RNA Express

Introduction 3
Run RNA Express 4
RNA Express App Output 6
RNA Express Workflow 12
Technical Assistance
This document and its contents are proprietary to Illumina, Inc. and its affiliates ("Illumina"), and are intended solely for the contractual use of its customer in connection with the use of the product(s) described herein and for no other purpose. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way whatsoever without the prior written consent of Illumina. Illumina does not convey any license under its patent, trademark, copyright, or common-law rights nor similar rights of any third parties by this document. The instructions in this document must be strictly and explicitly followed by qualified and properly trained personnel in order to ensure the proper and safe use of the product(s) described herein. All of the contents of this document must be fully read and understood prior to using such product(s).

FAILURE TO COMPLETELY READ AND EXPLICITLY FOLLOW ALL OF THE INSTRUCTIONS CONTAINED HEREIN MAY RESULT IN DAMAGE TO THE PRODUCT(S), INJURY TO PERSONS, INCLUDING TO USERS OR OTHERS, AND DAMAGE TO OTHER PROPERTY.

ILLUMINA DOES NOT ASSUME ANY LIABILITY ARISING OUT OF THE IMPROPER USE OF THE PRODUCT(S) DESCRIBED HEREIN (INCLUDING PARTS THEREOF OR SOFTWARE) OR ANY USE OF SUCH PRODUCT(S) OUTSIDE THE SCOPE OF THE EXPRESS WRITTEN LICENSES OR PERMISSIONS GRANTED BY ILLUMINA IN CONNECTION WITH CUSTOMER'S ACQUISITION OF SUCH PRODUCT(S).

FOR RESEARCH USE ONLY

© 2014 Illumina, Inc. All rights reserved.

Illumina, IlluminaDx, 24Sure, BaseSpace, BeadArray, BeadXpress, BlueFish, BlueFuse, BlueGnome, cBot, CSPro, CytoChip, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, HiSeq X, Infinium, iScan, iSelect, MiSeq, MiSeqDx, NeoPrep, Nextera, NextSeq, NuPCR, SeqMonitor, Solexa, TruGenome, TruSeq, TruSight, Understand Your Genome, UYG, VeraCode, VeriSeq, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks of Illumina, Inc. in the U.S. and/or other countries. All other names, logos, and other trademarks are the property of their respective owners.
Introduction

The BaseSpace® RNA Express app combines the capabilities of the STAR aligner and DE-Seq analysis tools in one simple workflow. The aim of this app is to provide the most commonly used set of RNA analysis features in a convenient and rapid analysis package.

Versions

The following module versions are used in the RNA Express apps:
- STAR 2.3.1s
- DESeq2 1.0.17

Current Limitations

Before running the RNA Express app, be aware of the following limitations:
- Reads must be at least 35 bp and no more than 500 bp in length.
- Individual samples must be between 100,000 and 400 million reads.
- The total read count across all samples must be less than 2 billion reads.
- Only UCSC hg19 (human), UCSC mm10 (mouse), and UCSC rn5 (rat) are currently supported.
Run RNA Express

1. Navigate to the project or sample that you want to analyze.
2. Click the Launch App button and select RNA Express from the drop-down list.
3. Read the End-User License Agreement and permissions, and click Accept if you agree.
4. Fill out the app session storage information:
   a. **App Session Name**: provide the app session name. Default name is the app name with the date and time the app session was started.
   b. **Save Results To**: select the project that stores the app results.
5. Fill out the sample criteria:
   a. **Reference Genome**: select the reference genome.
   b. **Stranded**: Indicate if samples were stranded.
   c. **Trim TruSeq Adapters**: If selected, the application attempts to trim TruSeq® adapters from the FASTQ sequence. Typically, this trimming is unnecessary as adapter trimming is performed as part of demultiplexing during sample upload. However, if the user did not specify adapter sequences in the sample sheet during upload, this option provides a second opportunity to trim the adapters.
6. Fill out the control group information:
   a. **Group Label**: provide the control group label. Default name is control.
   b. **Select Sample**: browse to the sample you want to use as control, and select the checkbox. You can use multiple samples as control.
7. Fill out the comparison group information:
   a. **Group Label**: provide the comparison group label. Default name is comparison.
   b. **Select Sample**: browse to the sample you want to use as comparison, and select the checkbox. You can use multiple samples as comparison.
8. Click Continue.

