

RNA-Seq Analysis

Introduction

This tutorial describes the RNA-Seq analysis workflow in Avadis NGS. It assumes that Avadis NGS is installed and the steps listed in the 'Getting Started' tutorial have been completed.

Dataset

This tutorial is based on an RNA-Seq study to find differences between CD4 and Jurkat cell lines (<http://www.ncbi.nlm.nih.gov/pubmed/19528076>). This is a single-end data generated from a GAlIx. The dataset zip file (`rnaseq-small-dataset-illumina.zip`) contains the SAM files for these samples. The reads were aligned against the hg18 assembly. For the purpose of this tutorial, we will consider data that aligned to chr1 of the genome.

Goals

In this tutorial we will perform gene quantification, differential expression and differential splicing steps as part of the RNA-Seq analysis.

Sample import

For this tutorial, we need to create an RNA-Seq analysis experiment with UCSC annotations and other specific parameters as shown in the image below.

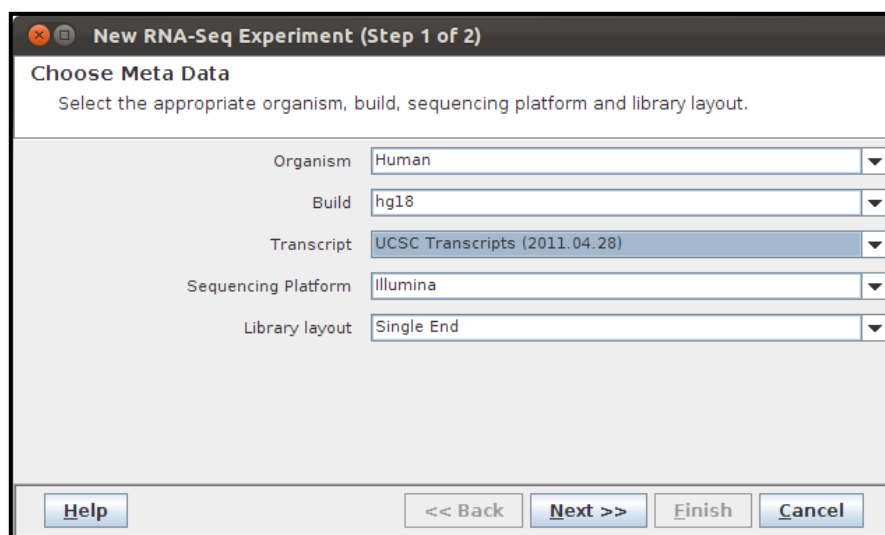


Figure 1: RNA-Seq experiment creation

While loading the SAM files present in the zipped file `rnaseq-small-dataset-illumina.zip` (using the 1st icon from the left), we have an option of renaming the samples to make them easily distinguishable.

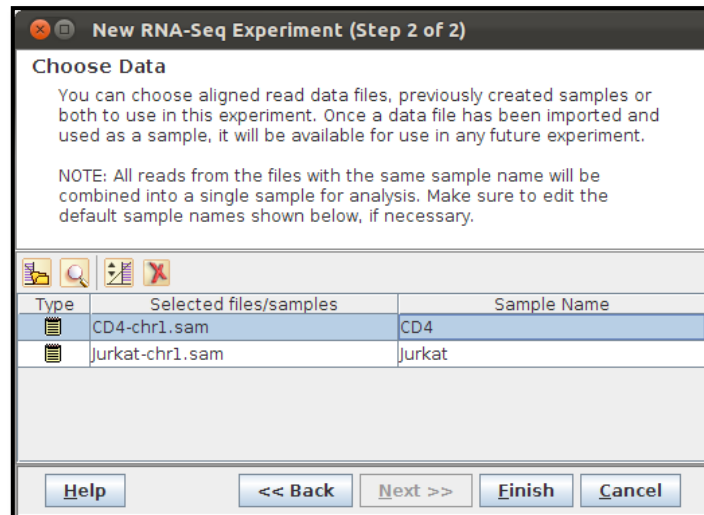


Figure 2: Renaming the samples

After the experiment is created, we can find out the read count in each sample by double-clicking on the 'All Aligned Reads' list in the experiment navigator. For this dataset, the number of reads in each sample is around 2.3 million. Also since it is non-directional RNA-Seq, there would be approximately equal number of +ve and -ve aligned reads.

