Analysis Datasheet
Exosome RNA-seq Analysis

Overview

RNA-seq is a high-throughput sequencing technology that provides a genome-wide assessment of the RNA content of an organism, tissue, or cell. Small RNA molecules have been increasingly found to play an important role in genetics and are likely critical for the extracellular signaling function of exosomes. Maverix Biomics Exosome RNA-seq analysis uses open-source tools and includes quality assessment, read alignment, transcript abundance estimation, and differential expression analysis (Figure 1). Small RNA molecules included in the analysis are ncRNAs (miRNAs, tRNAs, rRNAs, lincRNAs, piRNAs, snoRNAs), anti-sense transcripts, coding genes and repeat elements (LTR, LINE, SINE, and tandem repeats). Results are provided as visual representations, including interactive tabular and heat map views linked to an integrated, private version of the UCSC Genome Browser.

This datasheet describes how to launch an exosome RNA-seq analysis and reviews the results produced by the pipeline. For an overview of the platform interface, please see the Quick Start Guide.

Configuration

After uploading data and choosing Exosome RNA-seq Analysis 1.8 you will be ready to configure and launch an analysis (Figure 2).

Analysis Name

Create a unique, informative name for this instance of the analysis. The name will appear in the Analysis Execution History on the Explore tab and should be a name that will distinguish it from the other analyses in your history.

Add Samples

Click Add Samples to begin choosing files for the analysis. You will be taken to the sample entry interface, which will allow you to name the sample and pick the associated FASTQ files including paired-end files and additional replicates. Once you have added a sample to the analysis, click Add Samples again and continue adding samples until the dataset for analysis is complete. The name you select will become the base name for all of the replicates of that sample and will allow

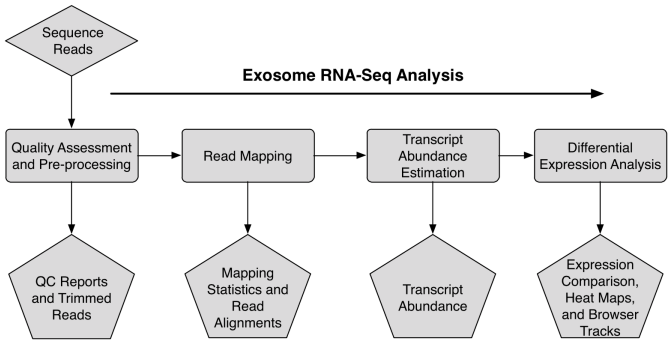

Figure 1. Overview of the exosome RNA-seq analysis pipeline.

them to remain grouped through the analysis. The sample name will also become the root name for output files generated during analysis, allowing you to distinguish them in the File Library for later download. Additionally, the sample name will be the name that distinguishes genome browser tracks and other analysis output visualizations, such as tables, plots and heat maps. The characters in a name are limited to alphanumeric, underscore, and period.

Edit

Click Edit to change the sample name and input files before starting the analysis run.

Organism

Select the organism and reference genome to use for sequence alignment.

Spike-in

For calibration and quality assessment, sequencing runs often use a spike-in, which is a control library generated from an organism with a small, well-defined genome, such as the PhiX or a bacterium. If a spike-in was used during sequencing, use the Spike-in pulldown to select the appropriate control library from the menu. These reads will be counted and filtered out during quality assessment.

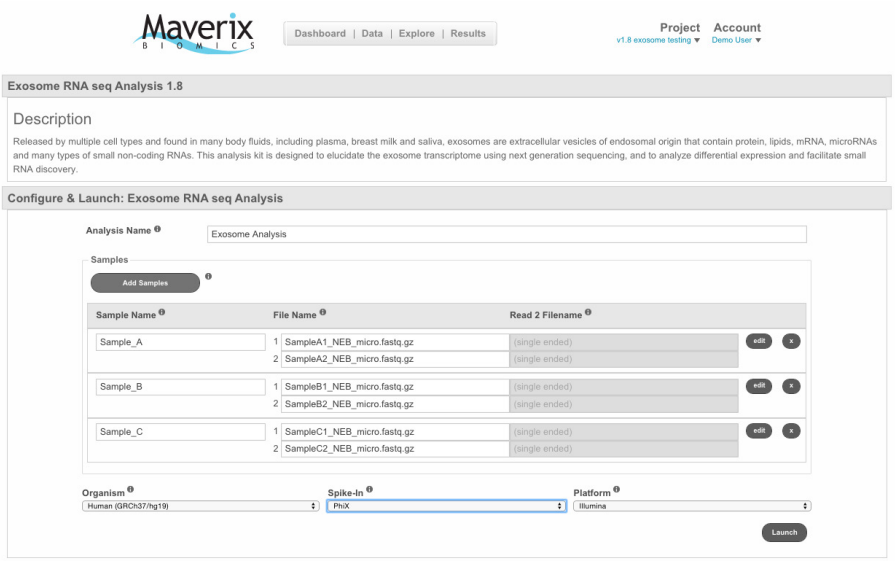

Figure 2. The main configuration screen for the exosome RNA-seq analysis.

