Analyzing RNA-seq data with Qlucore Omics Explorer

Introduction

This document describes a workflow for analyzing RNA sequencing (RNA-seq) data with Qlucore Omics Explorer (QOE). We first give a brief introduction to RNA-seq data. Then, we discuss the recommended pre-processing steps that should be applied before the RNA-seq data are imported into Qlucore Omics Explorer and the suitable normalization of the data within the software.

RNA-seq data

In recent years, RNA sequencing has emerged as an alternative to microarrays for quantification of gene expression. RNA-seq measures relative transcript abundances by high-throughput sequencing of complementary DNA (cDNA) which is generated from the RNA of interest through reverse transcription. The number of sequenced reads that align to a specific transcript is used as a measure of the abundance of that transcript in the sample.

A typical RNA-seq experiment follows the steps outlined in Figure 1. The starting point is an extracted sample of RNA, which is possibly enriched for mRNA using poly(A) selection. The isolated RNA is reverse transcribed to cDNA which is fragmented and sequenced in parallel with a next generation sequencing technique. The result from the sequencing is a large number of *reads*. Each of these reads is aligned to a reference genome and further mapped to a particular transcript. Then, the number of reads mapping to each transcript (commonly called the *count* for the transcript) is determined and used as a measure of the abundance of the transcript in the sample. After this step, the data are visualized and analyzed, for example by searching for genes that are differentially expressed between different conditions.

In the rest of this document, we focus on the analysis steps (7 and onwards), and therefore assume that the data have been appropriately quality checked and that they are given as counts for a collection of genes in a number of samples.

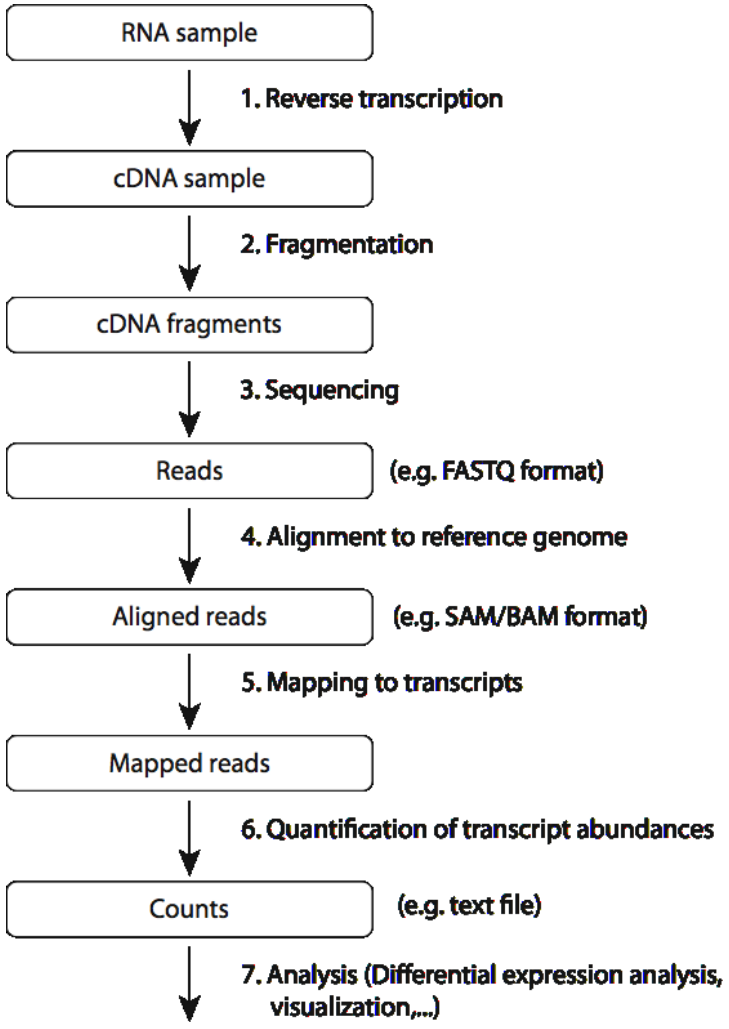

Figure 1. The steps in a typical RNA-seq experiment.

Required data pre-processing

Most likely, the total number of mapped reads (the sequencing depth) is different for the different samples. In order to make the observed counts for a gene comparable across samples, this needs to be accounted for. Moreover, the number of reads mapping to a longer gene is higher than the number of reads mapping to a shorter gene which is equally highly expressed. Therefore, the length of the gene (the number of nucleotides) should be taken into account in order to give the genes equal influence in multivariate methods such as Principal Components Analysis (PCA).

As discussed above, we assume that the data are given as counts and collected into a matrix in R or as text file where each row represents one transcript and each column represents one sample. The element in the i'th row and the j'th column corresponds to the number of counts for the i'th transcript in the j'th sample.

We recommend that the observed gene counts are pre-processed according to the following steps before they are imported into Qlucore Omics Explorer:

1. Normalize the counts for each sample by dividing with a correction factor based on the sequencing depth of the sample. The most straight-forward estimate of sequencing depth is obtained by the total number of mapped reads. However, this measure can have serious drawbacks if the pool of expressed RNA differs between two samples and an additional correction factor, called "TMM" (trimmed mean of M values) can be estimated as described by Robinson and Oshlack (2010).

2. Normalize the observed count for each gene by dividing with the length of the gene (the number of nucleotides).

Given that the data is available as a matrix in R, these two normalization steps can be applied using the normalization script available from www.qlucore.com (see the support pages).

Note that if the correction factor in step 1 is defined by the total number of mapped reads for a sample, the data normalized by steps 1 and 2 are equivalent to the RPKM values discussed by Mortazavi et al (2008).

Normalization in Qlucore Omics Explorer

Since the data from an RNA-seq experiment come in the form of counts, they are usually assumed to follow either a Poisson or a Negative Binomial distribution. For PCA, the underlying distribution of the data does not affect the interpretation of the principal components as the variable combinations with highest variances in the data. The statistical tests implemented in Qlucore Omics Explorer, on the other hand, are based on the assumption that the data are drawn from a Normal distribution. For large sample sizes, the statistical tests are robust against the underlying distribution of the data. However, also for smaller sample sizes, we can often approximate the Poisson and the Negative Binomial distributions reasonably well by Normal distributions, and the tests provided by Qlucore Omics Explorer can be used for approximate differential expression analysis. In general, the approximations are better for genes with high mean count and low variance (for the Negative Binomial distribution). To further increase the accordance with a normal distribution the normalized counts can be log transformed in Qlucore Omics Explorer. One should, however, note that this transformation is not applicable if there are zero counts in the data.

As for microarray data, the RNA-seq data can be normalized to zero mean and unit variance in the PCA. This means that focus is put on the correlation structure between the variables instead of their individual variances. As for microarray data, it is important to note that variables with very low variances (typically noise variables) tend to be enhanced in such standardization. This can be remedied by applying a variance filter to the data, which removes the variables with the lowest variances.

Analysis

At this point we have imported data into Qlucore Omics Explorer and also described how the data should be pre-processed and normalized. From this point and onwards all the tools in Qlucore Omics Explorer can be used for RNA-seq data.

References

A Mortazavi, B A Williams, K McCue, L Schaeffer and B Wold: Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 2008, 5(7):621-628

M D Robinson and A Oshlack: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 2010, 11:R25.

Disclaimer

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing.

Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document.

Qlucore Omics Explorer is only intended for research purposes.