

Statistical Microarray Data Analysis

A6M33BIN

cw.felk.cvut.cz/doku.php/courses/a6m33bin/start

Jiří Kléma

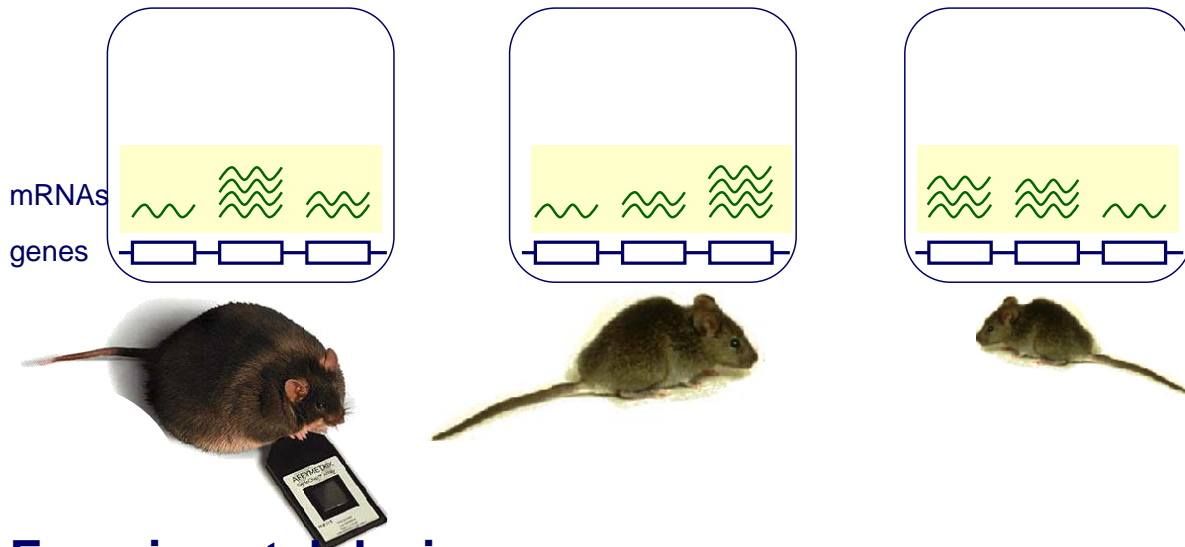
klema@labe.felk.cvut.cz

Spring 2019

Outline

- High-throughput screening
 - microarray data – origin, aims of analysis
- Hypothesis generation
 - traditional statistics vs learning patterns
- Finding differentially expressed ...
 - genes
 - often an ill-posed problem
 - gene sets
 - apriori defined,
 - Prior knowledge makes the analysis robust
- Methods (so far without annotations)
 - gene significance, clustering

