a pattern**
- **Step 1: Use freq pattern mining to enumerate large sub-populations and collect their statistics**
  - Stored in the CFP-tree structure, which supports efficient subset/superset/exact search
- **Step 2: Pair sub-populations up to form hypotheses, and then calculate their p-values**
  - Use each freq pattern as a context
  - Search for immediate supersets of the context patterns, and then pair these supersets up to form hypotheses

Algo for hypothesis analysis

- **Given a hypothesis H**
  - To check whether H forms a Simpson's Paradox with an attribute A,
    - **add values of A to context of H**
    - **re-calculate the diff betw the two sub-populations**
  - To calculate DiffLift and Contribution of an attribute-value pair A=v,
    - **add A=v to context of H**
    - **re-calculate the diff**

- **All can be done via immediate superset search**

Experiment settings

- **PC configurations**
  - 2.33Ghz CPU, 3.25GB memory, Windows XP
- **Datasets:**
  - mushroom, adult: UCI repository
  - DrugTestI, DrugTestII: study assoc betw SNPs in several genes & drug responses.

Datasets	#instances	#continuous attributes	#categorical attributes	A_{target}/v_{target}
adult	48842	6	9	class=>50K (nominal)
mushroom	8124	0	23	class=poisonous (nominal)
DrugTestI	141	13	74	logAUCT (continuous)
DrugTestII	138	13	74	logAUCT (continuous)

NUS National University of Singapore

Running time

- **Three phases**
  - Frequent pattern mining
  - Hypothesis generation
  - Hypothesis analysis

Datasets	min_sup	min_diff	GenH	AnalyzeH	AvgAnalyzeT	#tests	#signH
adult	500	0.05	0.42 s	6.30 s	0.0015 s	5593	4258
adult	100	0.05	2.69 s	37.39 s	0.0014 s	41738	26095
mushroom	500	0.1	0.67 s	19.00 s	0.0020 s	16400	9323
mushroom	200	0.1	5.45 s	123.47 s	0.0020 s	103025	61429
DrugTestI	20	0.5	0.06 s	0.06 s	0.0031 s	3627	20
DrugTestII	20	0.5	0.08 s	0.30 s	0.0031 s	4441	97

max_pvalue = 0.05

Case study: Adult dataset

- **Simpson's paradox**

Context	Comparing Groups	sup	$P_{class=>50K}$	p-value
Race =White	Occupation = Craft-repair	3694	22.84%	1.00×10^{-19}
	Occupation = Adm-clerical	3084	14.23%	

Context	Extra attribute	Comparing Groups	sup	$P_{class=>50K}$
Race =White	Sex = Male	Occupation = Craft-repair	3524	23.5%
		Occupation = Adm-clerical	1038	24.2%
	Sex = Female	Occupation = Craft-repair	107	8.8%
		Occupation = Adm-clerical	2046	9.2%

A few suggestions

- Look deeper into your $\chi 2$ test statistic
- Explore more stratifications of your data
- **Know when to discard the 1st PC in PCA**
- Make sure you get the null hypothesis right, and exploit domain knowledge properly

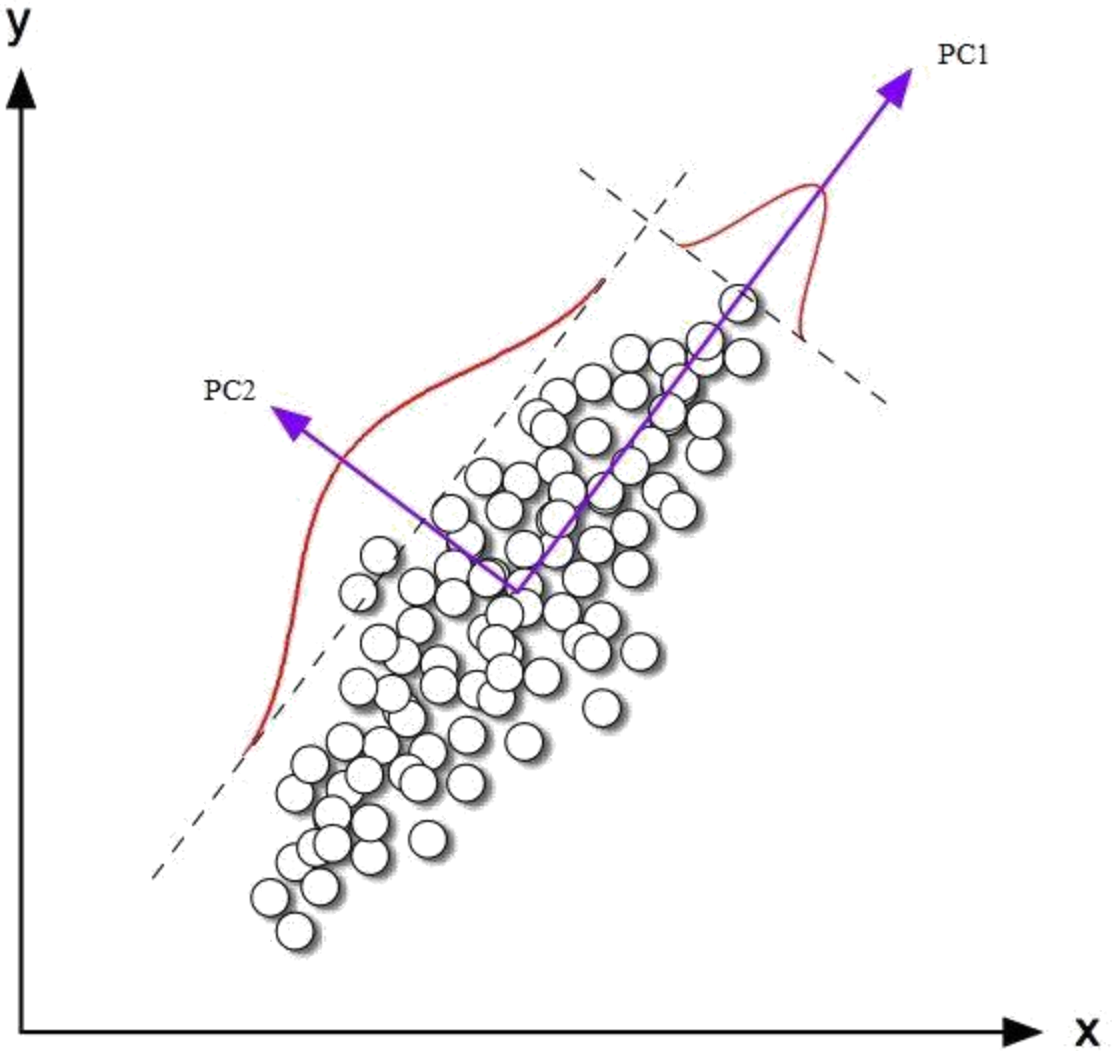
Uses of PCA

- Dimension reduction
  - Summarize the data with a smaller number of variables, losing as little info as possible
  - Graphical representations of data

- Input for regression analysis
  - Highly correlated explanatory variables are problematic in regression analysis
  - One can replace them by their principal components, which are uncorrelated by definition

Credit: Marloes Maathuis

Principal component analysis



Credit: Alessandro Giuliani

PCA, a la Pearson (1901)

)(98)(

SULLE FUNZIONI BILINEARI

DI

E. BELTRAMI

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London* *.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$
$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