Quantification

A key step in progressing further is to quantify expression levels of genes, exon partitions and transcripts, in preparation for downstream analysis. The expression values obtained by quantifying the association of reads to genes and exons are called 'raw counts', and can be normalized using one of multiple normalization options available: DESeq, TMM, Quantile, or Sample Read Count. For this tutorial, we will run quantification (using the Analysis → Quantification workflow step) with the default options. The experiment navigator after quantification would look like this:

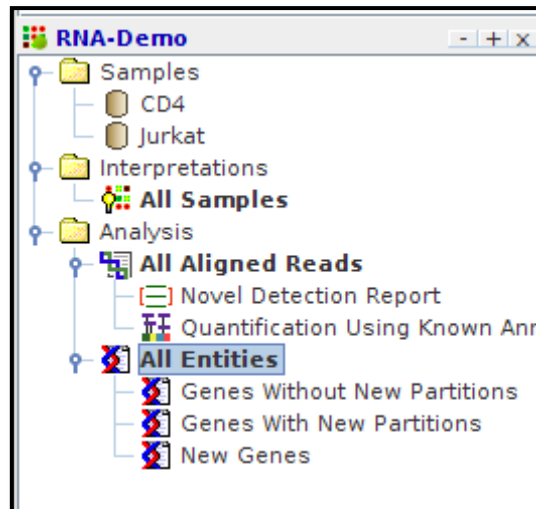


Figure 3: Post-quantification navigator

Novel exon

Expressed regions that do not overlap exonic regions (of the chosen transcript model) are shown in the Novel Detection Report. We can prioritize these regions using the following criteria

- regions with good conservation (this requires the conservation score annotations to be downloaded from the Annotations Manager)
- sufficient number of reads

Known gene quantification results

We can look at the normalized values associated with the genes using the View -> Normalized Signal Values menu. In addition one can also examine the raw counts, drill-downed to the level of individual exonic regions using the View → Raw Counts menu option. These options can also be accessed by right clicking on the quantification node in the experiment navigator.

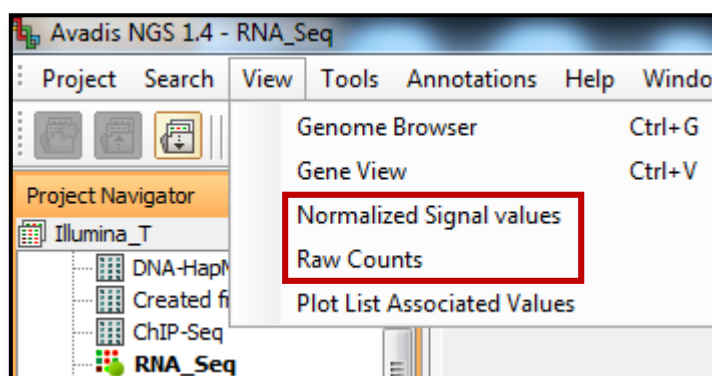


Figure 4: View Options

Filter Genes

Before looking for genes differentially expressed across these two conditions, we could use a filter to retain only those genes which have at least 20 reads mapping to them in at least one of the two conditions (using the Quantification → Filter genes by Expression workflow step).

Figure 5: Filter by expression

This would leave us with 1518 genes out of the total of 33218 genes. Note that the profile plot will automatically update itself to show only the genes in the active entity list. The active entity list is the entity list that is currently selected in the experiment navigator on the left.

Differential expression

For differential expression, we use the Experimental Setup → Experiment Grouping workflow step to specify the parameters for the samples.

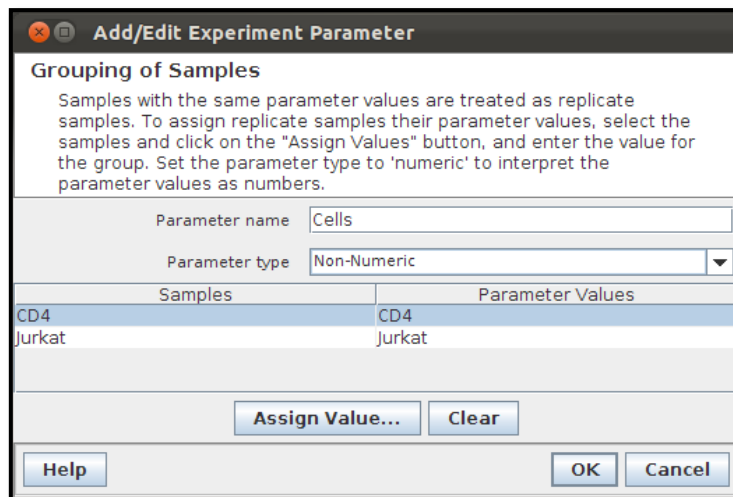


Figure 6: Sample parameters

We can then create an interpretation based on this parameter.

