MiSeq Reporter Generate FASTQ Workflow Guide

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

ILLUMINA PROPRIETARY
Document # 15042322 v01
September 2015

illumina®

Revision History

Document #	Date	Description of Change
Document # 15042322 v01	September 2015	Changed the name of the guide from the MiSeq Reporter Generate FASTQ Workflow Reference Guide to the MiSeq Reporter Generate FASTQ Workflow Guide. Updated the read stitching description to include information on what occurs when the Q-score is the same in an overlap region, and information on alignment in the BAM file for stitched reads. Updated the default setting for ReverseComplement from 1 to 0 in the Sample Sheet Settings for Analysis section.
Part # 15042322 Rev. C	December 2014	Added a note in the Demultiplexing section about the default index recognition for index pairs that differ by < 3 bases.
Part # 15042322 Rev. B	August 2013	• Updated to MiSeq Reporter v2.3: added sample sheet setting StitchReads and description of read stitching. • Added descriptions of Adapter settings and ReverseComplement.
Part # 15042322 Rev. 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 Generate FASTQ workflow is provided in this guide.

Introduction

The Generate FASTQ workflow generates intermediate analysis files in the FASTQ file format, and then exists the workflow. No alignment is performed.

In the MiSeq Reporter Analyses tab, a run folder associated with the Generate FASTQ workflow is represented with the letter **G**. For more information about the software interface, see the *MiSeq Reporter Software Guide (document # 15042295)*.

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

Generate FASTQ Workflow

The Generate FASTQ workflow performs a demultiplexing step for runs with index reads and multiple samples. After demultiplexing, the workflow generates analysis output in the FASTQ file format.

After FASTQ file generation, the workflow exits analysis. The FASTQ files are suitable for secondary analysis using third-party analysis tools.

Demultiplexing

Demultiplexing separates data from pooled samples based on short index sequences that tag samples from different libraries. Index reads are identified using the following steps:

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

NOTE

Illumina indexes are designed so that any index pair differs by ≥ 3 bases, allowing for a single mismatch in index recognition. Index sets that are not from Illumina can include pairs of indexes that differ by < 3 bases. In such cases, the software detects the insufficient difference and modifies the default index recognition (mismatch=1). Instead, the software performs demultiplexing using only perfect index matches (mismatch=0).

When demultiplexing is complete, 1 demultiplexing file named DemultiplexSummaryF1L1.txt is written to the Alignment folder, and summarizes the following information:

- In the file name, **F1** represents the flow cell number.
- In the file name, **L1** represents the lane number, which is always L1 for MiSeq.
- Reports demultiplexing results in a table with 1 row per tile and 1 column per sample, including sample 0.
- Reports the most commonly occurring sequences for the index reads.

FASTQ File Generation

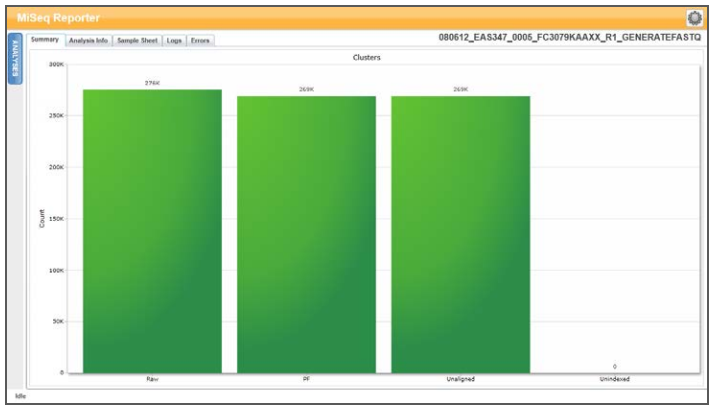
MiSeq Reporter generates intermediate analysis files in the FASTQ format, which is a text format used to represent sequences. FASTQ files contain reads for each sample and their quality scores, excluding reads identified as inline controls and clusters that did not pass filter.

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 1 sample, and the name of that sample is included in the FASTQ file name. For more information about FASTQ files, see the *MiSeq Reporter Software Guide (document # 15042295)*.

Generate FASTQ Summary Tab

Results written to FASTQ files appear in the clusters graph on the Summary tab for the run. The clusters graph shows numbers of raw clusters, clusters passing filter, clusters that did not align, and clusters not associated with an index.

Figure 1 Example Clusters Graph on Summary Tab



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.

Optional Settings for the Generate FASTQ Workflow

Sample sheet settings are optional commands that control various analysis parameters.

Settings are used in the Settings section of the sample sheet and require a setting name and a setting value.

If you are viewing or editing the sample sheet in Excel, the setting name resides in the first column and the setting value in the second column.

If you are viewing or editing the sample sheet in a text editor such as Notepad, follow the setting name is by a comma and a setting value. Do not include a space between the comma and the setting value.

Example: Adapter,CTGTCTCTTATACACATCT

The following optional settings are compatible with the Generate FASTQ workflow.