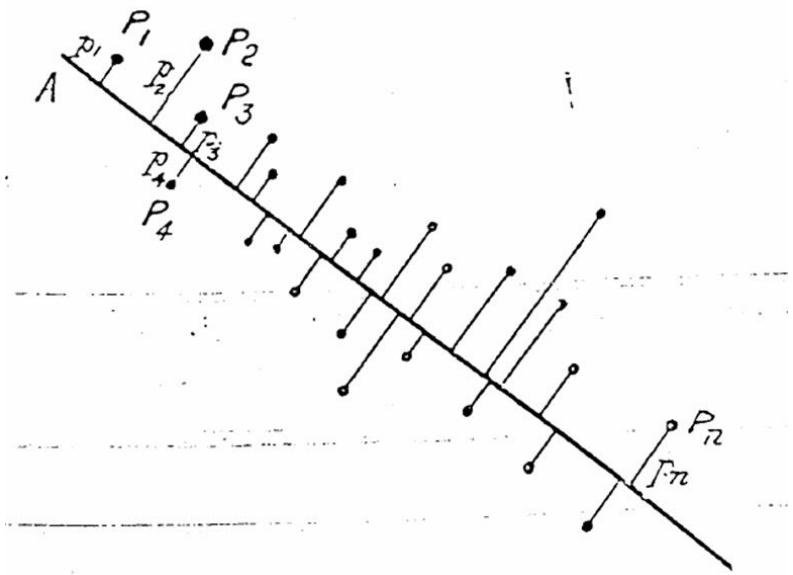
where $y, x, z, x_1, x_2, \ldots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n$

For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U = S(p^2) = \text{a minimum.}$$

If y were the dependent variable, we should have made

$$S(y' - y)^2 = \text{a minimum}$$



Credit: Alessandro Giuliani

Credit: Marloes Maathuis

PCA, in modern English ☺

Introduction

- Technique quite old: Pearson (1901) and Hotelling (1933), but still one of the most used multivariate techniques today
- Main idea:
  - Start with variables $X_1, \ldots, X_p$
  - Find a *rotation* of these variables, say $Y_1, \ldots, Y_p$ (called principal components), so that:
    - $Y_1, \ldots, Y_p$ are uncorrelated. Idea: they measure different dimensions of the data.
    - $\text{Var}(Y_1) \geq \text{Var}(Y_2) \geq \ldots \text{Var}(Y_p)$. Idea: Y_1 is most important, then Y_2, etc.

9 / 33

Definition of PCA

- Given $X = (X_1, \ldots, X_p)'$
- We call $a'X$ a standard linear combination (SLC) if $\sum a_i^2 = 1$
- Find the SLC $a'_{(1)} = (a_{11}, \ldots, a_{p1})$ so that $Y_1 = a'_{(1)}X$ has maximal variance
- Find the SLC $a'_{(2)} = (a_{12}, \ldots, a_{p2})$ so that $Y_2 = a'_{(2)}X$ has maximal variance, subject to the constraint that Y_2 is uncorrelated to Y_1.
- Find the SLC $a'_{(3)} = (a_{13}, \ldots, a_{p3})$ so that $Y_3 = a'_{(3)}X$ has maximal variance, subject to the constraint that Y_3 is uncorrelated to Y_1 and Y_2
- Etc...

10 / 33

1st principal component

- How to combine the scores on 5 different exams to a total score? One could simply take the average. But it may be better to use the first principal component

- How to combine different cost factors into a cost of living index? Use first principal component

- The first principal component maximizes the variance, it spreads out the scores as much as possible

Credit: Marloes Maathuis

2nd and other principal components

- **When all measurements are positively correlated,** the 1st principal component is often some kind of average of the measurements
  - Size of birds
  - Severity index of psychiatric symptoms, …
- **The 2nd and other principal components give important info about the remaining pattern**
  - Shape of birds
  - Pattern of psychiatric symptoms, …

Credit: Marloes Maathuis

Growth, 1960, **24**, 339-354.

SIZE AND SHAPE VARIATION IN THE PAINTED TURTLE.[1] A PRINCIPAL COMPONENT ANALYSIS

PIERRE JOLICOEUR AND JAMES E. MOSIMANN[2]

Walker Museum, University of Chicago
and
Institut de Biologie, Université de Montréal

(Received for publication July 11, 1960)

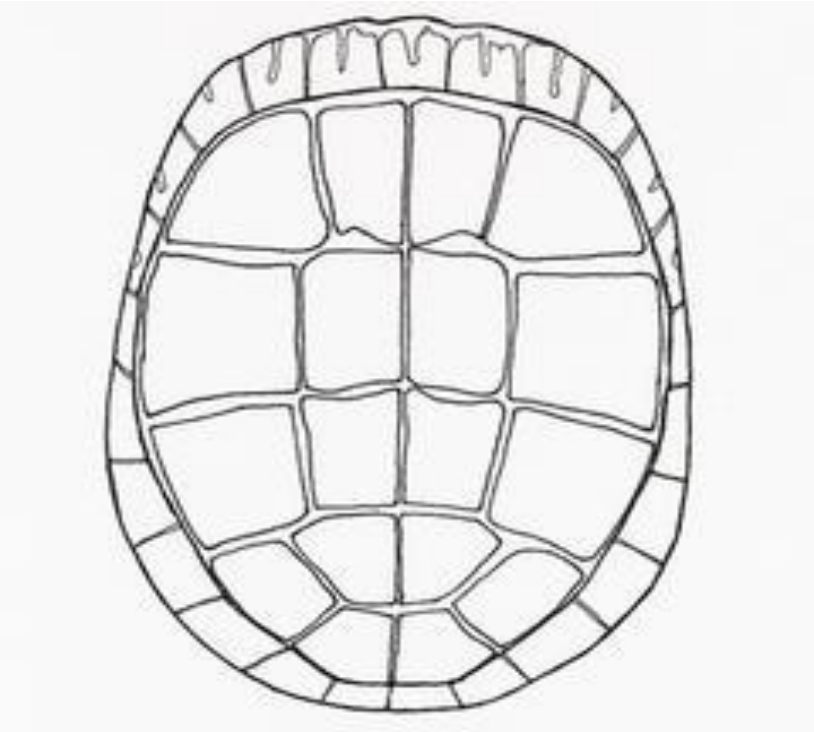
Credit: Alessandro Giuliani

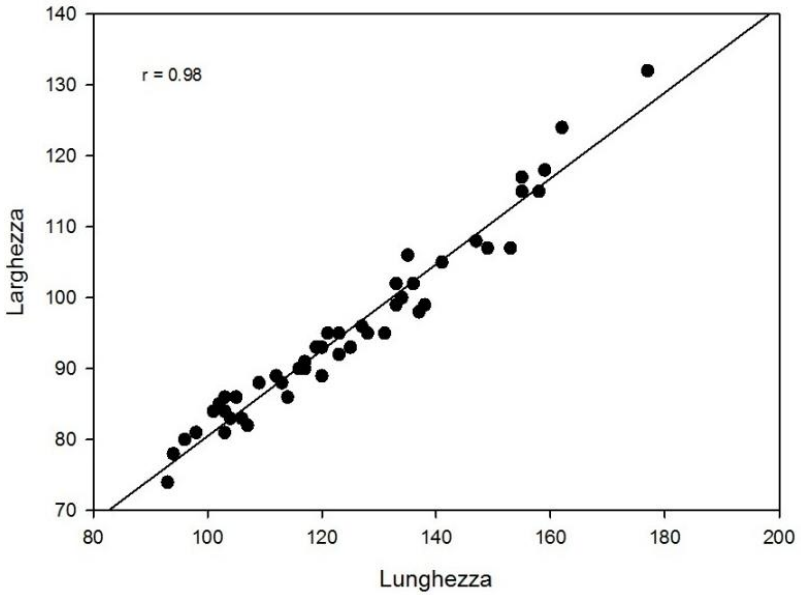
TABLE 1

CARAPACE DIMENSIONS OF PAINTED TURTLES (*Chrysemys picta marginata*) IN MM.

