Small RNA Analysis

Introduction

This tutorial describes the small RNA analysis workflow in Avadis NGS using data generated from a MiSeq run. 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 a small RNA study generated from a MiSeq run. It contains 4 samples from the brain, kidney and lung tissues – with two replicates from the brain. The dataset file (`smallrna-miseq-dataset-illumina.zip`) contains the MiSeq run folder for the data of this experiment. Note that it has been trimmed to contain only the raw files needed for this tutorial.

Goals

In this tutorial, we will do small RNA alignment and then use the aligned data for subsequent analysis including quantification of the small RNA genes, differential expression and target prediction. For the alignment, we will be requiring a 64-bit machine, so in case you have a 32-bit machine, you will be able to do only the analysis part.

Experiment creation

A new small RNA alignment experiment can be created from Project→ New experiment. In this wizard select the experiment type as Small RNA alignment.

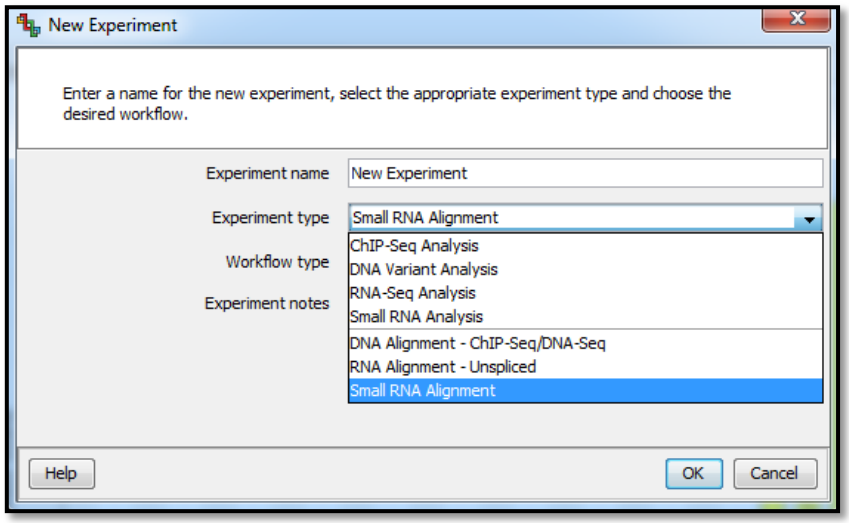


Figure 1: Experiment selection

In the next step, the hg19 build along with other parameters needs to be selected for experiment creation.

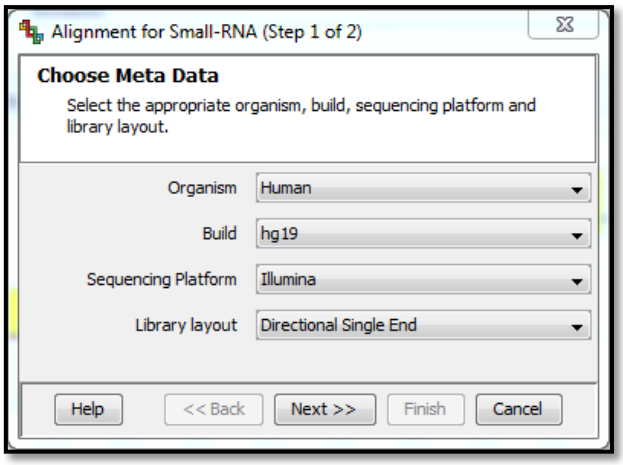


Figure 2: Experiment creation parameters

In the sample selection wizard, choose the Miseq button and provide the file path to the folder generated after the MiSeq run (`110512_BoltM8_0073_AFCA010H_smallRNA`). Then the sample present in the MiSeq folder will be listed below and clicking on 'Finish' would create a small RNA alignment experiment.

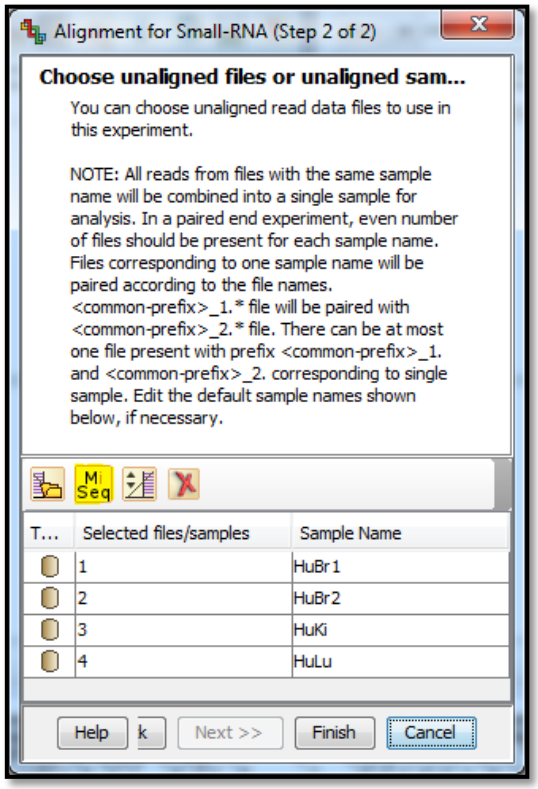


Figure 3: Sample selection

Pre-Alignment QC

Before proceeding with alignment, we can look at various QC plots (using the Pre-Alignment QC workflow steps). In this tutorial, when we look at the 'Base Quality by Position', a dip is seen in the quality at the 3' end of the reads.

