Differentially Expressed Genes in Breast Cancer Subtypes

A comparative study of DESeq2 and Limma Voom

Feb 2021

Table of Contents

1. Introduction

1.1. Research Background

The female breast is composed of skin, fibrous tissue, breast and fat. Breast cancer is a malignant tumour that occurs in breast epithelial tissue, which has been the leading cause of cancer deaths in women worldwide, accounting for approximately 23% of all cancer cases..Though breast cancer in situ is normally non-fatal, metastasis of the breast cancer cells to other parts of the body can lead to death. 99% of breast cancers occur in women, while men account for only 1%.

Traditional medical research is unable to fully understand breast cancer and formulate treatment plans due to the shortcomings of the medical model and the insufficient clinical experience of individual doctors. The application of data analysis makes up for this deficiency, increases human knowledge and understanding of breast cancer, and is more conducive to comprehensive evaluation and comparison of existing treatment options, thereby changing clinical practice guidelines.

Breast cancer can be divided into different subtypes based on the invasive, histologic or molecular subtype. The dataset used in this study follows the molecular subtypes[5]:

A. **Luminal A (HR+/HER2-)**: This subtype's postoperative immunohistochemistry showed estrogen receptor (ER)/progesterone receptor (PR) positive, KI67 positive rate was <14%, human epidermal growth factor receptor 2 (HER2)/Cerbb2 was negative or weakly positive. Luminal type A is the most common molecular pressure type of breast cancer. This subtype is usually early breast cancer, with a low risk of recurrence, and is sensitive to endocrine therapy, but not to chemotherapy, and has a good prognosis.

B. **Luminal B (HR+/HER2+):** This subtype is hormone-receptor (HR) positive (ER and/or PR positive), and either HER2 positive or HER2 negative with high levels of Ki-67. Luminal B cancers generally grow slightly faster than luminal A cancers and their prognosis is slightly worse.

C. **Basal-like (HR-/HER2-):** This subtype is HR negative (ER and PR negative) and HER2 negative. This type of cancer is more common in women with BRCA1 gene mutations.

D. **HER2-enriched (HR-/HER2+):** This subtype is HR negative (ER and PRnegative) and HER2 positive. HER2-enriched cancers tend to grow faster than luminal cancers and can have a worse prognosis, but they are often successfully treated with targeted therapies aimed at the HER2 protein.

1.2. Introduction of dataset

The dataset used in this study was obtained from EMBL-EBI Expression Atlas, with the title of "RNA-seq of 17 breast tumor samples of three different subtypes and normal human breast organoids samples" [10]. EMBL-EBI Expression Atlas is a free open-source database for life science experiments, focusing on information about gene and protein expression with basic data analysis results from popular processing pipelines.

Using RNA-sequencing technology, the dataset aims to define the digital transcriptome of three breast cancer subtypes based on immuno-histochemical and RT-qPCR classification:

- Triple negative breast cancer (TNBC) with ER -, PR - and HER2 -
- Non-triple negative breast cancer (NTNBC)
- HER2 positive breast cancer (HER2)

Subtypes specified in this dataset were not one-to-one correspondence of the molecular subtypes mentioned in section 1.1. Some of the subtypes mentioned in this study can be non-exclusive, and one sample could potentially be categorized under multiple subtypes. The dataset also contains healthy samples as control group. All the human samples and data were used in accordance with the IRB procedures of Baylor College of Medicine.

2. Methodology

In this study, raw gene counts of 3 subtypes of breast cancer (i.e. TNBC, NTNBC, HER2) were compared against those of healthy samples to identify differentially expressed genes (DEGs) using two R libraries of DESeq2 and Limma-Voom. The difference in results were compared and explained based on the difference of algorithms used by the 2 packages. The DEGs were then extracted as features for sample clustering, where the gene counts were normalised and transformed to log2 scale by regularised log (rlog) to minimise the differences between samples. The clustering result was evaluated against the original labels of the samples.

2.1. Library Introduction - DESeq2 & Limma-Voom

2.1.1. DESeq2

DeSeq2 is an R/Bioconductor software package that provides methods for expression strength and differential analysis of count data, including RNA-Seq by using shrinkage estimation for dispersions and fold changes to improve stability and interpretability of estimates. The task of differential analysis is to find a list of genes that are differentially expressed across groups of samples. However, the accuracy of the result may suffer from small replicate numbers,

discreteness, large dynamic range and the presence of outliers, making the estimation of dispersion intractable.

The null hypothesis of the differential analysis with DESeq2 is that the gene is not differential expressed across the groups. If the between-group variance is greater than the within-group variance (error effect) and is statistically significant, it is considered differential expressed.

2.1.2.Limma-Voom

Limma is an R/Bioconductor software package that provides an integrated solution for analyzing data from genome-wide gene expression experiments. This statistical method provides the most popular differential expression analysis pipeline for the last 2 decades; however, it was developed for microarrays data (continuous numerical) rather than RNA-seq read counts (discrete integer), that are different in the nature. It is problematic to apply Limma on integer RNA-seq count distribution, as it relies on use of the normal distribution of numerical data.

The mathematical theory of count distributions is originally hard to deal with than that of the normal distribution. Despite the use of probabilistic distributions, the estimated dispersions of counts distribution are still uncertain especially with small sample sizes. While Limma based on NB (negative binomial) distribution treat the estimated dispersions as if they were already known, and small.

