Genomic assays: Fighting the odds of being wrong

Maria “Ken” Figueroa, MD

2-23-10

The ‘Omics’ era

- ✓ 1995 – H. influenzae 1st cellular organism sequenced
- ✓ 1996 - 1st eukaryotic genome sequenced (S. cerevisiae)
- ✓ 1998 – 1st multicellular organism sequenced (C. elegans)
- ✓ 2001 – Human genome sequenced

The ‘Omics’ era

Functional genomics

Microarrays

Epigenomics

Next-Gen sequencing

GWAS studies

Proteomics

Pathway analysis

Motif discovery

The ‘Omics’ era

High-throughput technologies allow for us to simultaneously query tens of thousands (even millions) of targets

- Increased the amount of biology captured by one experiment

- Significant amount of noise

- Pose specific statistical problems

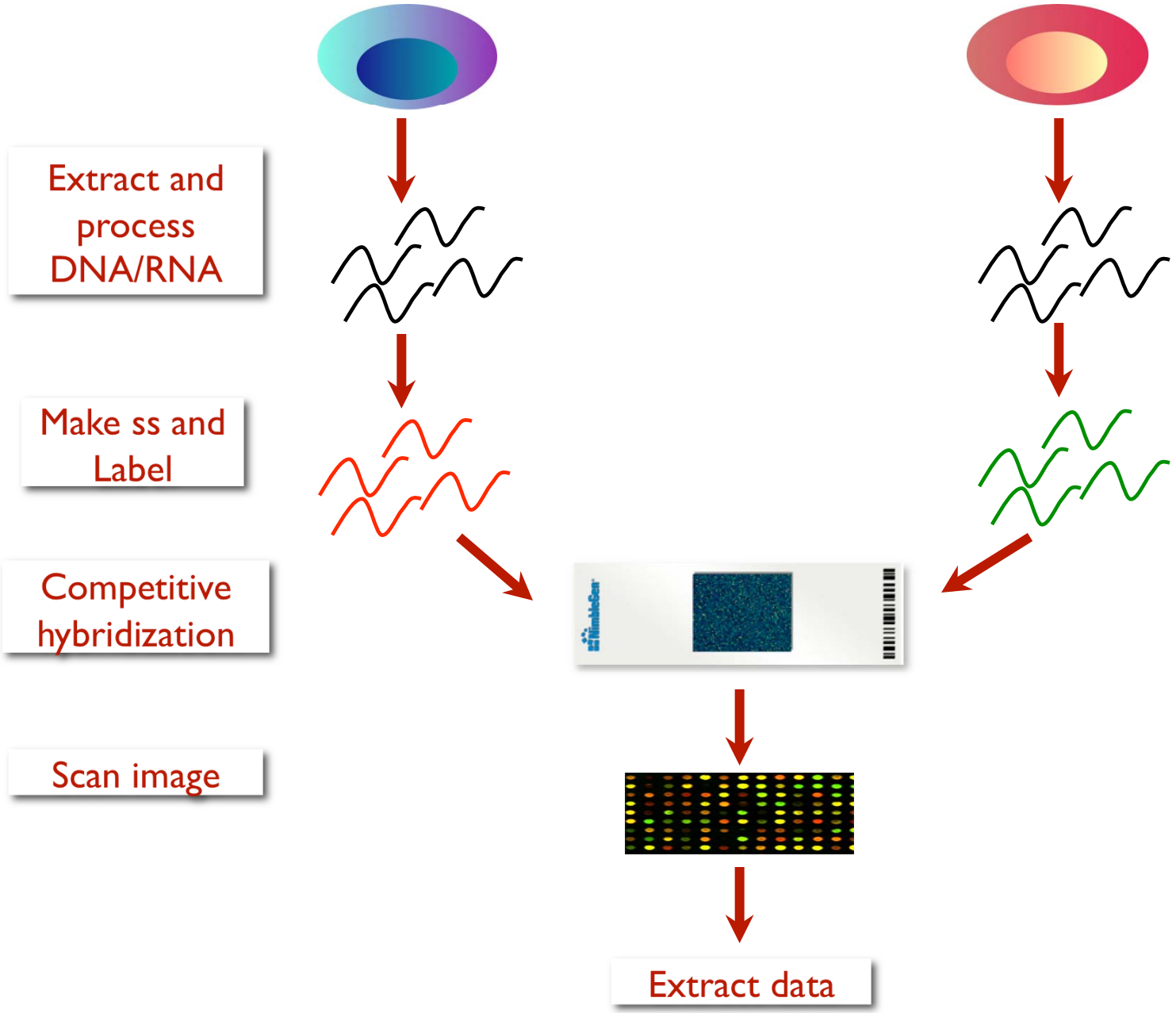
Basic concepts on microarray technology

- Collection of *known* ssDNA probes arrayed on a solid surface by covalent attachment to a chemically suitable matrix

- Quantitative and qualitative measurements of nucleic acids

- Rely on the ability of nucleic acids to hybridize to the DNA probes through base pair recognition under specific experimental conditions

Microarray-based experiments: General design



Different type of Biological platforms

- ✓ Gene Expression: Changes in gene expression levels
- ✓ Array-based Comparative Genomic Hybridization(aCGH): DNA copy number variations
- ✓ ChIP-on-chip/ChIP-seq: Genomic localization of DNA-Protein interactions
- ✓ DNA Methylation: Localization of 5-methyl-Cy
- ✓ Genotyping: sequence variants

Some statistical considerations

- Variables far exceed number of samples

- e.g.: Test clinical response to a new drug for treatment of high blood pressure on 200 pts.

vs. Identify gene expression changes associated with the same drug in 200 pts.

- Multiple comparisons

- i.e. in order to identify genes that change in a statistically significant manner with the drug we will need to *test each of the 37,000 genes* on the array in parallel and then select the significant ones

