Limma Package:

Limma is a package for the analysis of gene expression data arising from microarray or RNA-Seq technologies. It has features that make the analyses stable even for experiments with small number of arrays—this is achieved by borrowing information across genes. It is specially designed for analyzing complex experiments with a variety of experimental conditions and predictors. The linear model and differential expression functions are applicable to data from any quantitative gene expression technology including microarrays, RNA-seq and quantitative PCR. Limma can handle both single-channel and two-color microarrays.

Limma uses a range of normalization methods. The methods may be broadly classified into methods which normalize the M-values for each array separately i.e. within-array normalization and methods which normalize intensities or log-ratios to be comparable across arrays i.e. between-array normalization. The scale normalization to RNA-seq read counts, and the TMM normalization method in particular has been found to perform well in comparative studies.

The limma approach to RNA-seq uses the voom transformation which is applied to the read counts. This converts the counts to log-counts per million with associated precision weights. After this, the RNA-seq data can be analyzed as if it was microarray data. Limma trend accommodates the mean variance relationship as a part of the empirical Bayes procedure. Voom estimates the mean variance relationship robustly and it generates a precision weight for each individual normalized organization. Limma package is used to analyze the log-counts after normalization by sequencing depth. This means for example that any of the linear modelling or gene set testing methods in the limma package can be applied to RNA-seq data.

The package can now perform both differential expression and differential splicing analyses of RNA sequencing (RNA-seq) data. The features of package allow users to analyze both RNA-seq and microarray data. Also, the package is now able to go past the traditional gene-wise expression analyses in a variety of ways, analyzing expression profiles in terms of co-regulated sets of genes or in terms of higher-order expression signatures. This provides enhanced possibilities for biological interpretation of gene expression differences.

DESeq2

1. What does DESeq2 do?

2. What normalization methods does it use (does it use a library size normalization)?

3. How does it model RNAseq(limitingvariance,variancestabilizingtransform?,regularizedlogtransform? what does the package even do? I’m not sure, ﬁnd out....look at my code too, see what the functions do)?

4. What type of statistical test does it do?

5. Whats unique about it from the other packages?

6. Look at my code and see what the DESeq2 functions that I used do ... explain some of them... the important ones

DESeq2 package

The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions. The count data is obtained from RNA-seq or any sequencing experiment in the form of a matrix of integer values. The DESqeq 2 model internally corrects for library size, so transformed and normalized values so counts scaled by library size must be used as in input.

The featureCounts function in the Rsubread package, was used for quickly producing count matrices from alignment files. The PlotCounts function normalizes counts by sequencing depth and adds a pseudo count of 0.5 for a log scale plotting.

RNA-seq data sometimes contain isolated instances of very large counts that are apparently unrelated to the experimental or study design, and which may be considered outliers. There are many reasons why outliers can arise, including rare technical or experimental artifacts, read mapping problems in the case of genetically differing samples, and genuine, but rare biological events. The DESeq function calculates, for every gene and for every sample, a diagnostic test for outliers called Cook’s distance. Cook’s distance is a measure of how much a single sample is influencing the fitted coefficients for a gene, and a large value of Cook’s distance is intended to indicate an outlier count.

Plotting the dispersion estimates is a useful diagnostic. The dispersion plot is typical, with the final estimates shrunk from the gene-wise estimates towards the fitted estimates. Some gene-wise estimates are flagged as outliers and not shrunk towards the fitted value. The amount of shrinkage can be more or less than seen here, depending on the sample size, the number of coefficients, the row mean and the variability of the gene-wise estimates. A local smoothed dispersion fit is automatically substituted in the case that the parametric curve doesn’t fit the observed dispersion mean relationship.

The differential expression analysis is wrapped into a single function in DESeq 2, which gives the estimates of the logarithmic fold change by log2. The analysis is fast and also experiments with larger samples have parallelized computation. For that parallel = TRUE which distributes computation across cores specified by register function of Bioparallel. The results can be ordered by the smallest adjusted p- value.

EdgeR

edgeR provides statistical routines for assessing differential expression in RNA-Seq experiments or differential marking in ChIP-Seq experiments. edgeR is concerned with differential expression analysis rather than with the quantification of expression levels. It is concerned with relative changes in expression levels between conditions, but not directly with estimating absolute expression levels. This greatly simplifies the technical influences that need to be taken into account, because any technical factor that is unrelated to the experimental conditions should cancel out of any differential expression analysis.

edgeR adjusts any differential expression analysis for varying sequencing depths as represented by differing library sizes. This is part of the basic modeling procedure and flows automatically into fold-change or p-value calculations. In edgeR, normalization takes the form of correction factors that enter into the statistical model. Such correction factors are usually computed internally by edgeR functions, but it is also possible for a user to supply them. The correction factors may take the form of scaling factors for the library sizes. Normalization in edgeR is model-based, and the original read counts are not themselves transformed. This means that users should not transform the read counts in any way before inputting them to edgeR

Differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. As well as RNA-seq, it be applied to differential signal analysis of other types of genomic data that produce counts, including ChIP-seq, SAGE and CAGE. edgeR uses the quantile-adjusted conditional maximum likelihood (qCML) method for experiments with single factor. Compared against several other estimators (e.g. maximum likelihood estimator, Quasilikelihood estimator etc.) using an extensive simulation study, qCML is the most reliable in terms of bias on a wide range of conditions and specifically performs best in the situation of many small samples with a common dispersion, the model which is applicable to NextGen sequencing data. edgeR uses the Cox-Reid profile-adjusted likelihood (CR) method in estimating dispersions. The CR method is derived to overcome the limitations of the qCML method as mentioned above. It takes care of multiple factors by fitting generalized linear models (GLM) with a design matrix.

Generalized linear models (GLMs) are an extension of classical linear models to nonnormally distributed response data. GLMs specify probability distributions according to their mean-variance relationship, for example the quadratic mean-variance relationship specified above for read counts.