MiSeq Reporter Assembly

Workflow Guide

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

Revision History 3
Introduction 4
Assembly Workflow Overview 5
Assembly Summary Tab 6
Assembly Details Tab 7
Optional Settings for the Assembly Workflow 8
Analysis Output Files 10
Technical Assistance

ILLUMINA PROPRIETARY
Document # 15042313 v01
September 2015

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 # 15042313 v01	September 2015	Changed the name of the guide from MiSeq Reporter Assembly Workflow Reference Guide to MiSeq Reporter Assembly Workflow Guide. Updated the default setting for ReverseComplement from 1 to 0 in the Sample Sheet Settings for Analysis section.
Part # 15042313 Rev. B	August 2013	• Added descriptions of optional settings Adapter and ReverseComplement. • Noted that an optional reference genome can be provided with either *.fasta or *.fa extensions.
Part # 15042313 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 Assembly workflow is provided in this guide.

Introduction

The Assembly workflow assembles small genomes (< 20 Mb) without the use of a reference genome, and is best suited for the assembly of bacterial genomes, such as *E. coli*.

The Assembly workflow uses the Velvet software and writes assembly results in the FASTA format.

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

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

Assembly Workflow Overview

The Assembly workflow uses a *de Bruijn* graph methodology to assemble reads into contigs, which are consensus DNA sequences representing overlapping sets of reads. The resulting contigs are written to a FASTA file named contigs.fa in a subfolder of the Alignment folder named AssemblyN, where N is the sample number.

Reads are randomly subsampled from the total data output to produce Assemble_N_Rx.fastq.gz files, where N refers to the sample number and x refers to the read number. These files contain the reads used in the assembly process. The selection process is random but not stochastic, meaning the same subset of reads are selected each time that the Assembly workflow is run. The subsampling of reads prevents overloading of the RAM built into MiSeq instrument computer.

NOTE
A reference genome is optional. Reference genomes can use either a *.fasta or *.fa file extension.

If a reference genome is specified, the workflow performs the following steps:

- Compares contigs against the reference genome.
- Reorders contigs to match the order of the reference genome, as closely as possible.
- Generates the samples graph (dot-plot), which summarizes the match between contigs and the reference genome. For more information, see *Samples Graph* on page 7.

The assembly process uses the Velvet software. For a description of Velvet, see *Velvet: algorithms for de novo short read assembly using de Bruijn graphs*, Zerbino and Birney, Genome Research 2008.

Assembly Summary Tab

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

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.

High Percentages Graph

Y Axis	X Axis	Description
Percent	PF	The percentage of clusters passing filters.
	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.

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.

Assembly Details Tab

The Details tab for the Assembly workflow includes a samples graph and a samples table:

- **Samples graph**—Summarizes the match between contigs and the reference genome in a syntenic plot (dot-plot). This plot is available only if a reference genome was specified in the sample sheet.
- **Samples table**—Summarizes the sequencing results for each sample.

Samples Graph

Contigs are arranged end-to-end along the X axis and the reference chromosomes are arranged bottom-to-top along the Y axis. Each pixel of the plot is colored according to how many short sequences of the corresponding contig have a match in the corresponding portion of the reference genome.

An identical assembly results in a diagonal line. A vertical gap in the plot might indicate a portion of the reference that is absent in the assembly, such as a plasmid, which is found in some bacteria populations.

Y Axis	X Axis	Description
Reference	Assembly Position	A syntenic plot of assembled contigs compared to a reference. A reference genome must be specified in the sample sheet.

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.
Num Contigs	The number of contigs assembled for this sample.
MeanContigLength	The average contig length for this sample.
MedContigLength	The median contig length for this sample.
MinContigLength	The minimum contig length for this sample.
MaxContigLength	The maximum contig length for this sample.
Base Count	The total length of the resulting assembly.
N50	N50 length is the length of the shortest contig. The sum of contigs of equal length or longer is at least 50% of the total length of all contigs.

Optional Settings for the Assembly 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 Assembly 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.
Kmer	This setting overrides the k-mer size used by Velvet. Default is 31; odd-numbered values up to 255 are supported.
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 Assembly workflow with Nextera Mate Pair libraries.

Optional Configurable

MiSeq Reporter uses a maximum of 550 Mbp of sequence data for *de novo* assembly. This setting is controlled in the MiSeq Reporter.exe.config file using the configuration setting **MaximumMegabasesAssembly**.

Configuration Name	Values and Description
MaximumMegabasesAssembly	550 (default) The maximum number of megabases to assemble. Larger values require more RAM.

NOTE
Assembly of reads from longer runs requires more memory than assembly of reads from shorter runs. If the process terminates due to memory requirements, consider lowering the setting MaximumMegabasesAssembly.

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

Analysis Output Files

The following analysis output files are generated for the Assembly workflow and written to the MiSeqAnalysis folder.

File Name	Description
Contigs.fa	Contains the contigs for each assembly. Located in Data\Intensities\BaseCalls\Alignment.
DotPlot.png	Summarizes the match between contigs and the reference genome. 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. 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.
AssemblyRunStatistics.xml	Contains summary statistics specific to the run. 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.
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