Transcriptome/RNA experiments

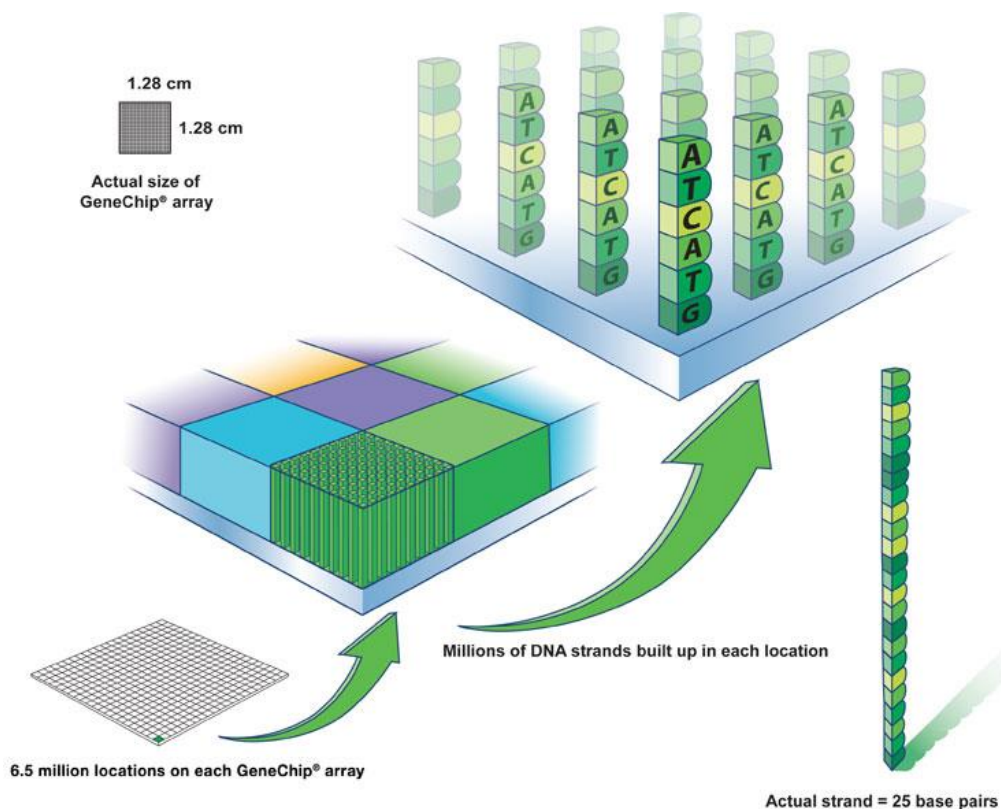


Experimental design

- *Independent variable (predictor)*: treatments, individuals, strains, cell types, environmental conditions, disease states, etc.
- *Dependent variable (response)*: RNA quantities for genes, exons or other transcribed sequences

3

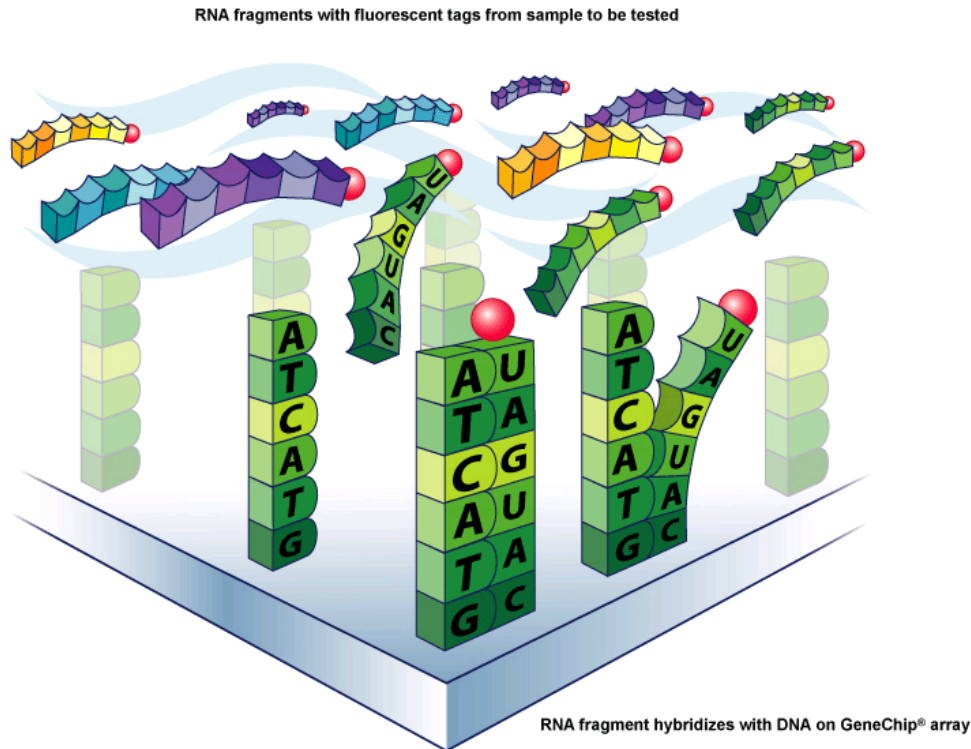
DNA microarrays (gene chips)



