UHASSELT

KNOWLEDGE IN ACTION

Next Generation Sequence Analysis

Mai Nguyen

Faculty of Science
Master of Biostatistics
Specialization in Bioinformatics

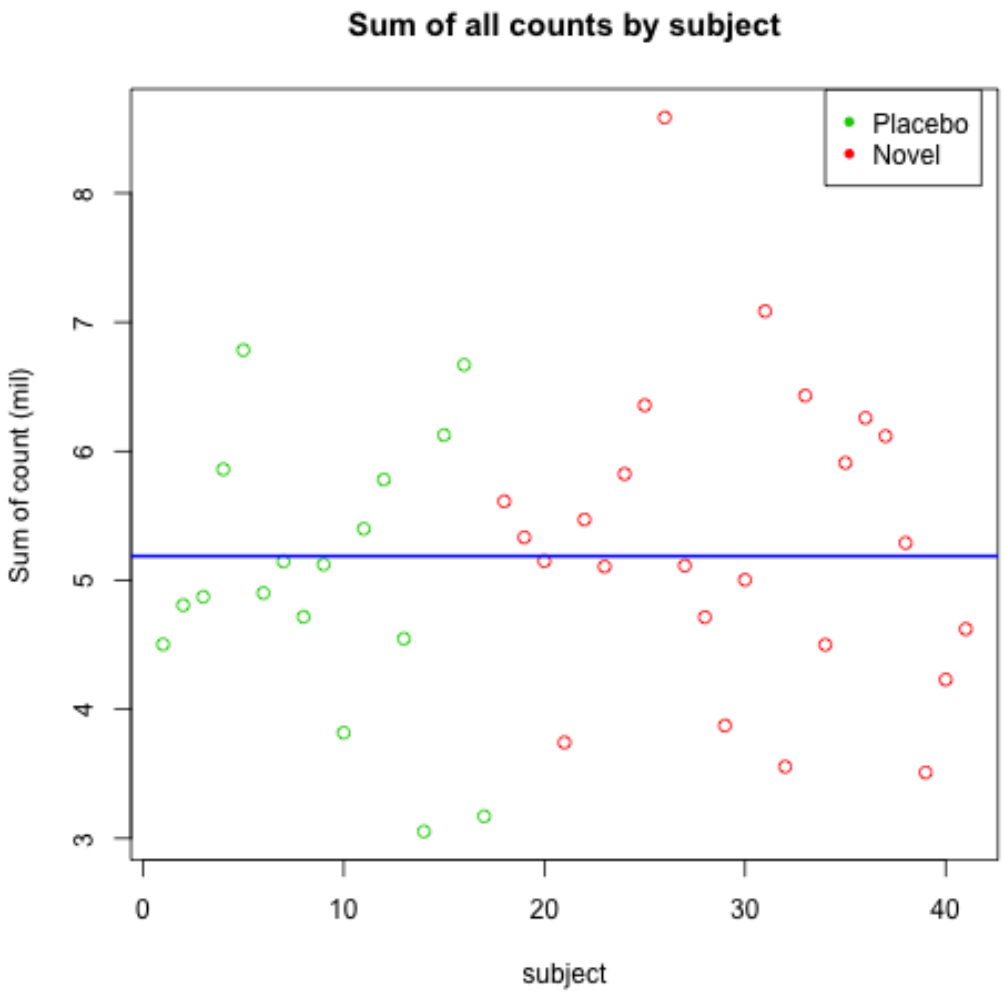
April 24, 2019

Data Overview

Study design

- 5258 genes are sequenced
- 41 subjects: 17 patients receive placebo treatment and 24 patients receive novel treatment