Multiple comparisons: a practical example

	Treatment (+)					Treatment (-)				
Gene 1	0.701365258	0.847689154	0.945472154	0.644555958	0.86880259	0.553831918	0.216928593	0.973412306	0.999717081	0.030686471
Gene 2	0.019693544	0.998953774	0.79541506	0.784368111	0.786279804	0.488011858	0.109621914	0.370060164	0.699715047	0.906833389
Gene 3	0.823234225	0.009390884	0.173507875	0.86814406	0.781284479	0.084611403	0.697088945	0.592397243	0.158629413	0.387556786
Gene 4	0.831201089	0.672332684	0.709812715	0.614309625	0.058084282	0.057314605	0.036616132	0.515439251	0.824838113	0.902083252
Gene 5	0.618048089	0.493722217	0.582979716	0.909020223	0.089930431	0.435987475	0.300954006	0.401800668	0.36287023	0.721856109
Gene 6	0.314244277	0.693208332	0.507662222	0.910433429	0.642351972	0.650730411	0.694156972	0.952770501	0.165252532	0.503087392
Gene 7	0.834701125	0.975953907	0.538782775	0.544151697	0.431703426	0.40012594	0.090574576	0.778406246	0.099311443	0.59307239
Gene 8	0.632542712	0.320787292	0.573479184	0.600636977	0.280344436	0.840668539	0.953859038	0.93067047	0.183795382	0.638818057
Gene 9	0.613812632	0.943127333	0.789148665	0.740696336	0.756161519	0.225290514	0.998161929	0.192950694	0.152709112	0.672583819
Gene 10	0.326036635	0.138067146	0.613095022	0.782722541	0.055087176	0.105971326	0.89495784	0.619088186	0.798195475	0.416937562
Gene 11	0.634973714	0.556111533	0.843606126	0.770987963	0.243204132	0.625448193	0.774528794	0.350605578	0.36276179	0.835054279
Gene 12	0.965398561	0.057168922	0.567125297	0.763013231	0.413766749	0.327217012	0.311494135	0.134875146	0.517469133	0.95852006
Gene 13	0.12216374	0.433638925	0.669994608	0.92908475	0.946953019	0.204031316	0.656656377	0.009321932	0.637010051	0.141680378
Gene 14	0.414223175	0.383942752	0.682146127	0.918495607	0.382467827	0.782112064	0.333122917	0.143586717	0.898119274	0.557894875
Gene 15	0.285974499	0.155930996	0.330072963	0.383671395	0.716907409	0.864141357	0.490873804	0.781127292	0.92330326	0.021729016
Gene 16	0.672888773	0.772635752	0.674517227	0.765489034	0.713345501	0.317341191	0.415206224	0.385831293	0.378462402	0.730507282
Gene 17	0.016216298	0.008760328	0.122856594	0.911411537	0.054231562	0.094487454	0.345526591	0.057715898	0.016620408	0.8738592
Gene 18	0.551922437	0.097837061	0.6162674	0.410259157	0.913703161	0.789701193	0.026344507	0.093459699	0.292196191	0.590586608
Gene 19	0.88922594	0.629840151	0.642071927	0.437341731	0.349580595	0.717605676	0.253664017	0.681060437	0.682633708	0.585084141
Gene 20	0.679047253	0.610385651	0.984636956	0.522444904	0.983714469	0.008354579	0.54121905	0.910983448	0.862391892	0.104260295

1- Gene by gene Two-tailed T test

2- Significance of $p < 0.05$

Multiple comparisons: a practical example

	Treatment (+)					Treatment (-)					P-value
Gene 1	0.701365258	0.847689154	0.945472154	0.644555958	0.86880259	0.553831918	0.216928593	0.973412306	0.999717081	0.030686471	0.258952072
Gene 2	0.019693544	0.998953774	0.79541506	0.784368111	0.786279804	0.488011858	0.109621914	0.370060164	0.699715047	0.906833389	0.477616141
Gene 3	0.823234225	0.009390884	0.173507875	0.86814406	0.781284479	0.084611403	0.697088945	0.592397243	0.158629413	0.387556786	0.517460405
Gene 4	0.831201089	0.672332684	0.709812715	0.614309625	0.058084282	0.057314605	0.036616132	0.515439251	0.824838113	0.902083252	0.641959022
Gene 5	0.618048089	0.493722217	0.582979716	0.909020223	0.089930431	0.435987475	0.300954006	0.401800668	0.36287023	0.721856109	0.550259337
Gene 6	0.314244277	0.693208332	0.507662222	0.910433429	0.642351972	0.650730411	0.694156972	0.952770501	0.165252532	0.503087392	0.903471832
Gene 7	0.834701125	0.975953907	0.538782775	0.544151697	0.431703426	0.40012594	0.090574576	0.778406246	0.099311443	0.59307239	0.14690471
Gene 8	0.632542712	0.320787292	0.573479184	0.600636977	0.280344436	0.840668539	0.953859038	0.93067047	0.183795382	0.638818057	0.194666534
Gene 9	0.613812632	0.943127333	0.789148665	0.740696336	0.756161519	0.225290514	0.998161929	0.192950694	0.152709112	0.672583819	0.104214494
Gene 10	0.326036635	0.138067146	0.613095022	0.782722541	0.055087176	0.105971326	0.89495784	0.619088186	0.798195475	0.416937562	0.379330623
Gene 11	0.634973714	0.556111533	0.843606126	0.770987963	0.243204132	0.625448193	0.774528794	0.350605578	0.36276179	0.835054279	0.893488236
Gene 12	0.965398561	0.057168922	0.567125297	0.763013231	0.413766749	0.327217012	0.311494135	0.134875146	0.517469133	0.95852006	0.634666711
Gene 13	0.12216374	0.433638925	0.669994608	0.929084475	0.946953019	0.204031316	0.656656377	0.009321932	0.637010051	0.141680378	0.194537816
Gene 14	0.414223175	0.383942752	0.682146127	0.918495607	0.382467827	0.782112064	0.333122917	0.143586717	0.898119274	0.557894875	0.941420469
Gene 15	0.285974499	0.155930996	0.330072963	0.383671395	0.716907409	0.864141357	0.490873804	0.78127292	0.92330326	0.021729016	0.240506468
Gene 16	0.672888773	0.772635752	0.674517227	0.765489034	0.713345501	0.317341191	0.415206224	0.385831293	0.378462402	0.730507282	0.00693229
Gene 17	0.016216296	0.008760328	0.122856594	0.911411537	0.054231562	0.094487454	0.345526591	0.057715898	0.016620408	0.8738592	0.821530697
Gene 18	0.551922437	0.097837061	0.6162674	0.410259157	0.913703161	0.789701193	0.026344507	0.093459699	0.292196191	0.590586608	0.44261104
Gene 19	0.88922594	0.629840151	0.642071927	0.437341731	0.349580595	0.717605676	0.253664017	0.681060437	0.682633708	0.585084141	0.965814376
Gene 20	0.679047253	0.610385651	0.984636956	0.522444904	0.983714469	0.008354579	0.54121905	0.910983448	0.862391892	0.104260295	0.23427917

Conclusion: Gene 16 is upregulated with the treatment

But… let's review a few things

✓ $p < 0.05$: This means we accept the risk of erroneously rejecting the null hypothesis in 5% of the cases i.e. we are willing to accept 5% false positive calls.

✓ In our example we did not do *1 comparison* (treated vs. untreated), we in fact did *20 comparisons in parallel*.

✓ Each time we had a 5% error, so if we repeat the test 20 times we are likely to get at least 1 false positive.

✓ Gene 16 may or may not change its expression level with the treatment, but we do not have enough evidence to claim that it does.

Our "example data set" was in fact generated with a random number generator.

Probabilities of 1 or more false positives by chance

If we set p-value at < 0.05

# genes tested (N)	False positives incidence	**Probability** of calling 1 or more false + by chance
1	1/20	5%
2	1/10	10%
20	1	64%
100	5	99.4%

$$1-(1-0.05^{N})$$

And on a genomics scale...