Therefore, the issue is the ability to adapt to different types of data with high or low dispersion heterogeneity. The goal becomes modelling the mean-variance relationship correctly, rather than specifying the exact probabilistic distribution of the counts.

The method of Voom (variance modeling at the observational level) is introduced, which incorporates the mean-variance trend into a precision weight for each individual normalized observation. Since different samples may be sequenced to different depths, the count sizes may vary considerably from sample to sample for the same gene. For this reason, we wish to model the mean-variance trend of the log-cpm values at the sample-level, rather than a gene-level variability estimate to all observations from the same gene.[2]

Voom is powerful in 3 aspects:

- It controls the type I error (false positive) rate correctly.
- Voom has the low false discovery rate, and is robust to outliers.
- Voom is faster than specialist RNA-seq methods.

2.2. Identify DEGs with DESeq2

DESeq was performed on the original dataset with 42,617 prefiltered genes with counts no less than 10 across all samples. The pipeline included estimating size factors, estimating dispersions, gene-wise dispersion estimates, mean-dispersion relationship, final dispersion estimates and fitting model and testing. To identify the DEGs in TNBC, result of DESeq was obtained by 'results' function with the corresponding comparison group specified (i.e. TNBC vs Healthy), which contained average normalised count value (baseMean), effect size estimate (log2FoldChange), standard error estimate (lfcSE), p-value from Wald test and FDR (padj, Benjamini-Hochberg adjustment). The result table was filtered by 10% of False Discovery Rate (FDR, or 'padj' column in the result table) and ranked by their log2FoldChange, where positive values indicate the DEGs to be up-regulated in the diseased samples, and negative values indicate down-regulated DEGs in the diseased samples.

The same procedure was repeated to identify the DEGs in NTNBC and HER2 samples.

2.3. Identify DEGs with Limma-Voom

To start the Limma-Voom pipeline, the original gene counts were log-transformed into CPM (count per million, LCPM), after which the lowly-expressed genes were excluded. The gene expression distribution was then normalised by TMM to eliminate any impact from non-biological factors and ensure the distributions of each sample was similar across the experiment [2]. A design matrix and contrasts were created from the datasets, and a contrast matrix was created to specify the comparison groups of interest (i.e. TNBC vs Healthy, NTNBC vs Healthy, HER2 vs Healthy). Next, the differential analysis was conducted, removing heteroscedascity from count data and fitting linear models for comparisons of interest. The number and list of DEGs identified could be extracted using 'decideTests' function.

2.4. Sample Clustering Using DEGs as Features

The DEGs identified from section 2.1 and 2.2 were used as clustering features, respectively. There were 8,549 DEGs and 2,874 DEGs identified by the 2 packages, respectively. Counts of these selected DEGs were normalised and blog transformed. Then, to avoid the curse of dimensionality and any collinearity among the genes (which could be caused by gene co-expression), principle component analysis was conducted. The top principle components (PCs) with at leat 95% of covariance was selected as the final features for clustering. Hierarchical clustering was performed to cluster the samples with euclidean distance and Ward.D2 criterion. Number of clusters were determined by the silhouette plot with the highest average silhouette width. To evaluate the model performance, the clustering result was then compared against the original sample labels.

3. Result and Analysis

3.1. Identification of DEGs Against Healthy Samples

3.1.1. DEGs Identified per cancer subtype

From the 58,735 genes with non-zero counts originally in the datasets, the number of DEGs identified using DESeq2 and Limma-Voom in each breast cancer subtype are shown in Table 1 and Table 2, respectively. Total DEGs are all significant DEGs with FDR of 10% for each comparison group. Unique DEGs are DEGs are that only found in the corresponding cancer subtype, while common DEGs can be found in all cancer subtypes.

Table 1. Number of DEGs identified by DESeq2 per subtype of breast cancer

DESeq2 *(FDR = 0.1)*	TNBC vs Healthy	NTNBC vs Healthy	HER2 vs Healthy
Total DEGs	20,043	15,146	15,032
Unique DEGs	5,388	1,455	2,829
Unique DEGs (log2FoldChange > 2 or < -2)	4,227	498	1,572
Common DEGs (among all subtypes)	8,549		
Common DEGs (among all subtypes, log2FoldChange > 2 or < -2)	4,035		

Table 2. Number of DEGs identified by DESeq2 per subtype of breast cancer

Limma-Voom ***(FDR≤ 0.1, LFC ≥ 2)***	TNBC vs Healthy	NTNBC vs Healthy	HER2 vs Healthy
Total DEGs	5,302	4,946	5,490
Unique DEGs	826	532	1,330
Common DEGs (among all subtypes)	2,874		

To compare the DEGs found by the 2 pipelines, results were compared to examine if any genes were identified as DEGs by both pipelines. The overlapping DEGs are shown in Table 3, where percentage in each cell represents the overlapping percentage against the total DEGs per pipeline for each category.