Platform

Select the sequencing platform that was used to generate your raw RNA-seq reads. Choices include Illumina (NextSeq 500, HiSeq 2500, HiSeq 2000, MiSeq, Genome Analyzer IIx) or Ion Torrent (Proton). This information is used to choose the correct adaptor sequence to use for trimming and filtering as well as to select the correct quality thresholds for filtering and trimming.

Launch

Once configuration is complete, click Launch to begin your analysis. The progress of the analysis can be monitored in real-time by clicking on the Explore tab.

Analysis Overview and Outputs

Quality Assessment

Open source tools used in this step include FastQC, FastqMcf, PRINSEQ, Bowtie2, SAMtools, and Reaper.

Raw reads are trimmed and filtered based on quality as well as to remove adaptor sequence and spike-in controls. The scores used for trimming and filtering are specific to the sequencing platform. The processed reads are then again assessed for quality. Plots are generated of per base quality before and after trimming. Proportion of unique reads (i.e. non-duplicates) is also measured.

Per base quality

Plots of quality scores are generated by FastQC for each sample before and after trimming (Figure 3). The pre-trimming plots show the quality scores of raw reads for each FASTQ file included in the analysis. The post-trimming plots show the quality scores of the reads after pre-processing, which includes removal of adapters, removal of N's at the ends of reads, and filtering based on overall read quality and length.

Sequence Read Quality Assessment

Summary data is presented for the proportion of reads removed due

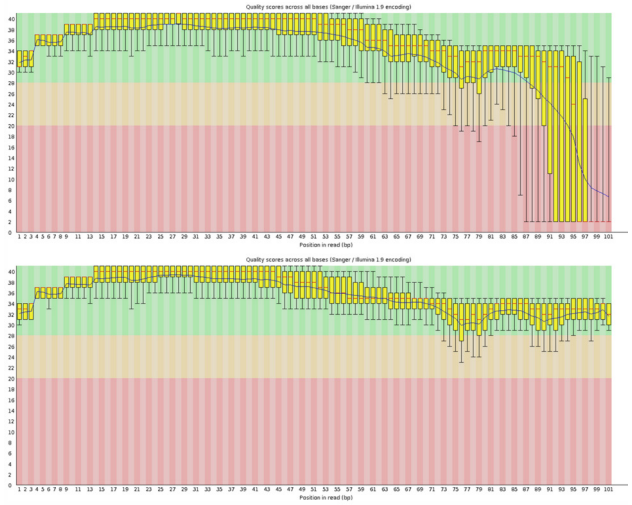

Figure 3. FastQC per base sequence quality score plots before and after filtering and trimming.

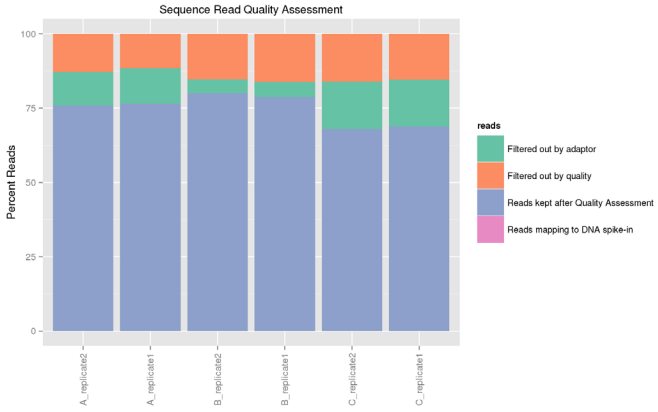

Figure 4. Sequence read quality assessment.

to the presence of adaptor, low quality, and spike-in (Figure 4).

Number of Reads Pre and Post-Quality Filtering

Number of reads remaining after quality assessment is presented (Figure 5). Number of unique reads is also measured and reported.

Read Alignment

Open source tools used in this step include Bowtie, SeqPrep, SAMtools, and BEDTools.

Trimmed reads are merged into pseudo single-end reads using SeqPrep and then mapped to the reference genome using Bowtie. Summary data and downloadable BAM files are then produced (Figures 6 and 7). The BAM files are deployed to the integrated UCSC Genome Browser for visualization in a genomic context with public datasets.