RNA Express starts analyzing your sample. When completed, the status of the app session is automatically updated, and you receive an email.
Figure 1  RNA Express Input Form

- **App Session Name:** RNA Express 02/05/2014 2:45:11
- **Save Results To:** Select a Project

### Sample Criteria
- **Reference Genome:** Homo sapiens/hg19 (RefSeq)
- **Stranded:**
- **Trim TruSeq Adapters:**

### Control Group
- **Group Label:** control
- **Control Samples:** Select Sample

### Comparison Group
- **Group Label:** comparison
- **Comparison Samples:** Select Sample
RNA Express App Output

This chapter describes the RNA Express output. To go to the results, click the Projects button, then the project, then the analysis.

Figure 2  RNA Express Output Navigation Bar

When the analysis is completed, you can access your output through the left navigation bar, which provides the following:

- **Analysis Info**: an overview of the analysis. See *Analysis Info* on page 9 for a description.
- **Inputs**: an overview of the input samples and settings. See *RNA Express Inputs Overview* on page 10 for a description.
- **Output Files**: access to the output files, organized by sample and app session. See *RNA Express Output Files* on page 10 for descriptions.
- **Analysis Reports: Summary**: access to analysis metrics for the aggregate results. See *RNA Express Report* on page 6 for a description.

RNA Express Report

The RNA Express app provides an overview for all samples on the Summary page. A brief description of the metrics is below.

**Primary Analysis Information**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Length</td>
<td>Number and length of reads.</td>
</tr>
<tr>
<td>Number of reads</td>
<td>Total number of reads passing filter for this sample.</td>
</tr>
</tbody>
</table>

**Alignment Information**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Total Aligned</td>
<td>Percentage of reads passing filter that aligns to the reference.</td>
</tr>
</tbody>
</table>
### Statistic | Definition
--- | ---
% Abundant | Percentage of reads that aligns to abundant transcripts, such as mitochondrial and ribosomal sequences.
% Unaligned | Percentage of reads that do not align to the reference.
Multi-mapped (% Aligned Reads) | The percentage of aligned reads that have more than one equally good alignment position in the genome.
Reads with spliced alignment (% Aligned Reads) | The percentage of aligned reads that map over splicing-events. Each case where a read-alignment skips over a known or discovered intron is counted.
Link to BAM File | Download link to BAM file for this sample.

#### Read Counts

| Statistic | Definition |
--- | --- |
Exonic Reads (%) | Reads mapping to exonic regions (% of uniquely aligned reads).
Non-exonic Reads (%) | Reads mapping to non-exonic regions (% of uniquely aligned reads).
Ambiguous Reads (%) | Reads aligning to more than one locus or to a locus overlapping multiple genes (% of uniquely aligned reads).

#### Differential Expression

| Statistic | Definition |
--- | --- |
Annotation Gene Count | Number of genes in annotation.
Assessed Gene Count | Number of genes tested for statistical significance.
Differentially Expressed Gene Count | Number of significantly differentially expressed genes.
Link to Merged Gene Counts | Download link to CSV file describing the number of reads mapped to each gene for each sample in the control and comparison groups.
Link to Results | Download link to CSV file describing the mean expression, log2 fold change, standard error of log2 fold change, p-value, adjusted p-value, and the expression status for each gene.

