MiSeq Reporter
Small RNA Workflow
Reference Guide

FOR RESEARCH USE ONLY

illumına®

Revision History

Part #	Revision	Date	Description of Change
15042320	A	June 2013	Initial release. The information provided within was previously included in the *MiSeq Reporter User Guide*. With this release, the *MiSeq Reporter User Guide* contains information about the interface, how to view run results, how to requeue a run, and how to install and configure the software. Information specific to the Small RNA workflow is provided in this guide.

Introduction

The Small RNA workflow aligns reads against databases for contaminants, mature miRNA, small RNA, and a genomic reference, in that order. Alignment is performed using Bowtie.

In the MiSeq Reporter Analyses tab, a run folder associated with the Small RNA workflow is represented with the letter **S**. For more information about the software interface, see the *MiSeq Reporter User Guide (part # 15042295)*.

This guide describes the analysis steps performed in the Small RNA workflow, the types of data that appear on the interface, and the analysis output files generated by the workflow.

Workflow Requirements

The Small RNA workflow requires reference databases for alignment and identification. Specify the path to the folder containing the references in the sample sheet.

- **Contaminants**—A list of contaminants in FASTA file format.
- **miRNA**—A list of mature miRNA in FASTA file format.
- **RNA**—A list of small RNA in FASTA file format.
- **Reference genome**—[Optional] If provided, reads are aligned against the full reference genome.

  For more information, see the *MiSeq Sample Sheet Quick Reference Guide (part # 15028392)*.

Small RNA Workflow Overview

The Small RNA workflow measures the abundance of various types of short RNA sequences, particularly miRNA. It is suitable for identifying and quantifying miRNA expression and for comparing abundance across samples.

The workflow generates FASTQ files, trims adapters, and aligns reads.

> NOTE
> The FASTQ (fastq.gz) files generated during the Small RNA workflow are intermediate files, and do not contain all of the reads generated during sequencing. These files only contain reads that were not removed by the duplicate filtering or alignment steps.

FASTQ File Generation

MiSeq Reporter 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 their quality scores, excluding reads identified as in-line controls and clusters that did not pass filter.

FASTQ files are the primary input for alignment. The files are written to the BaseCalls folder (Data\Intensities\BaseCalls) in the MiSeqAnalysis folder, and then copied to the BaseCalls folder in the MiSeqOutput folder. Each FASTQ file contains reads for only one sample, and the name of that sample is included in the FASTQ file name. For more information about FASTQ files, see the *MiSeq Reporter User Guide (part # 15042295)*.

Adapter Trimming

The adapter trimming step trims the adapter sequences from all reads. A histogram of read lengths is written to TrimmerHistogram.txt.

For the Small RNA workflow, the default adapter sequencing **TGGAATTCTCGGGTGCCAAGGC**. Optionally, use the Adapter sample sheet setting to specify an alternative adapter sequence. For more information, see *Optional Settings* on page 10.

Alignment

The Small RNA workflow uses Bowtie v0.12.8 to align each cluster against references specified in the sample sheet. Alignment to the references occurs in a specific order of precedence, shown below. All samples must use the same set of reference databases.

Database Name	Order of Precedence	Column in Sample Sheet
Contaminants	1	Specified in Contaminants column
Mature miRNA	2	Specified in miRNA column
RNA	3	Specified in RNA column
Reference genome	4	Specified in GenomeFolder column

Statistical Reporting

Clusters are aligned in order of precedence using the following criteria:
- A cluster that aligns to the contaminants database is considered to be a contaminant even if it also aligns to the reference genome.
- Clusters that match the miRNA database, but not contaminants, are counted as miRNA.

- By default, only exact matches to mature miRNA records are counted.
- Matches must align to the start of the reference sequence, on the same strand, with adapter-trimmed read lengths identical to the length of the reference sequence.
- If the same sequence maps to multiple database records with the same number of mismatches, counts are split across records.

Small RNA Summary Tab

The Summary for the Small RNA workflow includes a clusters graph and trimmed lengths graph.

- ▶ **Clusters Graph**—Shows numbers of raw clusters, clusters passing filter, clusters that did not align, clusters not associated with an index, and duplicates.
- ▶ **Trimmed Lengths Graph**—Shows a histogram of reads that were trimmed.

Clusters Graph

Y Axis	X Axis	Description
Clusters	Raw	The total number of clusters detected in the run.
	PF	The total number of clusters passing filter in the run.
	Unaligned	The total number of clusters passing filter that did not align to the reference genome, if applicable. Clusters that are unindexed are not included in the unaligned count.
	Unindexed	The total number of clusters passing filter that were not associated with any index sequence in the run.

Trimmed Lengths

Y Axis	X Axis	Description
Clusters	Trimmed Lengths	Histogram of reads indicating length at trimming because they reached adapter.

Small RNA Details Tab

The Details tab for the Small RNA workflow includes a samples table, pie chart, and graph.