⇒ Suppose no genes really changed (e.g. in random samples from the same population)

⇒ ~10,000 genes on an array

⇒ Each gene has a 5% chance of exceeding the threshold at a p-value of 0.05 (Type I error)

⇒ So by chance alone…

- the p-values for 500 genes should be significant!!

Corrections for multiple comparisons

- Most approaches for correcting for multiple comparisons work well for small number of parallel comparisons

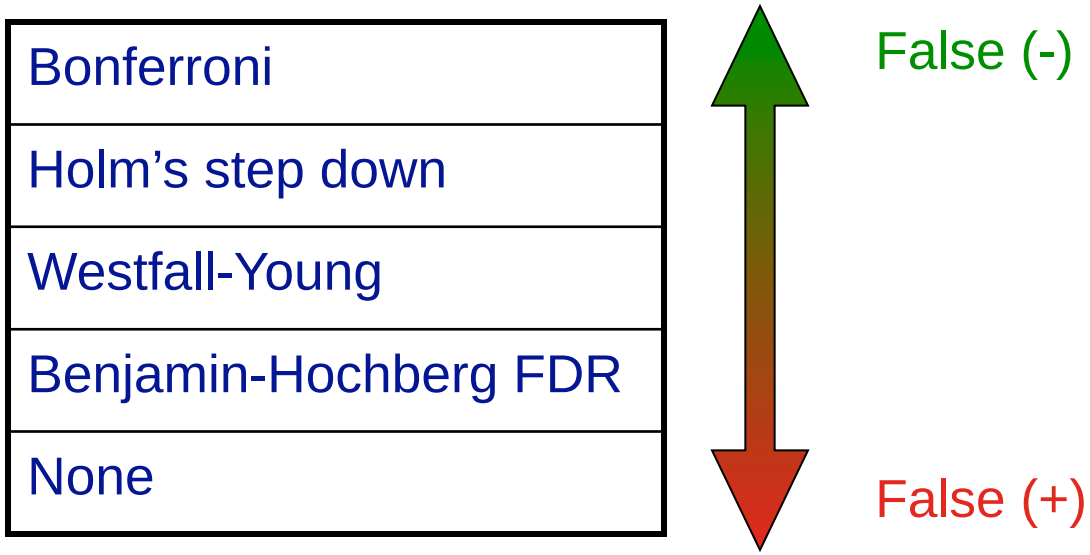
- But when tens of thousands of tests are performed most of these are too stringent (e.g. Bonferroni, Sidak, Holm's)

- The most accepted methods for multiple testing correction in the microarray field are:

    - the False Discovery Rate (FDR) determination (Benjamini-Hochberg)

    - the use of permutations (Westfall-Young, SAM)

The Sensitivity vs. Specificity trade off



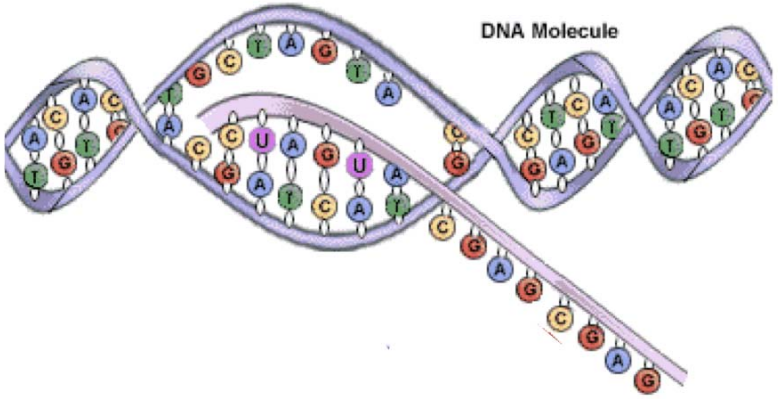
Cancer Cell
Article

Cell PRESS

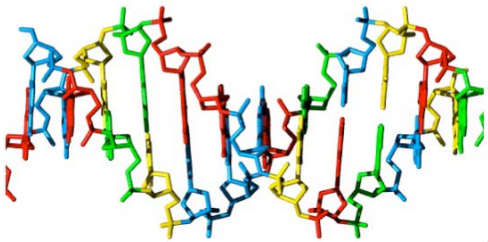
DNA Methylation Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia

Maria E. Figueroa,[1] Sanne Lugthart,[5] Yushan Li,[1] Claudia Erpelinck-Verschueren,[5] Xutao Deng,[2] Paul J. Christos,[3] Elizabeth Schifano,[7] James Booth,[7] Wim van Putten,[6] Lucy Skrabanek,[2,4] Fabien Campagne,[2,4] Madhu Mazumdar,[3] John M. Greally,[8] Peter J.M. Valk,[5] Bob Löwenberg,[5] Ruud Delwel,[5,*] and Ari Melnick[1,*]

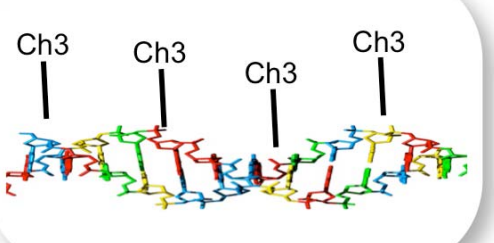
Genetic expression depends on multiple factors



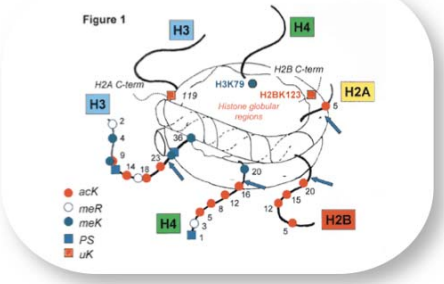
Gene Transcription



DNA Sequence



DNA Methylation



Histone Modifications

GENETIC REGULATION

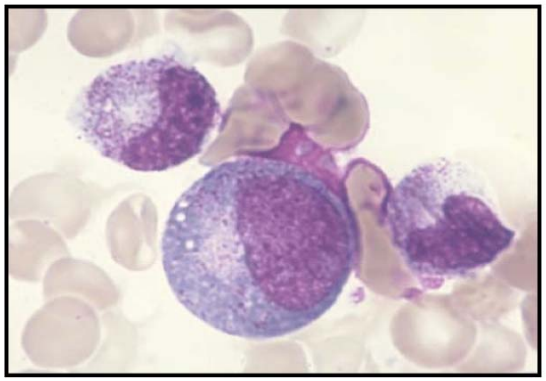
EPIGENETIC REGULATION

Gene expression profiling has limitations

- Gives only a snapshot of genes transcribed at the time, with no information on their availability for transcription.
- Does not detect epigenetic/copy number changes
- Only genes with high expression levels stand out above the noise level
- Sometimes biologically significant changes are lost within the noise signal

