illumina®

Local Run Manager
Library QC Analysis Module

Workflow Guide

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

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

Document # 1000000057749 v01

2

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

Overview

The Library QC analysis module evaluates the abundance, fragment length, and sample quality of DNA libraries. The analysis performed in the Library QC workflow is very similar to the Resequencing workflow. The Library QC workflow does not perform variant calling; instead, it provides a high-level report of the characteristics of each DNA sample.

The Local Run Manager Library QC v3.0.0 (or later) analysis module can only be run on Local Run Manager v3.0 (or later) running on a machine with Windows 10. The analysis module requires the library prep and index kits that are used in the run to be decoupled.

Compatible Library Types

The Library QC analysis module is compatible with specific library types represented by library kit categories on the Create Run screen. For a current list of compatible library kits, see the Local Run Manager support page on the Illumina website.

Input Requirements

In addition to sequencing data files generated during the sequencing run, such as base call files, the Library QC analysis module requires the following files.

▶ **Reference genome**—The Library QC analysis module requires a reference genome. The reference genome provides the chromosome and start coordinate in the BAM file output.

About This Guide

This guide provides instructions for setting up run parameters for sequencing and analysis parameters for the Library QC analysis module. For information about the Local Run Manager dashboard and system settings, see the *Local Run Manager v3 Software Guide (document #1000000111492)*.

Set Parameters

1   If needed, log in to Local Run Manager.

2   Select **Create Run**, and select **Library QC**.

3   Enter a run name that identifies the run from sequencing through analysis.
    The run name can contain alphanumeric characters, spaces, and the following special characters:
    `~.!@#$%-_{}.

4   **[Optional]** Enter a run description to identify the run.
    The run description can contain alphanumeric characters, spaces, and the following special characters:
    `~.!@#$%-_{}.

Specify Run Settings

1   Select the library prep kit from the Library Prep Kit drop-down list.

2   Select the index kit from the Index Kit drop-down list.

3   Specify the number of index reads.
    ▶ **0** for a run with no indexing
    ▶ **1** for a single-indexed run

> ▶ **2** for a dual-indexed run

If your index kit supports only one option, the index read is automatically selected.

4 Select the read type for the run.

If your index kit supports only one option, the read type is automatically selected.

5 Enter the number of cycles for the run.

6 **[Optional]** For Custom Primers, specify any custom primer information to be used for the run by selecting the appropriate checkboxes.

Custom primer options vary based on your instrument or Local Run Manager implementation.

Specify Module-Specific Settings

1 Select the **On/Off** toggle to enable or disable the following settings:

> ▶ **Flag PCR Duplicates**—On by default. When enabled, 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 position for each read.

> ▶ **Reverse Complement**—(Only available with Nextera Mate Pair library prep kits) Off by default. When enabled, all reads are reverse-complemented as they are written to FASTQ files.

Specify Samples for the Run

Specify samples for the run using the following options:

> ▶ **Enter samples manually**—Use the blank table at the bottom of the Create Run screen.

> ▶ **Import sample sheet**—Navigate to an external file in a comma-separated values (*.csv) format.

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

Enter Samples Manually

1 Adjust the samples table to an appropriate number of rows.

> ▶ In the Rows field, use the up/down arrows or enter a number to specify the number of rows to add to the table. Select ➕ to add the rows to the table.

> ▶ Select ✖ to delete a row.

> ▶ Right-click on a row in the table and use the commands in the contextual menu.

2 Enter a unique sample ID in the Sample ID field.

Use alphanumeric characters, dashes, or underscores. Spaces are not allowed in this field.

3 Enter a sample name in the Sample Name field.

Use alphanumeric characters, dashes, or underscores. Spaces are not allowed in this field.

4 If you have a plated kit, select an index plate well from the Index well drop-down list and skip to step .

Import Sample Sheet

1 If you do not have a sample sheet to import, see for instructions on how to create and export a sample sheet. Edit the file as follows.

   a Open the sample sheet in a text editor.

   b Enter the sample information in the [Data] section of the file.

   c Save the file. Make sure that the sample IDs are unique.

2   Select **Import Sample Sheet** at the top of the Create Run screen and browse to the location of the sample sheet.
     Make sure that the information in the sample sheet is correct. Incorrect information can impact the sequencing run.

3   When finished, select **Save Run**.

Sample Sheet Fields

Manual editing of the sample sheet is intended for technically advanced users. If settings are applied incorrectly, serious problems can occur.

Visit the Local Run Manager support page for available sample sheet settings. Settings must be entered as specified to avoid analysis failure.

Analysis Methods

The Library QC analysis module performs the following analysis steps and then writes analysis output files to the folder.

