Statistics for Bioinformatics

Detecting differentially expressed genes

Jacques van Helden

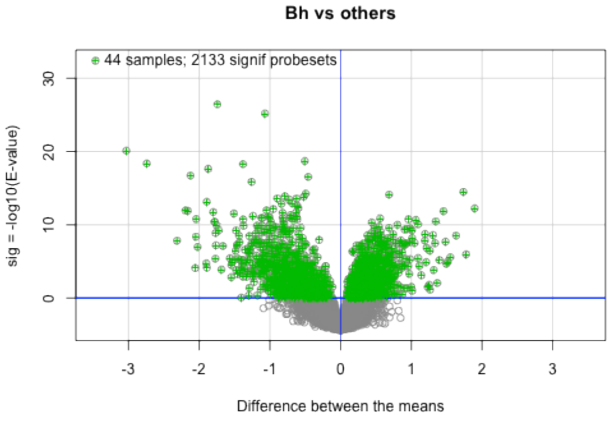
Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

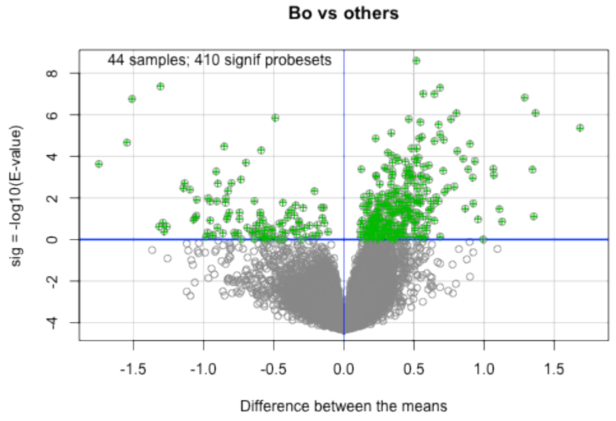
FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/

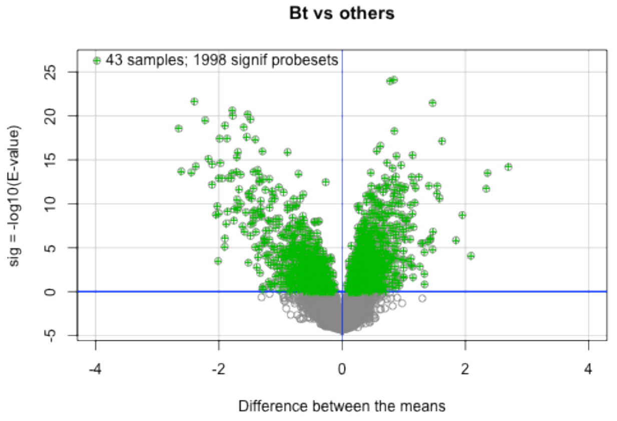
Principle of differential analysis

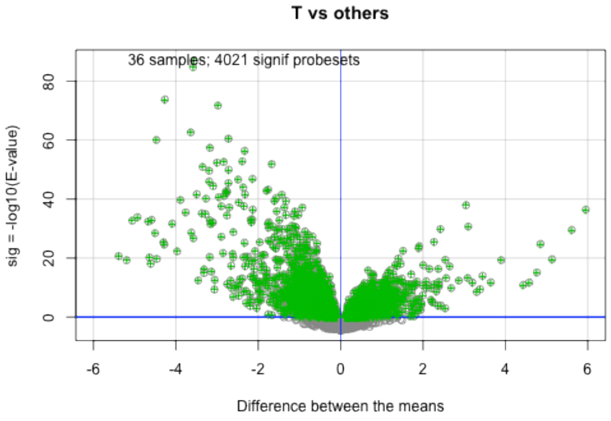
- **Two-groups differential analysis with Welch test**
  - Principle: define a group of interest ("goi", for example hyperdiploidy), and compare it to all other cancer subtypes.
  - For each gene *I,* test the null hypothesis of mean equality
    - H_0: $m_{i,goi} = m_{i,others}$
    - H_A: $m_{i,goi} <> m_{i,others}$
  - A priori, we expect that differential expression will cause a difference between group variances -> we apply Welch rather than Student test.
- **Multi-groups differential analysis with ANOVA**
  - Test the hypothesis of mean equality between all groups.
  - For each gene, analyze the variance and compare the inter-group variance with the intra-group (residual) variance.
- **Multiple testing corrections**
  - The data set from Den Boer (2009) contains 22,283 probes. We are thus challenging 22,283 times the risk of false positive (considering a gene as significant whereas it is "truly null").
  - Different methods have been proposed to control the number of false positives:
    - Bonferoni correction : decrease the significance threshold to alpha / N
    - E-value: compute the expected number of false positives: e-value = p-value * N
    - FWER: compute P(FP >= 1)
    - q-value: estimate the false discovery rate (proportion of FP among the genes declared significant).