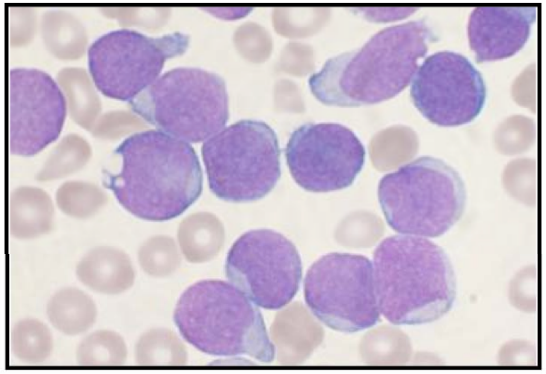
Aberrant DNA methylation is a hallmark of cancer

Normal



- Specific distribution of cytosine methylation
- Promoter CpG island hypomethylation
- Methylation of repetitive elements

Cancer



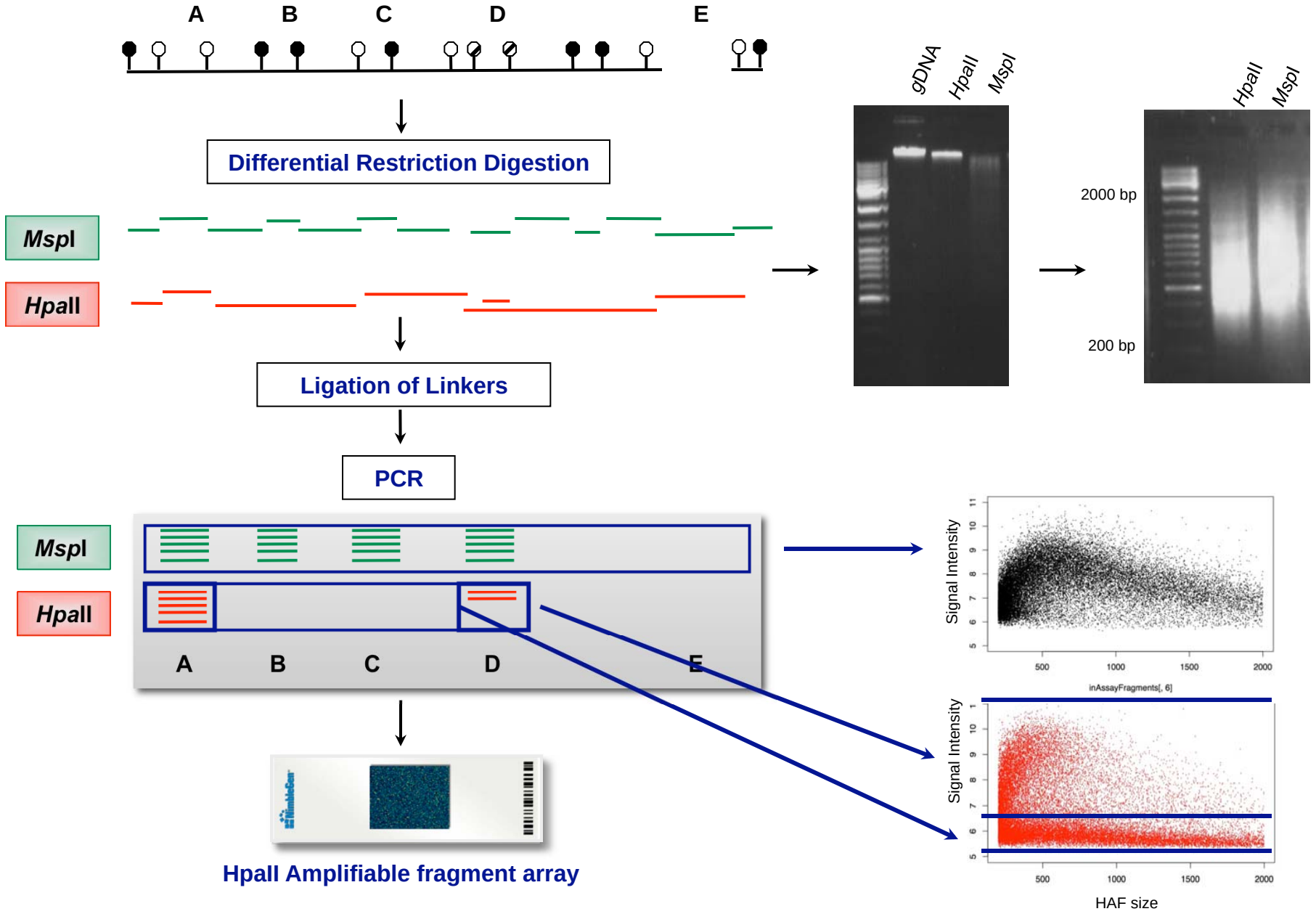
- Global hypomethylation
- Promoter CpG island hypermethylation
- Aberrant silencing of certain tumor suppressors
- Aberrant hypomethylation of certain oncogenes

Hypothesis

- DNA methylation in AML is not random, but rather specific and distinct patterns of DNA methylation characterize distinct forms of the disease.

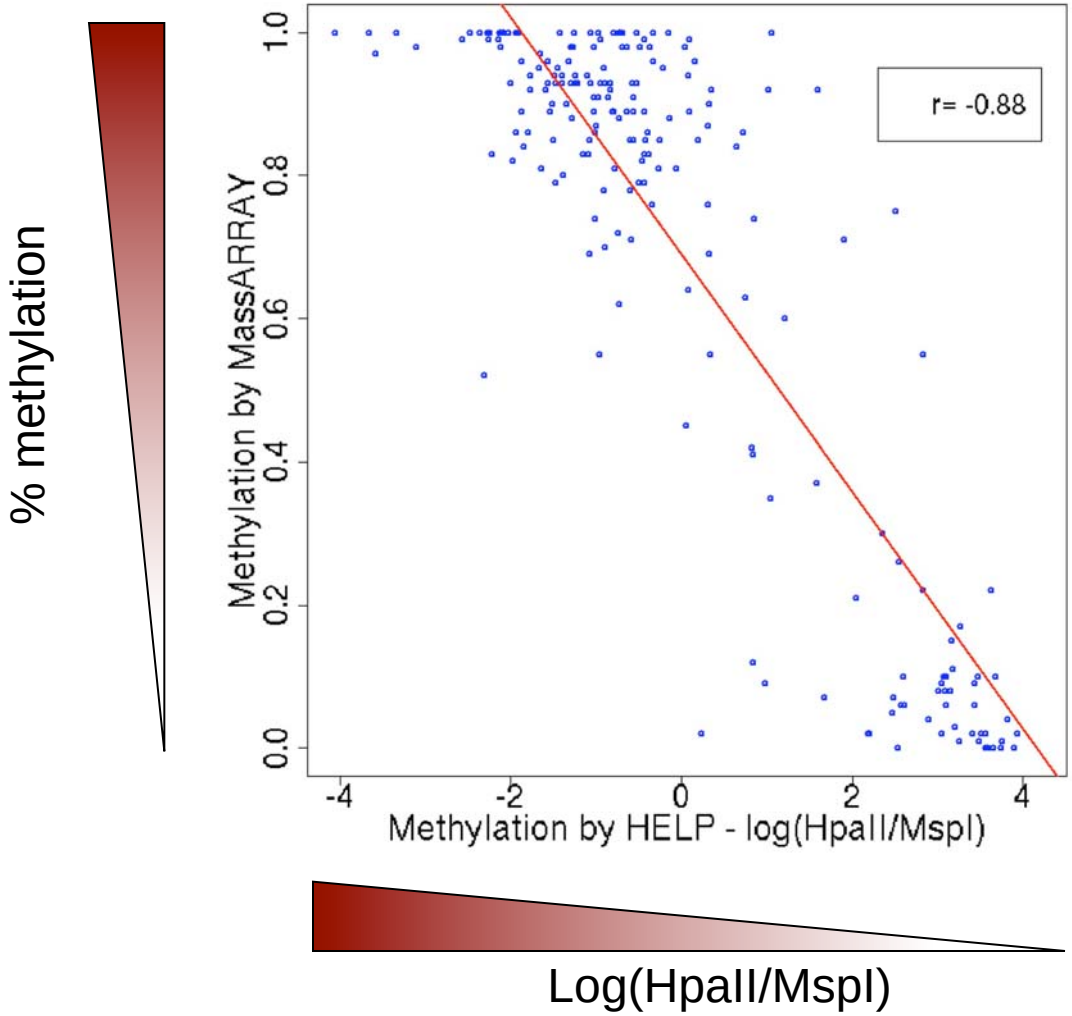
- Identifying aberrant epigenetic patterns in AML will:
  I. provide critical insight into the biological complexity of the disease
  II. help identify new and clinically relevant disease subtypes

The HELP Assay for Genome-wide 5me-Cy detection



Khulan et. al. Genome Res, 2006

Validation of HELP data by MassARRAY EpiTyper



Patients' characteristics

- ✓ 344 patients from Erasmus MC
- ✓ HOVON trials 04, 29, 32, 42 and 43
- ✓ Median follow-up: 87.4 months (0.1-214.5 m.)
- ✓ Median age: 48 years (15-77 years)
- ✓ Male: 188; Female: 156
- ✓ Molecular analysis available (cytogenetics, FISH, sequencing)
- ✓ CD34+ bone marrow cells from 8 healthy donors

Methods

Unsupervised analysis

- Hierarchical clustering to ***explore internal complexity*** of the data
- Class discovery approach

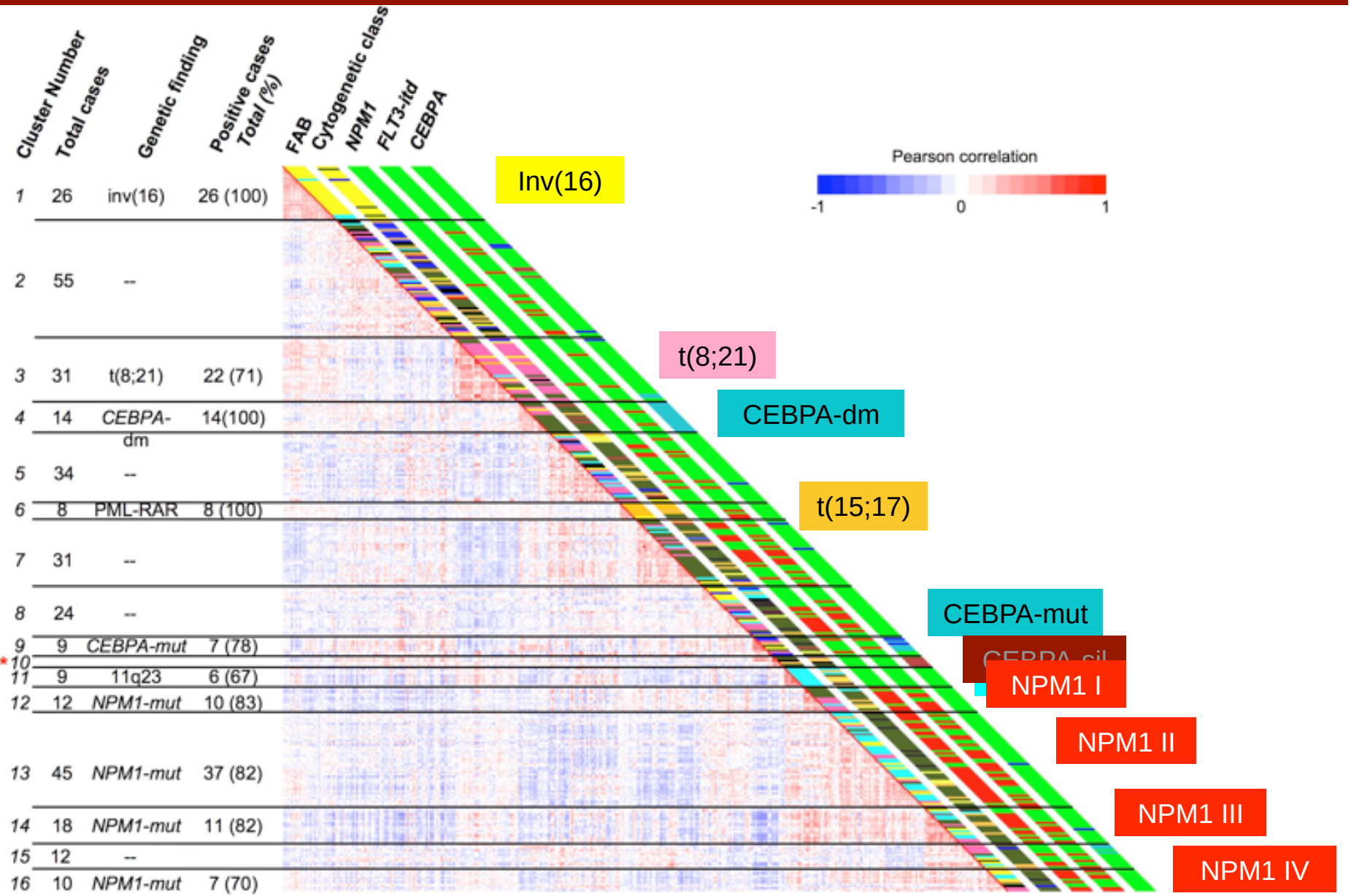
Supervised analysis

- ***Understand the biology*** associated with each methylation cluster
- Identify ***common epigenetic abnormalities***