▶ Demultiplexes index reads

▶ Generates FASTQ files

▶ Alignment

Demultiplexing

For runs with multiple samples and index reads, demultiplexing compares each Index Read sequence to the index sequences specified in the sample sheet. No quality values are considered in this step.

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, one demultiplexing file named **DemultiplexSummaryF1L1.txt** is written to the Alignment folder with 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 system.

▶ A table of demultiplexing results with one row per tile and one column per sample, including sample 0.

▶ The most commonly occurring sequences for the index reads.

Other demultiplexing files are generated for each tile of the flow cell. For more information, see *Demultiplexing File Format* on page 11.

FASTQ File Generation

Local Run Manager 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 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 Analysis folder, and then copied to the BaseCalls folder in the output 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, see *FASTQ File Names* on page 10.

Quality Trimming

FASTQ file generation optionally performs quality trimming of the 3' portion of nonindex reads with low quality scores. This step is performed by default during alignment using BWA. For workflows that do not use BWA, use the **QualityScoreTrim** sample sheet setting to include trimming during FASTQ file generation.

Alignment

Alignment is a way of identifying optimal matches between read sequences and the sequence of a reference genome. Aligned sequences are assigned a score based on their similarity to the reference.

Alignment results are written to Binary Alignment/Map (BAM) files. BAM files are the primary input for variant calling. For more information, see *BAM File Format* on page 1.

Alignment Methods

For workflows that include alignment, reads are aligned against the reference specified in the sample sheet or in a manifest file. Local Run Manager uses the BWA alignment method.

BWA-Backtrack

Formerly referred to as BWA, BWA-backtrack is an earlier version of the BWA (Burrows-Wheeler Aligner) algorithm that aligns sequencing read lengths in < 70 bp segments. Use this version for very short reads, or when consistency is required with previous study data.

BWA aligns relatively short nucleotide sequences against a long reference sequence. BWA automatically adjusts parameters based on read lengths and error rates, and then estimates insert size distribution.

View Analysis Results

1 From the Local Run Manager dashboard, select the run name.

2 From the Run Overview tab, review the sequencing run metrics.

3 To change the analysis data file location for future requeues of the selected run, select the **Edit** ✎ icon, and edit the output run folder file path.
The file path leading up to the output run folder is editable. The output run folder name cannot be changed.

4 **[Optional]** Select the **Copy to Clipboard** 📋 icon to copy the output run folder file path.

5 Select the Sequencing Information tab to review run parameters and consumables information.

6 Select the Samples & Results tab to view the analysis report.
   ▶ If analysis was requeued, select the appropriate analysis from the Select Analysis drop-down list.

7   [Optional] Select the Copy to Clipboard 📋 icon to copy the Analysis Folder file path.

Analysis Report

Sample Information

Column	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.
Percent Q30 Bases	The percentage of bases called with a quality score ≥ Q30.

Read Level Statistics

Column	Description
Read	The read number.
Total Aligned Reads	The total number of reads passing filter that aligned to the reference genome.
Percent Aligned Reads	The percentage of reads that aligned to the reference genome [100*(Total Aligned Reads/Total PF Reads)].

Base Level Statistics

Column	Description
Read	The read number.
Total Aligned Bases	The total number of bases passing filter that aligned to the reference genome.
Percent Aligned Bases	The percentage of bases that aligned to the reference genome.
Mismatch Rate	The average percentage of mismatches over all cycles.

Coverage Histogram

Graph	Description
Coverage Histogram	A histogram showing the number of bases covered by the depth of sequencing coverage.

Small Variants Summary

Row	Description
Total Passing	The total number of Single Nucleotide Variants, insertions, and deletions passing the quality filters.
Het/Hom Ratio	Number of heterozygous SNVs/Number of homozygous SNVs, insertions, and deletions.
Ts/Tv Ratio	The number of Transition SNVs, insertions, and deletions that pass the quality filters divided by the number of Transversion SNVs, insertions, and deletions that pass the quality filters. Transitions are interchanges of purines (A, G) or of pyrimidines (C, T). Transversions are interchanges of purine and pyrimidine bases (for example, A to T).

Fragment Length Summary

Column	Description
Fragment Length Median	Median length of the sequenced fragment. The fragment length is calculated based on the locations at which a read pair aligns to the reference. The read mapping information is parsed from the BAM files.
Minimum	Minimum length of the sequenced fragment.
Maximum	Maximum length of the sequenced fragment.
Standard Deviation	Standard deviation of the sequenced fragment length.