#### Sample Correlation Matrix

A heat map showing the relative similarity between all replicates in this analysis-run. Each row and column represents one replicate, ordered by similarity (hierarchical clustering). The color of each field indicates the Spearman Rho correlation between these replicates.
Figure 3  Sample Correlation Matrix

Control vs Comparison

The control vs comparison plot shows an interactive scatter plot of the log2(Fold Change) against the mean count for a gene. You can filter the results by the following metrics:

- Test status:
  - OK: test successful
  - Low: low average expression across samples (mean normalized count across all samples less than 10)
  - Outlier: a single (outlier) replicate strongly affects the result

- Significance: Genes with a multiple-testing adjusted p-value (q-value) for differential expression of less than 0.05

- Gene: allows you to search for a particular gene in the plot and the gene table below the scatter plot

The gene table below the scatter plot shows those metrics for individual gene results, in addition to the standard error of the Log2(Fold Change). If you click a gene, the corresponding dot is circled in the scatter plot. Likewise, if you click any of the dots in the scatter plot, the gene is highlighted in the gene table.

The following additional metrics are reported in the gene table:

- Std. err. log(Fold Change): Standard error of the Fold Change estimate
- q value: Multiple-testing adjusted p-value for differential expression (used for Significance filter)
Analysis Info

This app provides an overview of the analysis on the Analysis Info page.

A brief description of the metrics is below.

<table>
<thead>
<tr>
<th>Row</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Name of the app session.</td>
</tr>
<tr>
<td>Application</td>
<td>App that generated this analysis.</td>
</tr>
<tr>
<td>Date started</td>
<td>Date the app session started.</td>
</tr>
<tr>
<td>Date completed</td>
<td>Date the app session completed.</td>
</tr>
<tr>
<td>Duration</td>
<td>Duration of analysis.</td>
</tr>
<tr>
<td>Session Type</td>
<td>The number of nodes used.</td>
</tr>
<tr>
<td>Size</td>
<td>Total size of all output files.</td>
</tr>
<tr>
<td>Status</td>
<td>Status of the app session.</td>
</tr>
</tbody>
</table>

Log Files

Clicking the Log Files link at the bottom of the Analysis Info page provides access to RNA Express app log files.

The following files log information to help follow data processing and debugging:

- **WorkflowLog.txt**: Workflow standard output (contains details about workflow steps, command line calls with parameters, timing, and progress).
- **WorkflowError.txt**: Workflow standard error output (contains errors messages created while running the workflow).
Logging.zip: Contains all detailed workflow log files for each step of the workflow.
IlluminaAppsService.log: Wrapper log file containing information about communication (get and post requests) between BaseSpace and AWS.
CompletedJobInfo.xml: Contains information about the completed job.
SampleSheet.csv: Sample sheet.

The following files contain additional information in case things (like mono) do not work as expected:

- monoErr.txt: Wrapper mono call error output (contains anything that WorkflowError.txt does not catch; in most cases empty, except one line).
- monoOut.txt: Wrapper mono call standard output (contains command calling the workflow and anything that WorkflowLog.txt does not catch).

NOTE
For explanation about mono, see www.mono-project.com.

RNA Express Status
The status of the RNA Express app session can have the following values (in order):
- Downloading data
- Aligning
- Post-alignment processing
- Read counting
- Differential expression analysis
- Generating report
- Finalizing results

Depending on the size and number of samples, the complete analysis can take between a few hours and several days.

RNA Express Inputs Overview
The RNA Express app provides an overview of the input app results and settings on the Inputs page. A brief description of the metrics is below.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group Label</td>
<td>The group label for the comparison or control group.</td>
</tr>
<tr>
<td>Comparison Samples</td>
<td>Samples selected for comparison group.</td>
</tr>
<tr>
<td>Control Samples</td>
<td>Samples selected for control group.</td>
</tr>
<tr>
<td>Reference Genome</td>
<td>Reference genome selected.</td>
</tr>
<tr>
<td>Save Results To</td>
<td>The project that stores the app results.</td>
</tr>
<tr>
<td>Stranded</td>
<td>Indicates if samples were stranded.</td>
</tr>
<tr>
<td>Trim TruSeq Adapters</td>
<td>If selected, the application trims TruSeq adapters.</td>
</tr>
</tbody>
</table>