Abundance Determination

Open source tools used in this step include SAMtools and R.

Following alignment annotation files are used to determine how many reads map to each genomic feature. Custom scripts are used to produce raw numbers of reads mapping to each transcript, which are used as input for DESeq differential analysis and are available for download. The proportion of reads that map to different types of small RNAs is displayed in pie charts (Figure 8).

Differential Expression

Open source tools used in this step include R and Bioconductor package DESeq.

DESeq is used to calculate fold change and statistical significance for all differentially expressed genes across all samples. Results are shown as volcano plots as well as in an interactive heat map or table linked to the UCSC genome browser. DESeq results can also be downloaded directly. Short Interspersed Nuclear Elements (SINE), Long Interspersed Nuclear Elements (LINE), and Long Terminal Repeats (LTR) LTR RNAs are excluded from differential expression analysis. P-values are corrected for multiple comparisons.

Volcano Plot

A volcano plot shows fold-change and significance for every gene (Figure 9). Plotting points in this way allows the user to identify both the magnitude and significance of differential expression for all genes for a particular sample pair. P-values are corrected for multiple comparisons and are considered significant if $p < 0.05$.

UCSC Genome Browser

Expression data can be viewed within the integrated UCSC Genome Browser, which allows for comparison to public databases (Figure 10). RefSeq gene annotations, miRNAs, and other data tracks available in the public version of UCSC Genome Browser are pre-loaded for the target organism. Gene coverage can be visualized on a per treatment or per replicate basis. Additional documentation can be found at http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html.

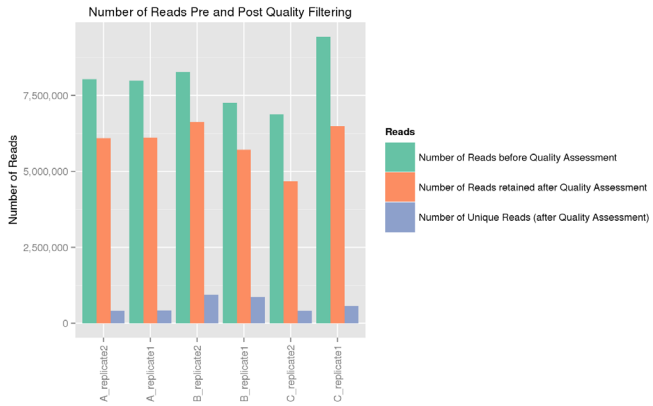

Figure 5. Number of reads pre- and post-quality filtering.

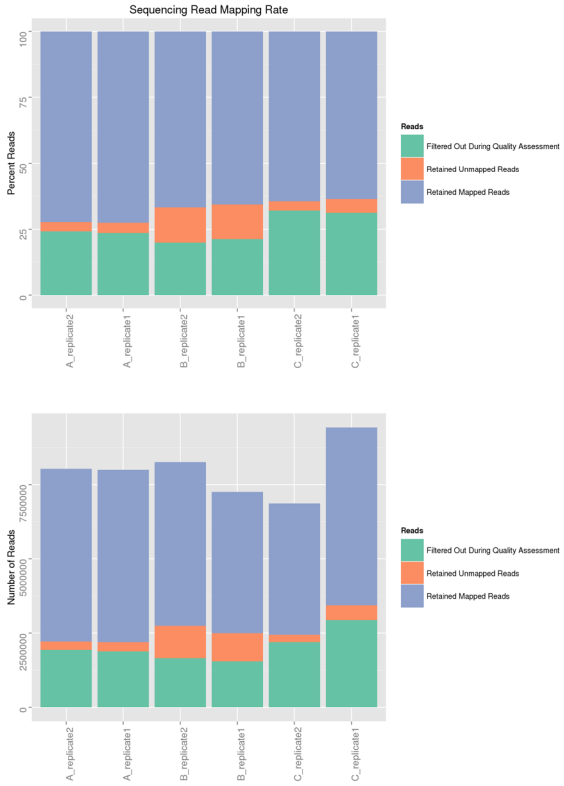

Figure 6. Read alignment summary results.