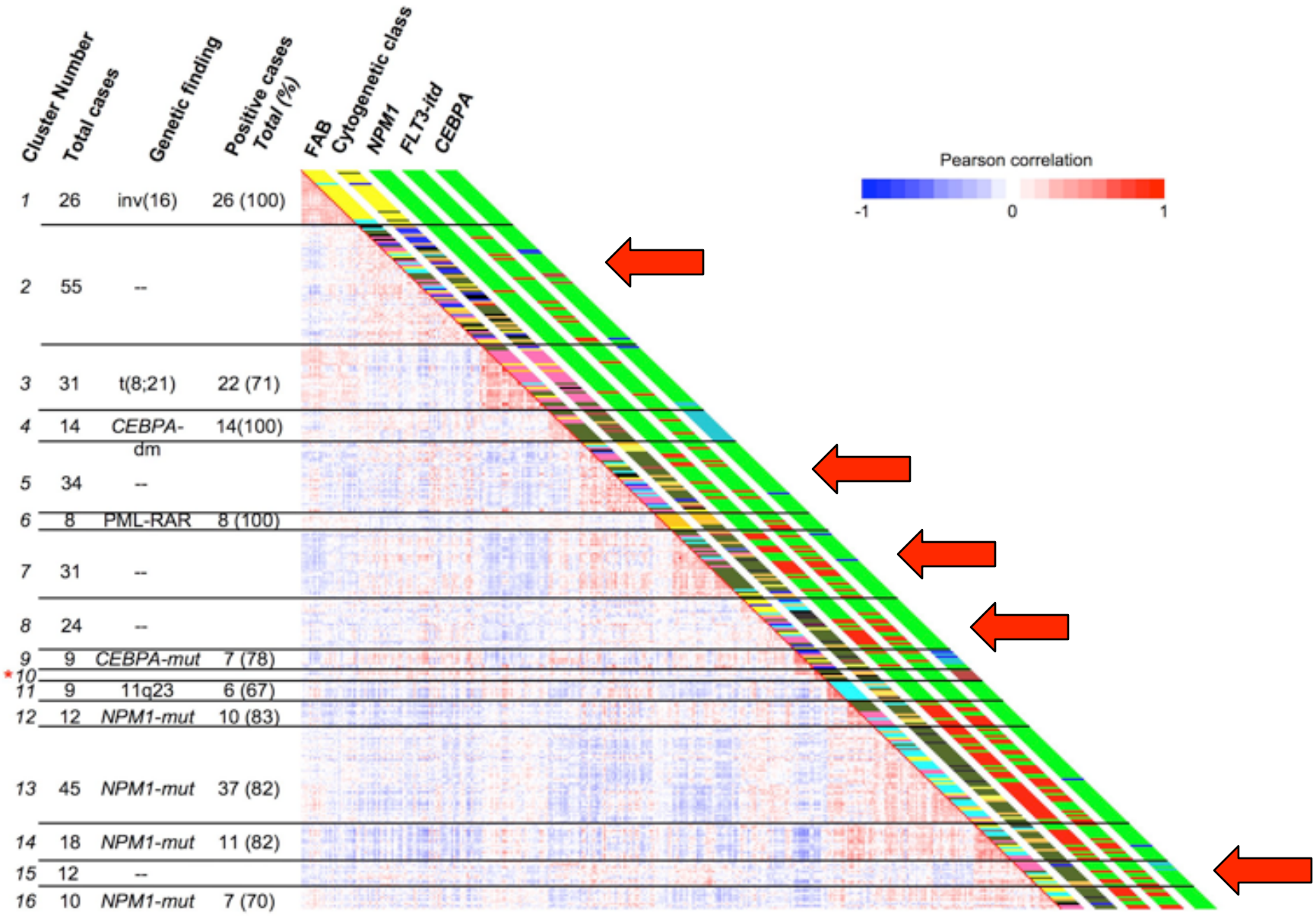
Survival analysis

- ***Determine associated risk*** with different methylation clusters

AMLs cluster into sixteen unique subtypes



DNA methylation profiling identifies five novel AML subtypes



Methods

Supervised analysis → - Understand the biology associated with each methylation cluster

Comparison of each cluster to normal CD34 + cells

↓

Identify aberrant DNA methylation signature for each cluster

↓

Pathway and Gene ontology analysis to understand associated biology

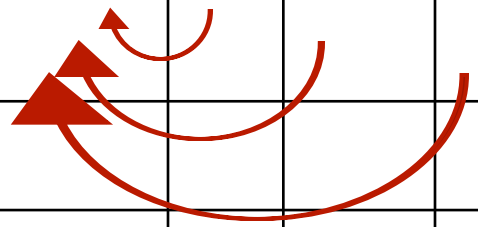
Methods

Supervised analysis → - Understand the biology associated with each methylation cluster

Multiple testing problem #2 →

Multiple testing problem #1 ↓

	K0 = Normals	K1	(K…)	K16
Gene 1				
Gene 2				
(Gene…)				
Gene 25,626				



Methods

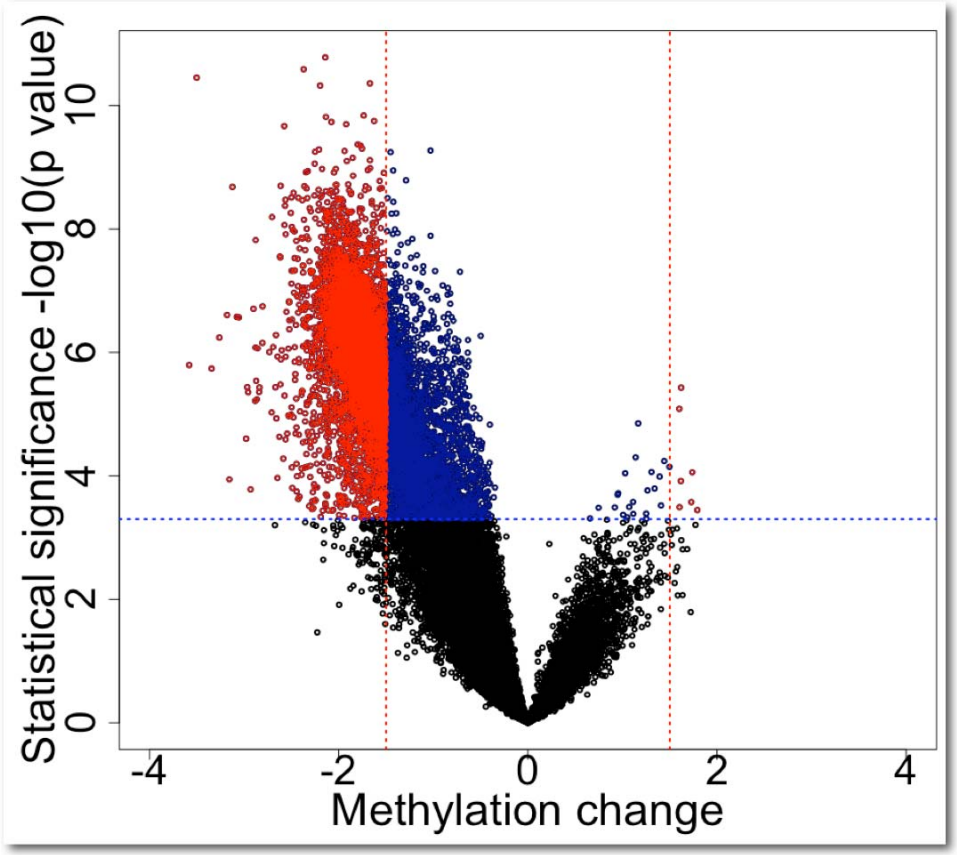
Supervised analysis → - Understand the biology associated with each methylation cluster