Courtesy of Affymetrix

4

Hybridization

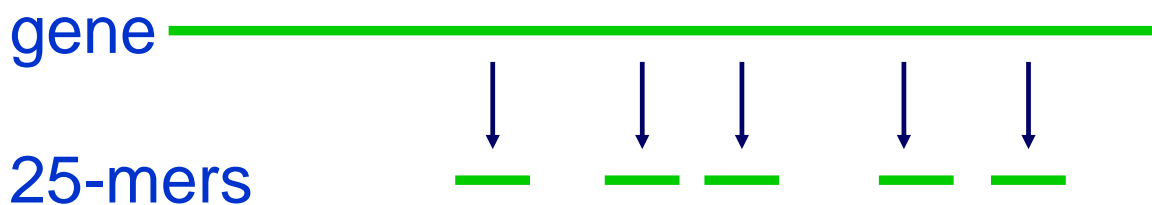


Courtesy of Affymetrix

5

Oligonucleotide arrays

- given a gene to be measured, select different n -mers for the gene



- can also select n -mers for noncoding regions of the genome
- selection criteria
 - specificity
 - hybridization properties
 - ease of manufacturing

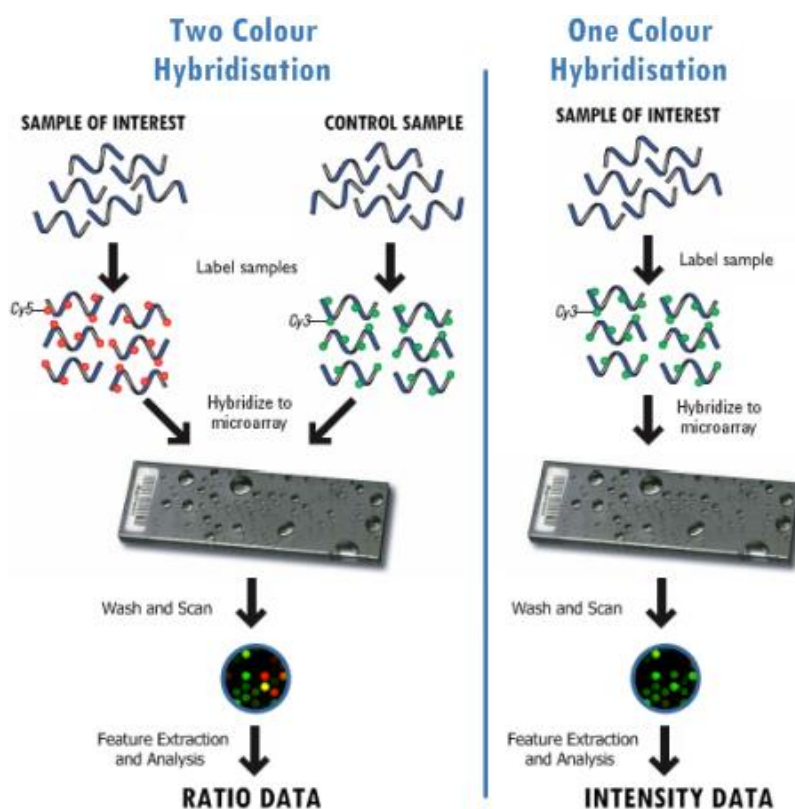
6

Microarrays



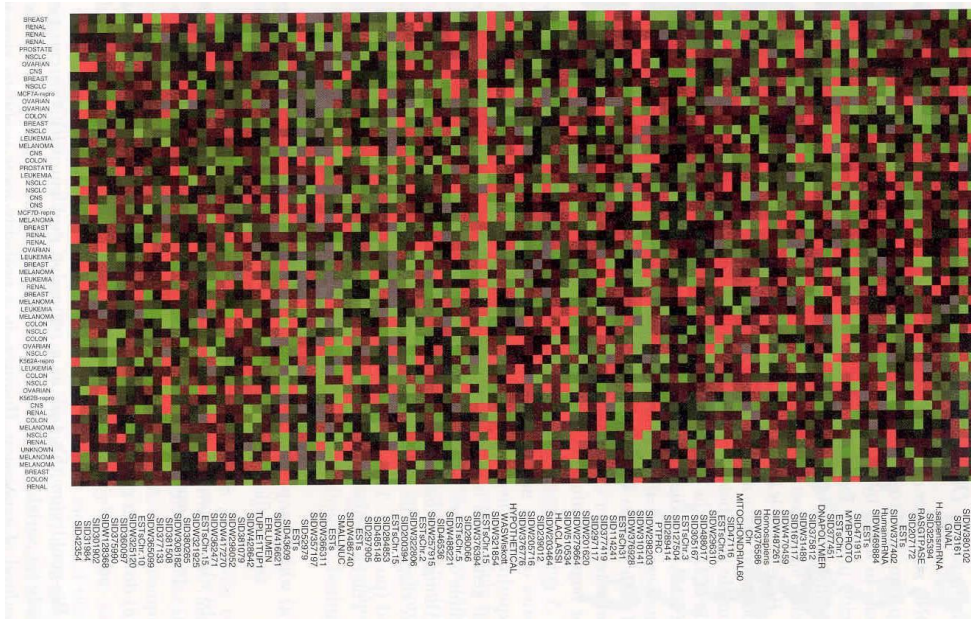
7

One-color vs two-color microarray



8

Microarray data



9

Goals of transcriptomic data analysis

- Human disease diagnostics and treatment
 - disease predispositions/risk factors
 - monitor disease stage and treatment progress
- Agricultural diagnostics and development
 - find plant pathogens to improve plant protection
 - efficiency and economy in plant biotechnology
- Analysis of food and GMOs
 - determine the integrity of food
 - detect alterations and contaminations
 - quantify GMOs
- Drug discovery and drug development

10

Other omics measurements

- RNA-sequencing: direct sequencing of RNA sequences to quantify transcript abundance
- Profiles of non-coding RNAs, including microRNAs, lncRNAs, ...
- Proteome: all proteins in a sample
- Metabolome: all metabolites (small molecules) in a sample
- Profiles of single nucleotide polymorphism (SNP) in a sample
- Epigenome: All modifications to DNA, such as DNA methylation arrays

11

Ways of MA data analysis

- **predictive modeling: molecular classifiers**
 - large potential applicability
 - but risk of low reliability and comprehensibility
 - e.g., 70% accuracy is not enough when explanation is missing
 - decision based on a large number of genes is expensive
 - SVM, RF, kNN, classification rules etc.
 - *classifying samples*: to which class a given sample belongs
 - *classifying genes*: to which functional class a given gene belongs

12

Transcriptomic data analysis

- rather simpler tasks of **descriptive modeling**
 - any genes with similar expression profiles?
 - clustering, bi-clustering
 - the genes potentially regulated together
 - any genes potentially discriminating among classes?
 - t-tests, SAM
 - potential risk factors
 - can we characterize these genes?
 - significant GO terms, pathways, locations (chromosomes)
- focus on human disease diagnostics and treatment.