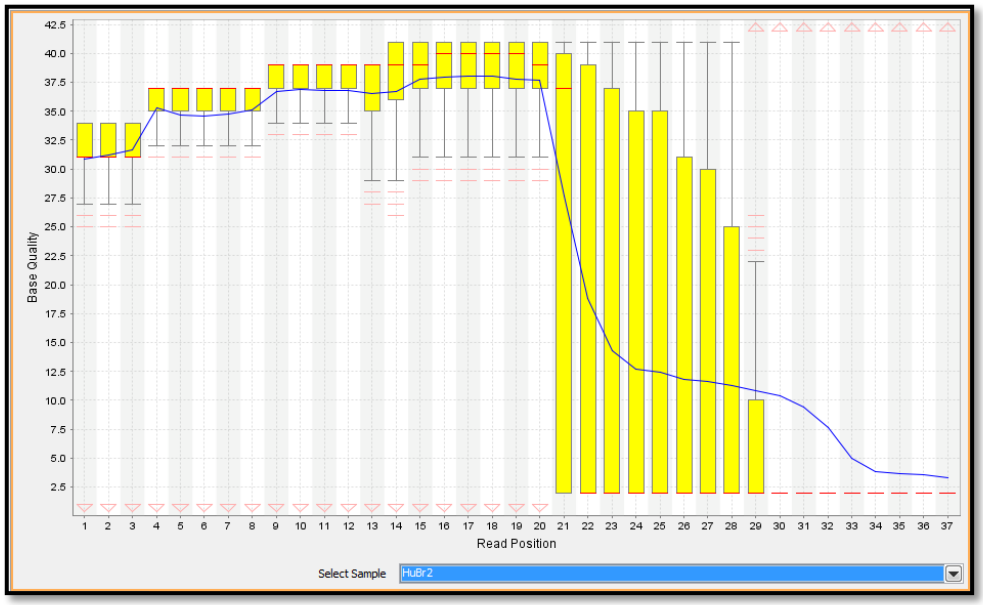


Figure 4: Base quality by position

When we look at the sample files in the MiSeq folder, we realise that the adapter sequences have been trimmed and therefore the dip in quality. This is a crucial piece of information as it would help us fine tuning the alignment parameters for optimum alignment.

Alignment

Alignment is done using the in-built alignment algorithm COBWeb (invoked from Alignment → Run Alignment from the workflow navigator). The name loosely comes from the approach taken to developing the algorithm - a computationally-optimized Burrows Wheeler transform (BWT).

In the first step of the wizard, we select all the samples for alignment. In the second step we will go with the default prameters for the number of mismatches allowed.

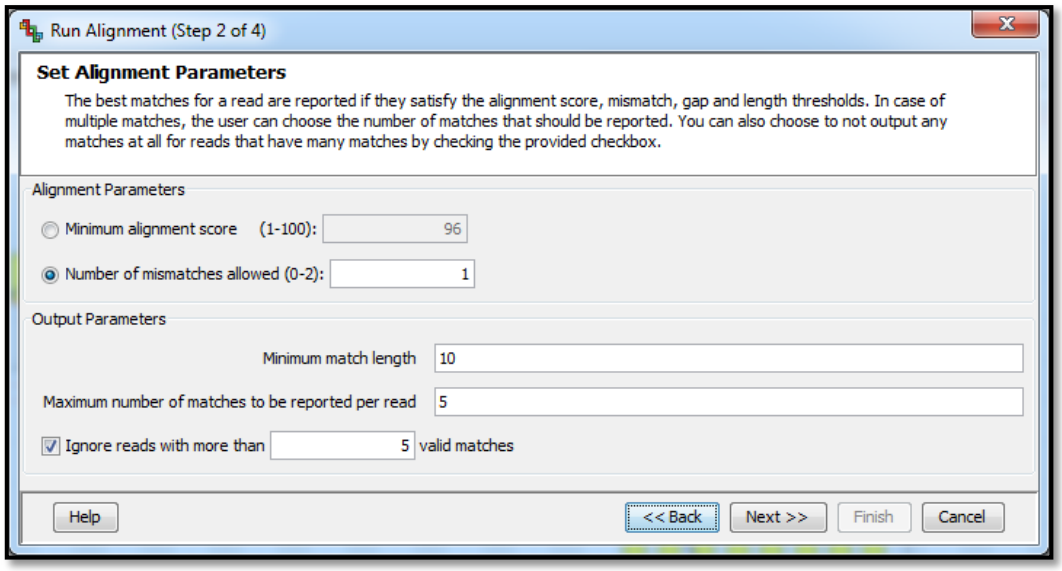


Figure 5: Alignment parameters

In the next step, we will have to specify the trimming parameters. If an adapter sequence is present we can specify that in this step. For this dataset, since the adapter has been trimmed and replaced by Ns, there is a dip in quality at the 3' end of the read. So we can do quality trimming at the 3' end by specifying the minimum quality to be 5.

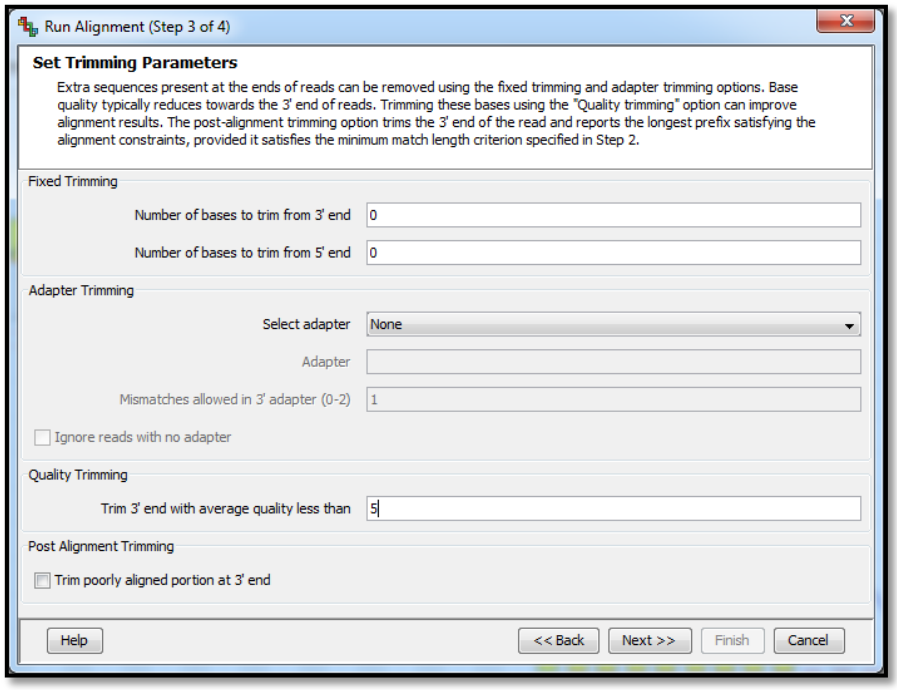
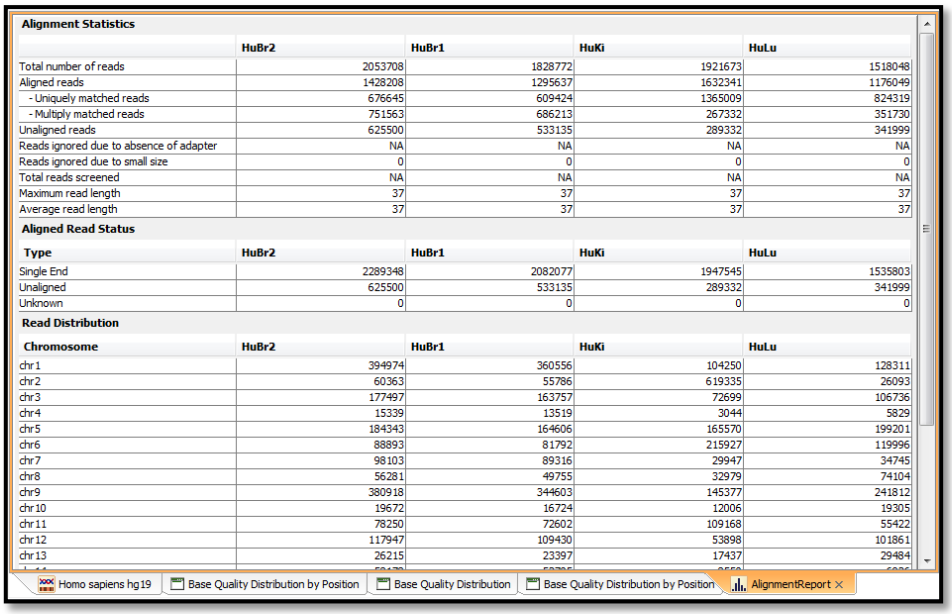


Figure 6: Trimming parameters

The next step provides us an option to select a screening database. Reads matching the screening database will be dropped from alignment. For this dataset, we can ignore this option and click on 'Finish'. After the alignment is finished, we can see the alignment report in the main view and an 'All Aligned Reads' list gets created in the experiment navigator.



Alignment Statistics

	HuBr2	HuBr1	HuKi	HuLu
Total number of reads	2053708	1828772	1921673	1518048
Aligned reads	1428208	1295637	1632341	1176049
- Uniquely matched reads	676645	609424	1365009	824319
- Multiply matched reads	751563	686213	267332	351730
Unaligned reads	625500	533135	289332	341999
Reads ignored due to absence of adapter	NA	NA	NA	NA
Reads ignored due to small size	0	0	0	0
Total reads screened	NA	NA	NA	NA
Maximum read length	37	37	37	37
Average read length	37	37	37	37

Aligned Read Status

Type	HuBr2	HuBr1	HuKi	HuLu
Single End	2289348	2082077	1947545	1535803
Unaligned	625500	533135	289332	341999
Unknown	0	0	0	0

Read Distribution

Chromosome	HuBr2	HuBr1	HuKi	HuLu
chr1	394974	360556	104250	128311
chr2	60363	55786	619335	26093
chr3	177497	163757	72699	106736
chr4	15339	13519	3044	5829
chr5	184343	164606	165570	199201
chr6	88893	81792	215927	119996
chr7	98103	89316	29947	34745
chr8	56281	49755	32979	74104
chr9	380918	344603	145377	241812
chr10	19672	16724	12006	19305
chr11	78250	72602	109168	55422
chr12	117947	109430	53898	101861
chr13	26215	23397	17437	29484

Figure 7: Alignment report

Post-Alignment QC

After the alignment, we can also take a look at the post alignment QC options such as the 'Base Quality by Tile'. The plot shows average base qualities rendererd as a heatmap with user controls for choosing the sample and lane of interest. Rows in the heatmap correspond to tiles of the lane under consideration, while columns represent flow cycles. Each cell is colored according to the average quality of all the bases that fall in the specific cycle of the specific tile.

The below figure shows the Base Quality QC plot for this dataset in which the qualities are good in the earlier flow cycles and degrade marginally towards the end.

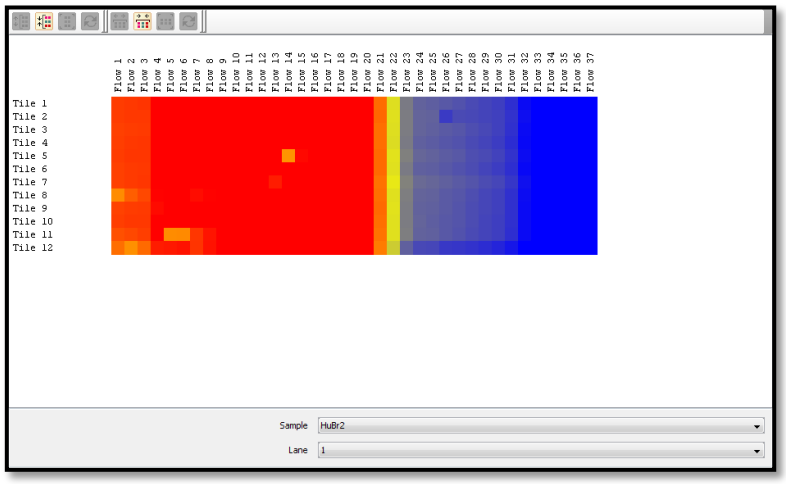


Figure 8: Base quality by tile

For doing further downstream analysis, a small RNA analysis experiment needs to be created (using Utilities → Create small RNA experiment workflow step).

Analysis Experiment

The analysis experiment for this dataset can either be done after the alignment experiment or directly from the BAM files (present in `smallrna-sam-dataset-illumina.zip`). The latter approach would involve selecting the new experiment option from the 'Project' tab in the menu bar, selecting the type as 'Small RNA Analysis Experiment' and loading the BAM files instead. The subsequent steps will be similar to both approaches.

In the first step, we can specify the build as hg19 and the transcript level as RefSeq. As mentioned in the 'Getting started' tutorial, we need to download small RNA annotations for the hg19 build from the 'Annotations Manager'.

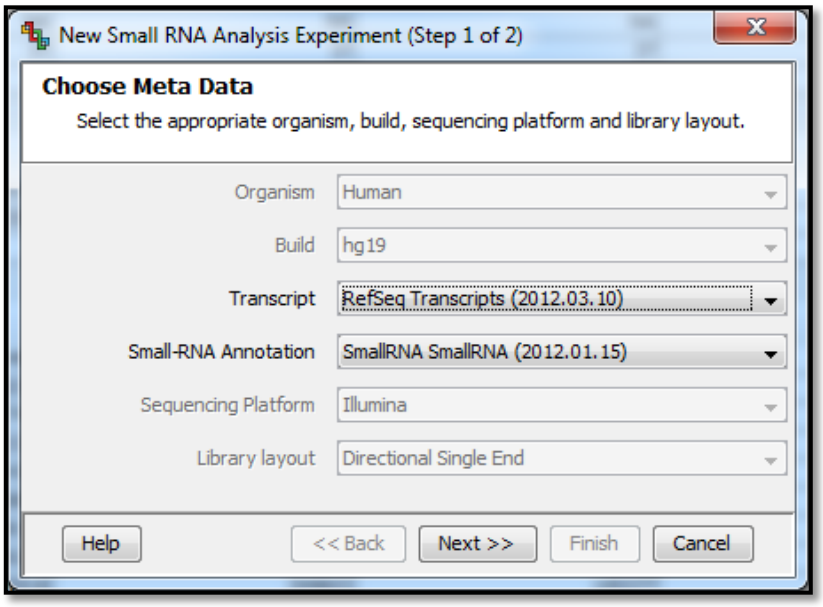
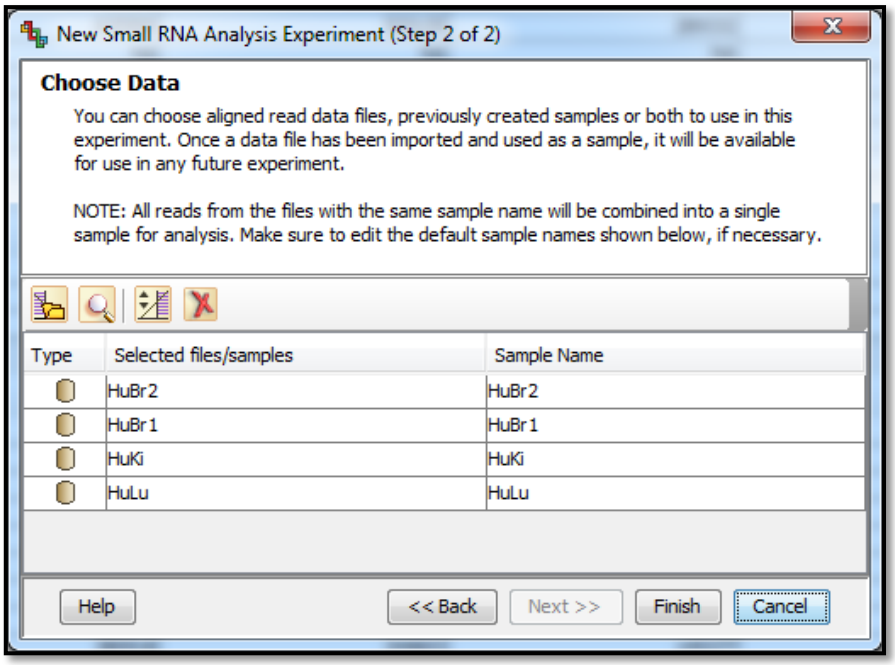


Figure 9: Experiment creation

If this experiment is created from the alignment experiment, then the samples will already be present otherwise the samples can be selected using the first icon on the toolbar.



Once the experiment is created, the genome browser view is launched with all the samples and the small RNA annotation track.