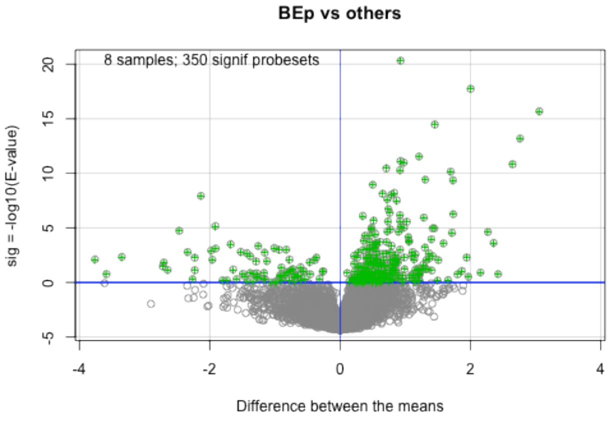
Welch test results for two-groups differential analysis

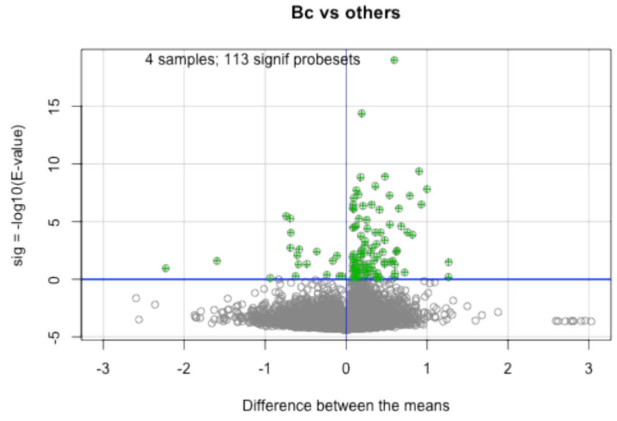


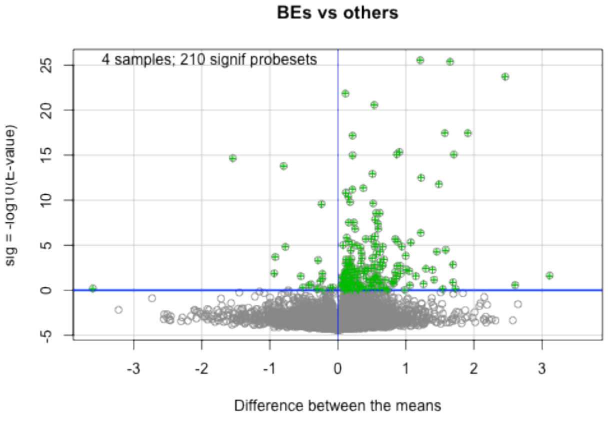


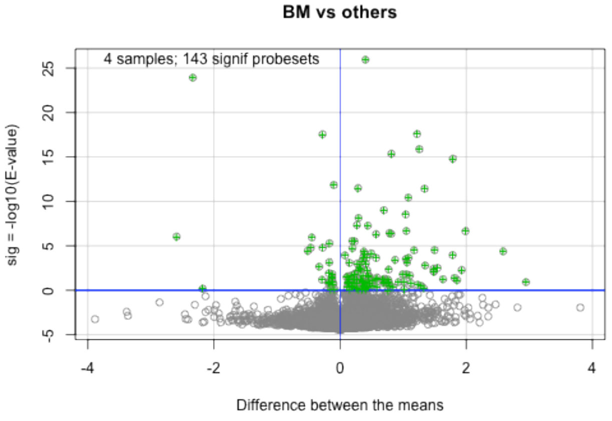








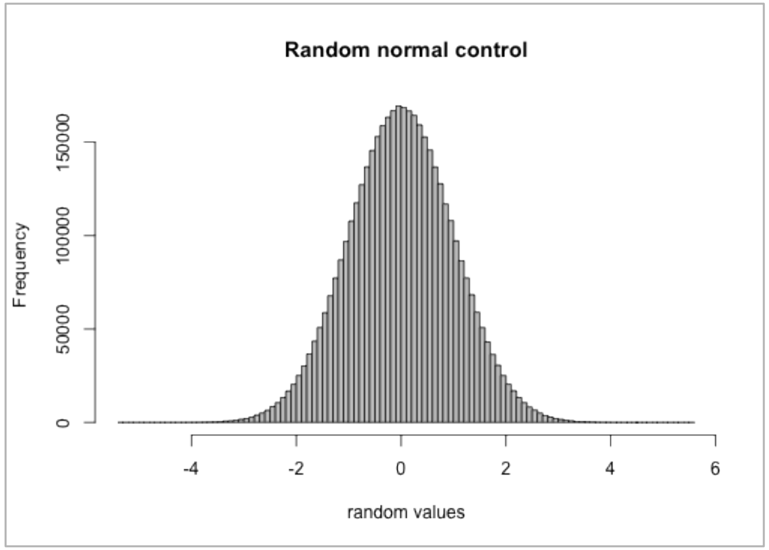