Dunnett's method →

ANOVA x 25,626
+
BH correction ↓

	K0 = Normals	K1	(K…)	K16
Gene 1				
Gene 2				
(gene …)				
Gene 25,626				

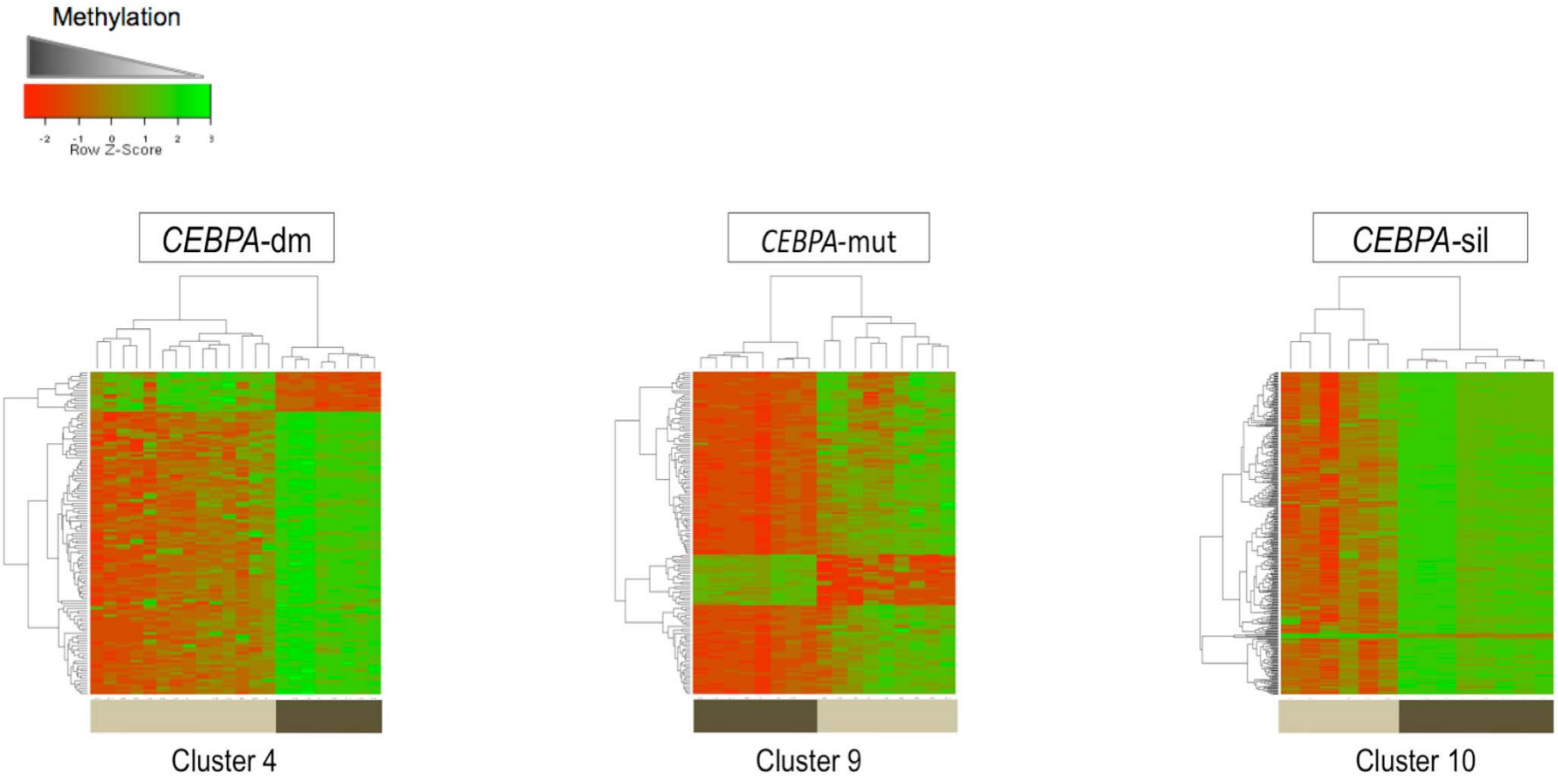
Combining statistical and biological significance



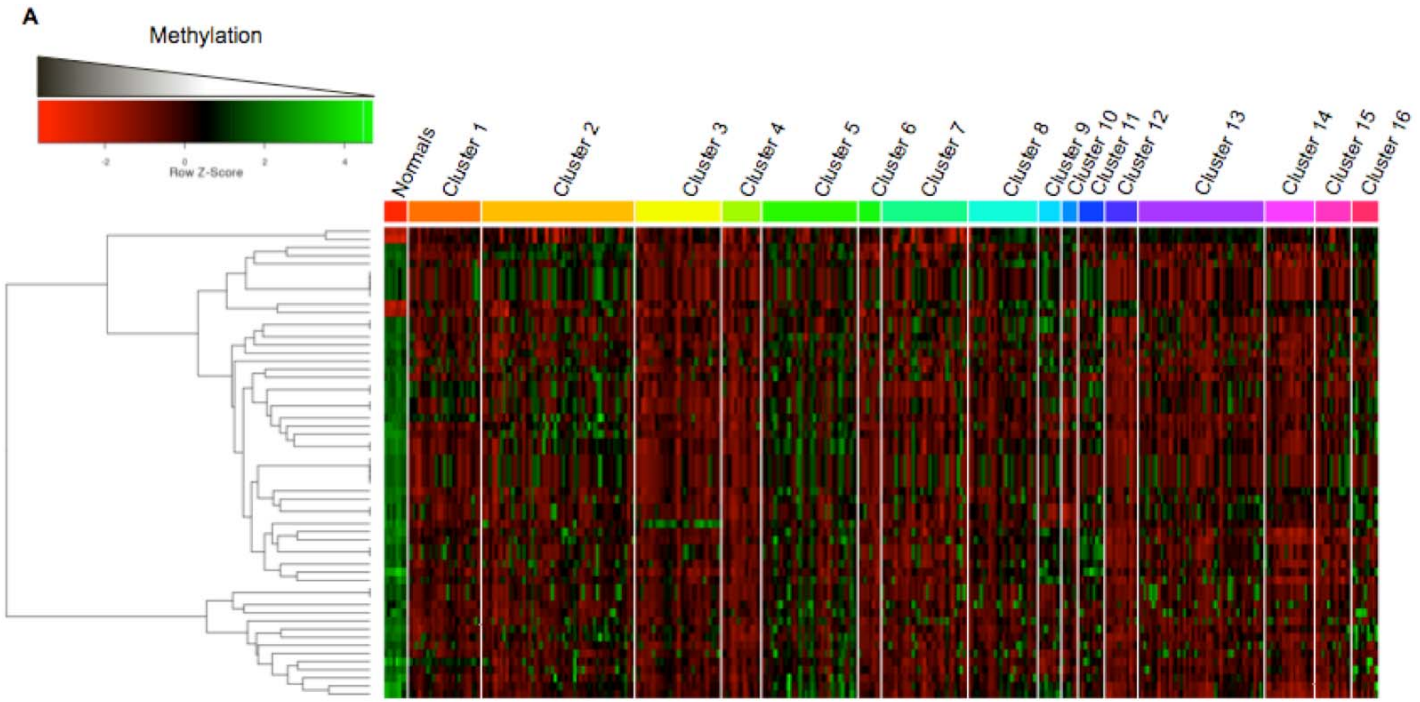
- Increases our chances of capturing biologically significant changes