24 Males			24 Females		
length	width	height	length	width	height
93	74	37	98	81	38
94	78	35	103	84	38
96	80	35	103	86	42
101	84	39	105	86	40
102	85	38	109	88	44
103	81	37	123	92	50
104	83	39	123	95	46
106	83	39	133	99	51
107	82	38	133	102	51
112	89	40	133	102	51
113	88	40	134	100	48
114	86	40	136	102	49
116	90	43	137	98	51
117	90	41	138	99	51
117	91	41	141	105	53
119	93	41	147	108	57
120	89	40	149	107	55
120	93	44	153	107	56
121	95	42	155	115	63
125	93	45	155	117	60
127	96	45	158	115	62
128	95	45	159	118	63
131	95	46	162	124	61
135	106	47	177	132	67

Credit: Alessandro Giuliani





Pearson Correlation Coefficients,

	length	width	height
length	1.00000	0.97831	0.96469
width	0.97831	1.00000	0.96057
height	0.96469	0.96057	1.00000

Width = 19,94 + 0,605*Length

Credit: Alessandro Giuliani

Credit: Alessandro Giuliani

Interesting info are often in the 2nd principal component

	PC1 (98%)	PC2 (1.4%)
Length	0,992	-0,067
Width	0,990	-0,100
Height	0,986	0,168

$$PC1 = 33.78*Length + 33.73*Width + 33.57*Height$$

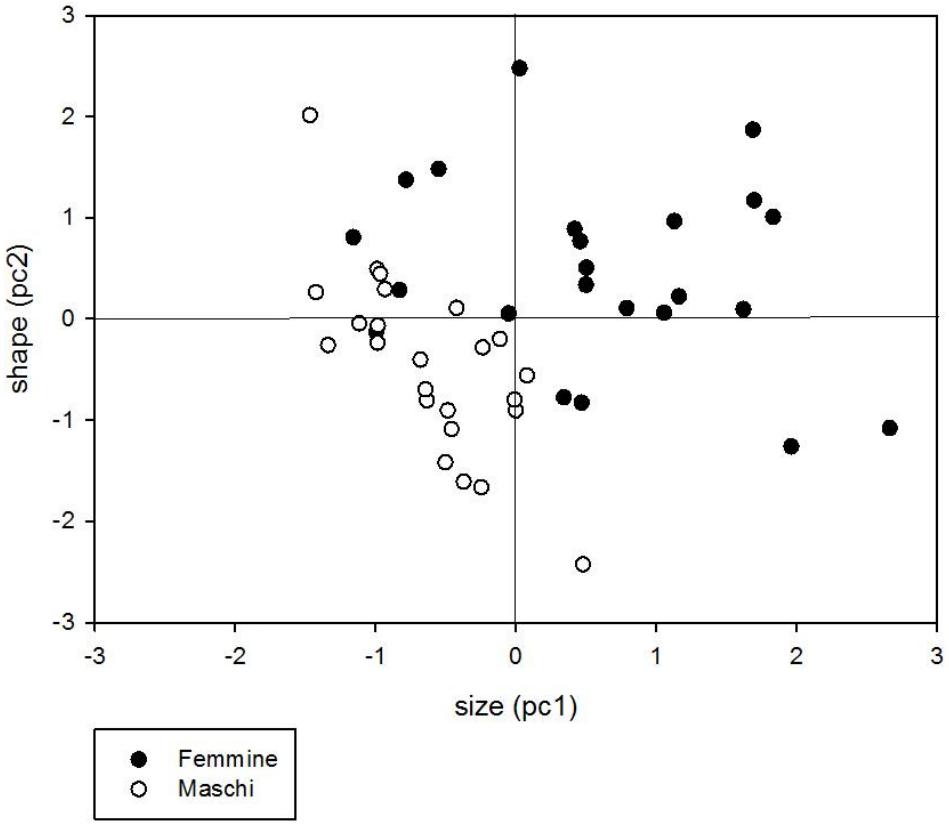
$$PC2 = -1.57*Length - 2.33*Width + 3.93*Height$$

- Presence of an overwhelming size component explaining system variance comes from the presence of a 'typical' common shape
- Displacement along pc = size variation (all positive terms)
- Displacement along pc2 = shape deformation (both positive and negative terms)

unit	sex	Length	Width	Height	PC1(size)	PC2(shape)
T25	F	98	81	38	-1,15774	0,80754832
T26	F	103	84	38	-0,99544	-0,1285916
T27	F	103	86	42	-0,7822	1,37433475
T28	F	105	86	40	-0,82922	0,28526912
T29	F	109	88	44	-0,55001	1,4815252
T30	F	123	92	50	0,027368	2,47830153
T31	F	123	95	46	-0,05281	0,05403839
T32	F	133	99	51	0,418589	0,88961967
T33	F	133	102	51	0,498425	0,33681756
T34	F	133	102	51	0,498425	0,33681756
T35	F	134	100	48	0,341684	-0,774911
T36	F	136	102	49	0,467898	-0,8289156
T37	F	137	98	51	0,457949	0,76721682
T38	F	138	99	51	0,501055	0,50628189
T39	F	141	105	53	0,790215	0,10640554
T40	F	147	108	57	1,129025	0,96505915
T41	F	149	107	55	1,055392	0,06026089
T42	F	153	107	56	1,161368	0,22145593
T43	F	155	115	63	1,687277	1,86903869
T44	F	158	115	62	1,696753	1,17117077
T45	F	159	118	63	1,833086	1,00956637
T46	F	162	124	61	1,962232	-1,261771
T47	F	177	132	67	2,662548	-1,0787317
T48	F	155	117	60	1,620491	0,09690818
T1	M	93	74	37	-1,46649	2,01289241
T2	M	94	78	35	-1,42356	0,26342486
T3	M	96	80	35	-1,33735	-0,258445
T4	M	101	84	39	-0,98842	0,49260881
T5	M	102	85	38	-0,98532	-0,2361914
T6	M	103	81	37	-1,11528	-0,0436547
T7	M	104	83	39	-0,96555	0,44687352
T8	M	106	83	39	-0,93257	0,29353841
T9	M	107	82	38	-0,98269	-0,066727
T10	M	112	89	40	-0,63393	-0,8042059
T11	M	113	88	40	-0,64405	-0,6966061
T12	M	114	86	40	-0,68078	-0,4047389
T13	M	116	90	43	-0,42133	0,10845233
T14	M	117	90	41	-0,48485	-0,9039457
T15	M	117	91	41	-0,45824	-1,0882131
T16	M	119	93	41	-0,37202	-1,610083
T17	M	120	89	40	-0,50198	-1,4175463
T18	M	120	93	44	-0,23552	-0,2831547
T19	M	121	95	42	-0,24581	-1,6640875
T20	M	125	93	45	-0,11305	-0,1986272
T21	M	127	96	45	-0,00023	-0,9047645
T22	M	128	95	45	-0,01035	-0,7971646
T23	M	131	95	46	0,079136	-0,559302
T24	M	135	106	47	0,477846	-2,4250481

Credit: Alessandro Giuliani

Female turtles are larger and have more exaggerated height ☺



Caution: PCA is not scale invariant

- Suppose we have measurements in kg and meters, and we want to have principal components expressed in grams and hectometers
- Option 1: multiply measurements in kg by 1000, multiply measurements in meters by 1/100, and then apply PCA
- Option 2: apply PCA on original measurements, and then re-scale to the appropriate units
- These two options generally give different results!

Credit: Marloes Maathuis

Caution: PCA is sensitive to outliers

Formal definition of PCA - population case

- We first consider the population case: Let $X \in \mathbb{R}^p$ be a random vector with mean μ and covariance matrix Σ (note that we don't make any assumptions about the distribution of X).
- Then the principal component transformation is the transformation

$$X \to Y = \Gamma'(X - \mu)$$

where Γ is the orthogonal matrix consisting of the standardized eigenvectors corresponding to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of Σ. Thus, $\Sigma = \Gamma\Lambda\Gamma'$, or equivalently, $\Gamma'\Sigma\Gamma = \Lambda$.