Sample Sheet Settings for Analysis

Parameter	Description
Adapter	Specify the 5' portion of the adapter sequence to prevent reporting sequence beyond the sample DNA. Illumina recommends adapter trimming for Nextera libraries and Nextera Mate Pair libraries. To specify 2 or more adapter sequences, separate the sequences by a plus (+) sign. For example: CTGTCTCTTATACACATCT+AGATGTGTATAAGAGACAG
AdapterRead2	Specify the 5' portion of the Read 2 adapter sequence to prevent reporting sequence beyond the sample DNA. Use this setting to specify a different adapter other than the one specified in the **Adapter** setting.
ReverseComplement	Settings are 0 or 1. Default is 0. If set to true (1), all reads are reverse-complemented as they are written to FASTQ files. Set this setting to 1 when using the Generate FASTQ workflow with Nextera Mate Pair libraries.
StitchReads	Settings are 0 or 1. Default is 0, paired-end reads are not stitched. If set to true (1), paired-end reads that overlap are stitched to form a single read. To be stitched, a minimum of 10 bases must overlap between Read 1 and Read 2. Paired-end reads that cannot be stitched are converted to 2 single reads. This setting requires MiSeq Reporter v2.3, or later.

Read Stitching

MiSeq Reporter v2.3, or later, is required to use the optional StitchReads setting.

When set to true (1), paired-end reads that overlap are stitched to form a single read in the FASTQ file. At each overlap position, the consensus stitched read has the base call and quality score of the read with higher Q-score.

Read stitching can only be applied to alignment and variant calling using the TruSeq Amplicon workflow, TruSight Tumor (15 Genes) workflow, and Amplicon DS workflow, but might be allowable input with some third-party analysis tools using the FASTQ files.

For each paired read, a minimum of 10 bases must overlap between Read 1 and Read 2 to be a candidate for read stitching. The minimum threshold of 10 bases minimizes the number of reads that are stitched incorrectly due to a chance match. Candidates for read stitching are scored as follows:

- For each possible overlap of 10 base pairs or more, a score of 1 – MismatchRate is calculated.
- Perfectly matched overlaps have a MismatchRate of 0, resulting in a score of 1.
- Random sequences have an expected score of 0.25.
- If the best overlap has a score of ≥ 0.9 ***and*** the score is ≥ 0.1 higher than any other candidate, then the reads are stitched together at this overlap.

Paired-end reads that cannot be stitched are converted to 2 single reads in the FASTQ file.

Optional Configurable

By default, MiSeq Reporter does not generate FASTQ files for index reads. To change this setting, use the **CreateFastqForIndexReads** setting in the MiSeq Reporter.exe.config file.

Setting Name	Values and Description
CreateFastqForIndexReads	0 (false; default) 1 (true) If set to false, FASTQ files are not generated for index reads. If set to true, FASTQ files are generated for index reads.

For more information about configuration settings, see the *MiSeq Reporter Software Guide (document # 15042295)*.

Analysis Output Files

The Generate FASTQ workflow generates analysis results for alignment in the FASTQ file format.

File Name	Description
*.fastq.gz	Contains base calls and quality values for each read per sample. Located in Data\Intensities\BaseCalls.

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** as shown in the following example:

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

FASTQ File Names

FASTQ files are named with the sample name and the sample number. The sample number is a numeric assignment based on the order that the sample is listed in the sample sheet. For example:

Data\Intensities\BaseCalls\samplename_S1_L001_R1_001.fastq.gz

- **samplename**—The sample name provided in the sample sheet. If a sample name is not provided, the file name includes the sample ID.
- **S1**—The sample number based on the order that samples are listed in the sample sheet starting with 1. In this example, S1 indicates that this sample is the first sample listed in the sample sheet.

  NOTE
  Reads that cannot be assigned to any sample are written to a FASTQ file for sample number 0, and excluded from downstream analysis.

- **L001**—The lane number. This segment is always L001 with the single-lane flow cell.
- **R1**—The read. In this example, R1 means Read 1. For a paired-end run, a file from Read 2 includes R2 in the file name.
- **001**—The last segment is always 001.

FASTQ files are compressed in the GNU zip format, as indicated by *.gz in the file name. FASTQ files can be uncompressed using tools such as gzip (command-line) or 7-zip (GUI).

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. This file is present only if adapter trimming was specified for the run. 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.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with 1 row per tile and 1 column per sample. Located in Data\Intensities\BaseCalls\Alignment.
ErrorsAndNoCallsByLaneTile ReadCycle.csv	A comma-separated values file that contains the percentage of errors and no-calls for each tile, read, and cycle. Located in Data\Intensities\BaseCalls\Alignment.
Mismatch.htm	Contains histograms of mismatches per cycle and no-calls per cycle for each tile. Located in Data\Intensities\BaseCalls\Alignment.
ResequencingRunStatistics.xml	Contains summary statistics specific to the run. Located in the root level of the run folder.
Summary.xml	Contains a summary of mismatch rates and other base calling results. Located in Data\Intensities\BaseCalls\Alignment.
Summary.htm	Contains a summary web page generated from Summary.xml. Located in Data\Intensities\BaseCalls\Alignment.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 1 Illumina General Contact Information

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
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
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

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
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