RNA Express Output Files
RNA Express produces the following output files in the indicated folders:
<AppResult>/differential/global
  ‣ deseq.corr.pdf—PDF file showing a heat map of the sample correlation matrix.
  ‣ deseq.corr.png—PNG file showing a heat map of the sample correlation matrix.
  ‣ deseq.corr.csv—CSV file describing the sample correlation matrix.
  ‣ gene.counts.csv—CSV file describing the number of reads mapped to each gene for each sample.

<AppResult>/differential/<control>_vs_<comparison>
  ‣ <control>_vs_<comparison>.deseq.ma.pdf—PDF file showing a scatter plot of log2 (fold change) versus mean of normalized counts. This file is not available when DESeq2 fails to converge.
  ‣ <control>_vs_<comparison>.deseq.counts.csv—CSV file describing the number of reads mapped to each gene for each sample in the control and comparison groups.
  ‣ <control>_vs_<comparison>.deseq.disp.pdf—PDF file showing a scatter plot of dispersion versus mean of normalized counts. This file is not available when DESeq2 fails to converge.
  ‣ <control>_vs_<comparison>.deseq.heatmap.pdf—PDF file showing a heat map of the expression of the differentially expressed genes with adjusted p-values < 0.05 for samples in the control and comparison groups. Only the top-5000 differentially expressed genes are used if there are more than 5000 differentially expressed genes. This file is not available when DESeq2 fails to converge or when there are no differentially expressed genes.
  ‣ <control>_vs_<comparison>.deseq.res.csv—CSV file describing the mean expression, log2 (fold change), standard error of log2 (fold change), p-value, adjusted p-value, and the expression status for each gene.

<AppResult>/samples/<group>/replicates/<sample>/alignments
  ‣ <sample>.alignments.sorted.bam—Alignments of reads against the genome (and transcriptome). For description, see also BAM Files on page 11.
  ‣ <sample>.coverage.bedGraph.gz—Genome coverage with aligned RNA-Seq reads.

<AppResult>/samples/<group>/replicates/<sample>/counts
  ‣ <sample>.counts.genes—Tab-delimited file describing the number of reads mapped to each gene. The last two lines in this file are not gene counts and should be removed.

BAM Files

The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments against reference sequences, supporting short and long reads (up to 128 Mb) produced by different sequencing platforms. SAM is a text format file that is human-readable. The Binary Alignment/Map (BAM) keeps the same information as SAM, but in a compressed, binary format that is only machine readable.

If you use an app in BaseSpace that uses BAM files as input, the app locates the file when launched. If using BAM files in other tools, download the file to use it in the external tool.

Go to samtools.sourceforge.net/SAM1.pdf to see the exact SAM specification.
RNA Express Workflow

This chapter describes the workflow and modules that are used in the RNA Express app.

1 **Alignment.**
   
   Reads for each sample are aligned against the corresponding genome using the Spliced Transcripts Alignment to a Reference (STAR) software\(^1\). STAR alignments are converted to BAM files in real time with samtools\(^2\).
   
   There is no pre-treatment (trimming or filtering) of the FASTQ files. Instead a trim5' and trim3' option is passed to STAR, which does the trimming. In addition, STAR performs local alignment, allowing it to softclip read ends automatically (e.g. low quality or missed splice-junctions). STAR is run in a mode looking for novel junctions. After the initial alignment, RNA Express filters the junction list by confidence and retains only alignments across high confidence junctions. Only correctly paired alignments are reported for paired-end runs.

2 **Post-Alignment.**
   
   After alignment, the BAM files are sorted, indexed, and bedGraph coverage files are created using bedtools\(^3\). Alignments to abundant sequences are determined from genomic alignments based on annotation of abundant regions of the genome.