18 / 33

- PCA is sensitive to outliers, since it is based on the sample covariance matrix Σ which is sensitive to outliers

Credit: Marloes Maathuis

A few suggestions

- Look deeper into your $\chi 2$ test statistic
- Explore more stratifications of your data
- Know when to discard the 1st PC in PCA
- **Make sure you get the null hypothesis right, and exploit domain knowledge properly**

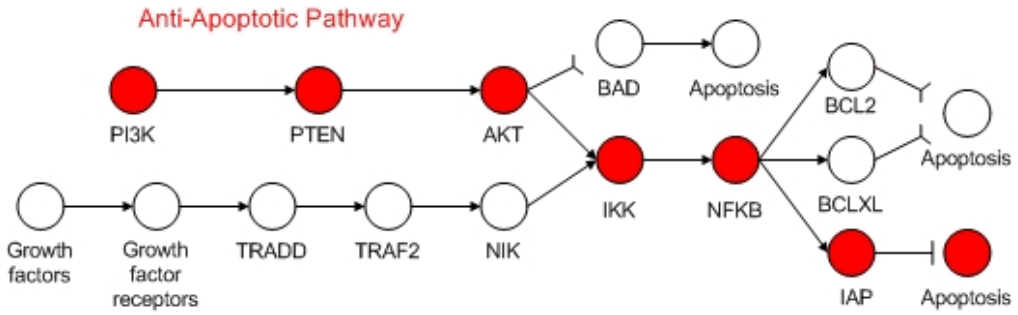
Percentage of overlapping genes

- **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

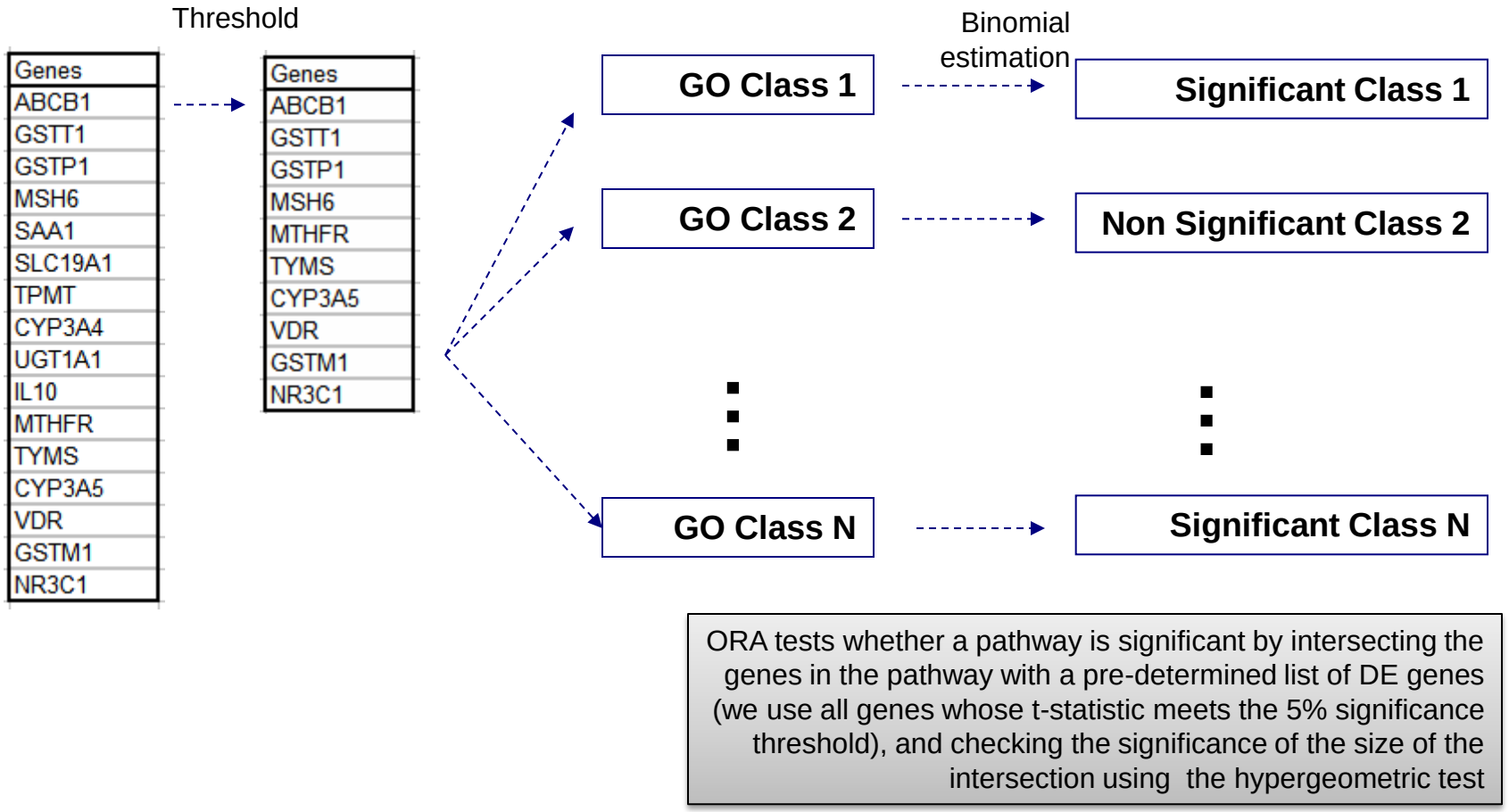
Gene regulatory circuits



- **Each disease phenotype has some underlying cause**
- **There is some unifying biological theme for genes that are truly associated with a disease subtype**

- **Uncertainty in selected genes can be reduced by considering biological processes of the genes**
- **The unifying biological theme is basis for inferring the underlying cause of disease subtype**

