Functional Genomics - Assignment 1
Due: November 11th 2016 23:45pm

Instructions

• The submission of your report must be done through the moodle site
• The report must be a single printable PDF (no other formats are accepted)
• The name of the PDF must follow the fga1XXX.pdf format, where XXX is your CRSid
• Clearly state which specific question you are addressing at each stage
• The report must be a *maximum* of 15 pages in total
• Try to keep the final PDF below 5 MB
• This course work will account for 30% towards your overall mark for this module

Part A - Microarray technology and analysis (1/3 of marks)

1) Describe the principles guiding the design of a microarray experiment.
2) Give a description of Illumina microarray platform including the advantages compared to other commercial platforms.
3) Describe at least two different normalization methods for single-channel microarrays. What could help in choosing the most appropriate method?
4) Explain why a probe filtering step is important and give an example of a meaningful filtering criterion.

Part B - Analysis of GSE51450 dataset (2/3 of marks)

Large collections of formalin-fixed paraffin-embedded (FFPE) samples are generated for diagnostic purposes and represent a valuable source of biological material in cancer translational research. Unfortunately, RNA extracted from FFPE samples has poor quality; however, strategies to obtain reliable gene expression data using microarrays have been developed. In a paper published on Plos One (PMC4386823), suitability of different microarray platforms was investigated in the context of breast cancer. Read the paper and then download the raw Affymetrix CEL files from GEO repository (GSE51450, n=12) and use it to answer the following questions. Please include the relevant code and plots that you generate.

1) Using the *affy* package, import the data in R and perform a sample quality control before and after summarization/normalization using appropriate diagnostic plots. Consider the presence of outliers and remove them from the analysis.

2) Retrieve information about estrogen receptor (ER) status of the samples (available in the Series Metrics file downloadable from the GEO entry). Create an appropriate design matrix and identify probesets differentially expressed between ER+ and ER- samples using the *limma* package. How many probesets are significantly differentially expressed (adjusted $p<0.05$)?

3) Repeat the differential analysis after filtering out non-informative probesets using either present call (similar to Illumina detection p-value) or IQR based filters. Is filtering affecting the number of differentially expressed probesets?