Study design



Normalisation

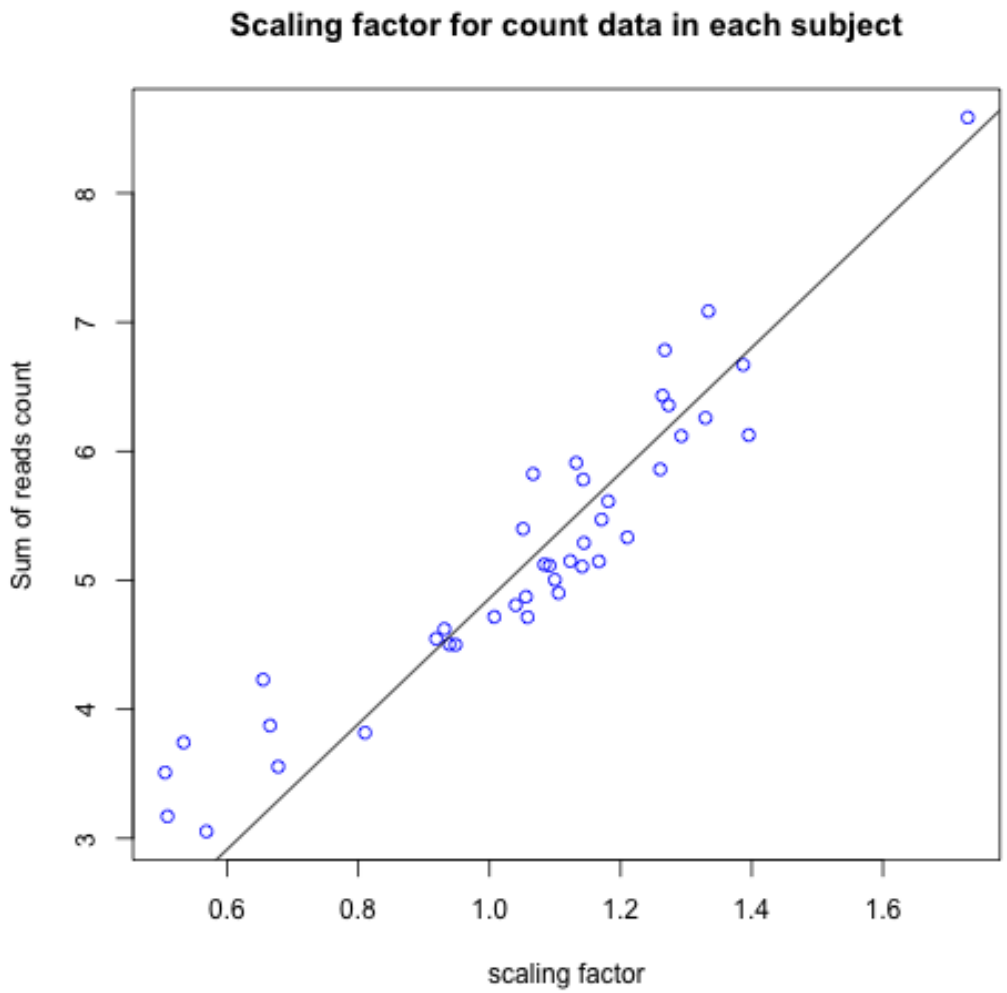
Median of gene expression ratio

$$\hat{s}_j = median \frac{k_{ij}}{(\prod_{v=1}^{m} k_{iv})^{1/m}} \quad (1)$$

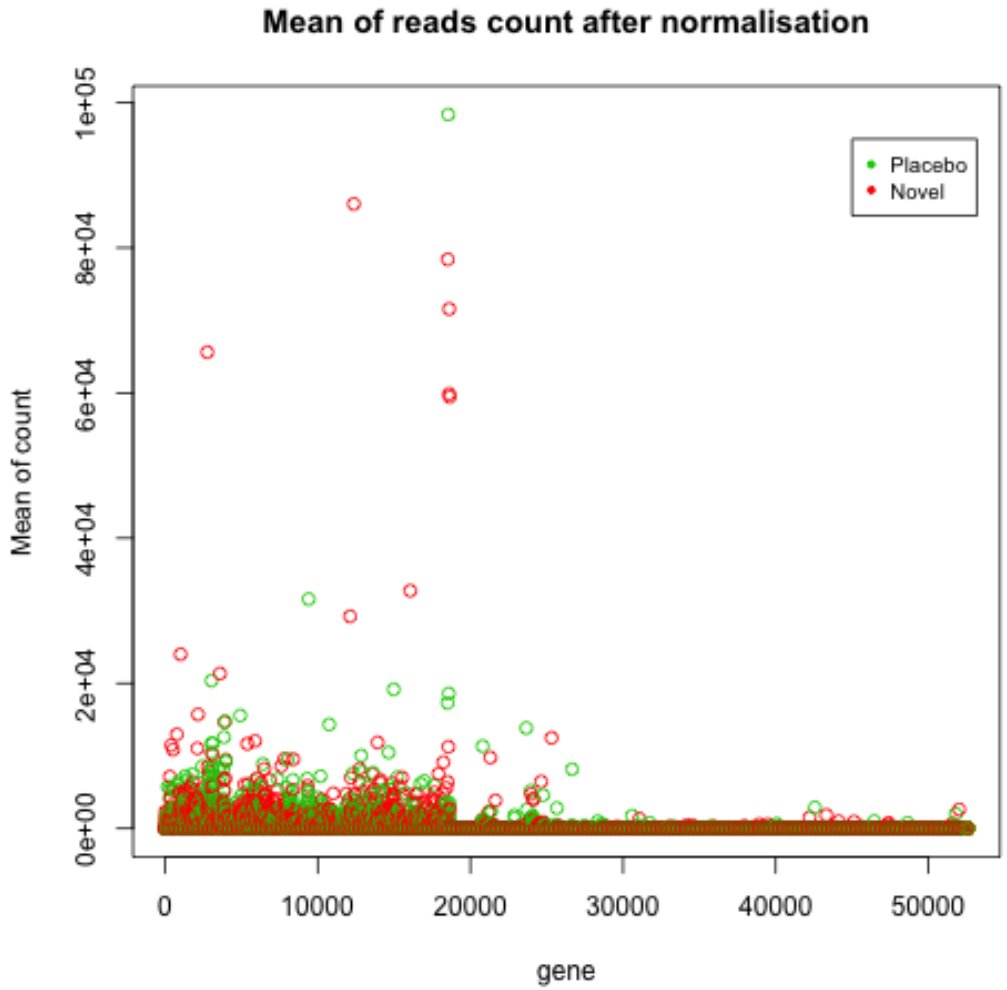
where:

- s_j is the size factor, median of the ratios of the j-th sample's counts to those of the pseudo-reference
- denominator is: pseudo-reference sample obtained by taking the geometric mean across samples