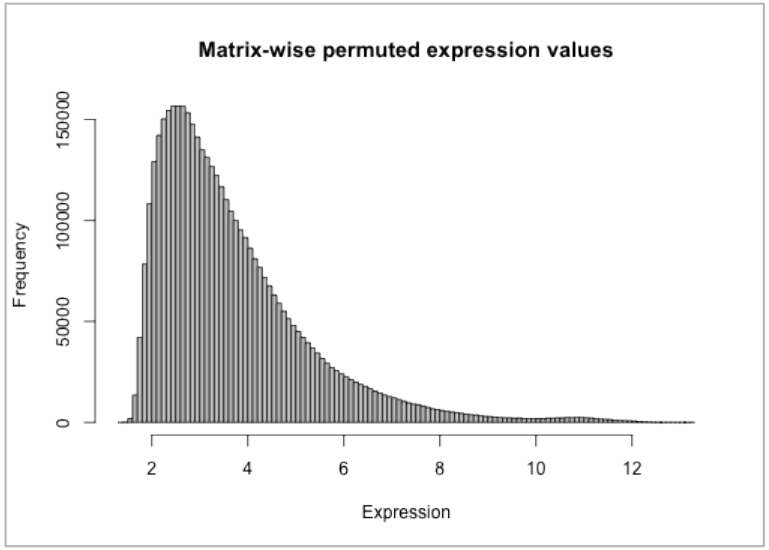


Bh	hyperdiploid	44
Bo	pre-B ALL	44
Bt	TEL-AML1	43
T	T-ALL	36
BEp	E2A-rearranged (EP)	8
Bc	BCR-ABL	4
BEs	E2A-rearranged (E-sub)	4
BM	MLL	4
Bch	BCR-ABL + hyperdiploidy	1
BE	E2A-rearranged (E)	1
Bth	TEL-AML1 + hyperdiploidy	1