- Still requires that we correct the p value for multiple testing

AML methylation profiles consist of both hyper and hypomethylation

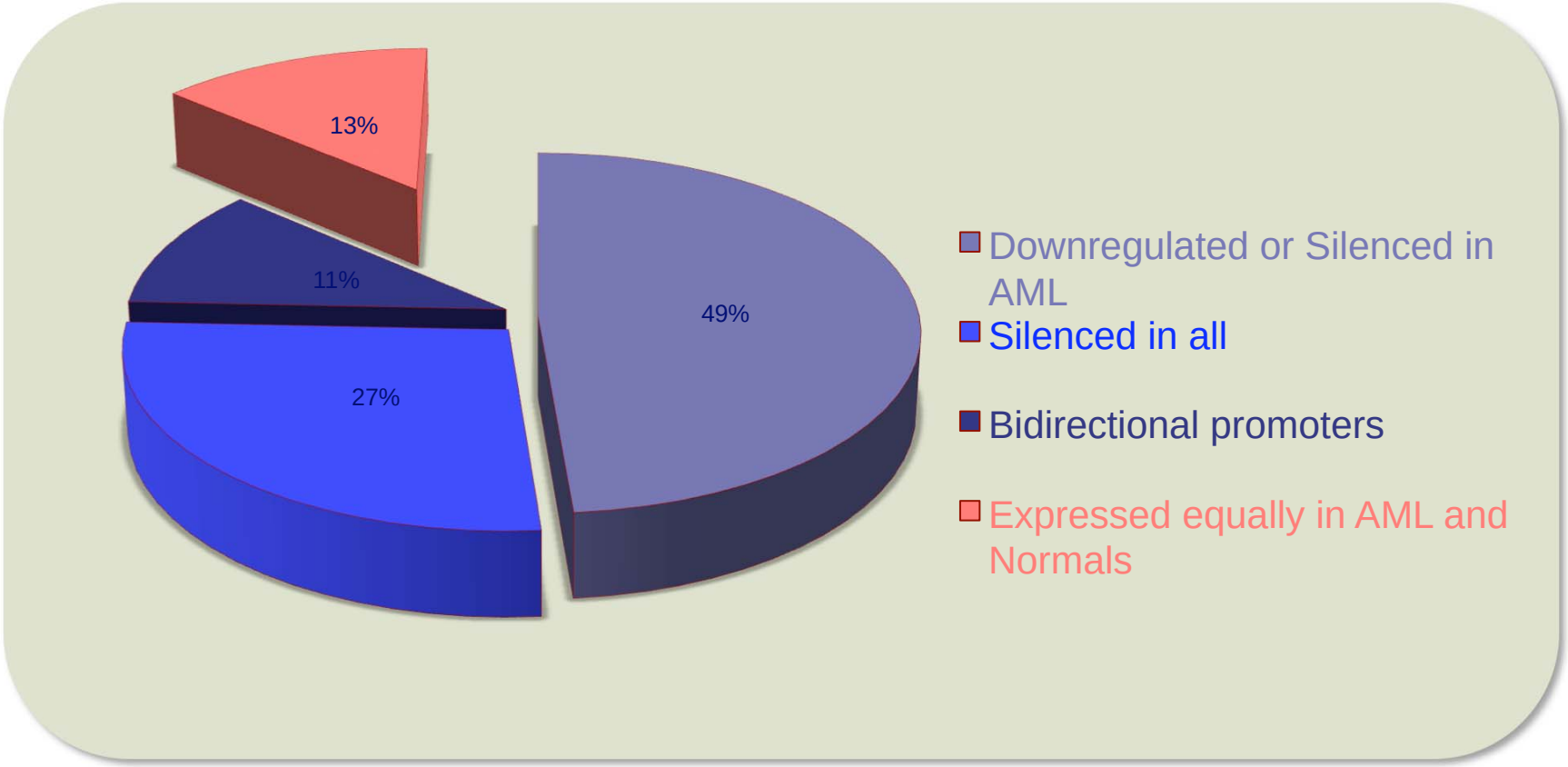


AML presents a common epigenetic signature



- ZNF proteins
- Nuclear import proteins
- Regulators of myeloid cytokines
- Members of the mediator complex
- Retinoic acid signaling
- Membrane anchor proteins
- Tumor suppressors
- Regulators of STAT signaling

AML presents a common epigenetic signature



Q-PCR in Different Patient Cohort

Unpublished data redacted

Summary

i) We demonstrated that unique and distinct DNA methylation patterns characterize distinct forms of AML

ii) identified novel, epigenetically defined subgroups of AML with distinct clinical behavior

iii) revealed the presence of a consistently aberrantly methylated signature across AML subtypes, with confirmed silencing of the genes involved

iv) report a 15-gene methylation classifier predictive of OS in an independent patient cohort, and confirmed as an independent risk factor when adjusted for known covariates.

So going back to the general issue of trying to be right against the odds…

Common concerns

“If I correct I do not get any significant genes, so I am better off not correcting”

Wrong! If you do not correct, your “*significant”* genes are probably not significant at all. This is like cheating your own self!

“My hypothesis was wrong because I do not have any significant genes after correction”

This may or may not be the case. You may just have insufficient power in your design to detect small changes. You can:

1- Increase the number of replicates/samples

2- Select a smaller number of genes to begin your analysis with (high variance genes, high SNR) and in this way the stringency of your correction will be reduced

In Summary

- High-throughput methods are very useful in biology.
- However, there is a risk for drawing the wrong conclusions if we are not careful.
- Conventional statistical approaches may not always be the most appropriate for these data sets.
- When selecting an analytical approach we need to remember the nature of the data we are analyzing (high number of correlated genes, lack of normality, etc)
- For multiple testing: B-H FDR and permutation-based methods are acceptable ways of dealing with this
- Nothing can replace experimental validations!!