Next Generation Sequencing Analysis

Exam Project
2018-2019

2nd year Master of Bioinformatics
Hasselt University

Student name:
Thi Ngoc Mai Nguyen (1746303)

Submission Date: 22nd April, 2019

Lecturers:
Dr. Jurgen Claesen
Dr. William Tholen
Prof. Oliver Thas

1 Introduction

High-throughput techniques provide the alternative way for gene expression analysis. With the discrete data obtained from NGS platforms, the statistical approach for analysing data from previous method such as microarray has to be changed. In this report, the analysis strategy is presented for discovering genes express differently by dealing with the count data obtained from two groups of treatment for pancreatic cancer patients.

1.1 Objectives

The main objective is of the sequencing analysis is to discover what gene expresses differently when patients receive the treatment for pancreatic cancer.

1.2 Data overview

There are two groups of patients receive the placebo and novel treatment for pancreatic cancer. The gene expression are measured for patients in these two groups, data generated are the number of reads that are aligned to the annotated genes. There are total 41 samples, placebo group has 17 patients and novel treatment group has 24 patients. 5258 genes are measured for all 41 subjects. Total counts of all genes by each patient is presented in figure 1.

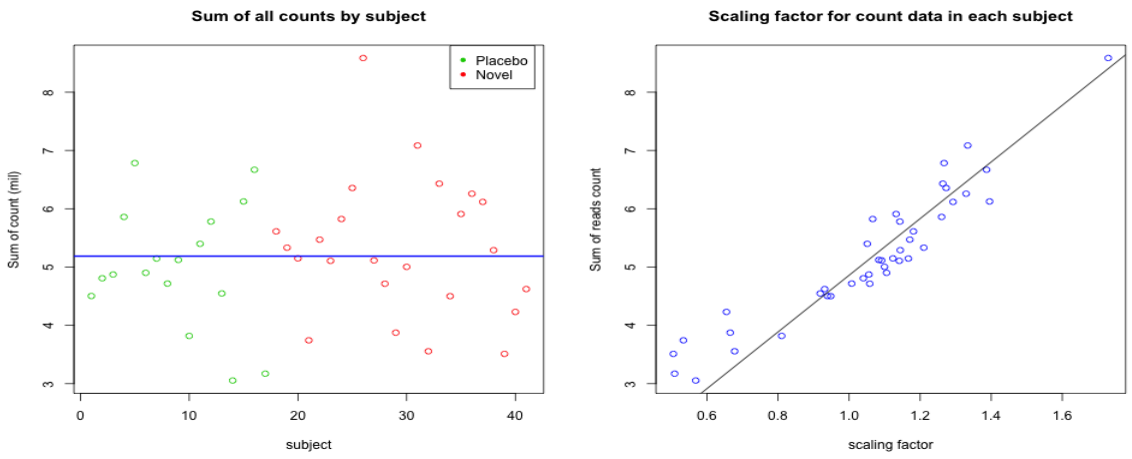

Figure 1: Total counts by each patient (left) - Normalisation by scaling factor (right)

1.3 Statistical software

R version 3.5.1 is used for analysing data set including packages: (7), (8), (9) and other visualisation packages.

2 Analysis approach

2.1 Normalisation

Before making comparison between samples, the raw data need to be to account for factors that prevent direct comparison of expression measures (1). Normalization the count data is important in downstream analysis. There are many techniques for normalising the counting of reads.

A normalisation technique that is used in package DESeq2 (2), Median of gene expression ratio. Dillies et al 2012 (3) recommended to use this method to distort the fraction of the total reads that contribute to genes with lower expression.

As Reddy A. et al (2016), using gene expression ratios presents a robust and novel method for constructing translatable biomarkers of compound response, which can also probe the underlying biology of treatment response. (4)

In this report, the normalisation technique by applying median gene expression ratio is proposed to adjust for the raw count data, the size factor for each subject will be calculated and the count data are scaled by that factor. Figure 1 (right side) presents the value of size factors adjusted for the raw count data.

2.2 Exploratory Data Analysis

2.2.1 Mean of reads count by gene and subjects

After normalisation the gene count data, the relationship between the variance and mean of reads count is visualised in figure 2.