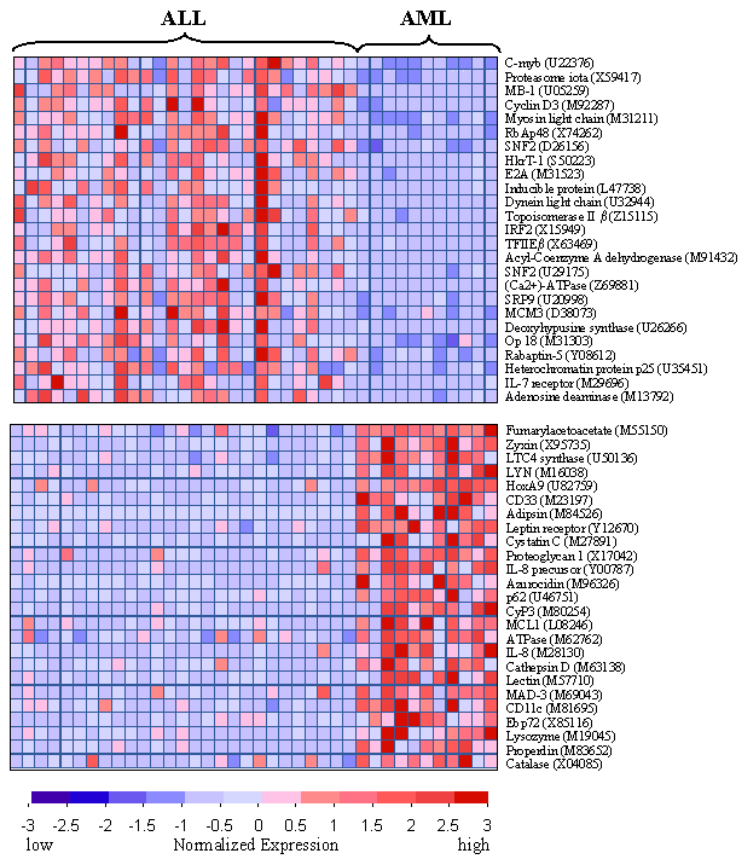
13

ALL/AML dataset

- distinguishing between two acute leukemia types
 - acute lymphoblastic leukemia (ALL)
 - largely a pediatric disease
 - acute myeloid leukemia (AML)
 - the most frequent leukemia form in adults
- first published in
 - Golub et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, pp. 531–537, 1999.
- Affymetrix HU6800 microarray chip
 - probes for 7129 genes, 72 class-labeled samples
 - 47 ALL (65%) and 25 AML (35%) samples

14

ALL/AML data analysis

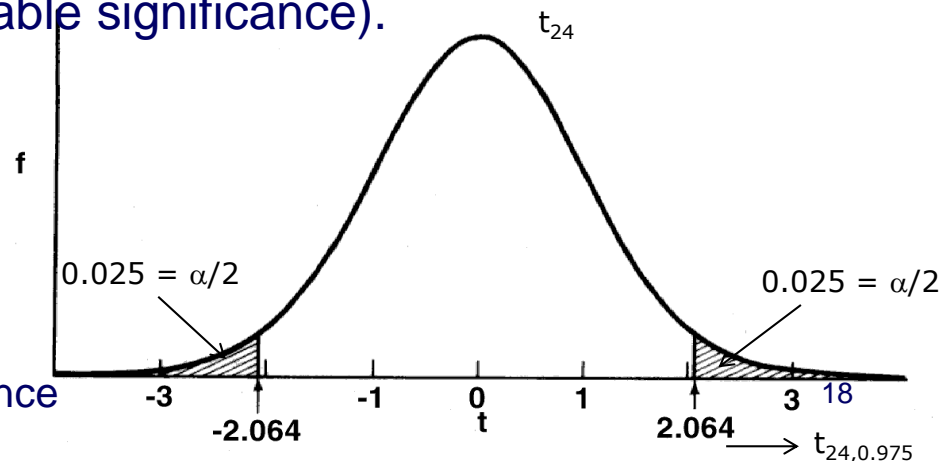


Differentially expressed genes (DEGs)

- standard t-test (or Wilcoxon test)
- for all the genes and their gene expression:
 - compute means (and standard deviation) in both groups,
 - Null hypothesis H_0 : the means are equal,
 - Alternative hypothesis H_a : the means disagree,
 - compute t, compare with T, determine p-value,
 - $p \leq \alpha$ (acceptable significance).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X1X2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