Negative controls

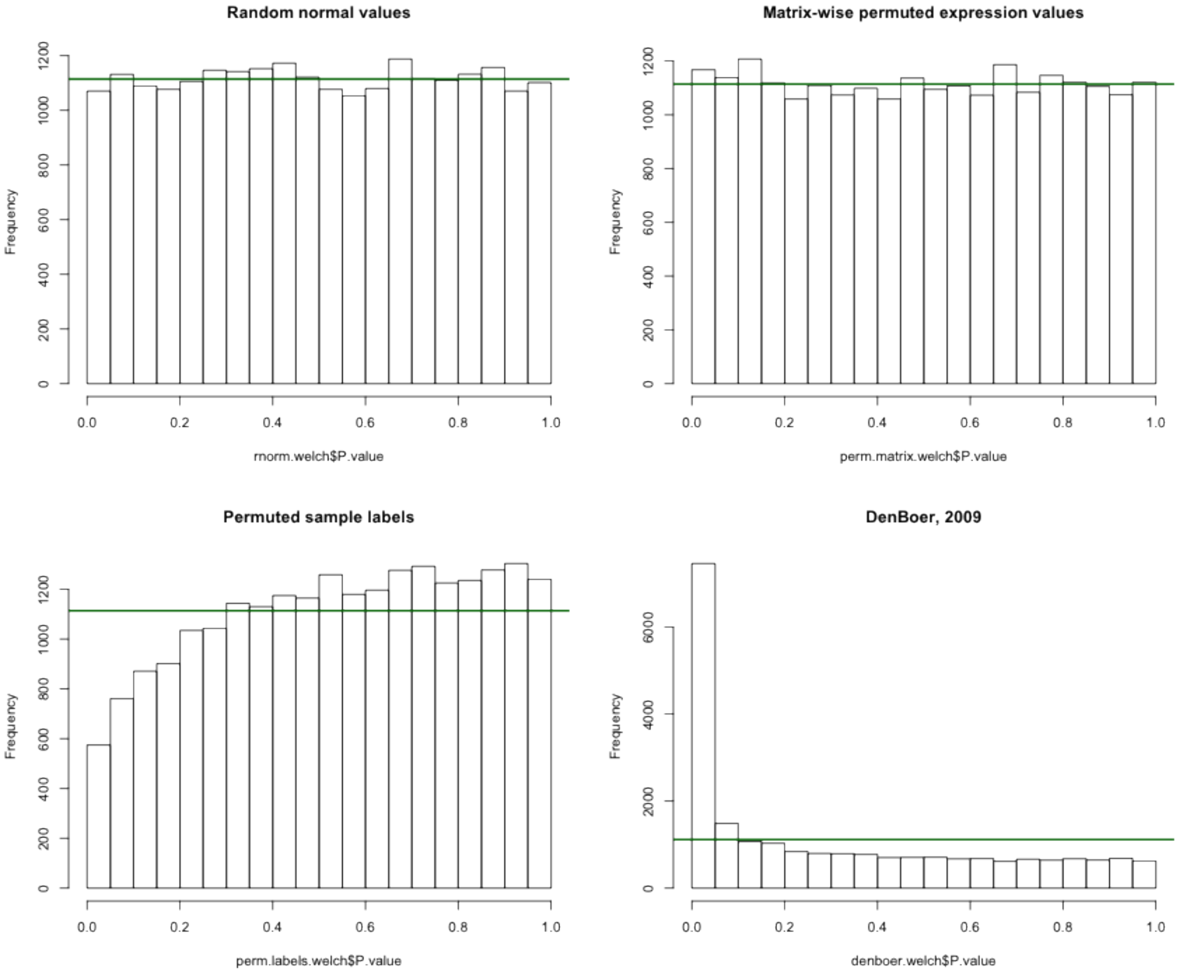
- It is always useful to check empirically the significance of a selection procedure.
- For this, we can build negative controls, i.e. datasets where no difference is expected between groups.
- 3 negative controls
  - **Random normal values**. We build a fake expression matrix by generating random numbers following a normal distribution. This perfectly fits the working hypotheses underlying statistical tests (Student, ANOVA, …) but is not a very realistic image of the biological data.
  - **Matrix-wise random permutation of expression values**. The distribution of values corresponds to the typical Affymetrix expression sets: left-skewed distribution.
  - **Permutation of sample labels**. We maintain the structure of the original expression matrix, but the sample labels are re-assigned at random. In principle, the labels are balanced between all the cancer subtypes, and there should be no significant difference between the randomized groups.





