**Onco-Wiki RNA-Sequencing**

This guide uses CLC Genomics Workbench 8.0 which requires a paid license. A popular free alternative is the Bowtie/Tophat/Cufflinks sequencing pipeline.

1. **Data Storage**

Prior to RNA-Sequencing, ensure that you save your data to an external hard drive with up to 20GB of available data for each sample. Also, ensure that you have adequate local disk space for reference genome tracks and live process runs of each sample (at least 20GB per sample).

1. **Import Data**

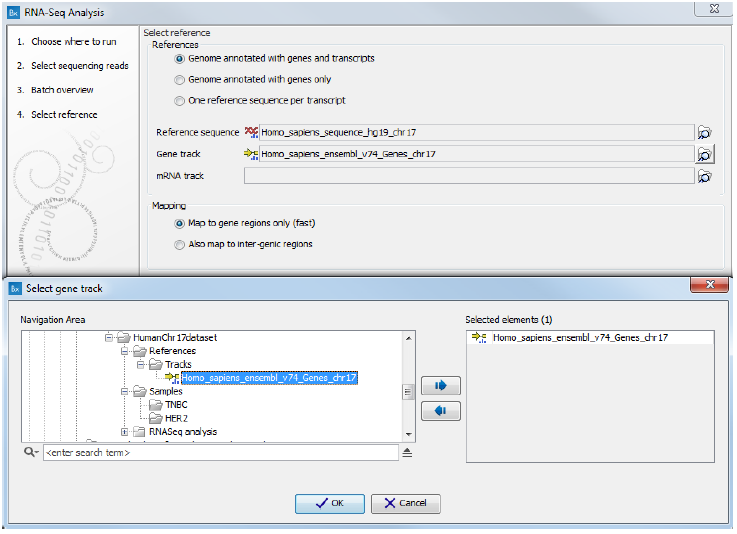
Import downloaded data via the toolbar File | Important | Standard Import or FASTA special import if downloaded data was in FASTA format. Ensure that the file extension of each file is either in .fastq.gz or .fasta.gz. Save imported data. It is possible to do multiple data sample imports if you have no other tasks. This process should take no longer than 10-15 minutes per sample.

1. **Genome Tracks**

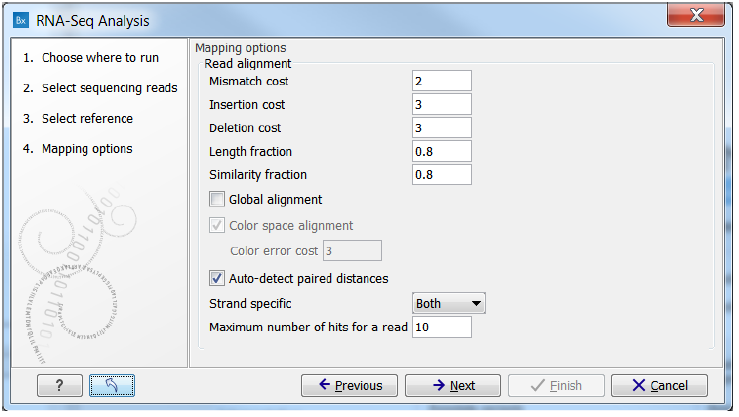
Select Download | Download Reference Genome Data | Animals (mammals) and Homo Sapiens (hg19) from the drop down list. Download the hg19 reference genome track from NCBI and the annotated gene track from Ensembl.