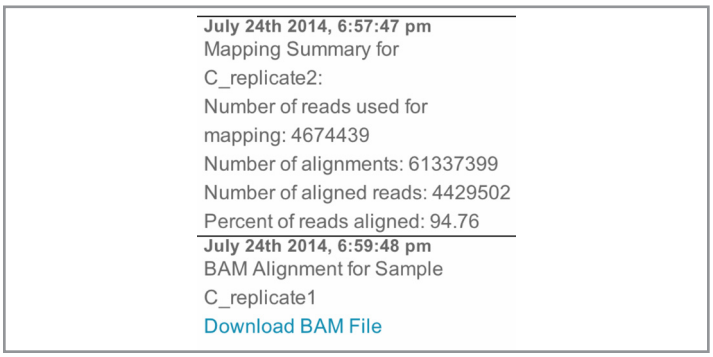

Figure 7. Read alignment summary results and BAM downloadable link.

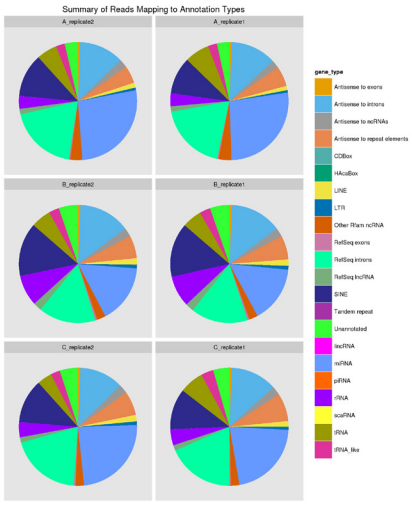

Figure 8. Percentage of reads mapping to a variety of small RNA types.

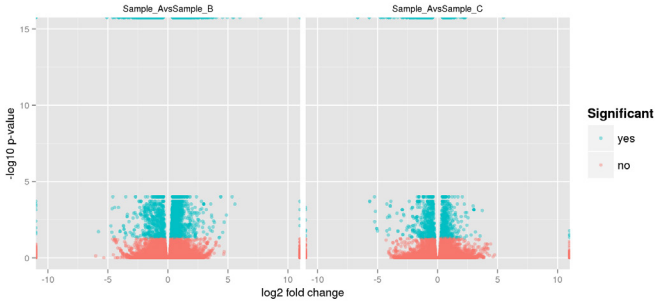

Figure 9. Volcano plot.

Heat Map

The heat map displays expression data and is integrated with the integrated UCSC Genome Browser (Figure 10). The magnitude of differential expression is distinguished by color with red representing increased expression relative to the control and blue showing decreased expression relative to the control. The intensity of color indicates the degree of the difference. Hovering over the heat map reveals the gene names, symbols, and chromosomal location in the reference genome. By clicking the gene symbol within the pop-up window the locus information can be displayed from a variety of sources including NCBI GenBank, miRBase, GtRNAdb, Human lincRNA catalog, and Rfam. Clicking on a gene feature in the heat map will load the gene locus in the integrated UCSC Genome Browser.

By clicking the settings tab the heat map display can be modified in variety of ways (Figure 11).

- Samples. Sample names can be modified and choice of control can be altered.

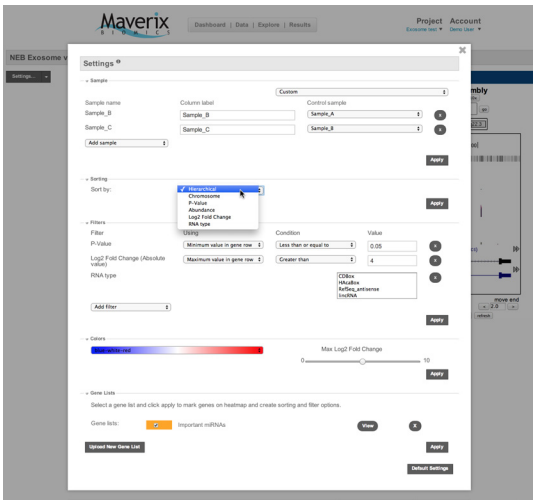

Figure 11. Display settings can be customized in the Heat Map or Table View display.

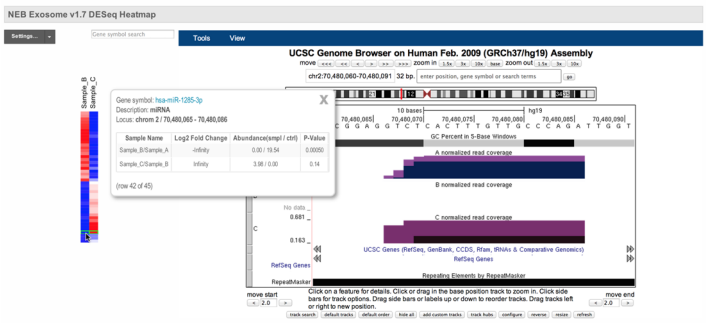

Figure 10. Visualize mapping data in the UCSC Genome Browser with integrated Heat Map.