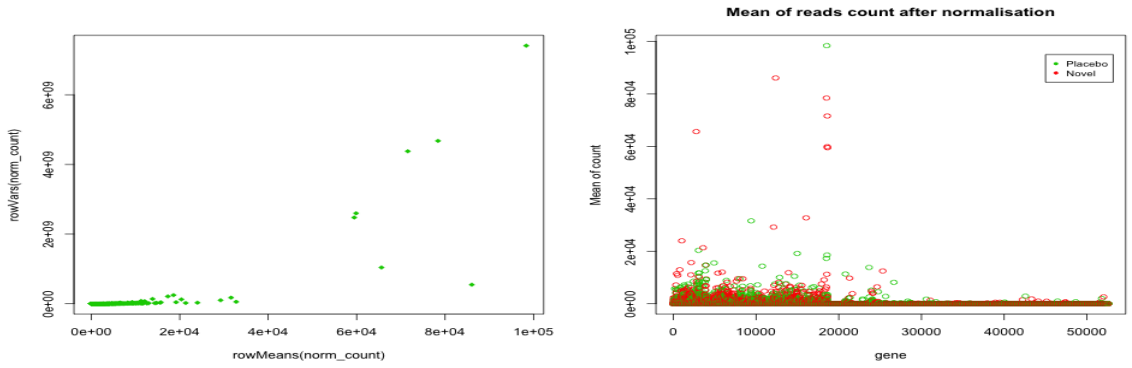

Figure 2: Mean of abundance after normalisation

2.2.2 Principle Component Analysis

Further exploratory data analysis and statistical model will be based on the normalised data to discover the genes that are differentially expressed. After normalising raw data, the mean of count by each subject and the mean of count by gene are presented in figure 2.

Due to extremely high dimensional data, more than 50,000 genes are measured between two groups of treatment, unsupervised technique, PCA, can be applied to get initial view on the structure of the data set. As in figure 3, first component can explain more than 80% of the variance in the data set. However, the classification of types of treatment is not clear.

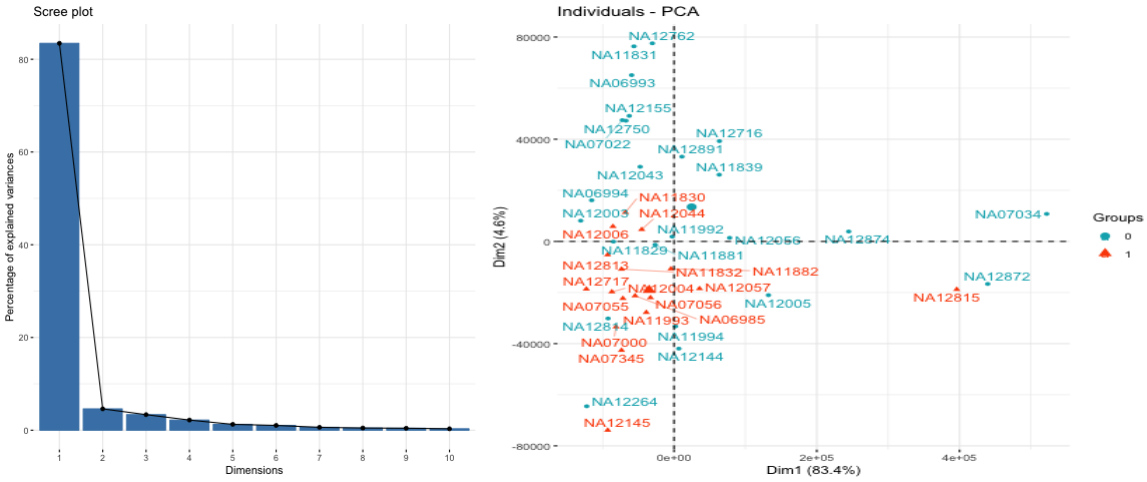

Figure 3: PCA analysis

2.3 Statistical methodology

2.3.1 Transforming the gene counts

In order to apply differential expression under the microarray context, gene counts data can be transformed by applying variance modeling at the observational level (VOOM).

VOOM (5) transforms count data to log2-counts per million (logCPM)

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

where n_{ig} is the number of gene counts in sample i. Then the mean-variance relationship is estimated and appropriate observation-level weights is computed. Then, linear regression can be applied for the count data and the multiplicity adjustment can be easily applied by using limma. The mean-variance relationship obtained by VOOM transformation after filtering low-expressed genes is presented in figure 4.