1. **RNA-Seq Analysis**

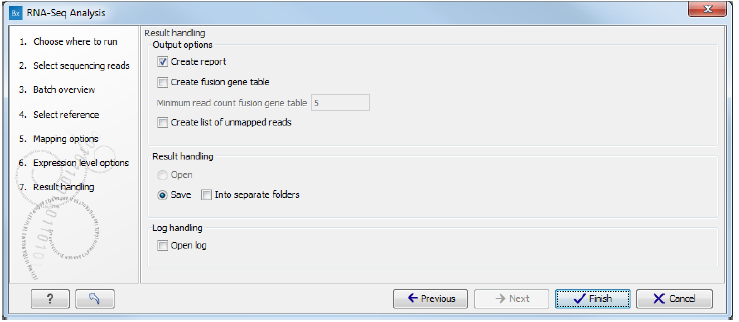
Go to Toolbox | Expression analysis | RNA-Seq analysis. Click the batch button if you plan to sequence multiple samples and select your recently imported data files (not your downloaded data files).



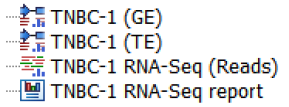
Select your downloaded hg19 reference genome and corresponding gene track from Ensembl and Map to gene regions only (fast).



Select default mapping settings by clicking the blue button as above. Change the maximum number of hits for a read to 1. This ensures that each read lies unique to the reference genome and non-specific matches are not accounted for.



Use default settings for the remaining screens and that you have selected ‘Create Report.’ After saving, RNA-Seq analysis is now running and will last up to 75-90 minutes per human sample. RNA-Sequencing is a CPU intensive process so expect other applications to run more slowly.



On completion, four files appear with (GE) indicating gene expression across 57,773 gene spliced variants and (TE) indicating transcript expression.

1. **Download Dataset**

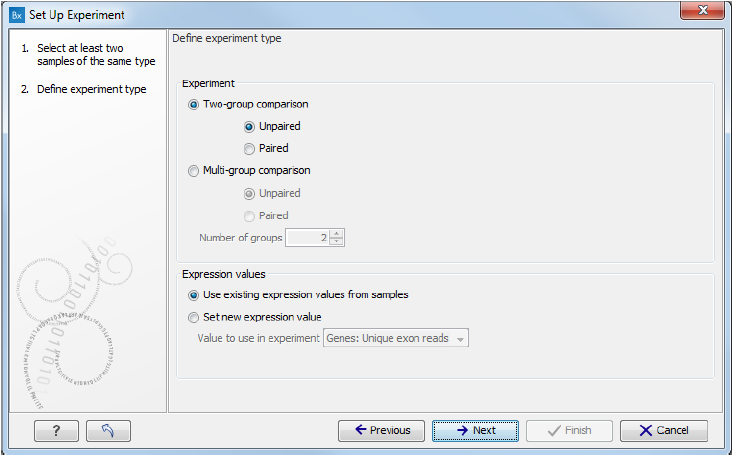
The Healthy dataset for CLC Genomics Workbench import can be downloaded from:

[**http://oncowiki.co.uk/support/healthyrawclc.clc**](http://oncowiki.co.uk/support/healthyrawclc.clc)

1. **Pair-wise Experiment**

In order to generate a fold change measurement between Tumour vs. Healthy, a pair-wise experiment is required.

Go to Toolbox | Expression Analysis | Set Up Experiment. Select your recently downloaded Healthy dataset first and then an individual RNA-Sequenced (GE) sample



Use default settings as above and assign names to your two groups such as Healthy and Tumour. Save and open your comparison. Column E (Experiment – Fold Change (original values)) is your fold change between the two samples. Export and save this file as an .xlsx file.

1. **Fold Changes Explanation**

Fold changes are measure of how many times bigger the mean expression value is. We have measured expression using total counts of reads mapped to the exons of genes and normalised through RPKM. If the mean expression value in:

The title of your file e.g. Healthy vs. Tumour correspond to Group 1 (Healthy) and Group 2 (Tumour).

Group 2 > Group 1: Group 2 / Group 1 (+ve sign)

Group 1 > Group 2: Group 1 / Group 2 (-ve sign)

If using the 'Healthy vs. Tumour' dataset, a positive fold change indicates that the tumour gene is upregulated and a negative fold change indicates that the tumour gene is downregulated with comparison to the healthy.