Table 3. Overlapping results of DEGs identified by both pipelines

Overlaps	TNBC vs Healthy	NTNBC vs Healthy	HER2 vs Healthy
Unique DEGs (log2FoldChange > 2 or < -2)	203 (5% of DESeq2, 25% of Limma-Voom)	148 (30% of DESeq2, 28% of Limma-Voom)	311 (20% of DESeq2, 24% of Limma-Voom)
Common DEGs (among all subtypes)	1,674 (41% of DESeq2, 58% of Limma-Voom)		

Clearly, an obvious difference in the DEGs identified was found using the 2 different libraries. By studying the algorithms employed by the libraries, some of the factors contributing to the difference have been identified as below (section 3.1.2 ~ section 3.1.3).

3.1.2. Comparison of Algorithms

Key components of the two algorithms were compared as below:

A. Exclusion of lowly-expressed genes

Normally, genes without any counts or very low counts across all samples tend not to be the target of interest. Even if it did has differential expression, the ratios across groups would be much more noisier (see Appendix Figure A1.1 - A1.3), making it almost impossible to get a reliable conclusion [1, 2]. This is also a common difficulty when analysing High-Throughput data.

To overcome this issue, a primary manual filtration was also applied. Before performing DESeq2, any genes with less than 10 total counts across all samples were excluded, which left 42,617 genes in the dataset. In addition, in DESeq2, an additional filtration of independent filtering using the mean of normalised counts irrespective of biological condition was applied after modelling. It was also to remove genes with low counts that are very unlikely to show significant evidence of differences.

On the other hand, when using Limma-Voom, the library contains a more stringent filtration criteria. The same threshold of exclusion was applied (i.e. filter out any genes with less than 10 genes). However, the actual filtration was done by filterByExpr() function from the library, which was based on LCPM (log counts per million) reads to avoid giving preference to samples with large library sizes. In this dataset, the median of library sizes for the 19 samples is around 68.6million. Hence, the filtration keeps any genes with a LCPM larger than 0.14 (= 19/68.6) in at least 3 replicates, which is the smallest number of replicates per condition group. This is much stricter than the manual filtration in DESeq2, as the latter does not consider the number of

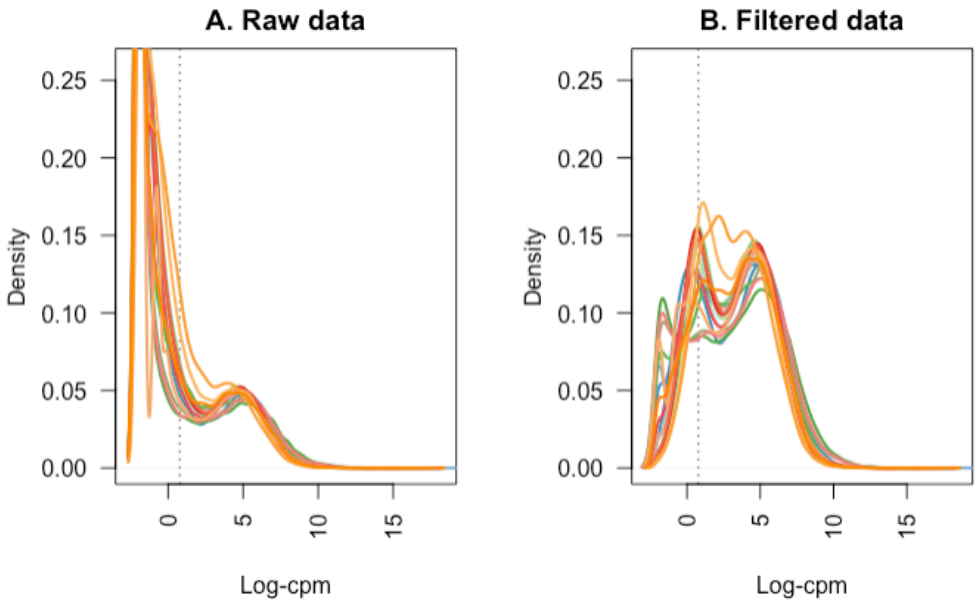

Figure 1. Density of LCMP per sample before and after applying filterByExpr() in Limma-Voom pipeline. Each line represents one sample.

samples with valid level of expression. In this case, only 21,023 genes were left. Figure 1 below shows the density of LCPM before and after applying the filtration function in Limma-Voom pipeline. An obvious reduction in density around low LCPM region can be observed after the filtration.

B. Counts normalisation

When identifying the DEGs, the raw read counts could not be compared across samples or condition groups directly. Because different sample may have different library size (i.e. total number of reads), leading to the same genes having read counts in completely incomparable ranges. Besides, due to the technical limitation or inconsistent nature of the samples, different group of genes may be detected across samples, making it even harder for comparison. Other factors like the impact of genes with dominant expression should also be considered before comparison. Therefore, counts normalisation is essential as part of differential expression analysis.

In DEseq2, median-to-ratio normalisation is one of the methods to normalise the counts [5]. Read counts are divided by the total count of their sample, then averaged across all samples in a condition for a given gene. This produces an average count-normalised value for each gene and each condition, and the median of the ratios of these values between conditions is taken. The original counts are then normalised by this median and their library size. DESeq2 scaling factor for a given lane is computed as the median of the ratio, for each gene, of its read count over its

geometric mean across all lanes, which is less sensitive to outliers. The underlying idea is that non-DE genes should have similar read counts across samples, leading to a ratio of 1. Assuming most genes are not differential gene, the median of this ratio for the lane provides an estimate of the correction factor that should be applied to all read counts of this lane to fulfil the hypothesis.