Additional Information

Report	Description
Duplicate Information	Percentage of duplicate paired reads. Duplicates are only reported for paired-end reads and if PCR duplicate flagging was selected in the settings.
Settings	Settings (reference genome, depth threshold, PCR duplicates, targeted regions, base padding, and Picard HS Metrics) specified during analysis set-up.
Software Versions	Software versions of Isis (Analysi Software), SAMtools, BWA (Aligner), Somatic Variant Caller, and IAS (Annotation Service).

Analysis Metrics

Quality Scores

The following table shows the relationship between the quality score and error probability.

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

ASCII Format for Quality Scores

During analysis, base call quality scores are written to FASTQ files in an encoded ASCII format (the value + 33). The ASCII format is illustrated in the following table.

Table 1   ASCII Codes for Q-Scores 0–40

Symbol	ASCII Code	Q-score	Symbol	ASCII Code	Q-score
!	33	0	6	54	21
"	34	1	7	55	22
#	35	2	8	56	23
$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	;	59	26

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

Table 1   ASCII Codes for Q-Scores 0–40

Symbol	ASCII Code	Q-score	Symbol	ASCII Code	Q-score
'	39	6	<	60	27
(	40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	**30**
+	43	10	@	64	31
,	44	11	A	65	32
-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	**40**
5	53	**20**			

Phasing and Prephasing

During the sequencing reaction, each DNA strand in a cluster extends by one base per cycle. A small portion of strands can become out of phase with the current incorporation cycle. Phasing occurs when a base falls behind. Prephasing occurs when a base jumps ahead. Phasing and prephasing rates indicate an estimate of the fraction of molecules that became phased or prephased in each cycle.

Figure 1   Phasing and Prephasing
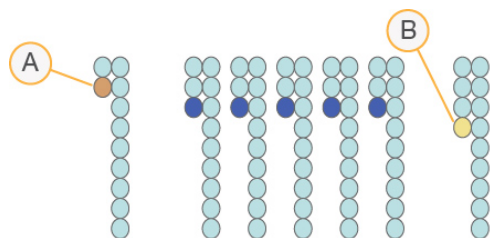
A   Read with a base that is phasing
B   Read with a base that is prephasing

The number of cycles performed in a read is one more cycle than the number of cycles analyzed. For example, a paired-end 150-cycle run performs two 151-cycle reads (2 x 151) for a total of 302 cycles. At the end of the run, 2 x 150 cycles are analyzed. The one extra cycle for Read 1 and Read 2 is required for prephasing calculations.

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.

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

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in the Alignment folder.

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 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.

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

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:

▶ Identifier

▶ Sequence

▶ Plus sign (+)

▶ 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=AAAAAAAAA9#:<#<;<<<????#=
```

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:

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.

▶ **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).

Demultiplexing File Format

For multiple sample indexed runs, 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.

Demultiplexing files are binary files written to the L001 folder in Data\Intensities\BaseCalls\L001. The 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.

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
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.
DemultiplexSummaryF1L#.txt	Reports demultiplexing results in a table with 1 row per tile and 1 column per sample. Located in the Alignment folder.
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 the Alignment folder.
Mismatch.htm	Contains histograms of mismatches per cycle and no-calls per cycle for each tile. Located in the Alignment folder.
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 the Alignment folder.
Summary.htm	Contains a summary web page generated from Summary.xml. Located in the Alignment folder.

Revision History

Document	Date	Description of Change
Document # 1000000057749 v01	January 2021	Updated software descriptions for Library QC v3.0, which includes updates to: • Software requirements. • Updated run name and description requirements. • Updated parameter settings for library prep kit and index decoupling.
Document # 1000000057749 v00	July 2018	Initial release.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

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

Illumina Customer Support Telephone Numbers

Region	Toll Free	Regional
North America	+1.800.809.4566	
Australia	+1.800.775.688	
Austria	+43 800006249	+43 19286540
Belgium	+32 80077160	+32 34002973
China	400.066.5835	
Denmark	+45 80820183	+45 89871156
Finland	+358 800918363	+358 974790110
France	+33 805102193	+33 170770446
Germany	+49 8001014940	+49 8938035677
Hong Kong, China	800960230	
Ireland	+353 1800936608	+353 016950506
Italy	+39 800985513	+39 236003759
Japan	0800.111.5011	
Netherlands	+31 8000222493	+31 207132960
New Zealand	0800.451.650	
Norway	+47 800 16836	+47 21939693
Singapore	+1.800.579.2745	
South Korea	+82 80 234 5300	
Spain	+34 911899417	+34 800300143
Sweden	+46 850619671	+46 200883979
Switzerland	+41 565800000	+41 800200442
Taiwan, China	00806651752	
United Kingdom	+44 8000126019	+44 2073057197
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 from support.illumina.com.

illumına®