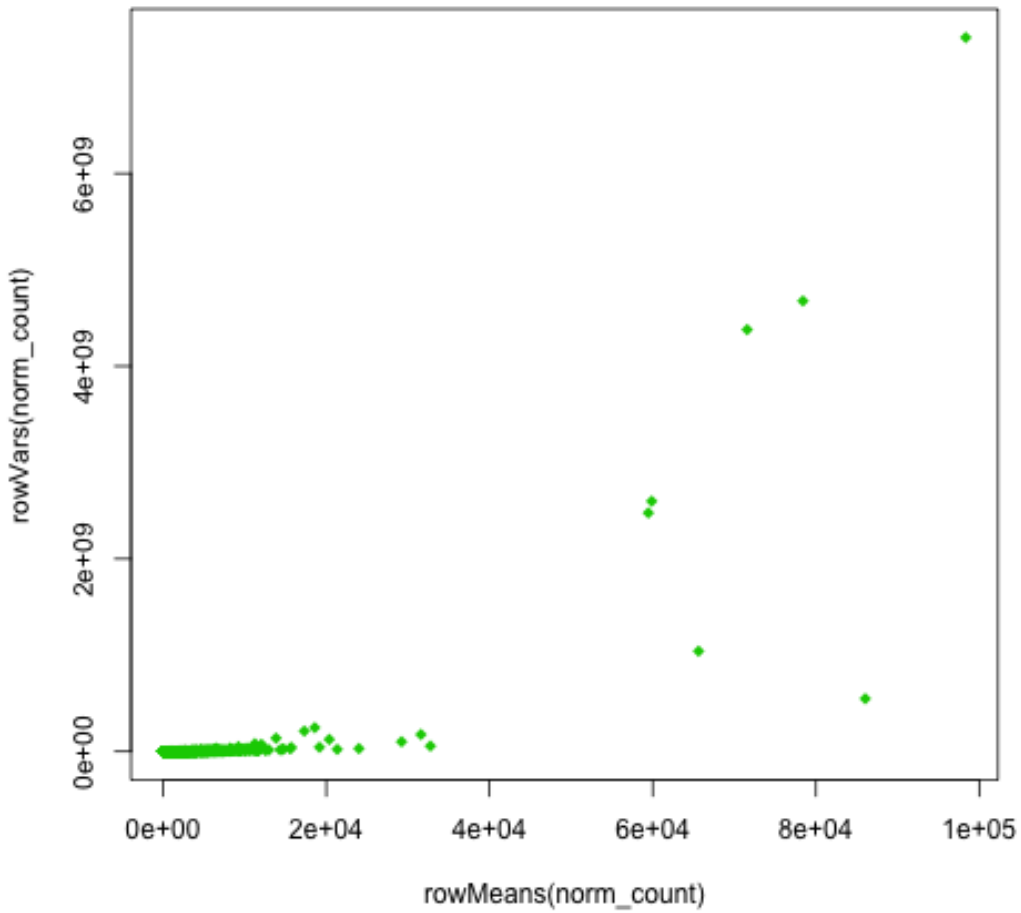
Normalisation



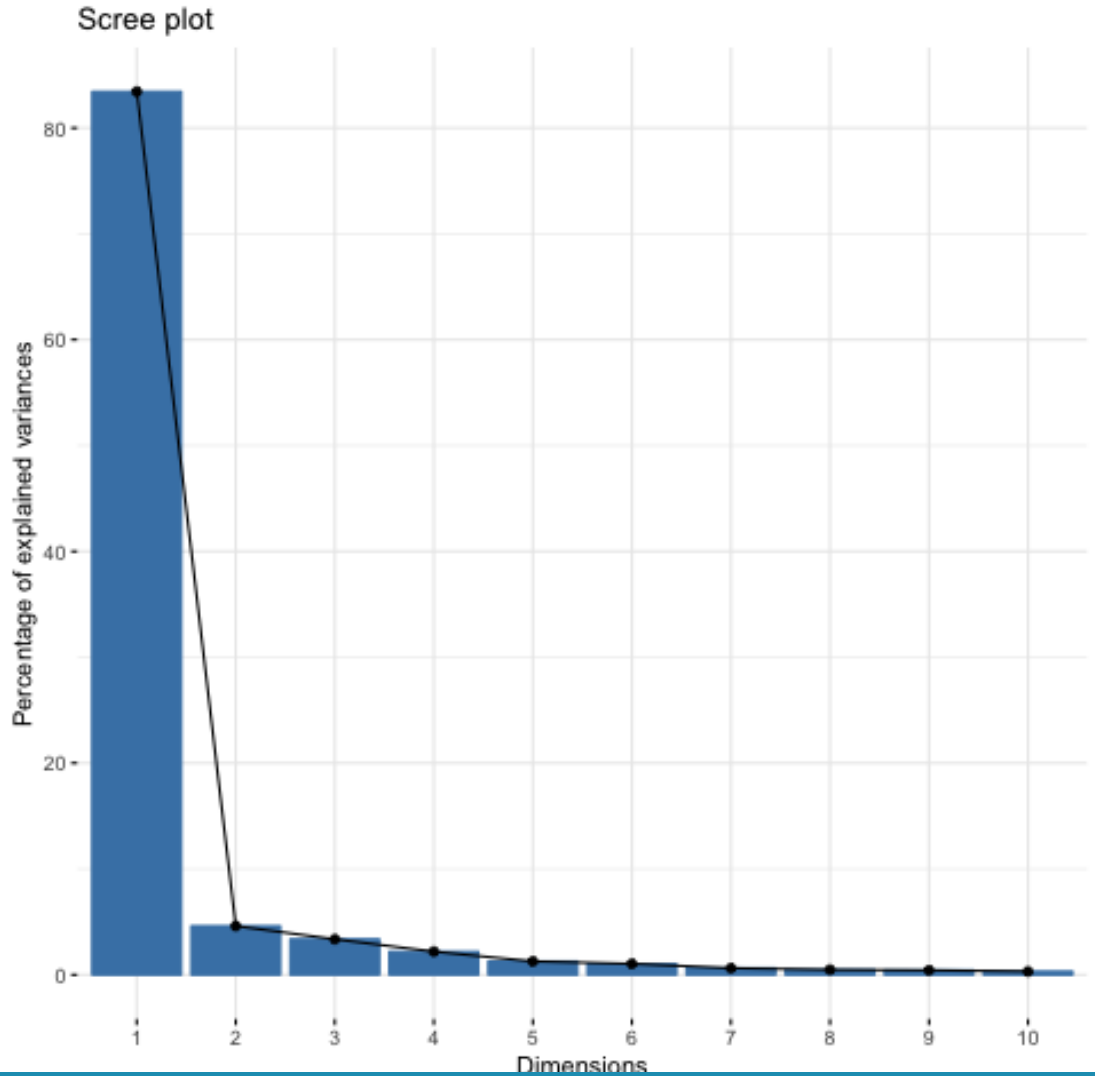
Exploratory Data Analysis