On the filtered entity list, fold change analysis can be used to find the subset having two fold differences between the two conditions (from Expression Analysis → Fold Change in the workflow navigator).

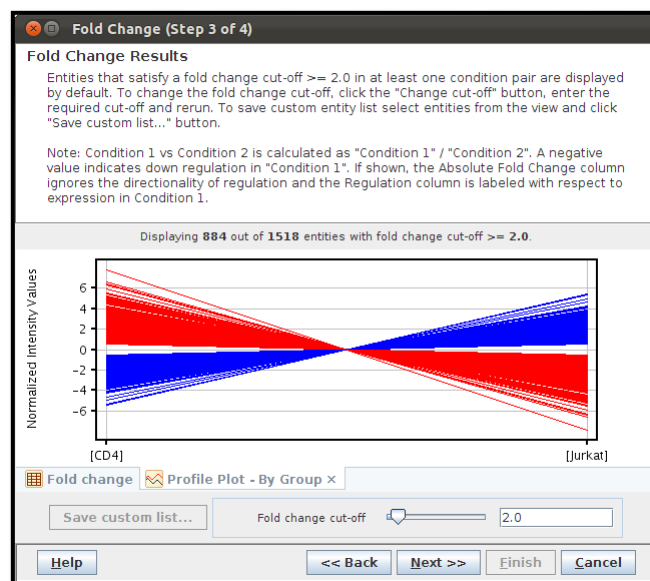




Figure 7: Fold change analysis

This would leave us with 884 genes out of the filtered set of 1518 genes.

Genes with more than two fold change, can be saved and GO analysis can be used on the resulting entity list to find processes/pathways which are different between the two cell lines.

Differential Splicing

Differential splicing analysis can be run on the filtered entity list (using the Expression Analysis → Differential Splicing step in the workflow navigator). The splicing index default cut-off of 0.25 can be changed to 0.5 to retain genes with a fair amount of differential splicing activity.

Some of the genes which are differentially spliced could have the major contributor be an isoform that is not in the transcript model. The novel isoform could either be a combination of known exons, or it might involve novel exons. As a first step, we can ignore the genes which have novel exons and which are differentially spliced. Launch the Venn diagram (using the icon  from the toolbar) and choose the 'Genes with New Partitions' and 'Splicing Analysis result: Splicing Index cut-off ≥ 0.50 ' entity lists as inputs. Then select the blue area in the Venn diagram to save a new list (using the Create Entity List from Selection toolbar button () on the Venn diagram view) containing only those genes that are differentially spliced within the known transcript model.

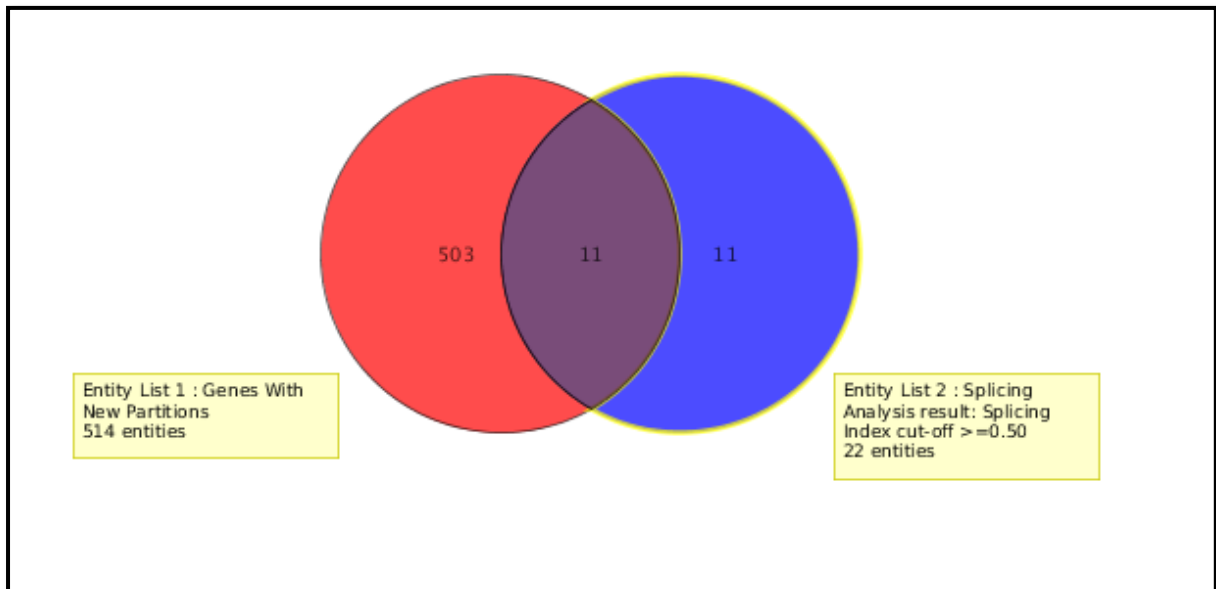


Figure 8: Venn diagram showing the novel and differentially spliced genes in the intersection