- ▸ **Samples Table**—Summarizes the sequencing results for each sample.
- ▸ **Small RNA Pie Chart**—Visualization of identified clusters.
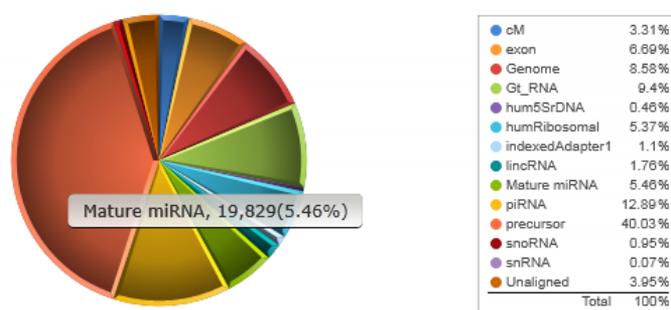- ▸ **Small RNA Graph**—Visualization of mature miRNA sequences for a sample.

Samples Table

Column	Description
#	An ordinal identification number in the table.
Sample ID	The sample ID from the sample sheet. Sample ID must always be a unique value.
Sample Name	The sample name from the sample sheet.
Cluster Raw	The number of raw clusters detected for the sample.
Cluster PF	The number of clusters passing filter for the sample.
Cluster Align Contam	The number of clusters that match records in the Contaminants database.
Cluster Align miRNA	The number of clusters that exactly match records in the Mature miRNA database.
Cluster Align RNA	The number of clusters that match records in the RNA database.
Cluster Align Genome	The number of clusters that match records in the genomic database.
Cluster Unaligned	The number of clusters that did not align against any reference database.

Small RNA Pie Chart

The Small RNA pie chart provides a visualization of clusters identified as mature miRNA, other forms of RNA, genomic sequence, or contaminants.

Figure 1   Small RNA Pie Chart

● cM	3.31%
● exon	6.69%
● Genome	8.58%
● Gt_RNA	9.4%
● hum5SrDNA	0.46%
● humRibosomal	5.37%
● indexedAdapter1	1.1%
● lincRNA	1.76%
● Mature miRNA	5.46%
● piRNA	12.89%
● precursor	40.03%
● snoRNA	0.95%
● snRNA	0.07%
● Unaligned	3.95%
Total	100%

Mature miRNA, 19,829(5.46%)

Common categories for the Small RNA pie chart are as follows:

- ▸ Unaligned clusters that did not align against any reference
- ▸ Genome clusters that aligned to the reference genome
- ▸ miRNA clusters that aligned to the mature miRNA database

Hits to the mature miRNA database are counted only if the cluster aligned to the correct strand and position for the mature miRNA.

The remaining category names in the Small RNA pie chart are taken from the FASTA file names in the databases. For example, if the RNA database contains a file named rRNA.fa, then matches to this file are reported as the category **rRNA**.

Small RNA Graph

The Small RNA graph provides a plot of the common mature miRNA sequences for a sample and their abundances. The most common miRNA sequences for the selected sample (up to ten records) are shown in proportion to the number of clusters matched.

Optional Settings

By default, adapter trimming is performed for the Small RNA workflow using the standard adapter sequence **TGGAATTCTCGGGTGCCAAGGC**. Use the sample sheet setting to optionally override the default sequence.

Sample Sheet Setting Name	Description
Adapter	Specify the 5' portion of the adapter sequence to prevent reporting sequence beyond the sample DNA.

For more information about sample sheet settings, see the *MiSeq Sample Sheet Quick Reference Guide (part # 15028392)*.

Analysis Output Files

The following analysis output files are generated for the Small RNA workflow and provide matches to the genomic databases.

File Name	Description
*.export.contam, *.contam	Lists the number of clusters that exactly match records in the Contaminants database. These files are used to populate the samples table and pie chart. Located in Data\Intensities\BaseCalls\Alignment.
*.export.genome, *.genome	Lists the number of clusters that exactly match records in the genomic database. These files are used to populate the samples table and pie chart. Located in Data\Intensities\BaseCalls\Alignment.
*.export.mirna, *.mirna	Lists the number of clusters that exactly match records in the Mature miRNA database. These files are used to populate the samples table and pie chart. Located in Data\Intensities\BaseCalls\Alignment.
*.export.rna, *.rna	Lists the number of clusters that exactly match records in the RNA database. These files are used to populate the samples table and pie chart. Located in Data\Intensities\BaseCalls\Alignment.
TrimmerHistogram.txt	Contains a histogram of masked read lengths. Located in Data\Intensities\BaseCalls\Alignment.

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.

File Name	Description
AdapterTrimming.txt	Lists the number of trimmed bases and percentage of bases for each tile. Located in Data\Intensities\BaseCalls\Alignment.
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.

File Name	Description
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one column per sample. Located in Data\Intensities\BaseCalls\Alignment.
SmallRNARunStatistics.xml	Contains summary statistics specific to the run. Located in the root level of the run folder.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 1   Illumina General Contact Information

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

Table 2   Illumina Customer Support Telephone Numbers

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

MSDSs

Material safety data sheets (MSDSs) 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**.