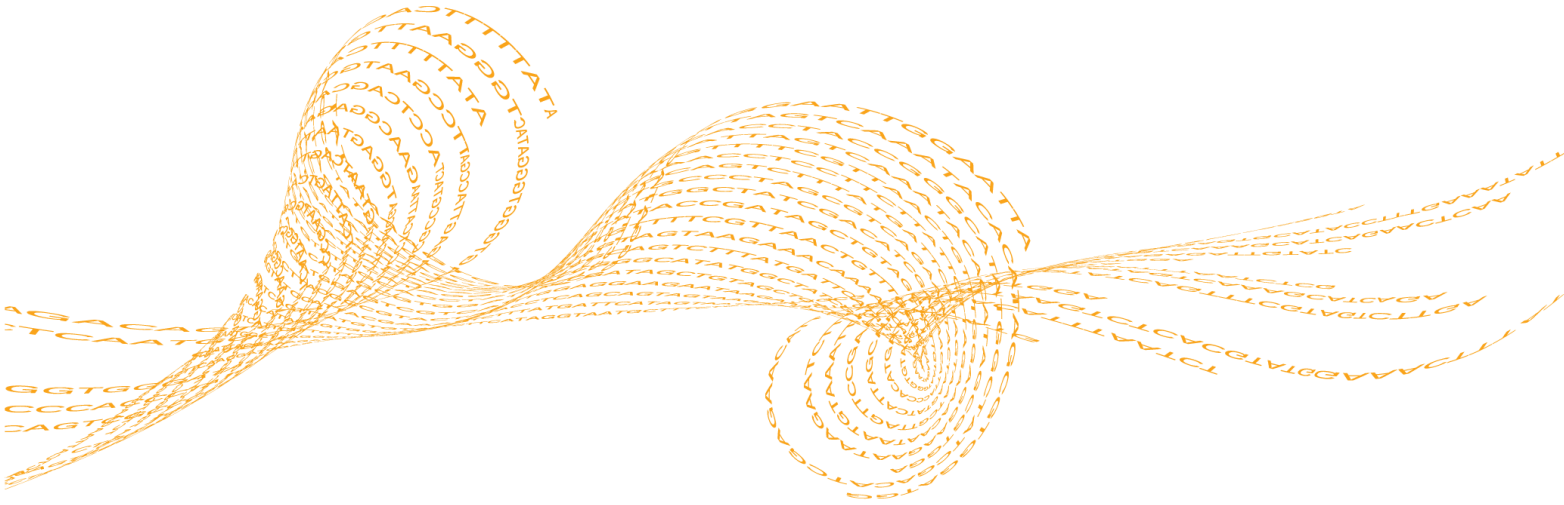
Local Run Manager
Small RNA Analysis Module
Workflow Guide

For Research Use Only. Not for use in diagnostic procedures.



ILLUMINA PROPRIETARY
Document # 1000000003345 v00
January 2016

illumina®

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).

© 2016 Illumina, Inc. All rights reserved.

Illumina, **24sure**, **BaseSpace**, **BeadArray**, **BlueFish**, **BlueFuse**, **BlueGnome**, **cBot**, **CSPro**, **CytoChip**, **DesignStudio**, **Epicentre**, **ForenSeq**, **Genetic Energy**, **GenomeStudio**, **GoldenGate**, **HiScan**, **HiSeq**, **HiSeq X**, **Infinium**, **iScan**, **iSelect**, **MiSeq**, **MiSeqDx**, **MiSeq FGx**, **NeoPrep**, **NextBio**, **Nextera**, **NextSeq**, **Powered by Illumina**, **SureMDA**, **TruGenome**, **TruSeq**, **TruSight**, **Understand Your Genome**, **UYG**, **VeraCode**, **verifi**, **VeriSeq**, the pumpkin orange color, and the streaming bases design are trademarks of Illumina, Inc. and/or its affiliate(s) in the U.S. and/or other countries. All other names, logos, and other trademarks are the property of their respective owners.

Overview

The Local Run Manager Small RNA analysis module uses Bowtie to align reads against databases for contaminants, mature miRNA, small RNA, and a genomic reference, in that order, and then measures abundance of short RNA sequences. It is suitable for identifying and quantifying miRNA expression and comparing abundance across samples. This workflow is designed specifically for RNA libraries prepared with the TruSeq Small RNA kit.

Input Requirements

In addition to sequencing data files generated during the sequencing run, such as base call files, the Small RNA analysis module requires the following references in FASTA file format. References are included with the installation of the Small RNA analysis module.

- A list of contaminants
- A list of mature miRNA
- A list of small RNA
- A reference genome

About This Guide

This guide provides instructions for setting up run parameters for sequencing and analysis parameters for the Small RNA analysis module. For information about the Local Run Manager dashboard and system settings, see the *Local Run Manager Software Guide (document # 1000000002702)*.

Set Parameters

1. Click **Create Run**, and select **Small RNA**.
2. Enter a run name that identifies the run from sequencing through analysis. Use alphanumeric characters, spaces, underscores, or dashes.
3. [Optional] Enter a run description to help identify the run. Use alphanumeric characters.

Specify Run Settings

1. Specify the number of index reads.
   - **0** for a run with no indexing
   - **1** for a single-indexed run
2. Enter the number of cycles for the run, if other than the default setting of 36 cycles.
3. [Optional] Specify custom primers to be used for the run.

NOTE
By default, the Small RNA analysis module is set to the library type TruSeq Small RNA and the read type Single Read. Read lengths are set to 36 cycles for Read 1 and 6 cycles for Index 1 Read.

Specify Module-Specific Settings

By default, the Small RNA analysis module uses Bowtie for alignment.

No module-specific settings are required for the Small RNA analysis module.

Specify a Genome Reference

1. To view available reference genomes, click the Genome Folder drop-down list and select a reference genome for the run.

Specify Samples for the Run

Specify samples for the run using the following options:

- **Enter samples manually**—Use the blank table on the Create Run screen.
- **Import samples**—Navigate to an external file in a comma-separated values (*.csv) format. A template is available for download on the Create Run screen.

After you have populated the samples table, you can export the sample information to an external file, and use the file as a reference when preparing libraries or import the file for another run.

Enter Samples Manually

1. Adjust the samples table to an appropriate number of rows.
   - Click the **+** icon to add a row.
   - Use the up/down arrows to add multiple rows. Click the **+** icon.
   - Click the **x** icon to delete a row.
   - Right-click on a row in the table and use the commands in the drop-down menu.
2. Enter a unique sample ID in the Sample ID field.

Use alphanumeric characters, dashes, or underscores.

3 [Optional] Enter a sample description in the Sample Description field. Use alphanumeric characters, dashes, underscores, or spaces.

4 Expand the Index 1 (i7) drop-down list and select an Index 1 adapter.

5 [Optional] Click the **Export** icon to export sample information in *.csv format.

6 When finished, click **Save Run**.

Import Samples

1 Click **Template**. The template file contains the correct column headings for import.

2 Enter the sample information in each column for the samples in the run, and then save the file.

3 Click **Import Samples** and browse to the location of the sample information file.

4 When finished, click **Save Run**.

Analysis Methods

The Small RNA analysis module performs the following analysis steps and then writes analysis output files to the Alignment folder.

- Demultiplexes index reads
- Generates FASTQ files
- Aligns to a reference
- Measures the abundance of various types of short RNA sequences

Demultiplexing

Demultiplexing compares each Index Read sequence to the index sequences specified for the run. No quality values are considered in this step.

Index reads are identified using the following steps:

- Samples are numbered starting from 1 based on the order they are listed for the run.
- Sample number 0 is reserved for clusters that were not assigned to a sample.
- Clusters are assigned to a sample when the index sequence matches exactly or when there is up to a single mismatch per Index Read.

FASTQ File Generation

After demultiplexing, the software generates intermediate analysis files in the FASTQ format, which is a text format used to represent sequences. FASTQ files contain reads for each sample and the associated quality scores. Any controls used for the run and clusters that did not pass filter are excluded.

Each FASTQ file contains reads for only 1 sample, and the name of that sample is included in the FASTQ file name. FASTQ files are the primary input for alignment.

Adapter Trimming

The Small RNA analysis module performs adapter trimming by default.

During longer runs, clusters can sequence beyond the sample DNA and read bases from a sequencing adapter. To prevent sequencing into the adapter, the adapter sequence is trimmed before the sequence is written to the FASTQ file. Trimming the adapter sequence avoids reporting false mismatches with the reference sequence and improves alignment accuracy and performance.

Adapter Sequences

When using the Small RNA analysis module, the following adapter sequence is trimmed:

```
TGGAATTCTCGGGTGCCAAGGC
```

Alignment

During the alignment step, the Bowtie alignment tool aligns clusters from each sample against references in the order of contaminants, mature miRNA, and RNA. If a reference genome was specified for the run, clusters are aligned to the full genome.

All samples are aligned to the same references. Alignment is performed using the following criteria:

- Clusters that aligns to the contaminants reference is considered to be a contaminant, even if the cluster also aligns to another reference.
- Clusters that match the miRNA reference, but not the contaminants reference, are considered to be miRNA. Only exact matches to the miRNA reference are counted.
- To be considered a match, alignment begins at the start of the reference sequence, must occur on the same strand, and be of an identical length. The length of a sequence is measured against the reference sequence after adapter trimming.
- If a sequence aligns to multiple references with the same number of mismatches, counts are divided equally across references.

View Analysis Results

1. From the Local Run Manager dashboard, click the run name.
2. From the Run Overview tab, review the sequencing run metrics.
3. [Optional] Click the **Copy to Clipboard** icon for access to the output run folder.
4. Click the Sequencing Information tab to review run parameters and consumables information.
5. Click the Samples and Results tab to view the analysis report.
   - If analysis was repeated, expand the Select Analysis drop-down and select the appropriate analysis.
   - From the left navigation bar, select a sample name to view the report for another sample.
6. [Optional] Click the **Copy to Clipboard** icon for access to the Analysis folder.

Analysis Report

Analysis results are summarized on the Samples and Results tab. The report is also available in a PDF file format for each sample and as an aggregate report in the Analysis folder.

Sample Information

Table 1 Sample Information Table

Column Heading	Description
Sample ID	The sample ID provided when the run was created.
Sample Name	The sample name provided when the run was created.
Run Folder	The name of the run folder.
Total PF Reads	The total number of reads passing filter.

Quality Control Statistics

Quality control statistics are provided in the form of graphs, such as a histogram or pie charts. The following graphs report quality control statistics:

- **Small RNA Length Distribution**—A histogram showing the number of reads for each read length after trimming.
- **Read Distribution**—A pie chart showing the number and percentage of reads in the categories of abundant, genome, miRNA, other RNA, and unaligned.
- **Abundant Distribution**—A pie chart showing the number and percentage of reads for abundant small RNA. Subcategories are species-dependent.
- **miRNA Distribution**—A pie chart showing the number and percentage of reads for miRNA in the categories of isomiR (known precursor) and known mature miRNA. Known precursors are listed in the miRNA database.
- **Other RNA Distribution**—A pie chart showing the number and percentage of reads for other RNA. Subcategories are species-dependent. Precursors are reads that slign to the precursor sequences, but are not mature miRNA or isomiRs.

Summary of Sequences

Summaries of the top-10 sequences with the most reads for the following small RNA marker types are provided in the analysis report for each sample.

- **Mature miRNA**—Only exact matches to miRNA records are listed. For the full list, see Hits.txt.
- **IsomiR (known precursor)**—Reads are counted if on the same strand of a precursor, but not on the 5' or 3' mature miRNA sequence. IsomiR hits are filtered to remove artifacts of PCR and sequencing errors. For the full list, see FilteredIsoHits.txt.
- **Precursor group**—Precursor reads are grouped based on the mature miRNA they produce. For the full list, see PrecursorHits.txt.
- **miRNA families**—Reads are counted according to sequences listed in the miRBase. For the full list, see FamilyHits.txt.
- **piRNA**—Reads are counted if they align to a piRNA sequence listed in the NCBI Nucleotide database, but have not been counted as a isomiR. For the full list, see PiRNAHits.txt.

For each category, data is provided in the following tables.

Table 2 Number of Marker Types Table

Column Heading	Description
Number of [marker type] with Reads	The number of marker types matched in the run.
Total Number of Reads	The total number of reads assigned to the marker type.