On the other hand, in Limma-Voom, LCPM read of the gene are used, which by definition normalised for sequencing depth. Other normalisation steps can be optional, depends on the data itself. In this study, in order to conduct DE analysis across samples, the library size should also be accounted for. Hence, another normalisation techniques of trimmed mean of M values (TMM) were used. With calNormFactors() function, it calculates a normalisation factor per sample to eliminate composition biases between libraries, after trimming off the extreme values in terms of LCPM counts, reducing the impact of outliers [7].

Both normalisation techniques are able to account for difference in sequencing depth and library composition, which is sufficient in our study. Besides, both normalisation techniques share the same assumption that most of the genes are not DEs. However, TMM also accounts for the difference in gene length. In case any gene expression were required to be compared within the same sample, normalised counts from Limma-voom pipeline would be more suitable. In addition, though both techniques work with log transformed data, the major difference lies with the calculation of normalisation factors, where DESeq2 relies on the exponential of sample median, and TMM in Limma-Voom uses weighted mean after trimming instead. This may not be a major contributor to the inconsistent DEGs identified in the result, but is still one of the worth-noticing factors.

C. Statistical modelling

DESeq2

In DESeq2 pipeline, the distribution of gene counts are assumed to follow negative binomial distribution, as the gene counts are positive discrete values, and the mean does not always equal to the variance. Mathematically, this can be written as:

$$K_{ij} \sim NB(s_j q_{ij}, \alpha_i)$$

where K is the count for gene i, sample j, α is the gene-specific dispersion (a measure of variance), s is the normalisation factor for sample j, and q is a parameter proportional to the expected true concentration of fragments for sample j.

The DEGs are tested by negative binomial generalised linear models, where the estimated standard error of a log2 fold change (LFC) is used to test if it is equal to zero via the Wald test. In other words, the null hypothesis is that the gene has no differential expression across the

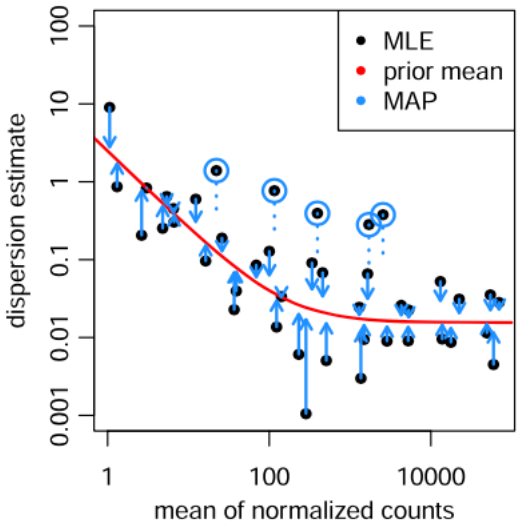

Figure 2. Shrinkage estimation of dispersion in DESeq2

groups (i.e. LFC = 0), and the alternative hypothesis is that the gene does have different expressions across the groups (LFC ≠ 0). As part of the Wald test, p-value is calculated for the product of LFC and its standard error on a normal distribution. As statistical test is conducted for thousands of genes in the sample, Benjamini-Hochberg procedure was also applied to eliminate the impact of multiple test problem, producing the adjusted p-value (padj).

However, this could be a very loose hypothesis test sometimes, especially when the comparison is made between samples from different tissues (e.g. liver vs thyroid), as many genes are differentially expressed in different tissues by nature. In our study, even though the samples are from the same tissue type, however, there's lack of information to determine if the disease condition is the only variables among the samples. Other than the disease condition, if there's any other features that are not well controlled (i.e. other type of unknown diseased in the breast tissue, race, age, previous conditions, etc.), the null hypothesis may be easily rejected with large number of False Positive DEGs identified. Moreover, due to the interconnectedness of gene expression, if there were indeed a biological difference between the condition groups, genes with 0 LFC as per null hypothesis would be perfectly decoupled from the biological difference. Hence, the null hypothesis with 0 LFC may be implausible for many of the genes, leading to variations in the DEGs identified.

In addition, when building the negative binomial model, the dispersion (α) is estimated based on all genes with the same average expression. Dispersion is used to measure the variability between replicates, accuracy of which is critical for the differential expression statistical inference. Due to the small number of replicates in High-throughput sequencing, dispersion of the gene expression tends to be more variable and noisy. To overcome this issue, the variance of gene counts is estimated based on all genes with the same level of average gene counts. In DESeq2 pipeline, it is assumed that all genes of similar average expression strength have similar dispersion. The common dispersion is predicted by a fitted curve between the actual dispersion

and average normalised counts for all genes (see Figure 2). Then, gene-wise shrinkage is done towards the fitted curve via empirical Bayes approach to avoid potential false positives, which could be caused by underestimates of dispersion [1].

However, this assumption may not always be true, as genes with the same average expression may not always have the same dispersion. It could also be overly liberal in some situations [3]. With fewer replicates available, the estimation of average expression for individual genes could be inaccurate and sensitive to noises, leading to wrong dispersion estimation. Hence, DESeq2 may have better performance and accuracy with larger sample sizes.

