BIOCONDUCTOR OBJECTS

Everything in r is an object but bioconductor objects are data structures containing many attributes
Both ExpSet and SumExp have the same structure :

- a matrix of data, rows are genomic features, and columns are samples
- a table of data about the samples (columns)
- a table of data about the features (rows)

ExpressionSet	SummarizedExperiment
Created to store microarray data→ column stores values from a particular microarray Came first The components of an ExpressionSet are the matrix of data, the table about the samples (phenotypic data) and the table about the features (feature data). ExperimentData(x) → Minimum Information About a Microarray Experiment	More for the sequencing era. Rows correspond to particular GRanges (number of RNA-seq reads that can be assigned to a particular gene, and the location of the gene in the genome.) This particularity enables faster data analysis In SummarizedExperiment, these are called assay, colData and rowData or rowRanges If we don't have ranges, we can just put a table on the side of the SummarizedExperiment by specifying rowData. we know metadata about the chromosomes, and the version of the genome

RNA-SEQ DATA ANALYSIS

<u>Clean the data before performing analysis:</u>

- Remove duplicates by counting samples per condition, observe the replicates and keep only the first
- Ensure the variables are of the appropriate class
- Metadata cleanup

Then we can continue to subset the information of interest

<u>Remove low expressed genes (CPM)</u>
with function in EdgeR !!
From a biological point of view, a gene must be expressed at some minimal level before it is likely to be translated into a protein or to be considered biologically relevant. From a statistical point of view, genes with consistently low counts are very unlikely assessed as significantly DE because low counts do not provide enough statistical evidence for a reliable judgement to be made. Such genes can therefore be removed from the analysis without any loss of information.

Data Normalisation

TMM (function in R) calculates a set of normalisation factors, one for each sample, to eliminate composition biases between libraries. The product of these factors and the library sizes defines the effective library size, which replaces the original library size in all downstream analyses. Result is a list

Exploratory data analysis

1) We can use PCA to reduce the dimensionality of our data, from thousands of genes to 2 principal components, and thus represent our individuals in 2D (losing some variability with respect to the original data, but gaining interpretability). The closer the individuals in this reduced space, the more similar they are with respect to the expression of their genes, and vice versa. PCA can help us to identify subgroups of individuals, outliers and batch effects.
2) Clustering and Heatmap
   What are rows/columns? What does the heatmap represent? Which are the main clusters? Is there a clear separation between clusters? What does this mean? Which variables separate the clusters? In addition to the variable of interest ("cohort"), are there any potential confounders? Do you see any outlier sample?
3) Outlier Removal
   Repeat the same process again

Differential Gene Expression analysis

1) Design matrix and contrast matrix.
   The design matrix has columns associated with the parameters and rows associated with samples. If the estimated parameters are not of direct interest, a contrast matrix can be used to calculate contests of the parameters.
   Indeed when looking at the PCA and clustering results, we identified some potential confounders, i.e. race and age. They should be included in the model too, so that their potential effects on gene expression are taken into account, even if we are not interested in building contrasts for these variables.
2) Removing mean-variance relationship from count data
   For RNA-seq count data, the variance is not independent of the mean
   Methods that model counts using a Negative Binomial distribution (*edgeR*, *DESeq2*) assume a quadratic mean-variance relationship. In *limma*, linear modelling is carried out on the log2(CPM) values, which are assumed to be normally distributed, and the mean-variance relationship is accommodated using precision weights calculated by the voom function. When operating on a DGEList-object, `voom` converts raw counts to log2(CPM) values by automatically extracting library sizes and normalisation factors from the object itself.
   High biological variation → flatter trends, where variance values plateau at high expression values.
   Low biological variation → sharp decreasing trends
   The voom-plot provides a visual check on the level of filtering performed.

3) Fitting linear models for comparisons of interest
   →lmFit - contrasts.fit
4) Examining the nb of DE genes
   topTable → top DE genes
5) Useful graphical representations of de results
   glMDPlot → interactive
   Volcano
   Boxplot
   Clustering

Biological SIgnificance analysis

1) GO analyses can be conveniently conducted using the goana function. The top most significantly enriched GO terms can then be viewed with topGO. However, the goana function uses the NCBI RefSeq annotation and requires the use of Entrez Gene IDs, while we are working with Ensembl IDs.
2) Kegg Pathways
   curated database of molecular pathways and disease signatures. A KEGG analysis can be done exactly as for GO, but using the kegga function: