MiSeq Reporter Library QC

Workflow Guide

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

ILLUMINA PROPRIETARY
Document # 15042316 v01
September 2015

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).

Revision History

Document #	Date	Description of Change
Document # 15042316 v01	September 2015	Changed the name of the guide from the MiSeq Reporter Library QC Workflow Reference Guide to the MiSeq Reporter Library QC Workflow Guide. In the BAM File Format section, revised the description of the alignment information in the file header, and updated the link for SAM format specifications. Added RunBwaAln parameter for setting the alignment method from BWA-MEM to BWA-backtrack. Updated the default setting for ReverseComplement from 1 to 0 in the Sample Sheet Settings for Analysis section.
Part # 15042316 Rev. C	February 2014	Updated to change introduced in MiSeq Reporter v2.4: • Added the alignment method to the description of the BAM file header.
Part # 15042316 Rev. B	August 2013	Added optional settings section with descriptions of Adapter, FlagPCRDuplicates, and ReverseComplement.
Part # 15042316 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 Library QC workflow is provided in this guide.

Introduction

Intended for evaluating the quality of DNA libraries, the Library QC workflow aligns reads against the reference genome specified in the sample sheet.

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

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

Library QC Workflow Overview

The Library QC workflow is intended for evaluating abundance, fragment length, and sample quality of DNA libraries.

The alignment step uses a faster, less sensitive setting with the Burrows-Wheeler Aligner (BWA) that provides an efficient turnaround time. After alignment, MiSeq Reporter calculates diversity and fragment lengths, and generates a sample report named LibraryQC.html that is written to the Alignment folder. The sample report lists the characteristics of each DNA sample in terms of percentage of reads aligned. Data written to the sample report appear in the samples table.

Library QC Summary Tab

The Summary tab for the Library QC workflow includes a low percentages graph, high percentages graph, clusters graph, and mismatch graph.

- **Low percentages graph**—Shows phasing, prephasing, and mismatches in percentages. Low percentages indicate good run statistics.
- **High percentages graph**—Shows clusters passing filter, alignment to a reference, and intensities in percentages. High percentages indicate good run statistics.
- **Clusters graph**—Shows numbers of raw clusters, clusters passing filter, clusters that did not align, clusters not associated with an index, and duplicates.
- **Mismatch graph**—Shows mismatches per cycle. A mismatch refers to any mismatch between the sequencing read and a reference genome after alignment.

Low Percentages Graph

Y Axis	X Axis	Description
Percent	Phasing 1	The percentage of molecules in a cluster that fall behind the current cycle within Read 1.
	Phasing 2	The percentage of molecules in a cluster that fall behind the current cycle within Read 2.
	Prephasing 1	The percentage of molecules in a cluster that run ahead of the current cycle within Read 1.
	Prephasing 2	The percentage of molecules in a cluster that run ahead of the current cycle within Read 2.
	Mismatch 1	The average percentage of mismatches for Read 1 over all cycles.
	Mismatch 2	The average percentage of mismatches for Read 2 over all cycles.

High Percentages Graph

Y Axis	X Axis	Description
Percent	PF	The percentage of clusters passing filters.
	Align 1	The percentage of clusters that aligned to the reference in Read 1.
	Align 2	The percentage of clusters that aligned to the reference in Read 2.
	I20 / I1 1	The ratio of intensities at cycle 20 to the intensities at cycle 1 for Read 1.
	I20 / I1 2	The ratio of intensities at cycle 20 to the intensities at cycle 1 for Read 2.
	PE Resynthesis	The ratio of first cycle intensities for Read 1 to first cycle intensities for Read 2.
	PE Orientation	The percentage of paired-end alignments with the expected orientation.

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.
	Duplicate	The total number of clusters for a paired-end sequencing run that are considered to be PCR duplicates. PCR duplicates are defined as two clusters from a paired-end run where both clusters have the exact same alignment positions for each read.

Mismatch Graph

Y Axis	X Axis	Description
Percent	Cycle	Plots the percentage of mismatches for all clusters in a run by cycle.

Library QC Details Tab

The Details tab for the Library QC workflow includes a samples table, targets table, coverage graph, and Q-score graph.

- **Samples table**—Summarizes the sequencing results for each sample.
- **Targets table**—Shows statistics for a particular sample and chromosome.
- **Coverage graph**—Shows read depth at a given position in the reference.
- **Q-score graph**—Shows the average quality score, which is the estimated probability of an error measured in $10^{-(Q/10)}$. For example, a score of Q30 has an error rate of 1 in 1000, or 0.1%.

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.
Clusters Raw	The number of clusters sequenced for this sample.
%Clusters	The percentage of the total cluster number matching the index for this sample.
%PF	The percentage of clusters passing filter for this sample.
%Aligned	The percentage of clusters successfully aligned.
%Mismatch	The percentage mismatch to reference averaged over cycles per read (Read 1/Read 2).
Median Len	The median fragment length for the sample.
Min Len	The low percentile of fragment lengths for this sample as they correspond to 3 standard deviations from the median.
Max Len	The high percentile of fragment lengths for this sample as they correspond to 3 standard deviations from the median.
Estimated Diversity	An estimate of the total library diversity derived from the observed diversity and the number of apparent PCR duplicates. This calculation is available for paired-end runs unless PCR duplicate flagging was disabled in the sample sheet.
Observed Diversity	Number of distinct aligned positions. Reads with the same aligned positions are assumed to be PCR duplicates. PCR duplicates are defined as sequences with identical Read 1 and Read 2 start sites.
Genome	The name of the reference genome.

