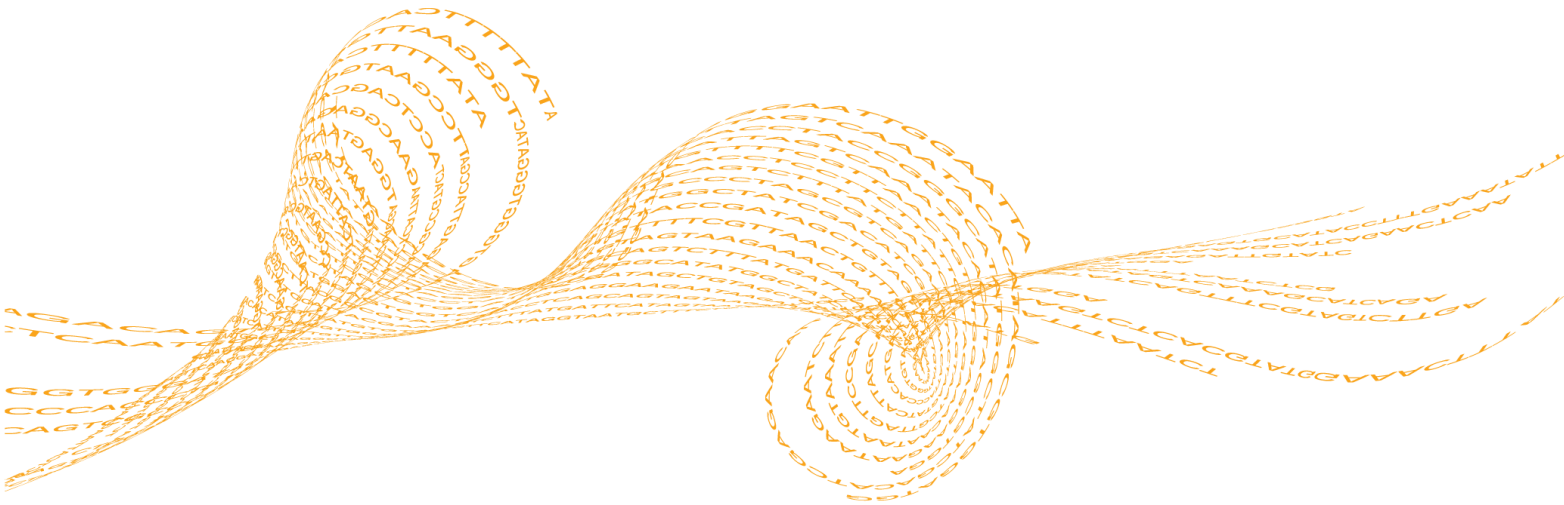
MiSeq Reporter Metagenomics Workflow Reference Guide

FOR RESEARCH USE ONLY



ILLUMINA PROPRIETARY
Part # 15042317 Rev. D
December 2014

illumina®

Revision History

Part #	Revision	Date	Description of Change
15042317	D	December 2014	Added a note in the Demultiplexing section about the default index recognition for index pairs that differ by < 3 bases.
15042317	C	February 2014	Added information on the rarefaction.txt and summary.txt analysis output files. Updated to changes introduced in MiSeq Reporter v2.4: Added information on classification reports in the alignment folder. Noted that species with abundance of <0.25% are grouped in the category *Other* in the samples table and clusters pie chart.
15042317	B	August 2013	• Updated to MiSeq Reporter v2.3, which includes species-level classification using an Illumina-proprietary classification algorithm, and update to an Illumina-curated version of the Greengenes 13.5 (May 2013) taxonomy database. • Updated TaxonomyFile setting to list override file for genus-level classification. • Added Adapter settings to optional settings section.
15042317	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 Metagenomics workflow is provided in this guide.

Introduction

The Metagenomics workflow classifies bacteria from a metagenomic sample by amplifying specific regions in 16S ribosomal RNA. Reads are classified using a database of 16S rRNA data.

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

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

Metagenomics Workflow Overview

The Metagenomics workflow is used to classify organisms from a metagenomic sample by amplifying specific regions in the 16S ribosomal RNA. This workflow is exclusive to Prokaryotes, which includes Bacteria and Archaea. The Metagenomics workflow generates a classification of reads at several taxonomic levels: kingdom, phylum, class, order, family, and genus or species.