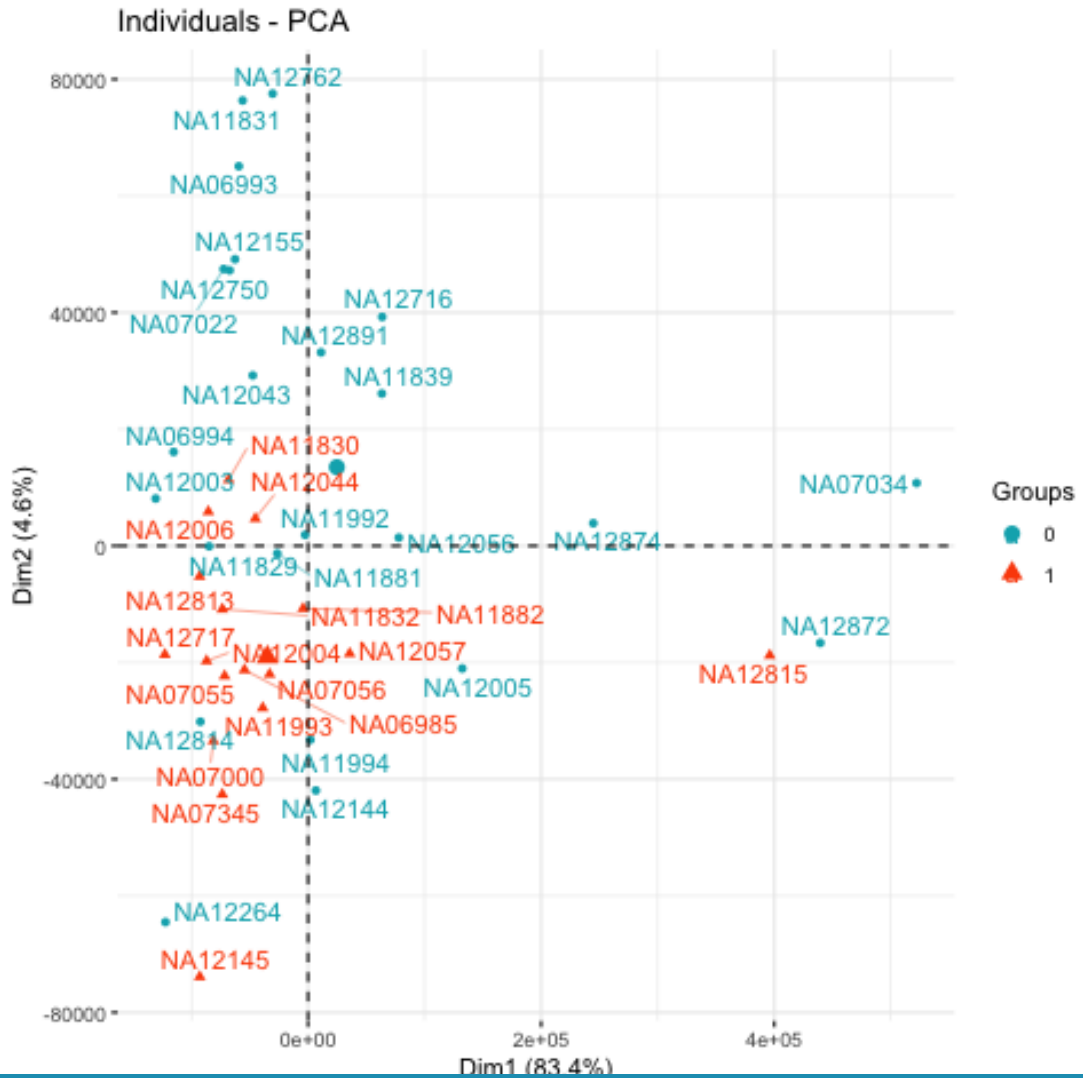
Exploratory Data Analysis



Principal Component Analysis



Principal Component Analysis



Statistical Approaches