Table 3 Top 10 of Marker Type Table

Column Heading	Description
Top 10 [marker type]	The top 10 marker type sequences with the most matching reads.
Number of Reads	The total number of reads assigned to each marker type sequence.

Analysis Output Files

The following analysis output files are generated for the Small RNA analysis module and provide analysis results for alignment and the number of reads counted for each data type. Analysis output files are located in the Alignment folder.

File Name	Description
Demultiplexing (*.demux)	Intermediate files containing demultiplexing results.
FASTQ (*.fastq.gz)	Intermediate files containing quality scored base calls. FASTQ files are the primary input for the alignment step.
Abundant alignment files in the BAM format (*abundant.bam)	Contains alignments against abundant sequences.
miRNA alignment files in the BAM format (*mirna.bam)	Contains alignments against mature miRNA sequences.
RNA alignment files in the BAM format (*rna.bam)	Contains alignments against other RNA sequences.
Genome alignment files in the BAM format (*genome.bam)	Contains alignments against the reference genome, if provided.
Hits.txt	Contains the number of reads counted for each mature miRNA, for each sample.
FilteredIsoHits.txt	Contains the number of reads counted as filtered isomiR listed, for each sample.
PrecursorHits.txt	Contains the number of reads counted for each precursor group for each sample.
FamilyHits.txt	Contains the number of reads counted for each miRNA family listed, for each sample.
PiRNAHits.txt	Contains the number of reads counted for each piRNA listed, for each sample.

Demultiplexing File Format

The process of demultiplexing reads the index sequence attached to each cluster to determine from which sample the cluster originated. The mapping between clusters and sample number are written to 1 demultiplexing (*.demux) file for each tile of the flow cell.

The demultiplexing file naming format is s_1_X.demux, where X is the tile number.

Demultiplexing files start with a header:

- Version (4 byte integer), currently 1
- Cluster count (4 byte integer)

The remainder of the file consists of sample numbers for each cluster from the tile.

When the demultiplexing step is complete, the software generates a demultiplexing file named DemultiplexSummaryF1L1.txt.

- In the file name, **F1** represents the flow cell number.
- In the file name, **L1** represents the lane number.
- Demultiplexing results in a table with 1 row per tile and 1 column per sample, including sample 0.
- The most commonly occurring sequences in index reads.

FASTQ File Format

FASTQ file is a text-based file format that contains base calls and quality values per read. Each record contains 4 lines:

- The identifier
- The sequence
- A plus sign (+)
- The quality scores in an ASCII encoded format

The identifier is formatted as:

@Instrument:RunID:FlowCellID:Lane:Tile:X:Y ReadNum:FilterFlag:0:SampleNumber

Example:

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

BAM File Format

A BAM file (*.bam) is the compressed binary version of a SAM file that is used to represent aligned sequences up to 128 Mb. SAM and BAM formats are described in detail at https://samtools.github.io/hts-specs/SAMv1.pdf.

BAM files use the file naming format of SampleName_S#.bam, where # is the sample number determined by the order that samples are listed for the run.

BAM files contain a header section and an alignments section:

- **Header**—Contains information about the entire file, such as sample name, sample length, and alignment method. Alignments in the alignments section are associated with specific information in the header section.
- **Alignments**—Contains read name, read sequence, read quality, alignment information, and custom tags. The read name includes the chromosome, start coordinate, alignment quality, and the match descriptor string.

The alignments section includes the following information for each or read pair:

- **RG:** Read group, which indicates the number of reads for a specific sample.
- **BC:** Barcode tag, which indicates the demultiplexed sample ID associated with the read.
- **SM:** Single-end alignment quality.
- **AS:** Paired-end alignment quality.
- **NM:** Edit distance tag, which records the Levenshtein distance between the read and the reference.
- **XN:** Amplicon name tag, which records the amplicon tile ID associated with the read.

BAM files are suitable for viewing with an external viewer such as IGV or the UCSC Genome Browser.

BAM index files (*.bam.bai) provide an index of the corresponding BAM file.