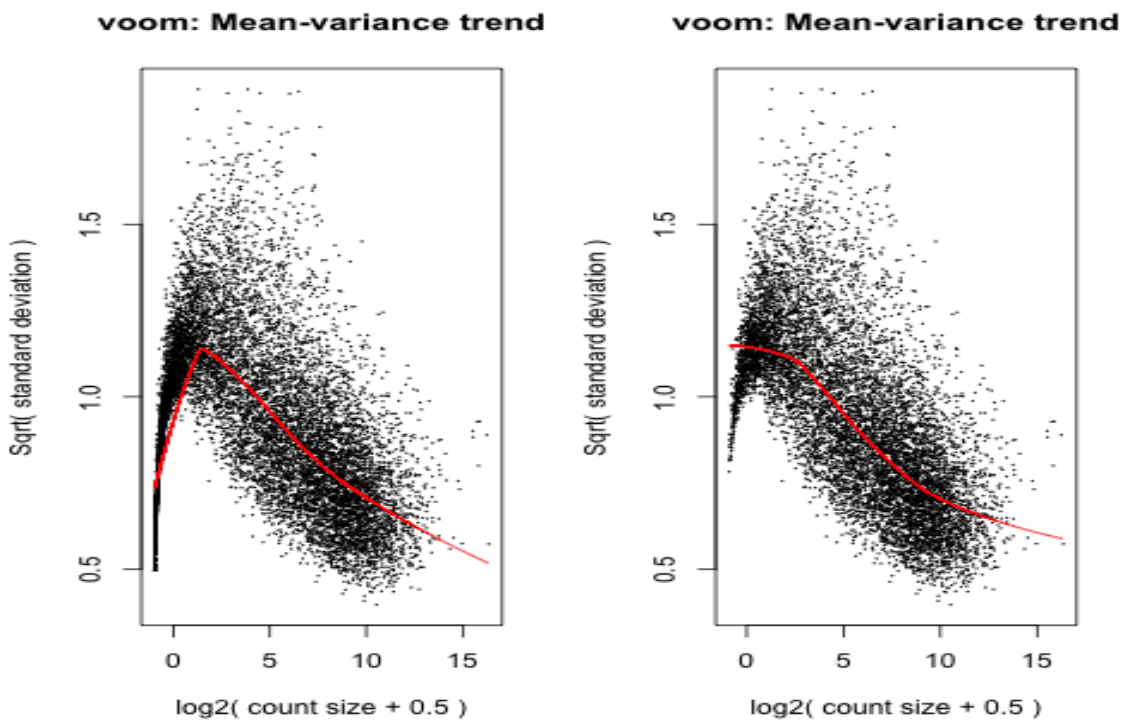

Figure 4: Mean-variance relationship by VOOM transformation:
before (left) and after (right) filtering low-expressed genes

2.3.2 Differential expression testing

As the count data is transformed by VOOM above, linear regression can be applied to test for differential gene expression then multiple comparison adjustment will be corrected by limma package. Limma is popular choice for gene discovery through differential expression analyses of microarray and high-throughput PCR data. By transforming reads count data by VOOM, limma can perform differential expression analyses of RNA sequencing (RNA-seq) data (6).

The null hypothesis $H_0 : n_{(g,placebo)} = n_{(g,novel)}$ for reads count of gene g are equal between two treatments. This hypothesis is tested by fitting linear regression model in limma. Gene that is expressed differentially is evaluated by adjusted p-value.

3 Results

By fitting linear regression in limma, the differential expression test results are shown in table 1. With p-value at 5%, there are only 7 genes expresses differentially between placebo and novel treatments for pancreatic cancer disease. Those genes are listed in table 2.

	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 1: Differential Expression Testing for top 10 Genes

In 7 genes in table 2, there are 4 genes have Y-linked. There are two genes that are not annotated in database yet.

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 2: Gene names expressed differentially between two treatments

4 Conclusion

Under the scope of this report, by normalisation the raw count data and transforming VOOM, the different gene expression analysis under the microarray context can be applied for discrete values between samples. Limma package is a powerful technique to discover the gene that is expressed differentially. This discovery is is crucial, especially in providing treatment to diseases. The pipeline of differential expression (DE) analysis in this report include following steps: normalising raw count data, transforming normalised count data by VOOM then testing DE by fitting linear regression in limma. The results fo DE analysis are seven genes listed table 2.

There are numerous statistical approaches can be applied to discover gene expresses differentially. Comparison between the approach in this report and other methods can be expanded in the future.

References

[1] Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2017;19(5):776792. doi:10.1093/bib/bbx008

[2] Anders S., Huber W., Differential expression analysis for sequence count data, Genome Biology 2010-11:R106 https://doi.org/10.1186/gb-2010-11-10-r106

[3] Dillies et. al (2012), A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis, Briefings in Bioinformatics, Volume 14, Issue 6, November 2013, Pages 671-683, https://doi.org/10.1093/bib/bbs046

[4] Reddy A, Growney JD, Wilson NS, et al. Gene Expression Ratios Lead to Accurate and Translatable Predictors of DR5 Agonism across Multiple Tumor Lineages [published correction appears in PLoS One. 2016;11(1):e0146635]. PLoS One. 2015;10(9):e0138486. Published 2015 Sep 17. doi:10.1371/journal.pone.0138486

[5] Charity W Law, Yunshun Chen, Wei Shi and Gordon K Smyth VOOM: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology2014 15:R29, https://doi.org/10.1186/gb-2014-15-2-r29

[6] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi:10.1093/nar/gkv007

[7] Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550 (2014)

[8] Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140. McCarthy DJ, Chen Y and Smyth GK (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40, 4288-4297

[9] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43(7), e47.