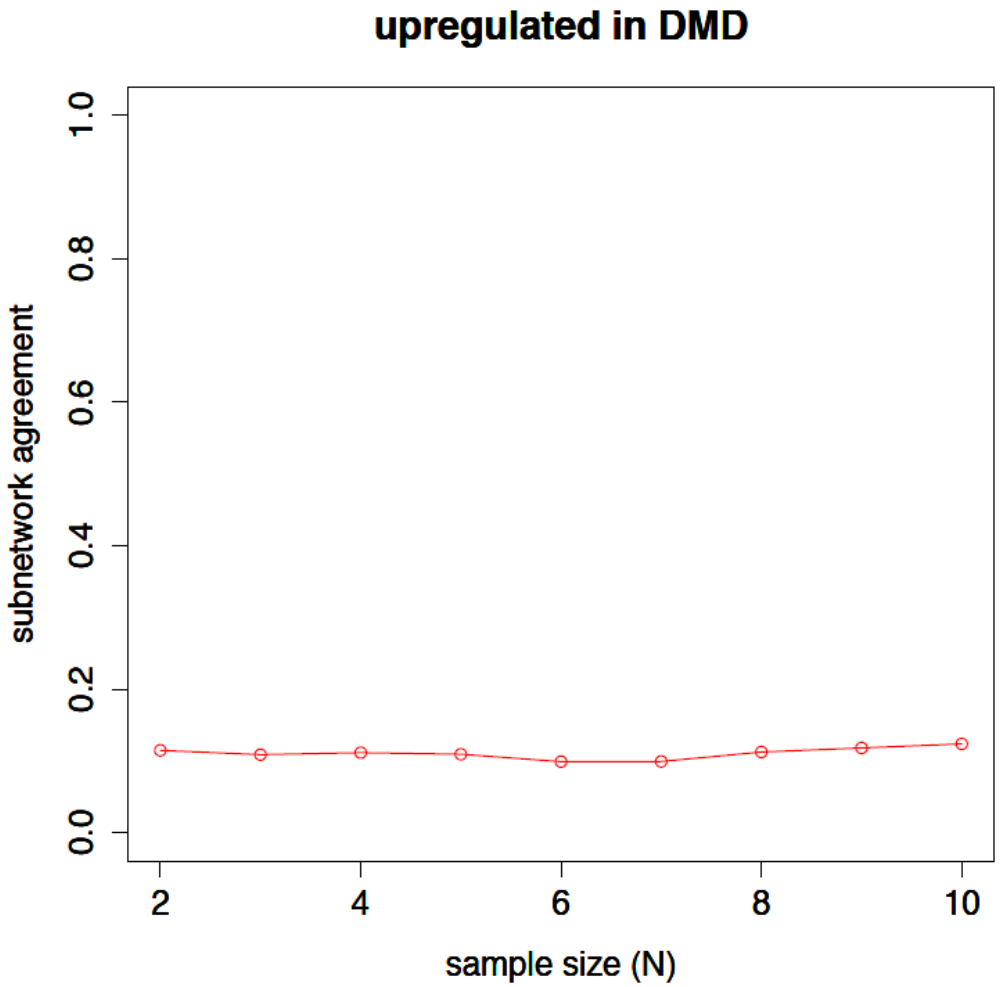
Overlap analysis: ORA



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

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

Disappointing performance



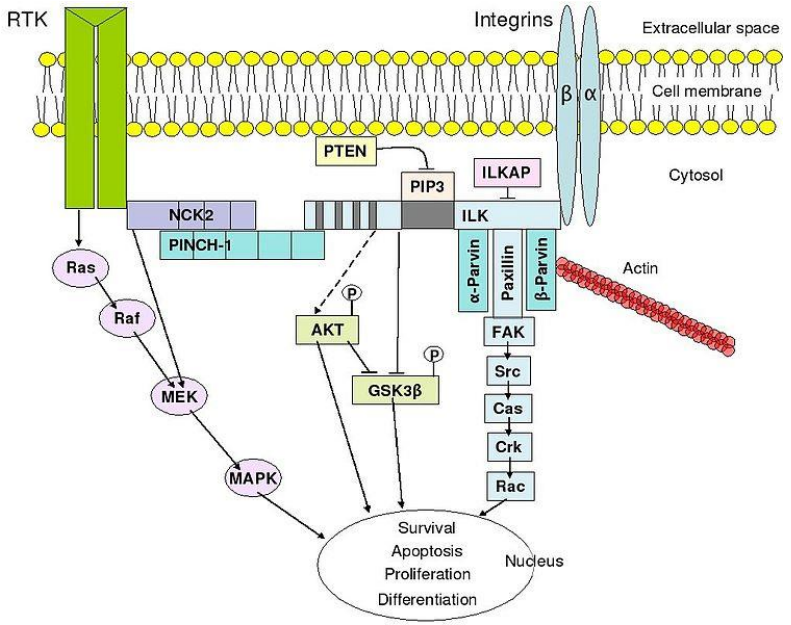
DMD gene expression data
- Pescatori et al., 2007
- Haslett et al., 2002

Pathway data
- PathwayAPI, Soh et al., 2010

Issue #1 with ORA

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

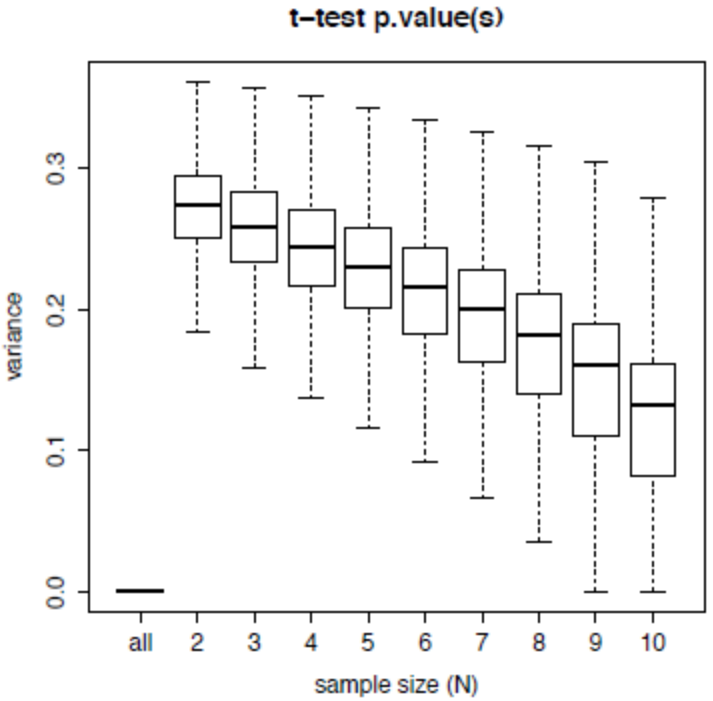
- **This may lead to lots of false positives**