Figure 3 shows the relationship between the variance and mean of gene counts for all HER2 samples. The variance tends to be more variable with smaller mean counts. With the increase of mean counts, there are fewer genes observed. The variance tends to converge and become

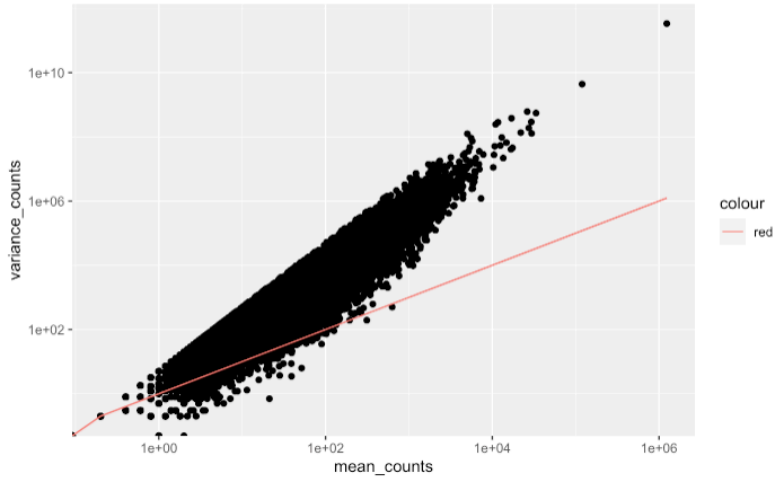

Figure 3. Correlation between average gene counts and its variance. Red line indicates when average and variance are equal.

larger than the average value (indicted by the red line).

Similarly, LFC estimates may also be inaccurately large in experiments with small sample sizes, due to the fact that ratios are inherently sensitive to noises [1]. Hence, LFC estimate from DESeq2 is also shrunken to zero, strength of which increases with higher diversion or fewer degrees.

Limma-Voom

On the other hand, in the Limma-Voom pipeline, the gene counts is analyzed using normal linear model, which has better type I error control with smaller sample size and more

complicated experimental design [2]. Linear models is applied to describe how the treatment factors such as batch effects or numerical covariates are assigned to the different RNA samples. The model is based on the assumption that:

$$E(y_{gi}) = x_i^T \beta_g$$

where x_i is a vector of covariates and β_g is a vector of unknown coefficients representing log 2-fold-changes between experimental conditions.

The linear model of Limma-Voom is fitted by ordinary least squares to the LCPM values for each gene. The regression coefficient estimates, fitted values, average log-cpm values and residual standard deviations are yielded as the results of model. Besides, square-root standard deviations are used as a statistic because of their roughly symmetrically distribution. A LOWESS curve (locally weighted regression) is fitted to the relationship between square-root standard deviations and mean log-counts to obtain a smooth mean-variance trend. The LOWESS curve provides an asymptotic line through most of the standard deviations with great statistical robusticity [2].

Generally speaking, with the linear-model based method, Limma-Voom makes fewer distributional assumptions than DESeq2, which assumes the gene expression follows negative binomial distribution. This makes Limma-Voom more robust in many of the scenarios, especially with unequal library size or complex experimental design [2].

Figure 4 shows the relationship between gene-wise square-root residual standard deviations and average log-count in our dataset, which is given by a robust LOWESS fit to the points. The standard deviation asymptotes at a moderate level corresponding to average log-count, which decreases steadily as a function of the LOWESS.

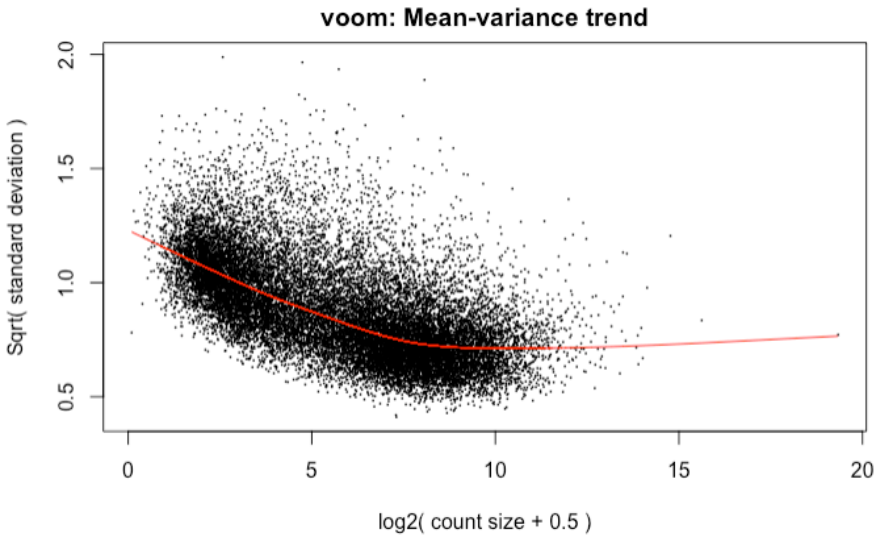

Figure 4. Mean-variance trend in Limma-voom pipeline

D. Handling of outliers

Presence of outliers may greatly distort the differential analysis results. Thus, to identify genes that actually have different expression levels, the outliers should be handled carefully. In DESeq2 pipeline, a standard outlier diagnostic is done using Cook's distance. Within each gene for each sample, the scaled distance of a linear model is measured and compared, if the model is to be refit with the gene excluded [1]. For the identified outliers, it cannot be simply ignored, especially with small sample sizes of no more than 6 replicates per condition. Instead, the whole gene is to be removed from the downstream analysis. In our datasets, 302 genes were tagged as outliers and removed from the subsequence analysis.