Introduced in MiSeq Reporter v2.3, the Metagenomics workflow uses a faster algorithm, which results in more than a two-fold reduction in analysis time, and an Illumina-curated version of the taxonomic database.

The Metagenomics workflow demultiplexes indexed reads, generates FASTQ files, and then classifies reads.

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, one 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 one row per tile and one 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 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)*.

Classification of Reads

The classification step uses ClassifyReads, a proprietary algorithm that provides species-level classification for paired-end reads. This process involves matching short subsequences of the reads (called words) to a set of 16S reference sequences. The accumulated word matches for each read are used to assign reads to a particular taxonomic classification. Analysis results list the total number of classified clusters for each sample at each taxonomic level. Statistics are written to the file Classification.txt.

Current Taxonomy

The current taxonomy is stored in Taxonomy.dat. As of MiSeq Reporter v2.3, the Metagenomics workflow generates classifications to the species level. For information about setting up analysis to genus level only, see *Sample Sheet Settings for Analysis* on page 10.

The taxonomy database for the Metagenomics workflow is an Illumina-curated version of the Greengenes database (greengenes.secondgenome.com/downloads/database/13_5). To generate species-level classifications, the following filters are applied:

1. Filter all entries where the 16S sequence length was below 1250 bp.
2. Filter all entries with more than 50 wobble bases (M, R, W, S, Y, K, V, H, D, B, and N).
3. Filter all entries that are partially classified with no classification for genus or species.

The following taxonomic counts are available for the Metagenomics workflow.

Taxonomy	Count
Kingdoms	3
Phyla	33
Classes	74
Orders	148
Families	321
Genera	1086
Species	6466

Alternative Taxonomy Database

You can prepare an alternative taxonomy database using the tool CreateTaxonomyDatabase distributed with MiSeq Reporter. This tool is located in the MiSeq Reporter install folder, typically on the C: drive:
C:\Illumina\MiSeq Reporter\Workflows\MetagenomicsWorker\CreateTaxonomyDatabase.exe.

CreateTaxonomyDatabase is a command-line tool; run it without arguments for a description of available options. For an example of a valid FASTA file, see:

greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/current_GREENGENES_gg16S_unaligned.fasta.gz

The Metagenomics workflow provides species-level classification. To configure the workflow for genus-level classification, use the TaxonomyFile sample sheet setting and specify gg_13_5_genus_32bp.dat. For more information, see *Sample Sheet Settings for Analysis* on page 10.

Metagenomics Summary Tab

The Summary tab for the Metagenomics workflow includes a clusters graph.

- **Clusters graph**—Shows numbers of raw clusters, clusters passing filter, clusters that did not align, clusters not associated with an index, and duplicates.

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.
	Unindexed	The total number of clusters passing filter that were not associated with any index sequence in the run.

Metagenomics Details Tab

The Details tab for the Metagenomics workflow includes a samples table and clusters pie chart.

- **Samples table**—Summarizes the sequencing results for each sample.
- **Clusters pie chart**—A graphical representation of the classification breakdown for each sample.

NOTE
Species with abundance of < 0.25% are grouped in the category ***Other***.

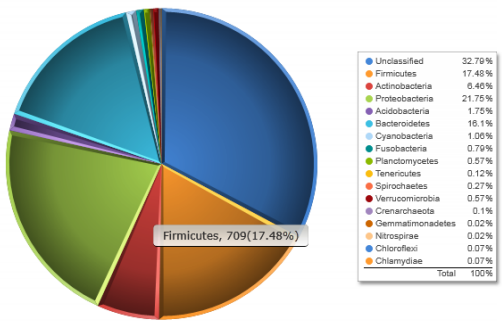
Samples Table

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 raw clusters detected for the sample.
Cluster PF	The number of clusters passing filter for the sample.
Taxonomic Level	The taxonomic level of classification. From broadest to most specific, the levels at which classification is done are as follows: Kingdom, Phylum, Class, Order, Family, and Genus or Species.
Clusters Classified	The number of clusters that were confidently classified at this taxonomic level.