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

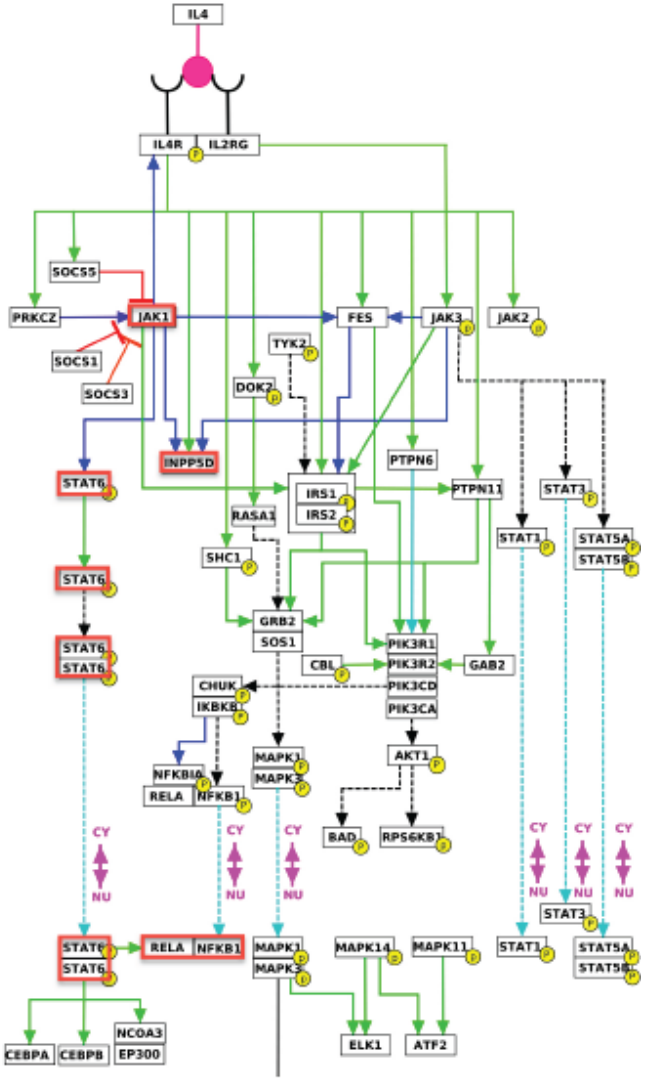
Issue #2 with ORA

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

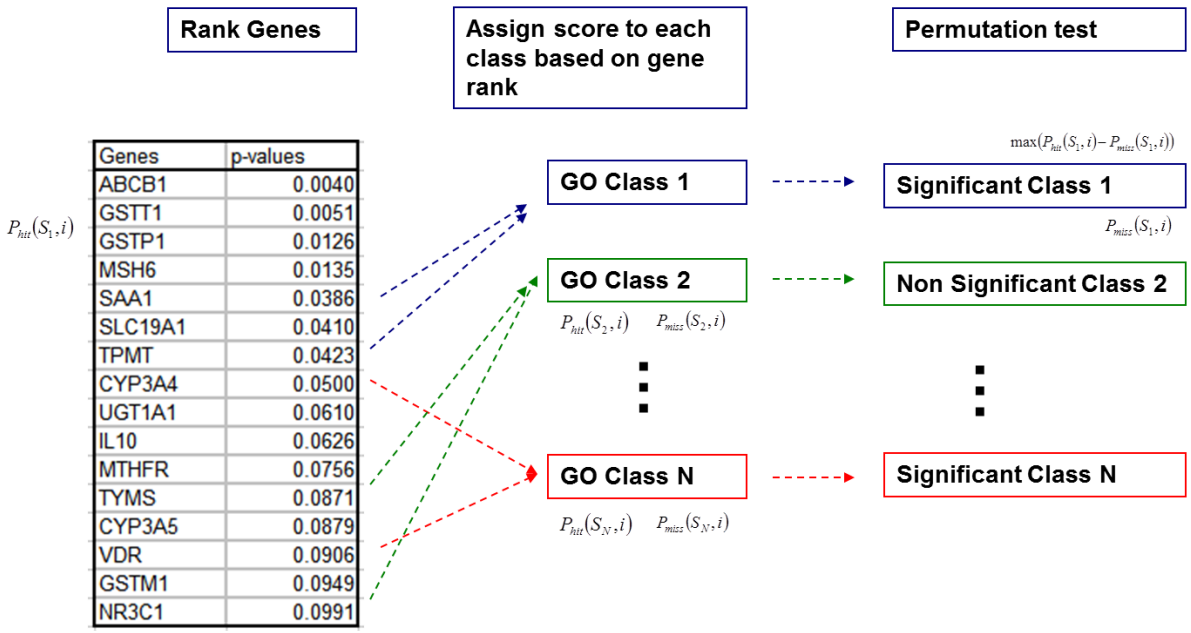


Issue #3 with ORA

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



GSEA

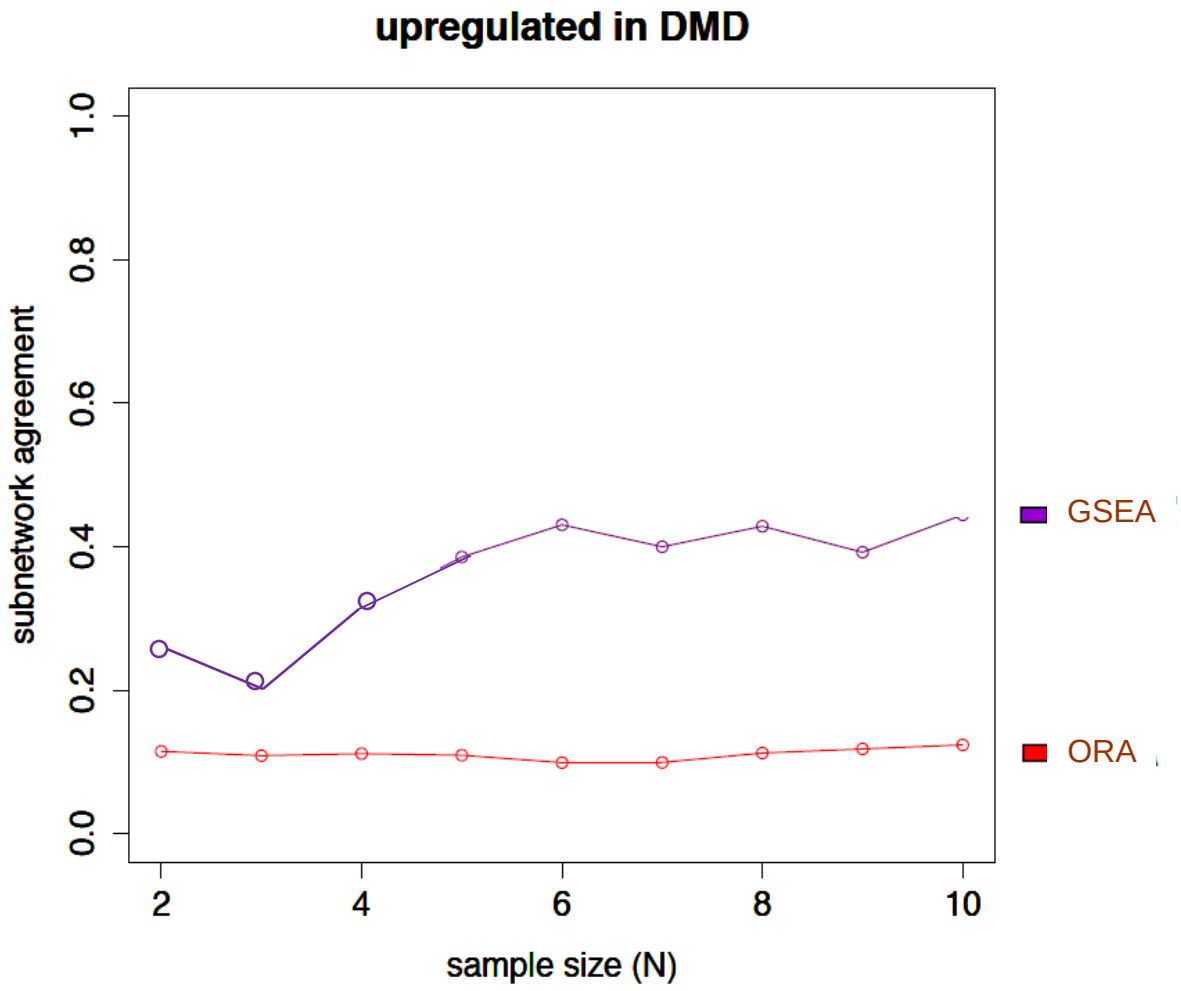


A Subramanian et al. "Gene set enrichment analysis: A knowledge-based approach for interpreting genome wide expression profiles". *PNAS*, 102(43):15545-15550, 2005

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

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

Better performance



ORA-Paired: Paired test and new null hypothesis

- Let g_i be genes in a given pathway P
- Let p_j be patients
- Let q_k be normals

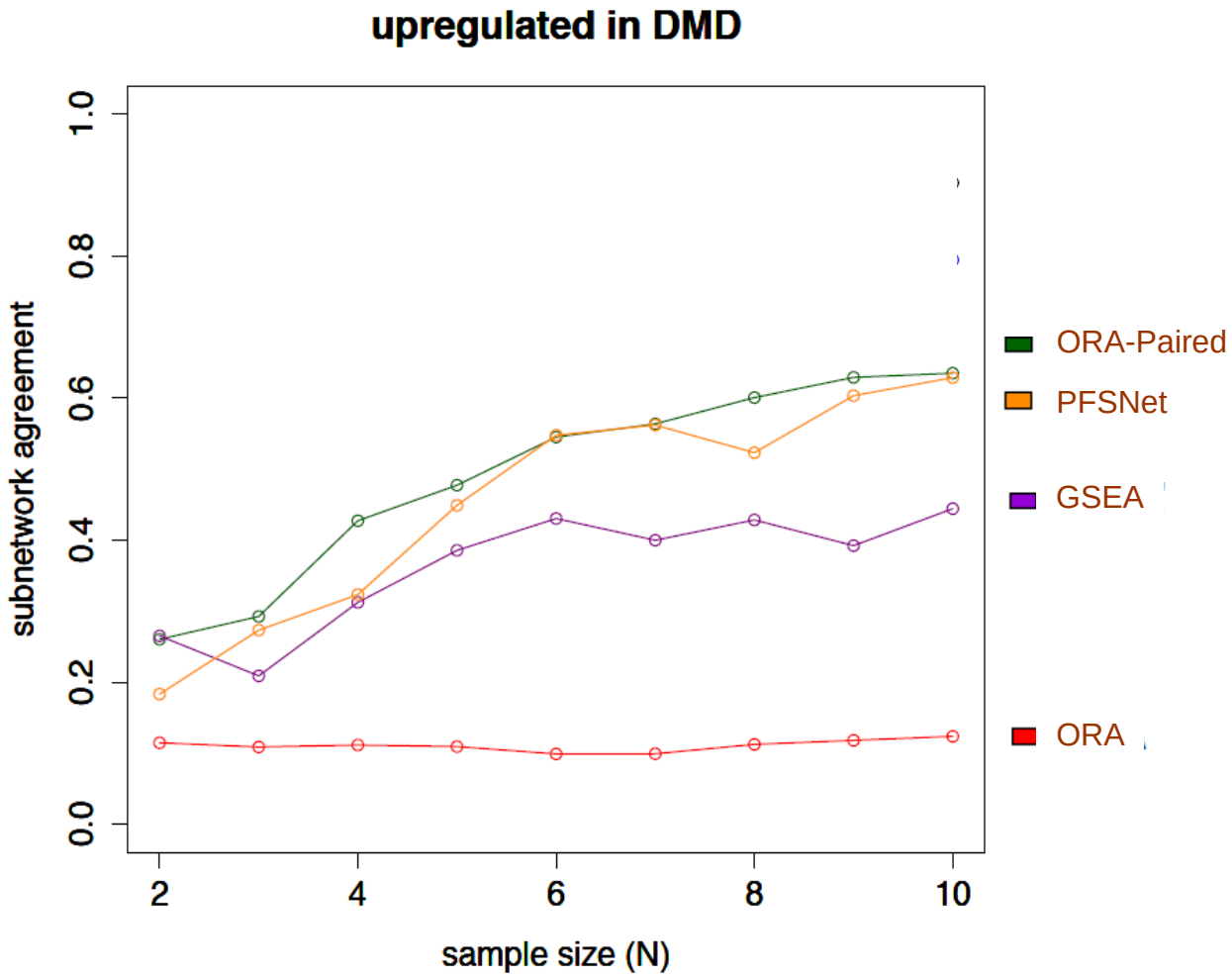
- Let $\Delta_{i,j,k} = Expr(g_i,p_j) – Expr(g_i,q_k)$