3 **Read Counting.**
   
   Gene expression is estimated at the gene level by counting the number of aligned reads that overlap each gene present in the annotation. The counting strategy is similar to htseq-count in the ‘union’ mode\(^4\).
   
   Ambiguous reads, either aligning to more than one locus or to a locus overlapping multiple genes, are not counted. Only if both reads of a pair overlap exons with the same unique gene_id, is the read counted towards that gene. The counts are written to .csv files per sample. The counter also reports basic statistics (number of reads filtered, assigned, unassigned).

4 **Global Expression.**
   
   The raw read counts are used as input for differential expression analysis using R and DESeq2\(^5\). The workflow writes an R script that loads all .csv files with read counts, generates a data frame from them and computes pairwise correlations. A Python script uses matplotlib to create a sample-to-sample correlation heat map. The correlations, a merged table with the read counts for all samples and the heat map are written to the output directory.

5 **Pairwise Differential Expression.**
   
   A new R script for the differential analysis is executed. This script loads the counts for all samples in this comparison and performs a pairwise differential expression analysis between them using DESeq2 (see online documentation for details of the model\(^6\)). The script filters low expressed genes (mean count across all samples less than 10) before testing to decrease the multiple testing burden. The DESeq2 variance model is used to detect outliers (based on extreme variation between replicates), which are also excluded.
   
   In the end, the status (filtered or passed) and the result of the analysis (mean expression, fold change, standard error, p-value, etc.) is reported for each gene. The script writes a table of raw counts across all replicates and plots a gene-level heat map sorted by hierarchical clustering. This heat map contains up to 5000 significantly differentially expressed genes, \(q < 0.05\).
Figure 5  RNA Express Workflow

References

2 SAMtools: samtools.sourceforge.net
3 Bedtools: bedtools.readthedocs.org
4 Htseq-count: www-huber.embl.de/users/anders/HTSeq/doc/count.html
5 DESeq2: www.bioconductor.org/packages/2.13/bioc/html/DESeq2.html
Notes
Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 1  Illumina General Contact Information

<table>
<thead>
<tr>
<th>Illumina Website</th>
<th><a href="http://www.illumina.com">www.illumina.com</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td><a href="mailto:techsupport@illumina.com">techsupport@illumina.com</a></td>
</tr>
</tbody>
</table>

Table 2  Illumina Customer Support Telephone Numbers

<table>
<thead>
<tr>
<th>Region</th>
<th>Contact Number</th>
<th>Region</th>
<th>Contact Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>North America</td>
<td>1.800.809.4566</td>
<td>Italy</td>
<td>800.874909</td>
</tr>
<tr>
<td>Austria</td>
<td>0800.296575</td>
<td>Netherlands</td>
<td>0800.0223859</td>
</tr>
<tr>
<td>Belgium</td>
<td>0800.81102</td>
<td>Norway</td>
<td>800.16836</td>
</tr>
<tr>
<td>Denmark</td>
<td>80882346</td>
<td>Spain</td>
<td>900.812168</td>
</tr>
<tr>
<td>Finland</td>
<td>0800.918363</td>
<td>Sweden</td>
<td>020790181</td>
</tr>
<tr>
<td>France</td>
<td>0800.911850</td>
<td>Switzerland</td>
<td>0800.563118</td>
</tr>
<tr>
<td>Germany</td>
<td>0800.180.8994</td>
<td>United Kingdom</td>
<td>0800.917.0041</td>
</tr>
<tr>
<td>Ireland</td>
<td>1.800.812949</td>
<td>Other countries</td>
<td>+44.1799.534000</td>
</tr>
</tbody>
</table>

Safety Data Sheets
Safety data sheets (SDSs) are available on the Illumina website at www.illumina.com/msds.

Product Documentation
Product documentation in PDF is available for download from the Illumina website. Go to www.illumina.com/support, select a product, then click Documentation & Literature.