S_{X1X2} : within group variance



Significantly diff. expressed genes

- bottleneck
 - p-value = probability that a difference occurred by chance
 - $p \leq \alpha_i = 0.01$ works when evaluating a small number of genes
 - a microarray experiment for 10,000 genes may identify up to 100 significant genes by chance
- multiple comparisons
 - familywise error rate α is the probability of rejecting at least one H_0 given that all H_0 are true
 - considering k independent comparisons:
 - $\alpha = 1 - (1 - \alpha_i)^k$
 - for $\alpha_i = 0.01$:

k	1	5	10	50	100	500	1000
α	0.01	0.05	0.10	0.39	0.63	0.99	1.00

19

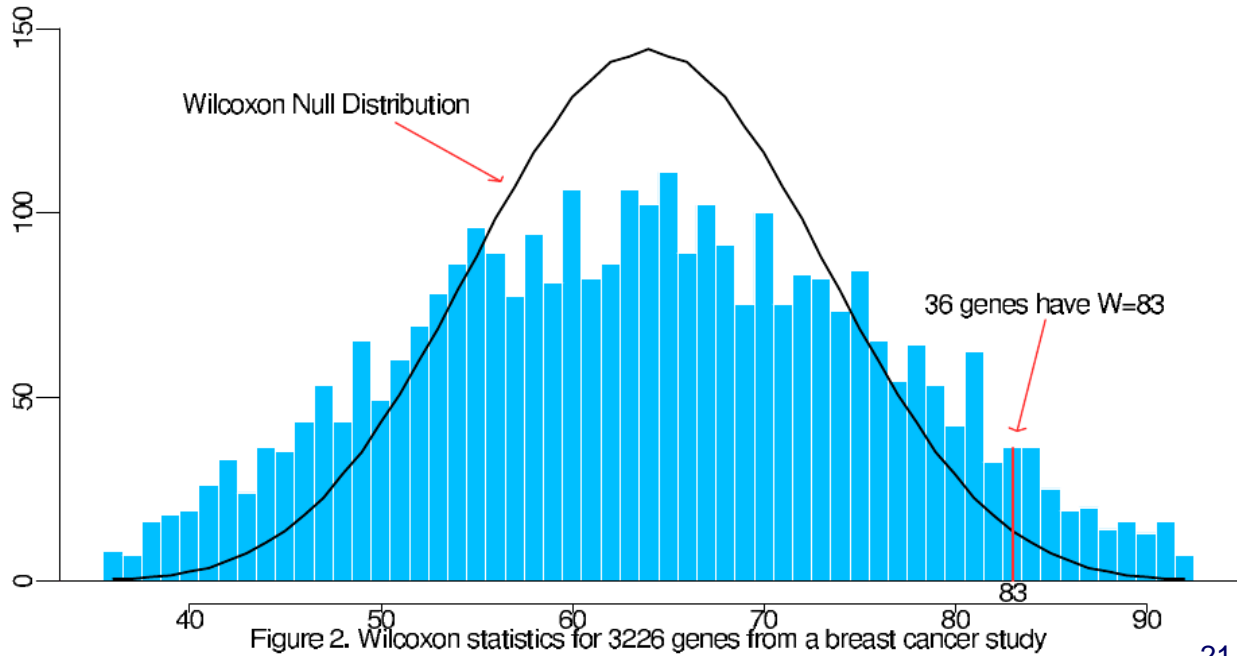
Multiple comparison strategies

- FWER – family-wise error rate
 - α value – prob that at least one comparison is FP,
- Bonferroni correction
 - the simplest (and most conservative) approach,
 - valid regardless correlation/dependence among comparisons,
 - α_i value for each comparison equals to α/k ,
 - too restrictive: 30.000 genes, $\alpha = 0.01 \rightarrow \alpha_i = 3 \cdot 10^{-7}$
- Holm–Bonferroni method
 - start by ordering the p-values in increasing order,
 - compare the smallest p-value to α/k ,
 - compare the second smallest p-value to $\alpha/(k-1)$ etc. ,
 - continue until the next hypothesis cannot be rejected,
 - stop and accept all hypotheses that have not been rejected yet,
 - step-wise method, has more power than Bonferroni.