In Limma-Voom pipeline, the impact of low-level outliers can be reduced by the log transformation of CPM reads, which tends to be robust except for high-level extreme outliers, and does not require any exclusion of genes.

However, both methods have their own shortcomings, especially with small sample size. For DESeq2, the Cook's distance can be hard to measure with extremely small size of less than 3 replicates per condition, as the information is so insufficient that the distance will always change during model refit with exclusion , and hence unable to determine what to be classified as outlier. For Limma-Voom pipeline, as mentioned above, the log transformation is only efficient with low-level outliers, whereas it is unable to lower the influence of high-level extreme outliers. In addition, the model becomes more sensitive with small sample sizes as well [4]. This could be further improved by a more robust voom transformation by β - divergence method [4].

3.1.3. Other Factors Impacting the DEG accuracy

Other than the pipeline-specific factors discussed above that might impact the accuracy of the DEGs identified, there are also some potential common factors.

First, from the experimental design of the RNA-seq, it could not be concluded if all samples were run in the same batch or separately. If the samples were divided into batches with the same arrangement of the subtypes (i.e. each subtype or condition of samples were run in the same batch, and one batch only contains one subtype or condition), then the DEGs identified could be either due to the actual biological difference between the subtypes or batch effect, which could not be differentiated. Hence, some of the result DEGs might not actually have differential expression biologically.

Besides, not much demographic information about the samples could be found from the experimental design. Samples taken from females from different races or with other breast-related conditions may also impact the truthfulness of the DEGs identified.

3.2. Sample Clustering Using Identified DEGs

Based on the DEGs identified by the two packaged, Figure 5 and Figure 6 shows the sample clustering results.

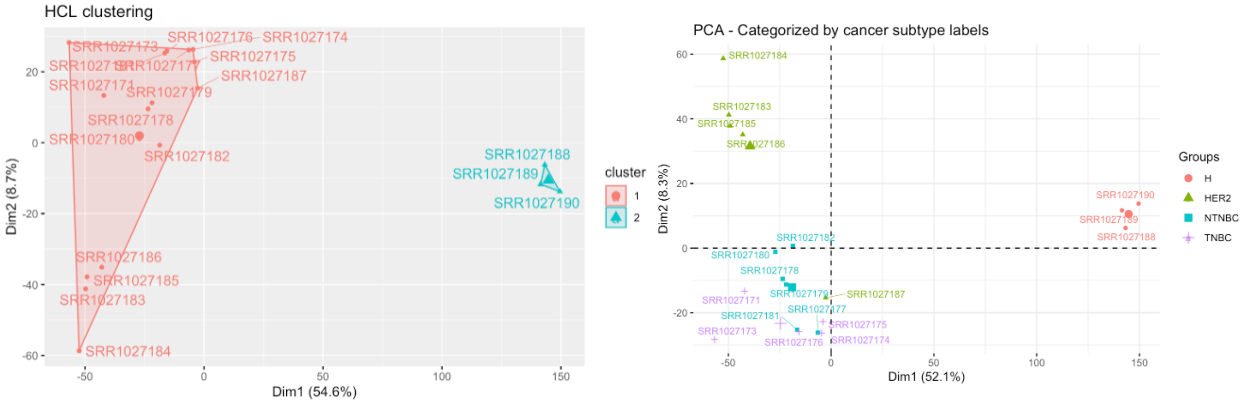

Figure 5. Sample clustering result using DEGs identified by DESeq2, compared against the original subtype labels

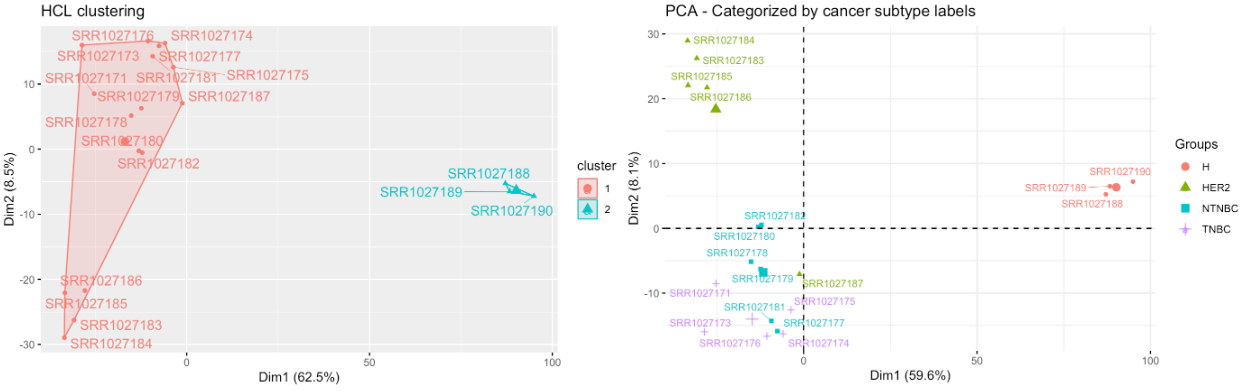

Figure 6. Sample clustering result using DEGs identified by Limma-Voom, compared against the original subtype labels

Two clusters were identified in both scenarios, where one cluster contains only the healthy samples, and the other contains diseased samples, regardless of cancer subtypes. This is because the feature genes were all identified by comparing one of the diseased subgroups with the healthy samples. They mostly carry information on the difference between healthy and diseased groups. Hence, the model is only effective to distinguish if the sample is healthy or diseased. Looking at the PCA plots on the right-hand side, this result is not surprising, as only healthy samples are well clustered on the right-hand side.