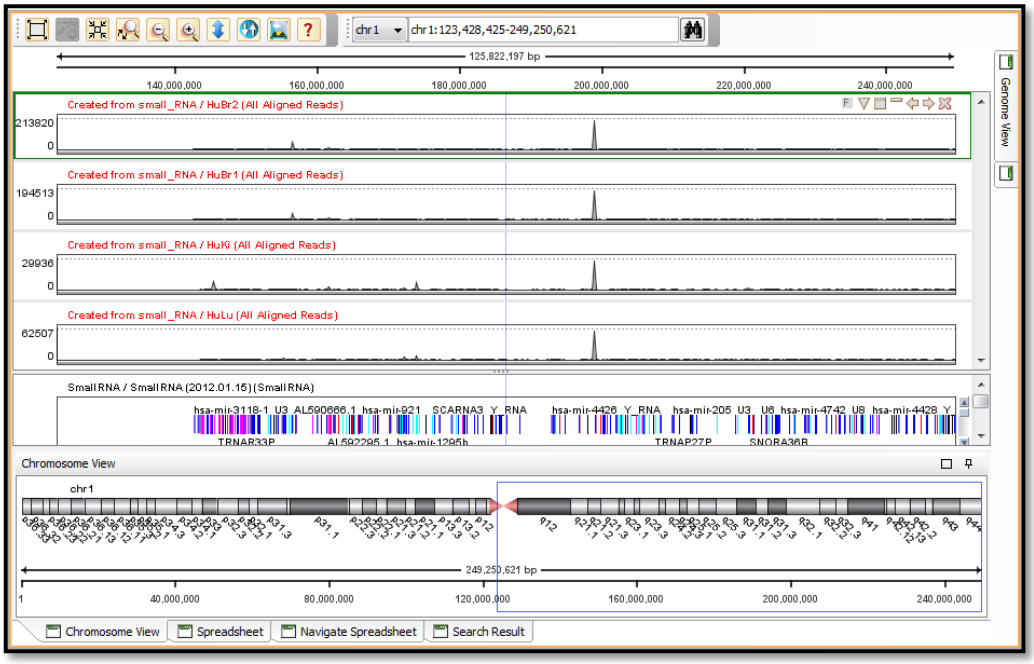


Figure 10: Default GB view

Quality Inspection and Filter

There are a variety of QC plots that give us an idea about the quality of the dataset. Most of these are common across the experiments (you will see them mentioned in other tutorials too) whereas some are specific to each experiment type. The small RNA specific QC plot can be launched from

Quality Inspection → Genic Region QC Plot in the workflow navigator. This shows a pie chart view of read distribution across different types of genic regions present in the annotations (miRNA, snoRNA, snRNA, scRNA, tRNA, exonic, intronic, etc.).

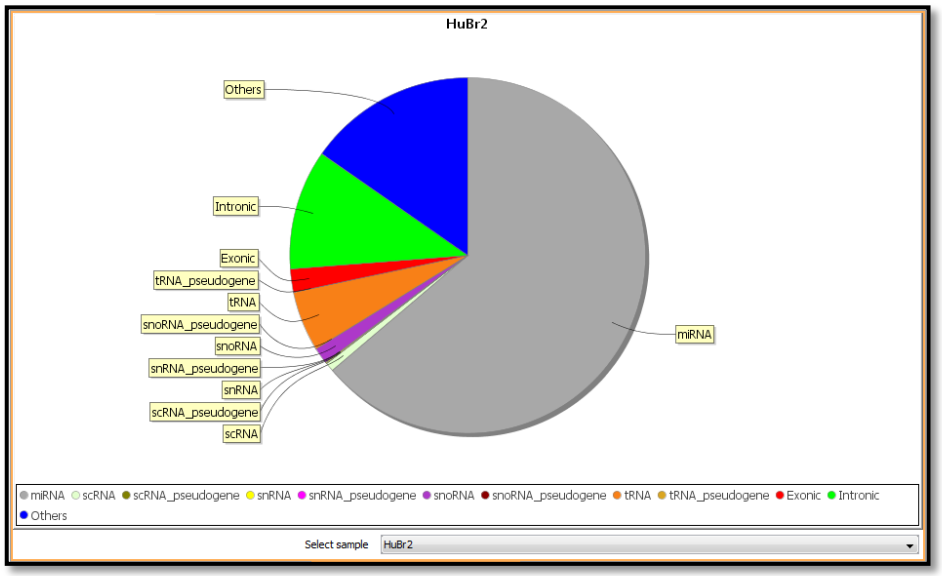


Figure 11: Genic region QC

Likewise the Filter steps also contains a mixture of common and experiment specific steps. For small RNA, there is a corresponding filter 'Filter by Genic Regions' in the Filters section using which you could filter out reads from specific genic regions.

For this dataset, we can filter with the default parameters which would exclude the reads in the exonic regions. This would help us in removing reads which might be due to RNA contamination. In this dataset, about 1,968,915 reads fall in the exonic regions and are removed.

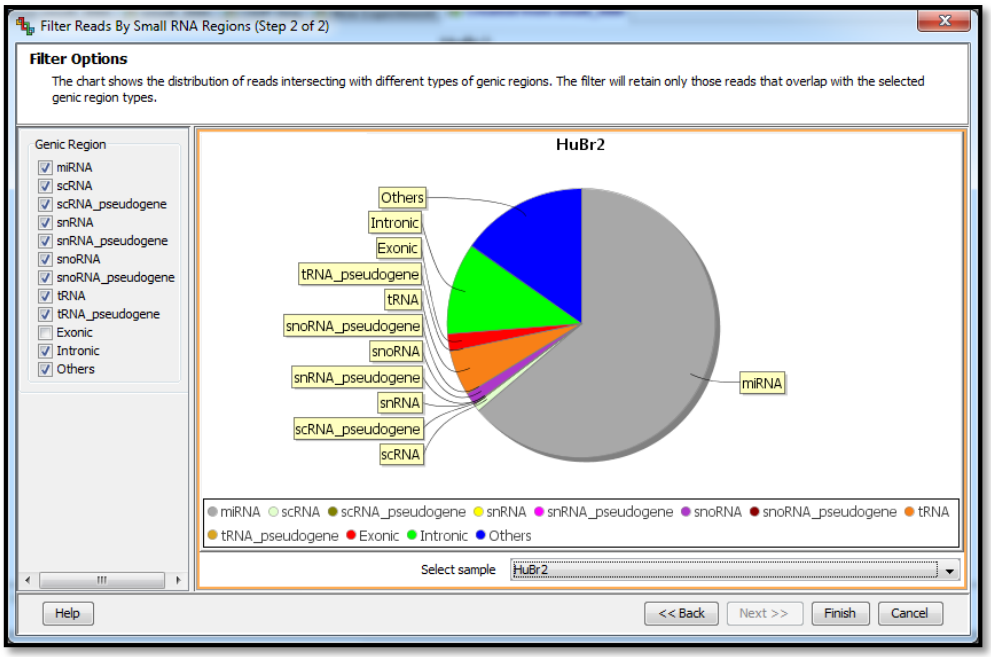


Figure 12: Filter by genic regions

Quantification

The next step after QC and filtering would be to quantify the expression levels of small RNAs (using the Analysis → Quantification workflow step).