Variance modeling at the observational level - VOOM

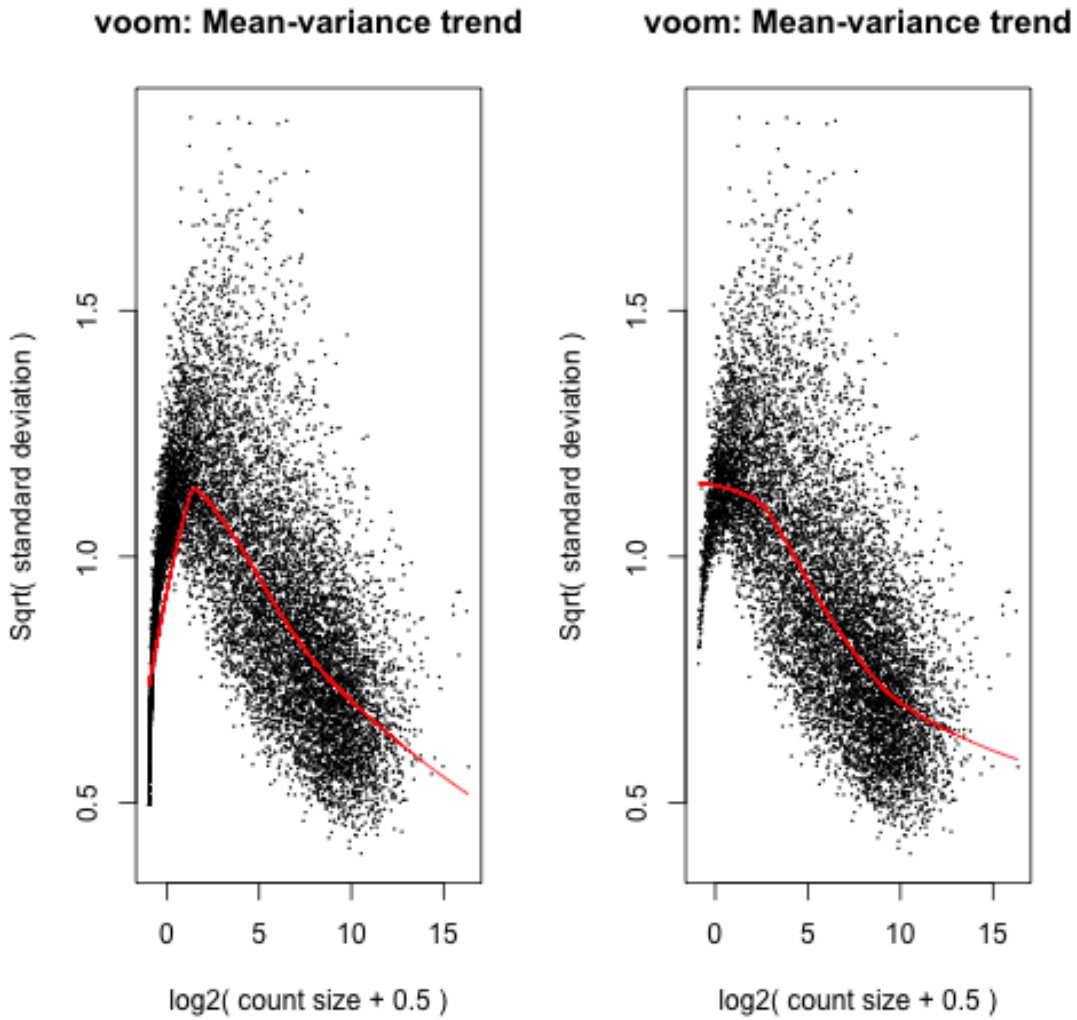
Transforms count data to log2-counts per million