The 'Gene View' can now be launched by right-clicking on the differentially spliced known genes. This view is a comprehensive one that shows via multiple views the change in transcript proportions across conditions shown by all the transcripts for a gene.

In the below image, we can see the "Transcript Read Densities" tab showing a case wherein the gene expression shows a very slight variation whereas one transcript show a very marked variation in its expression pattern across the two samples.

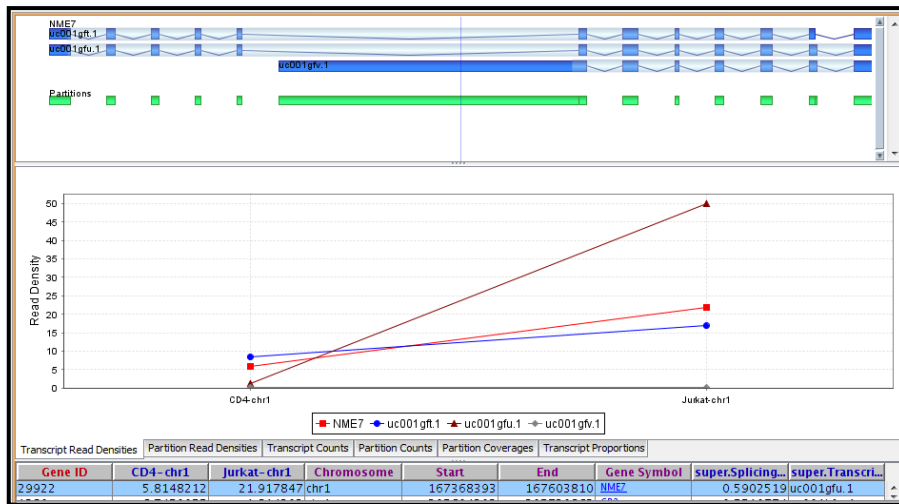


Figure 9: Transcript Read Densities for NME7 gene

Likewise the partition read densities tab would show the expression in each condition for the gene partitions across the transcripts.

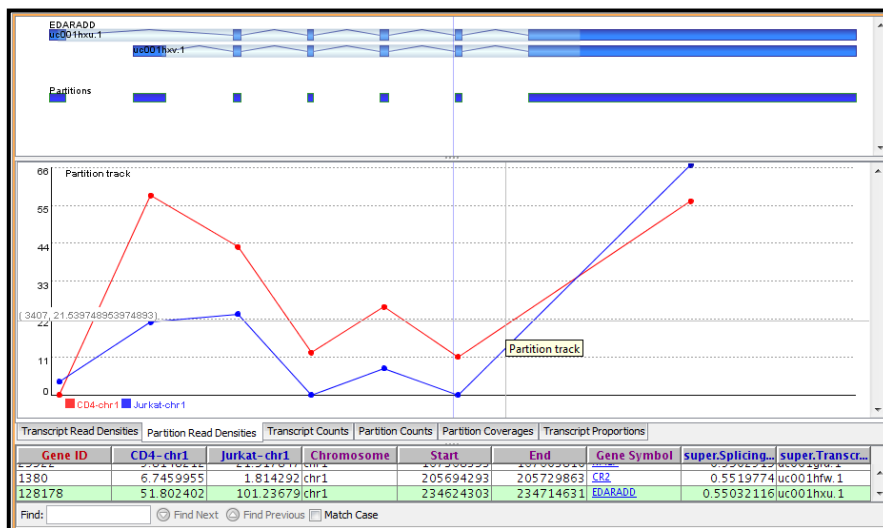


Figure 10: Partition Read Densities for EDARADD gene

The samples in the tutorial have been modified to contain only the reads from chr1. The whole genome RNA-Seq samples are also included in the `rnaseq-full-dataset-illumina.zip` file, should you want to try the analysis on the full dataset.

This is a very brief overview of the RNA-Seq experiment workflow in Avadis NGS. For more details or clarifications, please revert back (sales@avadisngs.com or support@avadisngs.com) and we will address your queries.