20

Wilcoxon test for DEGs

- genetic mutations BRCA1 and BRCA2 [Hedenfalk, Efron]
- BRCA1 and BRCA2 increase breast cancer risk
- are tumors with BRCA1 or BRCA2 observed genetically different?
- 15 samples (7/8), 3226 genes studied, Wilcoxon test used



21

Significant analysis of microarrays (SAM)

- computes false detection rate (FDR)
 - permutations of the repeated measurements to estimate the percentage of genes identified by chance

relative difference in gene exp.

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

gene-specific scatter $s(i)$

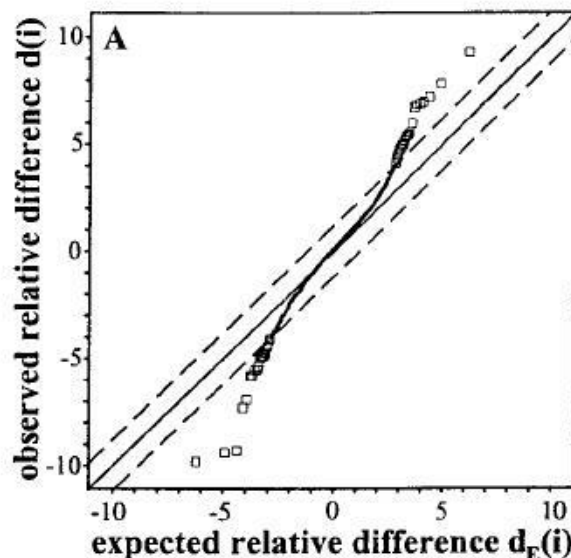
small constant s_0

t test $\sim d(i) > c, d(i) < -c$

instead compare with d_E :

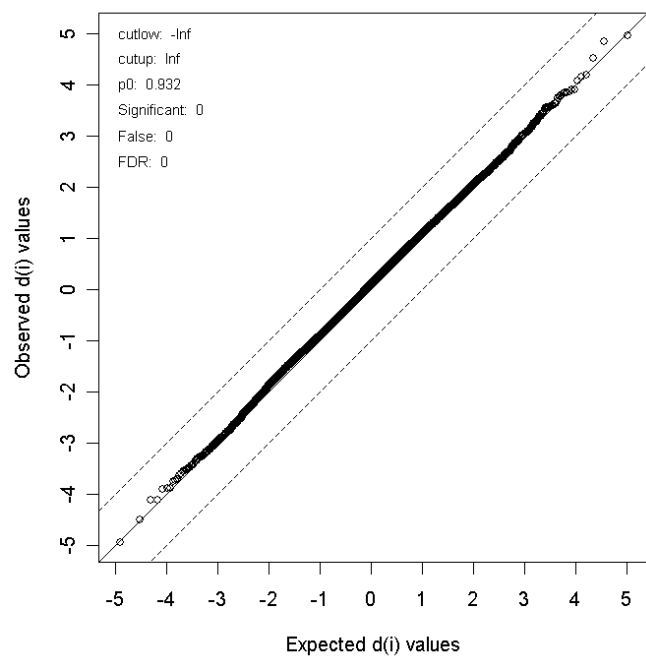
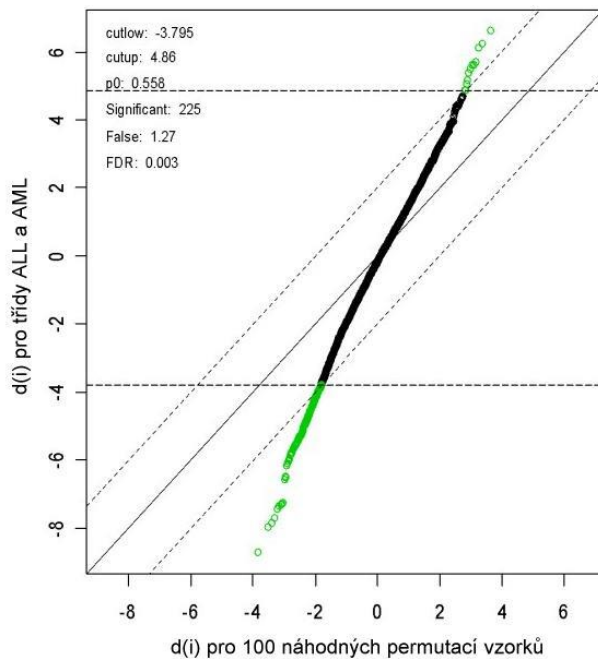
the same statistic averaged over multiple balanced random partitions