Small RNA annotations that are made available from the Avadis NGS server contain gene level annotations for multiple small RNA species such as - miRNA precursor genes, tRNAs, snoRNAs, scRNAs, snRNAs etc. In addition, the miRNA precursor genes are annotated with the locations of the mature miRNA sequences also. These locations are referred to as 'active regions' in Avadis. The quantification step, in addition to finding the expression for all the genes and active regions from the known annotations, has an option to detect novel small RNA genes.

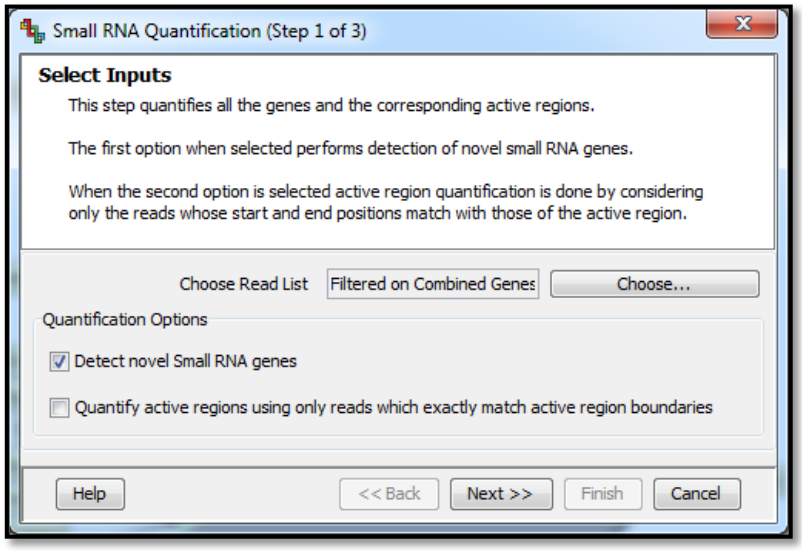


Figure 13: Input selection

For this dataset, we have used the read list generated after filtering exonic reads (as detailed in the QC and Filter section) and have used the default parameters for the other steps. Though we choose DESeq as the normalization option, other methods such as Sample Read Count and TMM are also available.

After quantification, we will see a quantification node, novel detection report and an all genes list (under All Entities) in the experiment navigator. There is also an active region list which is created as a child node of 'All Entities', and it contains only the active regions and not the genes.

Filter Genes

We also have a 'Filter by gene expression' step under the quantification tab. For this tutorial, select the 'All Active Regions' list and the 'All Samples' interpretation and run the filter with the parameters shown below. The resultant filter list containing 442 genes would then be saved under the 'All Active Regions' list in the experiment navigator.

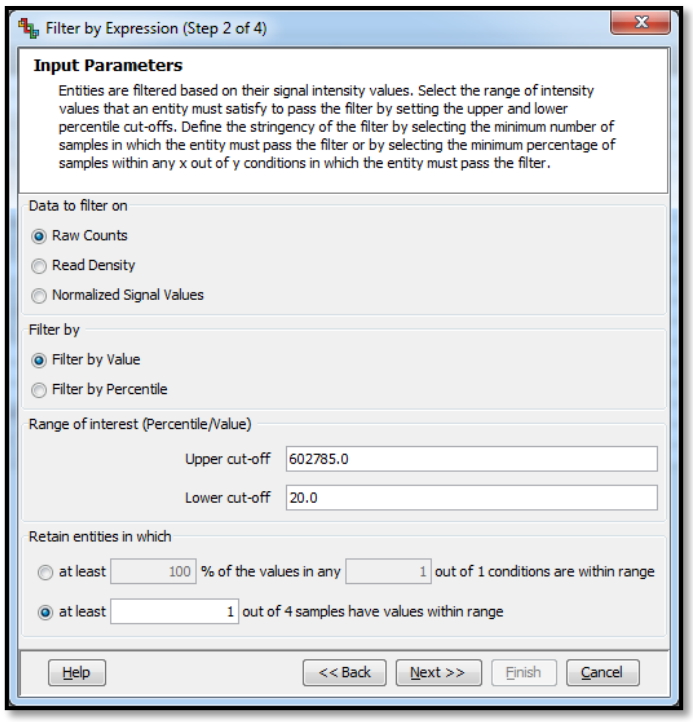


Figure 14: Filter by expression

Inspecting and visualizing the results of quantification

The raw counts and normalized signal values of the entities can be visualized by using the right-click options → 'View raw counts' and 'View normalized signal values' respectively on the quantification node. In addition we have a small-RNA specific ‘Gene View’ to visualize the expression values of different genes and active regions.

The Gene View can be launched either from the toolbar or by right-clicking on an entity list. The view consists of the following.

A table is shown at the bottom which has two tabs. The first tab `Read Densities' gives the read densities for all the entities from the selected entity list. The second tab `Entity List Data' shows all the associated values (other than the read densities) from the selected entity list.

In the top pane, a hairpin structure of the miRNA precursor is shown with flanking profile plots for read densities and counts.