Metagenomics Pie Chart

The Metagenomics pie chart provides a graphical representation of how many clusters from each sample were assigned to a category in each taxonomic level.

Figure 1 Metagenomics Pie Chart



At the Phylum level for a sample, the pie chart might include a wedge for Bacteroidetes and another for Firmicutes, among others. A label for each wedge appears when you hover your mouse over a wedge in the pie chart. Click another row in the samples table to change the pie chart to that sample or taxonomic level.

Optional Settings for the Metagenomics 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: TaxonomyFile,gg_13_5_genus_32bp.dat

The following optional settings are compatible with the Metagenomics 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 two 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.
TaxonomyFile	This setting overrides the taxonomy database; default is taxonomy.dat. As of MiSeq Reporter v2.3, species-level classification is enabled, by default. For faster, but less granular genus-level classification, specify gg_13_5_genus_32bp.dat.

Analysis Output Files

The analysis output file generated for the Metagenomics workflow provides a classification of reads for each sample.

File Name	Description
*.txt.gz file	A compressed text file that contains classification of reads from a given sample. Each entry provides classification at up to six taxonomic levels. Located in Data\Intensities\BaseCalls\Alignment.
Classification.txt	Contains the total number of classified clusters for each sample at each taxonomic level. Located in Data\Intensities\BaseCalls\Alignment.
Rarefaction.txt	Contains the number of unique genera discovered by the number of reads classified. Located in Data\Intensities\BaseCalls\Alignment.
Summary.txt	Contains the taxonomic classification, number of reads, and percent of the sample for all classifications in the top 95% by abundance. Located in Data\Intensities\BaseCalls\Alignment.

Classification.txt File

The classification.txt lists all samples and taxonomic results in a single file. The contents of the classification.txt file populate the Samples table and Metagenomics pie chart that appear on the Details tab.

Figure 2 Classification.txt File

SampleNumber	SampleName	Level	Group	Reads	Percentage
1	PhiX1	Kingdom	Unclassified	13	0.2639594
1	PhiX1	Kingdom	Bacteria	4902	99.533
1	PhiX1	Kingdom	Archaea	10	0.2030457
1	PhiX1	Phylum	Unclassified	1662	33.74619
1	PhiX1	Phylum	Firmicutes	896	18.19289
1	PhiX1	Phylum	Bacteroidetes	749	15.20812
1	PhiX1	Phylum	Proteobacteria	1154	23.43147
1	PhiX1	Phylum	Actinobacteria	226	4.588832
1	PhiX1	Phylum	Acidobacteria	77	1.563452
1	PhiX1	Phylum	Nitrospirae	3	0.0609137
1	PhiX1	Phylum	Cyanobacteria	38	0.7715736
1	PhiX1	Phylum	Verrucomicrobia	28	0.5685279

Column Heading	Description
Sample Number	The order that the sample is listed in the sample sheet.
Sample Name	The sample name from the sample sheet.
Level	The taxonomic level of classification.
Group	The taxonomic group name.
Reads	The number of reads from the sample assigned to the specific taxonomic level.
Percentage	The percentage of classified reads from the sample at the specific taxonomic level.

Classification Reports

Reports generated from the classification.txt and summary.txt files provides classification results for a single-sample.

Report	Description
16S Metagenomics Report (PDF)	Contains table charts of sample information, sequencing statistics, and a classification rate summary. The report also has a table chart and a pie chart for classification results at each taxonomic level. The file naming format is X_SY.report.pdf, where X is the sample ID and Y is the sample index. Located in Data\Intensities\BaseCalls\Alignment.
16S Metagenomics Report (HTML)	Contains all the information in the16S Metagenomics Report (PDF), and an interactive sunburst classification chart and a classification bar graph. The file naming format is X_SY.report.html, where X is the sample ID and Y is the sample index. 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
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 one row per tile and one column per sample. Located in Data\Intensities\BaseCalls\Alignment.
MetagenomicsRunStatistics.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

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

Safety data sheets (SDSs) are available on the Illumina website at support.illumina.com/sds.html.

Product Documentation

Product documentation in PDF is available for download from the Illumina website. Go to support.illumina.com, select a product, then click **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