$$logCPM_{ig} = log_2[(n_{ig} + 0.5)/10^6]$$

Where

- n_{ig}: number of read counts in gene g of sample i

Statistical Approaches



Linear regression by Limma

After transforming by VOOM, limma package can be applied for the data set to test for differential expression.

$$y = \beta_1 X + \epsilon_j \tag{2}$$

where

- y_j is the vector of gene expression
- x_{ij} is the design matrix for subjects with two treatments, placebo and novel
- β is the vector of parameter estimate

Null hypothesis for DE analysis:

$$\beta = 0$$

Multiplicity Adjustment

Apply linear regression in limma and adjusted p-value by Benjamin and Hochberg (1995) method.

Procedure

- Rank the p-value of each gene in order from the smallest to the largest.
- Multiply the largest p-value by the number of genes in test.
- Take the second largest p-value and multiply it by the total number of genes in gene list divided by its rank. If less than 0.05, it is significant. Corrected $p - value = p - value * (n/n - 1) < 0.05$, if so, gene is significant.

Results

	logFC	AveExpr	t	P.Value	adj.P.Val	B
ENSG00000129824	-9.93	3.55	-27.02	0.00	0.00	39.79
ENSG00000154620	-5.35	1.75	-23.95	0.00	0.00	36.71
ENSG00000157828	-5.94	0.39	-21.29	0.00	0.00	33.54
ENSG00000099749	-6.26	0.66	-17.28	0.00	0.00	29.48
ENSG00000198692	-6.03	0.39	-13.20	0.00	0.00	23.14
ENSG00000006757	0.93	4.99	6.27	0.00	0.00	7.41
ENSG00000183878	-2.04	-2.09	-6.08	0.00	0.00	5.96
ENSG00000105202	0.38	10.65	4.37	0.00	0.09	1.21
ENSG00000179094	0.95	4.73	3.96	0.00	0.28	0.30
ENSG00000102962	-0.74	13.50	-3.79	0.00	0.39	-0.63

Table: Differential Expression Testing for top 10 Genes

Results

Gene ID	Gene Name
ENSG00000129824	ribosomal protein S4 Y-linked 1
ENSG00000154620	thymosin beta 4 Y-linked
ENSG00000157828	NA
ENSG00000099749	NA
ENSG00000198692	eukaryotic translation initiation factor 1A Y-linked
ENSG00000006757	patatin like phospholipase domain containing 4
ENSG00000183878	ubiquitously transcribed tetratricopeptide repeat containing, Y-linked

Table: Gene names expressed differentially between two treatments

Reference Sources I

Li, Y., Tollefsbol, T.O. (2011)
DNA methylation detection: Bisulfite genomic sequencing analysis.
Methods in molecular biology (Clifton, N.J.),791,11-21, doi:10.1007/978-1-61779-316-52

Frommer M. et al. (1992)
A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands..
Proceedings of the National Academy of Sciences of the United States of America, 89(5), 18271831

Reference Sources II

Chen, Y. et al. (2018)
Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR.
F1000Research, 6, 2055. doi:10.12688/f1000research.13196.2

Shokoohi F. et. al. (2018
A Hidden Markov Model for Identifying Differentially Methylated Sites in Bisulfite Sequencing Data
Biometrics. doi:10.111/biom.12965

Dedeurwaerder S. et. al(2014)
A comprehensive overview of Infinium HumanMethylation450 data processing, Briefings in Bioinformatics
Volume 15, Issue 6, November 2014, Pages 929941,
https://doi.org/10.1093/bib/bbt054