Supplementary Output Files

The following output files provide supplementary information, or summarize run results and analysis errors. Although, these files are not required for assessing analysis results, they can be used for troubleshooting purposes. All files are located in the Alignment folder unless otherwise specified.

File Name	Description
AdapterTrimming.txt	Lists the number of trimmed bases and percentage of bases for each tile. This file is present only if adapter trimming was specified for the run.
AnalysisLog.txt	Processing log that describes every step that occurred during analysis of the current run folder. This file does not contain error messages. Located in the root level of the run folder.
AnalysisError.txt	Processing log that lists any errors that occurred during analysis. This file is present only if errors occurred. Located in the root level of the run folder.
CompletedJobInfo.xml	Written after analysis is complete, contains information about the run, such as date, flow cell ID, software version, and other parameters. Located in the root level of the run folder.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with 1 row per tile and 1 column per sample.
SmallRNARunStatistics.xml	Contains summary statistics specific to the run. Located in the root level of the run folder.

Analysis Folder

The analysis folder holds the files generated by the Local Run Manager software.

The relationship between the output folder and analysis folder is summarized as follows:

- During sequencing, Real-Time Analysis (RTA) populates the output folder with files generated during image analysis, base calling, and quality scoring.
- RTA copies files to the analysis folder in real time. After RTA assigns a quality score to each base for each cycle, the software writes the file RTAComplete.xml to both folders.
- When the file RTAComplete.xml is present, analysis begins.
- As analysis continues, Local Run Manager writes output files to the analysis folder, and then copies the files back to the output folder.

Folder Structure

Data

Intensities

BaseCalls

Alignment—Contains *.bam and *.vcf files, and files specific to the analysis module.

- **L001**—Contains one subfolder per cycle, each containing *.bcl files.
  - Sample1_S1_L001_R1_001.fastq.gz
  - Sample2_S2_L001_R1_001.fastq.gz
  - Undetermined_S0_L001_R1_001.fastq.gz
- **L001**—Contains *.locs files, 1 for each tile.
- **RTA Logs**—Contains log files from RTA software analysis.

InterOp—Contains binary files used by Sequencing Analysis Viewer (SAV).

Logs—Contains log files describing steps performed during sequencing.

Queued—A working folder for software; also called the copy folder.

AnalysisError.txt

AnalysisLog.txt

CompletedJobInfo.xml

QueuedForAnalysis.txt

[WorkflowName]RunStatistics

RTAComplete.xml

RunInfo.xml

runParameters.xml

Alignment Folders

Each time that analysis is requeued, the Local Run Manager creates an Alignment folder named **AlignmentN**, where N is a sequential number.

Custom Analysis Settings

Custom analysis settings are intended for technically advanced users. If settings are applied incorrectly, serious problems can occur.

Add a Custom Analysis Setting

1. From the Module-Specific Settings section of the Create Run screen, click **Show advanced module settings**.
2. Click **Add custom setting**.
3. In the custom setting field, enter the setting name as listed in the Available Analysis Settings section.
4. In the setting value field, enter the setting value.
5. To remove a setting, click the **x** icon.

Available Analysis Settings

- **Adapter Trimming**—By default, adapter trimming is enabled in the Small RNA analysis module. To specify a different adapter, use the Adapter setting. The same adapter sequence is trimmed for Read 1 and Read 2.
  - To specify 2 adapter sequences, separate the sequences with a plus (+) sign.
  - To specify a different adapter sequence for Read 2, use the AdapterRead2 setting.

Setting Name	Setting Value
Adapter	Enter the sequence of the adapter to be trimmed.
AdapterRead2	Enter the sequence of the adapter to be trimmed.

Notes

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 4 Illumina General Contact Information

Website	www.illumina.com
Email	techsupport@illumina.com

Table 5 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Japan	0800.111.5011
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
Belgium	0800.81102	Norway	800.16836
China	400.635.9898	Singapore	1.800.579.2745
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	Taiwan	00806651752
Hong Kong	800960230	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000
Italy	800.874909		

Safety data sheets (SDSs)—Available on the Illumina website at support.illumina.com/sds.html.

Product documentation—Available for download in PDF from the Illumina website. Go to support.illumina.com, select a product, then select **Documentation & Literature**.

Illumina
5200 Illumina Way
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com