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

- **Issue #1 is solved**
  - The null hypothesis is now "If a pathway P is irrelevant to the difference between patients and normals, then the genes in P are expected to behave similarly in patients and normals"

- **Issue #2 is solved**
  - No longer need a pre-determined list of DE genes
  - Sample size is now much larger
    - ~~# patients + # normals~~
    - # patients * # normals * # genes in P

Much better performance

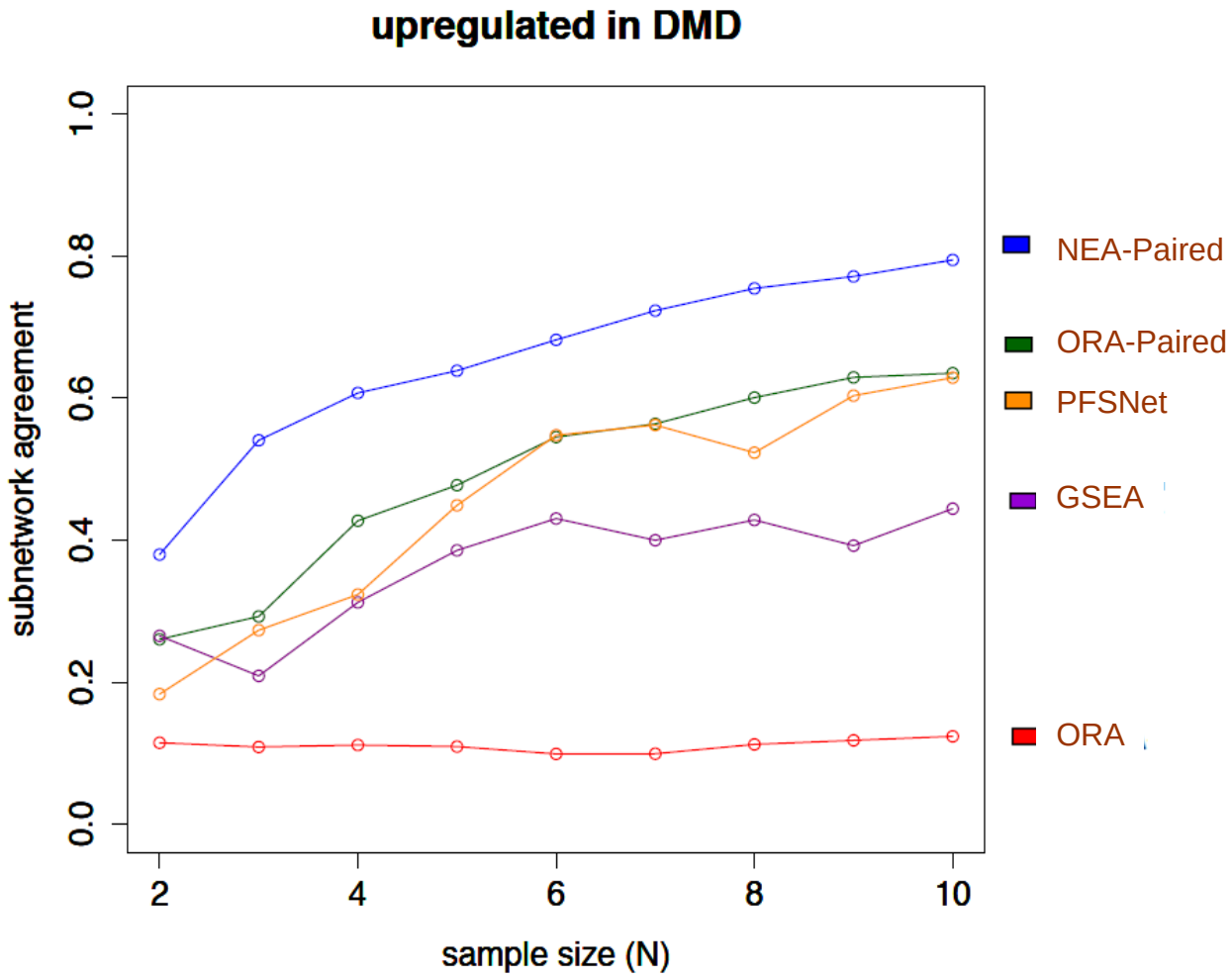


NEA-Paired: Paired test on subnetworks

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

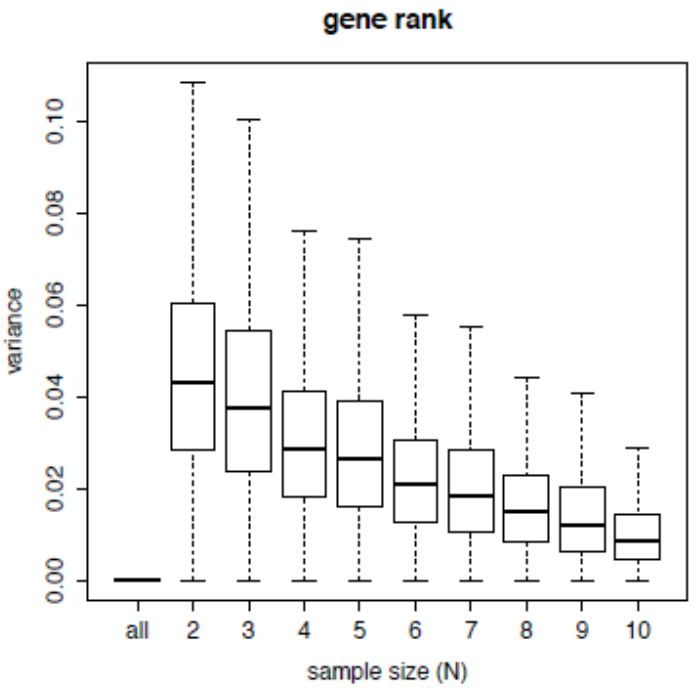
- **Issues #1 & #2 are solved as per ORA-Paired**
- **Issue #3 is partly solved**
  - Testing subnetworks instead of whole pathways
  - But subnetworks derived in fragmented way