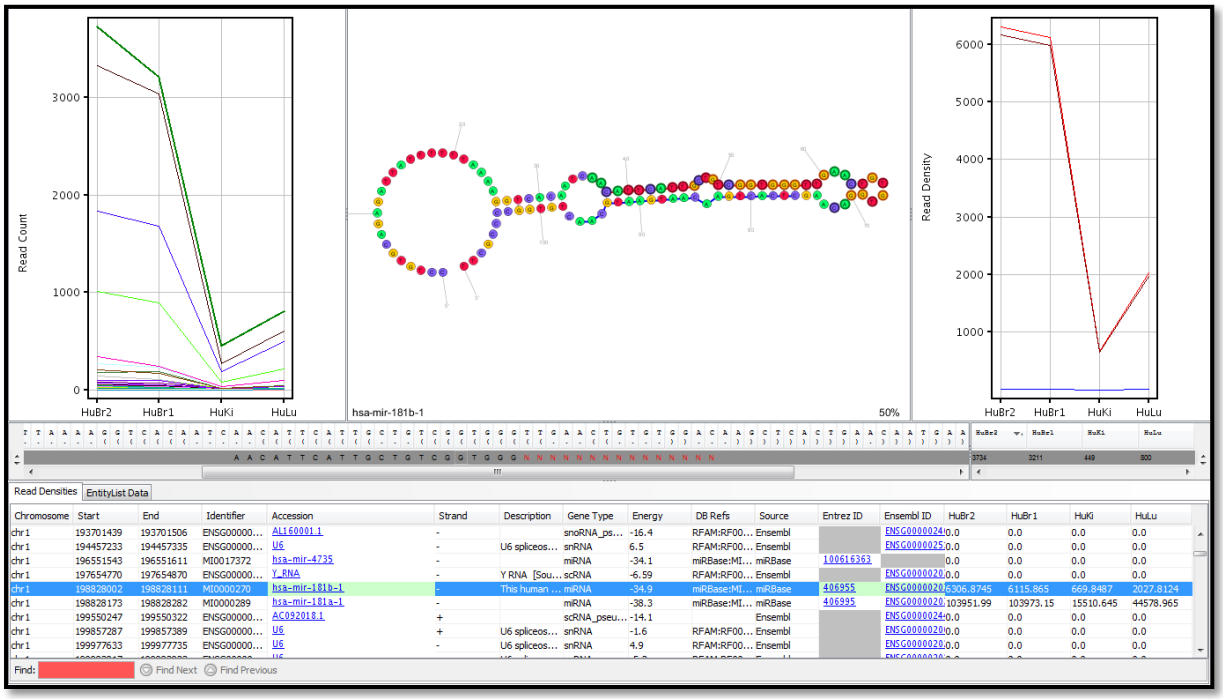


Figure 15: Gene view

Differential Expression

Differential expression, clustering and PCA can be performed on the quantification results using the workflow steps in the Analysis section of the workflow navigator.

In the present tutorial, the samples can be assigned the following parameters from Experiment Setup→Experiment Grouping from the workflow navigator.

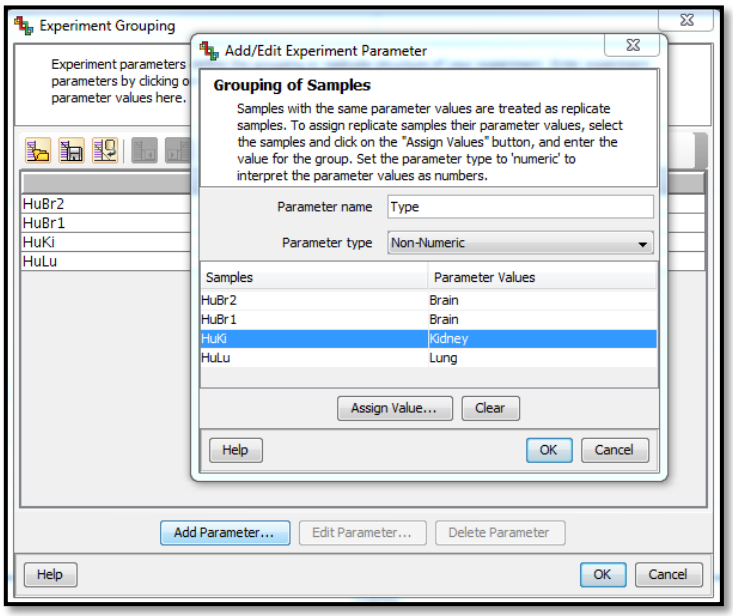
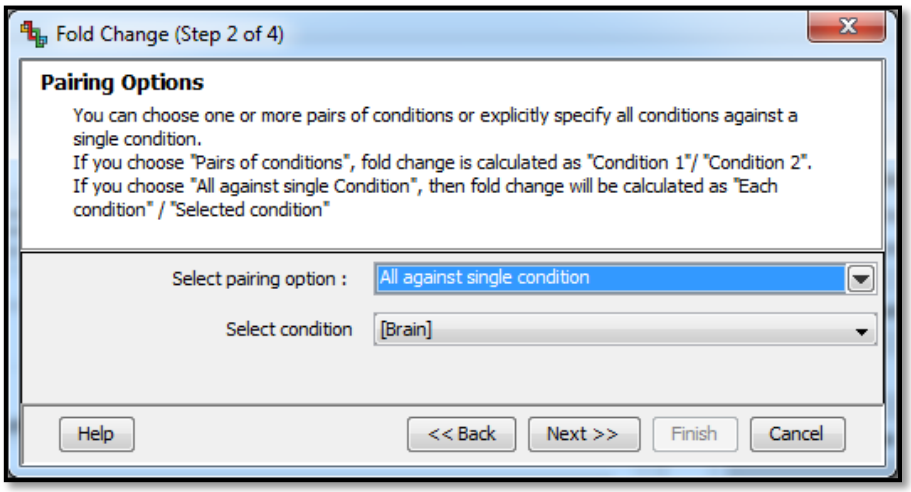


Figure 16: Experiment Grouping

After grouping the samples, we can create an interpretation based on these parameters from Experiment Setup→Create Interpretation from the workflow navigator. Once this is done, we can proceed to the differential expression from Expression Analysis→ Fold Change. For this tutorial, we can choose the list after filtering based on the expression with the below conditions for fold change.