Distribution of P-values from Welch test

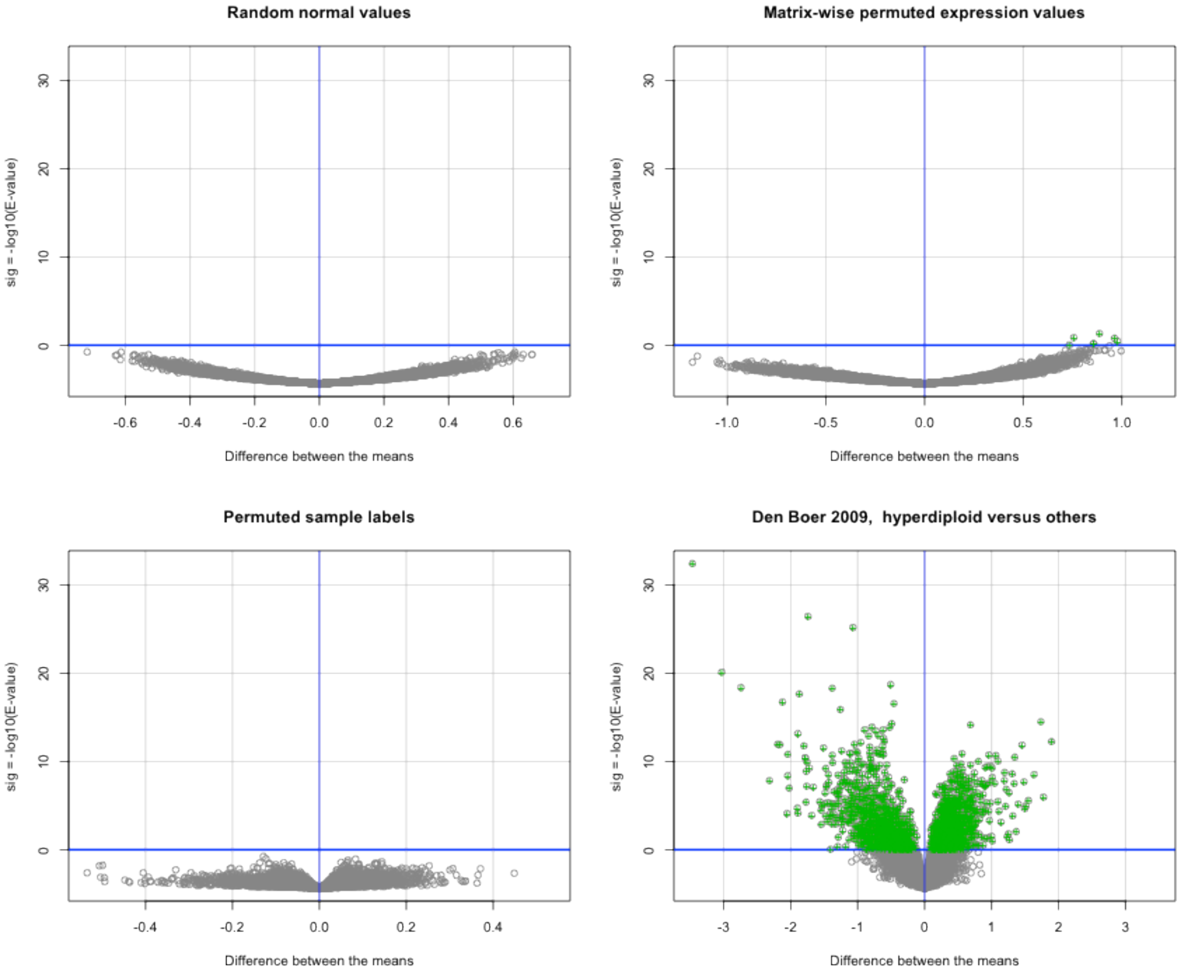
- Data set: Den Boer et al. (2009).
- Welch test: hyperdiploid versus other types of Acute Lymphoblastic Leukemia.
- P-value distribution
  - Abscissa: frequency class of the P-value.
  - Ordinate: number of genes falling in this class.
- 3 negative controls
  - Random normal values.
    - Flat distribution, as expected.
  - Matrix-wise random permutation of expression values.
    - Flat distribution, as expected.
  - Permutation of sample labels, analysis of the original expression matrix.
    - Under-representation of low P-values. Strange.
- Original expression matrix.
  - Striking over-representation of the low P-values. This likely corresponds to differentially expressed genes.



- Data source: Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.

Distribution of P-values from Welch test

- Data set: Den Boer et al. (2009).
- Welch test: hyperdiploid versus other types of Acute Lymphoblastic Leukemia.
- Volcano plots
  - Abscissa: difference between the means
  - Ordinate: significance of the test.
- 3 negative controls
  - Random normal values.
    - All significances are negative.
  - Matrix-wise random permutation of expression values.
    - 7 probesets are slightly significant.
  - Permutation of sample labels, analysis of the original expression matrix.
    - All significances are negative.
- Original expression matrix.
  - 2133 probesets are declared significant (differentially expressed) with E-value <= 1.



- Data source: Den Boer et al. 2009. A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol 10(2): 125-134.