$d(i) - d_E(i) \geq \Delta$ (image $\Delta = 1.2$)



Significant analysis of microarrays (SAM)

- truly significant genes (ALL/AML)
- no significant genes found (Motol – bladder relapse)



23

Understanding of gene groups

- web tools such as David, eGOn, Ingenuine pathways
- occurrence of specific subgroups (GO terms, pathways, diseases etc.)

TERM1 - **Structural molecule activity** (Molecular function) - active in nonrelapse

Relapse group

9118, INA, Internexin neuronal intermediate filament protein, alpha

Nonrelapse group

857, CAV1, Caveolin 1, caveolae protein, 22kDa; 1278, COL1A2, Collagen, type I, alpha 2; 1281, COL3A1, Collagen, type III, alpha 1; 1289, COL5A1, Collagen, type V, alpha 1; 1292, COL6A2, Collagen, type VI, alpha 2; 1293, COL6A3, Collagen, type VI, alpha 3; 1306, COL15A1, Collagen, type XV, alpha 1; 80781, COL18A1, Collagen, type XVIII, alpha 1; 11117, EMILIN1, Elastin microfibril interfacier 1; 2192, FBLN1, Fibulin 1; 25900, HOM-TES-103, Hypothetical protein LOC25900, isoform 3; 25984, KRT23, Keratin 23 (histone deacetylase inducible); 3908, LAMA2, Laminin, alpha 2 (merosin, congenital muscular dystrophy); 4131, MAP1B, Microtubule-associated protein 1B; 4629, MYH11, Myosin, heavy chain 11, smooth muscle; 10398, MYL9, Myosin, light chain 9, regulatory; 23037, PDZD2, PDZ domain containing 2; 64711, RPS2, Ribosomal protein S2; 7148, TNXB, Tenascin XB; 7461, WBSCR1, Williams-Beuren syndrome chromosome region 1

25

Gene-set enrichment analysis

- Find differentially expressed groups of genes rather than single genes, such as
 - A gene set on a pathway
 - A gene set with a GO term
- Overview of methods [Goeman, Buhlmann, 2007]
 - competitive vs self-contained tests
 - H_0^{comp} : The genes in the set G are at most as often differentially expressed as the genes in its complement G^c .
 - H_0^{self} : No genes in G are differentially expressed.
 - gene vs subject sampling
 - gs: study distributions where gene is the basic unit
 - ss: compare the actual subject with other randomly sampled subjects

26

Competitive gene sampling

Steps:

1. Apply t-test (or other) for diff. expression of genes.
2. Apply a cut-off to separate diff. expressed genes
 - either threshold p-values ($p < \alpha$),
 - or take k genes with smallest p-values.
3. Count frequencies in 2x2 table.
4. Do a test of independence
 - Chi-squared test
$$X^2 = \sum_{g \in \{G, G^c\}} \sum_{d \in \{D, D^c\}} \frac{(m_{gd} - m_g \times m_d)^2}{m_g \times m_d} < \chi_{df=1, \alpha}^2$$
 - Hypergeometric test

	Differentially expressed gene	Non-differentially expressed gene	Total
In gene set	m_{GD}	m_{GD^c}	m_G
Not in gene set	m_{G^cD}	$m_{G^cD^c}$	m_{G^c}
Total	m_D	m_{D^c}	m

27