Even better performance



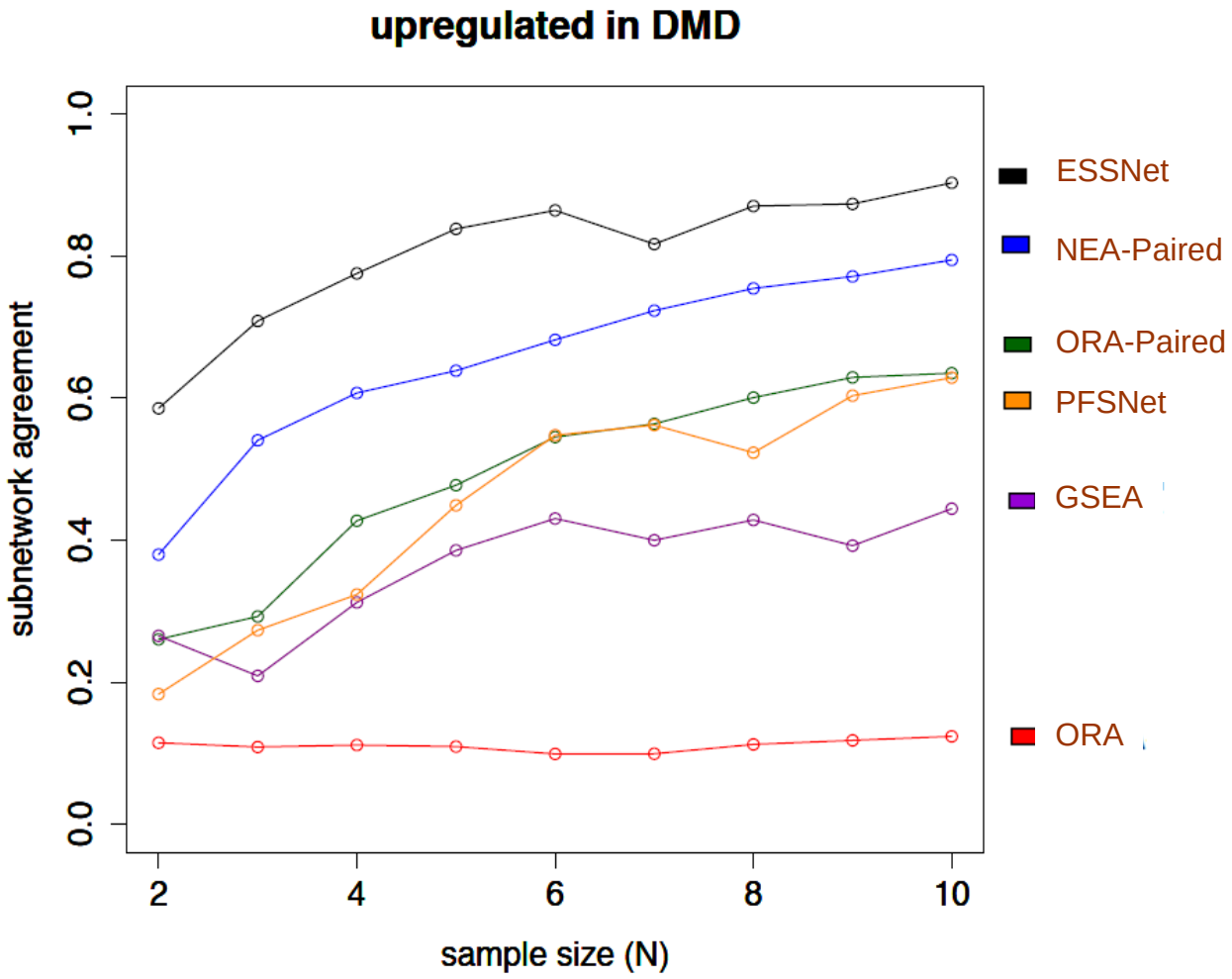
ESSNet: Larger subnetworks

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

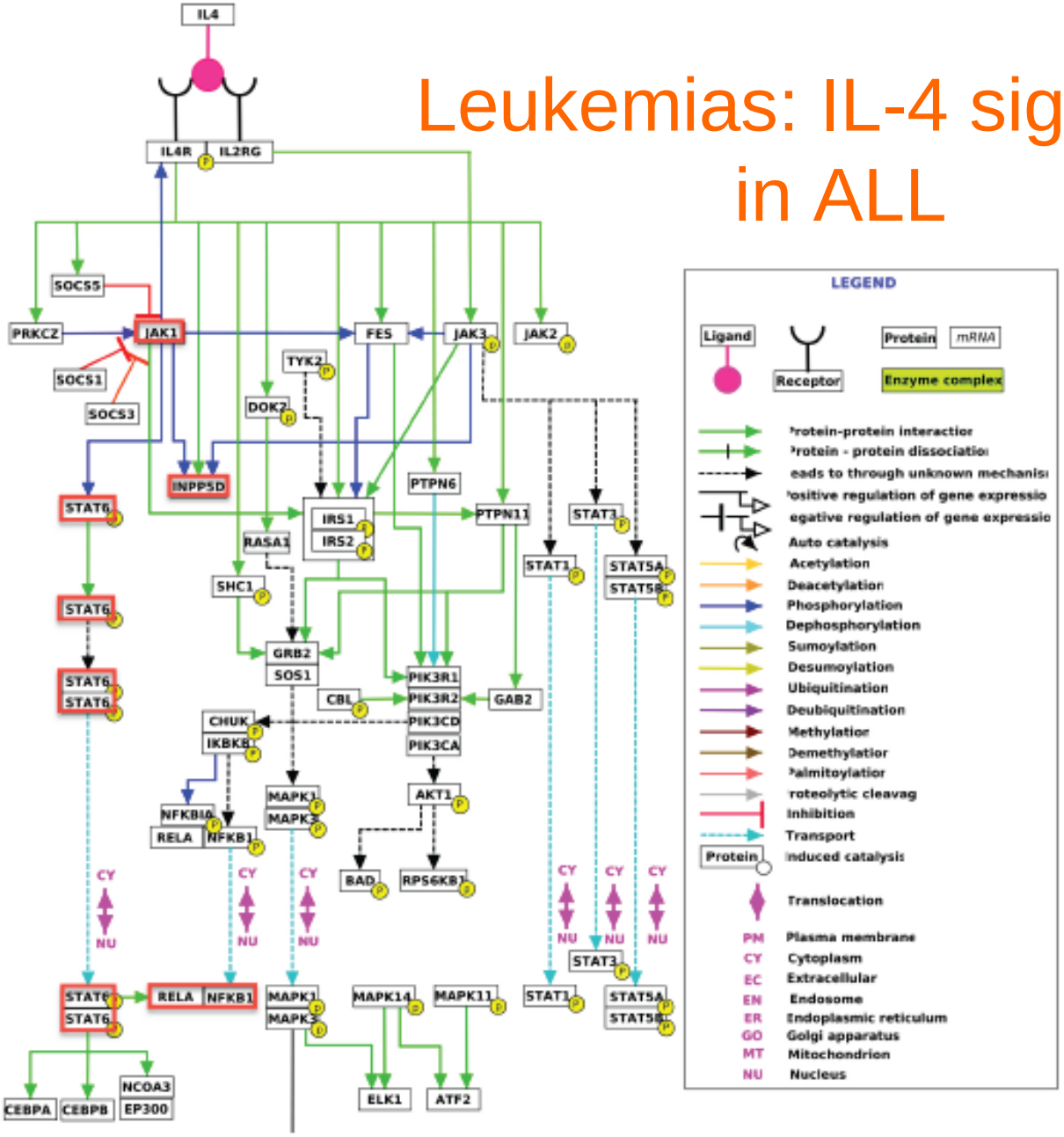


- Gene rank is very stable
- Issues #1 - #3 solved

Fantastic performance



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?

- **Mechanical application of statistical and data mining techniques often does not work**
- **Must understand statistical and data mining tools**
- **Must understand the problem domain**
- **Must know how to logically exploit both**