For diseased samples, most of HER2 samples are clustered on the top left corner, whereas NTNBC and TNBC samples tend to group together with small inter-cluster distance. This is different from our expectation that HER2 and NTNBC may be more similar. As the definitions of HER2 and NTNBC share the similarity where HER2 is positive in both subtypes, whereas TNBC and NTNBC are exclusive by definition. However, it is worth noticing that PC2 on y-axis only covers around 8% of the variance, while PC1 on x-axis covers more than 50% of variance. This means that the inter-cluster distance between HER2 and TNCB/NTNBC is much smaller than that between diseased and healthy samples.

To distinguish the subtypes among breast cancer samples, additional comparisons are required to determine DEGs among various cancer subtypes.

4. Future Works

In this study, DEGs were identified without differentiating the sign of LFC, which can be used to determine if the gene is up-regulated or down-regulated. Some common DEGs found in this study could have completely opposite expression level in different cancer subtypes. Alternatively, the same gene might have opposite LFC in the same cancer subtypes, which could help with deeper understanding of the difference between the two pipelines used.

In addition, in order to show the difference in DEGs identified in a more human-readable way, the identified DEGs could be used for Gene Ontology (GO) Analysis or Pathway Analysis. The enriched GO terms could be used to identify how the DEGs identified by the two pipelines are different from each other. GO analysis is also based on statistical test and can be done by various tools and libraries, the algorithms of which can be compared as well.

5. Conclusion

In this study, DEGs in three subtypes of breast cancers have been studied against healthy samples using two widely used packages, DESeq2 and Limma-Voom. It has been shown that due to the various differences found the algorithms of the two packages such as gene filtration, count normalisation, statistical modelling and handling of outliers, the result DEGs could be very different. This has also been verified by the code experiment. It has revealed the uncertainty in biological studies. Other than the unstable reproducibility of research result due to the difference in samples and sequencing techniques, the down-stream data analysis could also contribute to the low reproducibility of result. After all, many of the researches have been built upon statistical inference due to the technical challenges in measuring the raw signals. The impact may be particularly drastic for borderline cases, where the LFC of gene expression is small, or the overall expression level is low. Hence, the impact of statistical test or modelling

should always be taken into consideration, and based on the nature of the study, the most appropriate model should be selected to produce relatively more reliable result.

References

[1] Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**:550. 10.1186/s13059-014-0550-8

[2] Law, C.W., Chen, Y., Shi, W. *et al.* voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15,** R29 (2014). https://doi.org/10.1186/gb-2014-15-2-r29

[3] Soneson, C., Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14,** 91 (2013). https://doi.org/10.1186/1471-2105-14-91

[4] Shahjaman M, Manir Hossain Mollah M, Rezanur Rahman M, Islam SMS, Nurul Haque Mollah M. Robust identification of differentially expressed genes from RNA-seq data. Genomics. 2020 Mar;112(2):2000-2010. doi: 10.1016/j.ygeno.2019.11.012. Epub 2019 Nov 20. PMID: 31756426

[5] Evans, H. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics, 19(5), 776–792. https://doi.org/10.1093/bib/bbx008

[6] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F; French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013 Nov;14(6):671-83. doi: 10.1093/bib/bbs046. Epub 2012 Sep 17. PMID: 22988256

[7] Chen Y, Lun ATL and Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]. *F1000Research* 2016, **5**:1438 (https://doi.org/10.12688/f1000research.8987.2)

[8] American Cancer Society. Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society, Inc. 2019.

[9] Soneson, C., Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14, 91 (2013). https://doi.org/10.1186/1471-2105-14-91

[10] EMBL-EBI. (2013). RNA-seq of 17 breast tumor samples of three different subtypes and normal human breast organoids samples. EMBL-EBI Expression Atlas. https://www.ebi.ac.uk/gxa/experiments/E-GEOD-52194/Download