The resultant list would show 378 genes having a fold change greater than two for the selected conditions. One of the miRNA genes coming out as over expressed in the kidney sample would be the hsa-miR-196a gene and its over expression in this particular tissue type can also be verified in external databases.

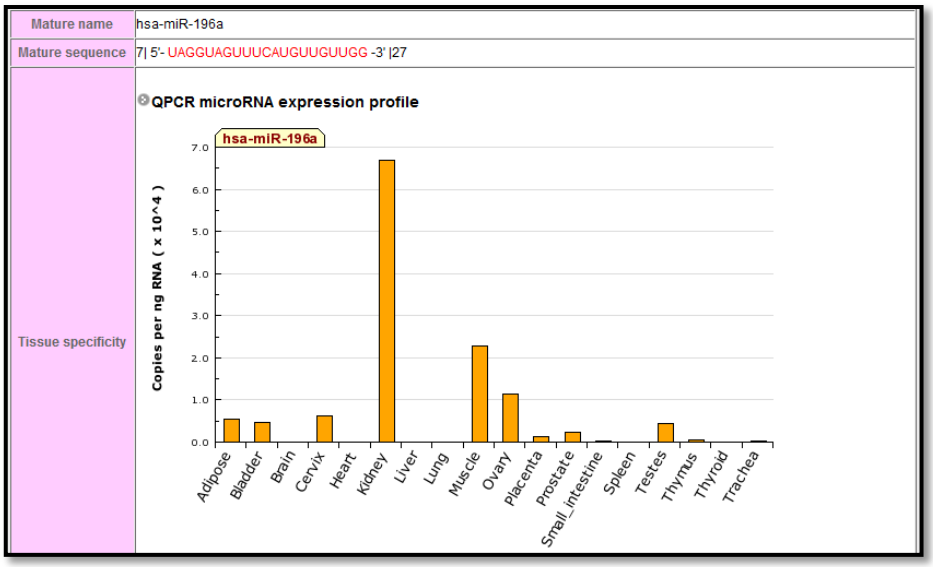


Figure 17: http://mirnamap.mbc.nctu.edu.tw/php/mirna_entry.php?acc=MI0000238

Results Interpretation

We will now proceed to interpreting the results obtained after quantification. We can predict the genes targeted by the miRNA's from external target prediction databases like TargetScan, PITA etc (using the Results Interpretation → Find Targeted Genes workflow step). For this dataset, use the region list generated after fold change and the TargetScan database (This can be downloaded from Annotations Manager as mentioned in the 'Getting Started' tutorial)

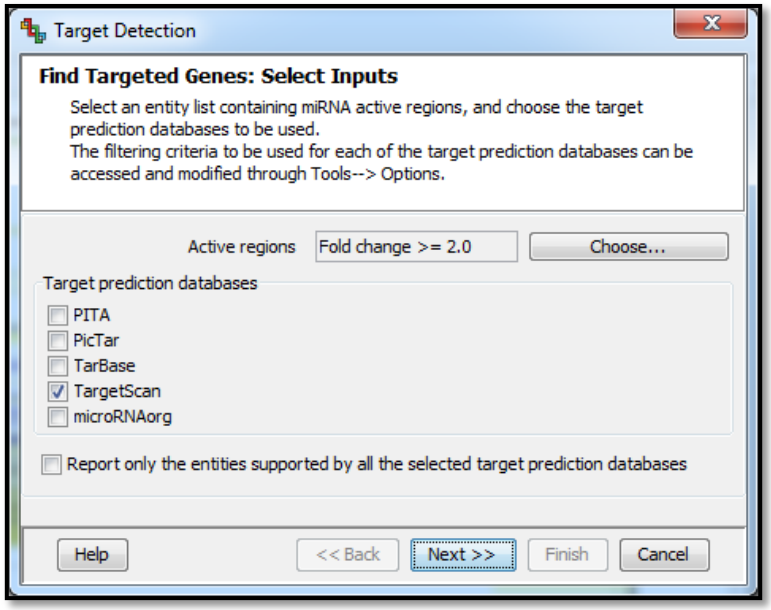


Figure 18: Target Detection

The resultant gene list containing 1042 genes can be used as input for further downstream analysis such as GO, Pathways etc. from the 'Results Interpretation' section in the workflow browser. The workflow also provides a step called 'Find Targeting miRNAs' to get the reverse mapping i.e., the list of all the miRNAs that target the genes in the given entity list.

Validating mature miRNA sequences

We can validate the small RNA annotations against the sample data (using the Utilities → Validate mature miRNA Annotations step in the workflow navigator). Ideally the small RNA reads should have an exact match with the active regions. Reads which do not have an exact match are considered discrepant. It takes a read list, minimum coverage, and minimum discrepant read percentage as inputs, and outputs an entity list of active regions for which the `minimum discrepant read percentage' cut-off is satisfied. After this step is done, the resultant list is stored under the parent read list in the experiment navigator and can be examined by Right click→ Inspect List. For this dataset, using the filtered list on genic regions, 573 entities get reported.

This tutorial was meant to give a brief overview of the features in small RNA alignment and analysis. For more details or clarifications, please revert back (sales@avadisngs.com or support@avadisngs.com) and we will address your queries.