Targets Table

Column	Description
#	An ordinal identification number in the table.
Chr	The reference target or chromosome name.
Cluster PF	The number of clusters passing filter for the sample that aligned to the reference genome.
Mismatch	The percentage mismatch to reference averaged over cycles per read (Read 1/Read 2).
No Call	The percentage of bases that could not be called (no-call) for the sample averaged over cycles per read (Read 1/Read 2).
Genome	The name of the reference genome.

Q-Score Graph

Y Axis	X Axis	Description
Q-Score	Position	The average quality score of bases at the given position of the reference.

Coverage Graph

Y Axis	X Axis	Description
Coverage	Position	The green curve is the number of aligned reads that cover each position in the reference. The red curve is the number of aligned reads that have a miscall at this position in the reference. SNPs and other variants show up as spikes in the red curve.

Optional Settings for the Library QC 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 Library QC 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
FlagPCRDuplicates	Settings are 0 or 1. Default is 1, filtering. If set to 1, PCR duplicates are flagged in the BAM files and not used for variant calling. PCR duplicates are defined as 2 clusters from a paired-end run where both clusters have the exact same alignment positions for each read. Duplicates are not flagged for single-read runs, including PCR duplicates. *(Formerly FilterPCRDuplicates. FilterPCRDuplicates is acceptable for backward compatibility.)*
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. This setting is necessary in certain cases, such as processing of mate pair data using BWA, which expects paired-end data. This setting can disrupt per-cycle metrics. Set this setting to 1 when using the Library QC workflow with Nextera Mate Pair libraries.
RunBwaAln	Settings are 0 or 1. Default is 0, BWA-MEM alignment method. BWA-MEM is for ≥ 70 bp read lengths. If set to 1, BWA-backtrack is used for alignment. Formerly referred to as BWA, BWA-backtrack is an earlier version of the BWA. Use BWA-backtrack for < 70 bp read lengths, or when consistency is required with previous study data.

Analysis Output Files

The following analysis output files are generated for the Library QC workflow and provide analysis results for alignment and a sample report.

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in Data\Intensities\BaseCalls\Alignment.
LibraryQC.html	Lists the characteristics of each sample in terms of percentage of reads aligned. Located in Data\Intensities\BaseCalls\Alignment.

BAM File Format

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

BAM files are written to the alignment folder in Data\Intensities\BaseCalls\Alignment. BAM files use the file naming format of SampleName_S#.bam, where # is the sample number determined by the order that samples are listed in the sample sheet.

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.
  Alignment methods include banded Smith-Waterman, Burrows-Wheeler Aligner (BWA), and Bowtie. The term Isis indicates that an Illumina alignment method is in use, which is the banded Smith-Waterman method.
- **Alignments**—Contains read name, read sequence, read quality, alignment information, and custom tags.

  ```
  GA23_40:8:1:10271:11781 64 chr22 17552189 8 35M * 0 0
  TACAGACATCCACCACCACACCCAGCTAATTTTTG
  IIIII>FA?C::B=:GGGB>GGGEGIIIHI3EEE#
  BC:Z:ATCACG XD:Z:55 SM:I:8
  ```

The read name maps to the chromosome and start coordinate **chr22 17552189**, with alignment quality **8**, and the match descriptor CIGAR string **35M**.

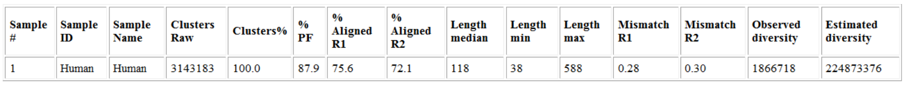
BAM files are suitable for viewing with an external viewer such as IGV or the UCSC Genome Browser.

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

Library QC Analysis File

The sample report, LibraryQC.html, lists cluster counts, cluster quality, alignment information, fragment length, and diversity for each sample. For a description of each column in the sample report, see *Targets Table* on page 9.

Figure 1 LibraryQC.html Analysis File



Sample #	Sample ID	Sample Name	Clusters Raw	Clusters%	% PF	% Aligned R1	% Aligned R2	Length median	Length min	Length max	Mismatch R1	Mismatch R2	Observed diversity	Estimated diversity
1	Human	Human	3143183	100.0	87.9	75.6	72.1	118	38	588	0.28	0.30	1866718	224873376

Data in the sample report is generated by calculating diversity and then computing fragment length.

- **Diversity calculation**—Considers all clusters that are proper pairs, meaning both mates map to the same chromosome, and then computes the number of clusters. The number of distinct alignment positions is reported as the observed diversity. Clusters with the same alignment positions are assumed to be PCR duplicates. From this information, the estimated diversity is calculated. This calculation is available for paired-end sequencing runs, unless PCR duplicate flagging is disabled in the sample sheet.
- **Fragment length**—Lengths are computed for clusters where both reads successfully aligned to the same chromosome. This workflow reports lengths in median (50th percentile), minimum (0.15th percentile, corresponding to 3 standard deviations below the mean for a normal distribution), and maximum (99.85th percentile). Fragment lengths of ≥ 10000 bases are discarded as possible chimeras.

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.
ErrorsAndNoCallsByLaneTileReadCycle.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.

File Name	Description
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.

Notes

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