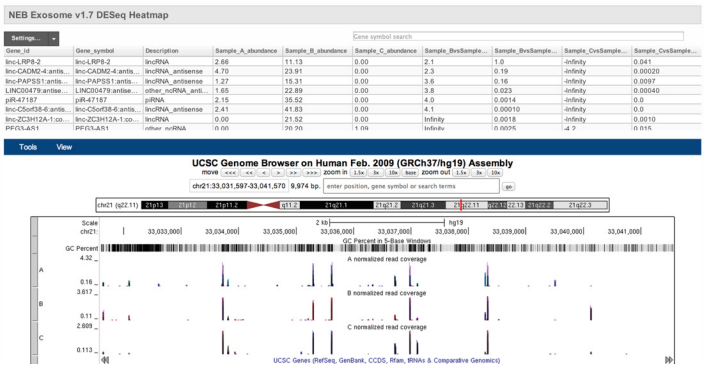

Figure 12. Table View of differencial expression results.

- Sorting. The genes in the heat map can be sorted based on hierarchical clustering, chromosome, p-value, abundance, and log2 fold change.
- Filtering. Genes can be filtered based on p-value, abundance, log2 fold change, absolute log2 fold change, chromosome, and gene list.
- Colors. Colors used to represent under and over expression can be changed from the default blue and red to green and red, blue and yellow, or purple and yellow.
- Gene Lists. Gene lists can be supplied to either mark or filter the heat map. The file should be a simple text file containing official gene symbols, each on a new line or separated by spaces. Genes displayed with red font are not present in the current data set.

Table View

The heat map can be replaced with a table view by clicking the arrow next to the Settings tab and selecting Table View (Figure 12). Using this same dropdown menu you are also able to download the table for off-line manipulation. The table view, like the heat map, is integrated with the UCSC Genome Browser so the same sorting and filtering options discussed in association with the heat map display are also available in the table view.

Conclusion

The exosome RNA-seq pipeline allows users to perform differential expression analysis starting from fastq input files. The results are provided as visual representations, including interactive tabular and heat map views linked to an integrated, private version of the UCSC Genome Browser (Figure 1).

References

1. Quinian, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bionformatics 26, 6, 841-842.
2. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, 3, R25.
3. Langmead, B., & Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 4, 357-359.
4. Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology 11, R106, 1-12.
5. Aronesty, B. (2011). ea-utils : Command-line tools for processing biological sequencing data. Retrieved from http://code.google.com/p/ea-utils
6. Andrews, S. FastQC. Retrieved from http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc
7. Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics 27(6): 863-864.
8. Davis, M. P. A., van Dongen, S, Abreu-Goodger C, Bartonicek N, and Enrigh AJ. (2013). Kraken: A set of tools for quality control and analysis of high-throughput sequence data. Methods 63(1): 41-49.
9. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
10. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 16, 2078-2079.
11. St. John, J. SeqPrep. Retrieved from (https://github.com/jstjohn/SeqPrep)
12. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, & Haussler D. (2002). The human genome browser at UCSC. Genome Research 12, 6, 996-1006.

Benchmarking

Our benchmark dataset consists of FASTQ files of Illumina HiSeq 2000 sequence reads derived from two technical replicates each from three samples of human plasma exosomes. Profiling analysis of exosomal RNA was carried out using several small RNA library preparation kits, with differential expression analysis between samples and between prep kits.

Huang, X, Yuan, T., Tschannen, M., Sun, Z., Jacob, H., Du, M., Liang, M., Dittmar, R. L., Liu, Y., Liang, M., Kohli, M., Thibodeau, S. N., Boardman, L., & Wang, L. (2013). Characterization of human plasma-derived exosomal RNAs by deep sequencing. BMC Genomics 14, 319, 1-14.

Open Source Tools Used

BEDTools[1] version 2.17.0
Bowtie[2] version 0.12.9
Bowtie2[3] version 2.1.0
DESeq[4] version 1.10.1
ea-utils[5] version 1.1.2-537
FastQC[6] version 0.10.1
PRINSEQ[7] version 0.20.3
Reaper[8] version 13-100
R[9] version 2.15.2
SAMtools[10] version 0.1.18
SeqPrep[11] version 1.0
UCSC Genome Browser[12] version 0cd5d4 (commercial license)

Maverix BIOMICS

1875 South Grant Street
Suite 700
San Mateo CA 94402

1.650.388.9277
www.maverixbio.com
info@maverixbio.com

Version: DS-exosome-seq-003
Pipeline Version: Exosome RNA-seq 1.8