Appendix A

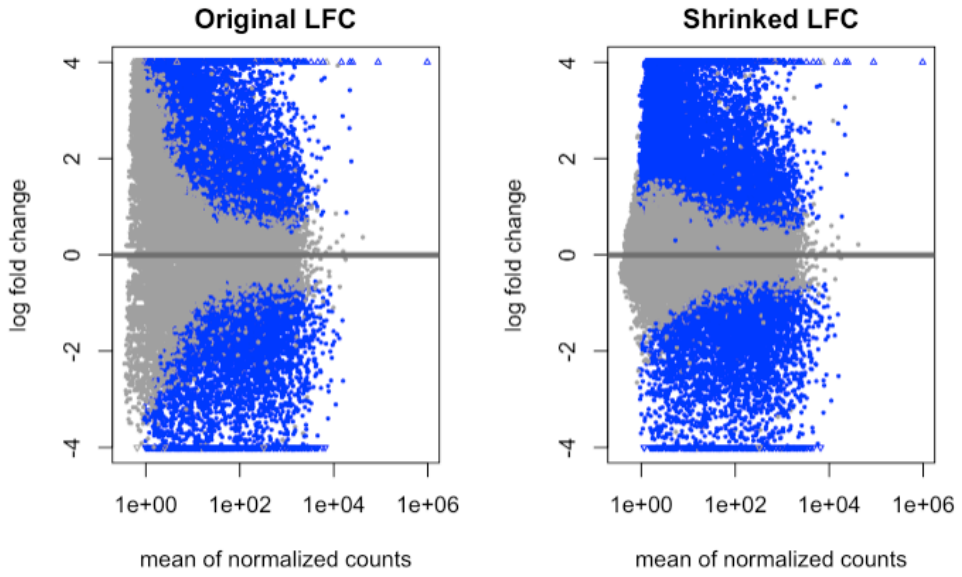

Figure A1.1 MA plot of HER2 against Healthy samples before and after LFC shrinkage. Blue dots represent significant DEGs with adjusted p-value < 0.1

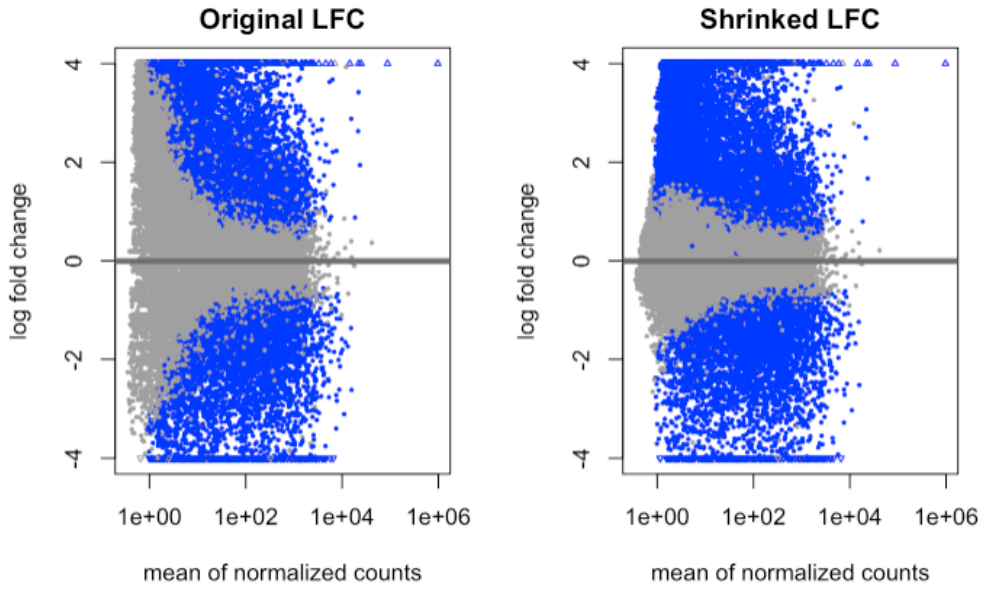

Figure A1.2 MA plot of TNBC against Healthy samples before and after LFC shrinkage. Blue dots represent significant DEGs with adjusted p-value < 0.1

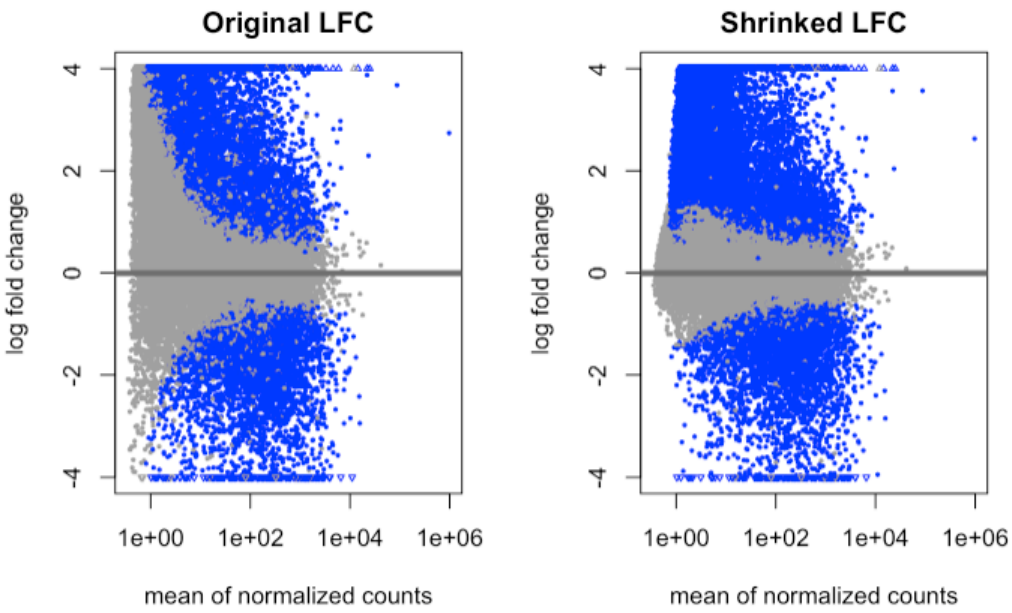

Figure A1.3 MA plot of NTNBC against Healthy samples before and after LFC shrinkage. Blue dots represent significant DEGs with adjusted p-value < 0.1