Report for assingment 03: "Bootstrap for gene differential expression analysis of the prostate cancer data"

Alexey Stupnikov
Computational Biology and Machine Learning Laboratory
Center for Cancer Research and Cell Biology
School of Medicine, Dentistry and Biomedical Sciences
Faculty of Medicine, Health and Life Sciences
Queen's University Belfast
97 Lisburn Road, Belfast, BT9 7BL, UK, astupnikov01@qub.ac.uk

12.11.2013

1 Summary

This report describes the bootstrap results for gene differential expression analysis performed under various sampling and adjustment techniques.

2 Aims and Objectives

Using Affymetrix gene expression data achieved in [1] explore the influence of sample size and adjustment technique on the gene differential expression results.

3 Methods

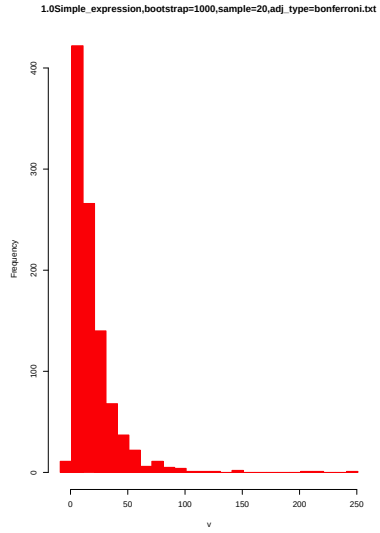
Data used in this assignment consists of 52 *normal* and 50 *tumour* samples. RMA was used for data preprocessing and normalisation. Then, expression matrix was processed to unite probesets corresponding to the same genes in order to implement *"genewise"* analysis instead of *"probewise"*. This also allowed to reduce the amount of hypothesises. After that bootstrap in two variations was performed. In the first approach, 20 samples from all samples pool of the same condition(normal/tumour) were selected randomly without replacement (randomisation was done by using standard R functions). Second approach involved random selection of 50 samples with replacement from same pools. Two obtained sample arrays were analysed for differential expression. Significance level was considered $p - value \leq 0.05$. Different multiple hypothesises testing adjustments were used: Bonferroni and less strict Benjamini-Hochberg FDR rate. Combined to 2 bootstrap variations it gives 4 different cases to compare.

4 Results

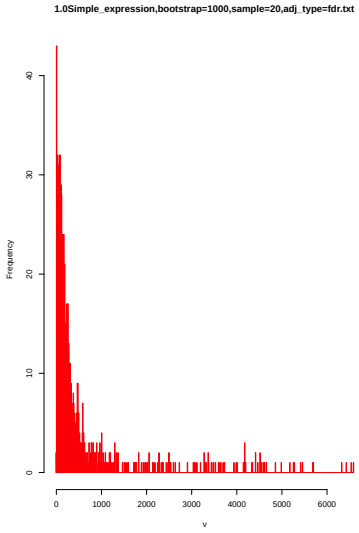
The results are introduced in a form of histograms. The horizontal axis represents the amount of significant genes found in a bootstrap iteration. The vertical axis represents the amount of iterations where such amount was observed.

5 Conclusion

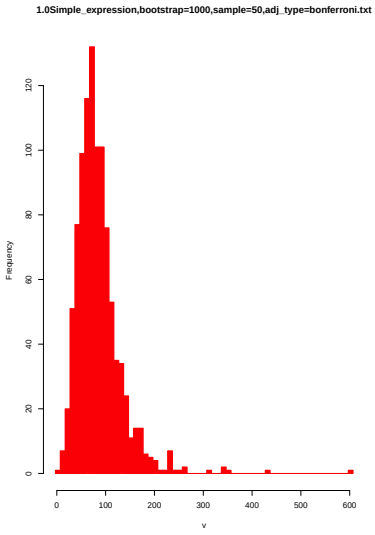
From the histograms it is clear that the amount of samples, as well as adjustment technique used for differential expression analysis has a drastic impact on the result. Bonferroni adjustment method excludes many significant genes, whereas Benjamini-Hochberg method allows to obtain much more significant genes. However, fdr method does not exclude certain unprobable



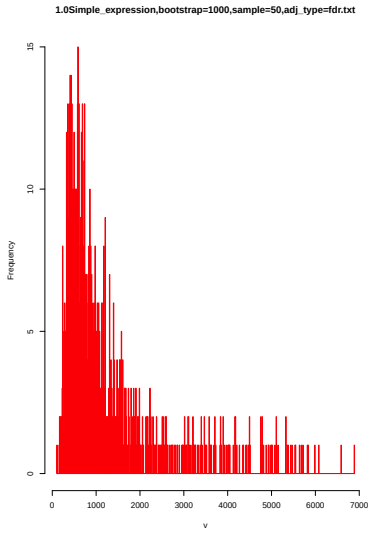
(a) Bootstrap sample size 20 Adjustment method Bonferroni



(b) Bootstrap sample size 20 Adjustment method Benjamini-Hochberg



(c) Bootstrap sample size 50 Adjustment method Bonferroni



(d) Bootstrap sample size 50 Adjustment method Benjamini-Hochberg

Figure 1: Significant genes amounts achieved in bootstrap inerations

from practical point of view cases with amount of significant genes of 7000. Bootstrap with sample size 20 gives much less significant genes than one with the size of 50. In a regular non-bootstrap analysis involving all samples, there were discovered 2263 significant genes and 7457 non-significant.

References

[1] Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.