MiSeq Reporter User Guide

FOR RESEARCH USE ONLY

ILLUMINA PROPRIETARY
Part # 15028784 Rev. J
March 2013

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) OR ANY USE OF SUCH PRODUCT(S) OUTSIDE THE SCOPE OF THE EXPRESS WRITTEN LICENSES OR PERMISSIONS GRANTED BY ILLUMINA IN CONNECTION WITH CUSTOMER'S ACQUISITION OF SUCH PRODUCT(S).

FOR RESEARCH USE ONLY

Note regarding biomarker patents and other patents unique to specific uses of products.

Some genomic variants, including some nucleic acid sequences, and their use in specific applications may be protected by patents. Customers are advised to determine whether they are required to obtain licenses from the party that owns or controls such patents in order to use the product in customer's specific application.

Revision History

Part #	Revision	Date	Description of Change
15028784	J	March 2013	Reorganized workflow-specific information into dedicated chapters, on for each analysis workflow. Added the following information introduced in MiSeq Reporter v2.2: • Added new Targeted RNA workflow: description of workflow, analysis reports, output files, and manifest file format • For the Small RNA workflow, updated the default aligner from Eland to Bowtie v0.12.8 • For the Enrichment workflow, added two additional output files: SampleName.enrichment_Summary.csv and SampleName_regions_Manifest_intervals.txt; added sample sheet setting EnrichmentMaxRegionStatisticsCount • For the Custom Amplicon workflow, added one additional output file: AmpliconCoverage_M#.tsv • For the Custom Amplicon workflow, noted that alignments containing more than three indels are filtered • For the Assembly workflow, updated the maximum k-mer size used by Velvet to 255 • For Resequencing and LibraryQC workflows, removed Eland as an alternative alignment method • For Metagenomics workflow, added description of current taxonomy and CreateTaxonomyDatabase tool • Noted that the GATK sub-sampling limit is 5000, raised from 250 in previous versions • Removed the sample sheet setting VariantFilterQualityCutoff; VariantMinimumGQCutoff is preferred
15028784	H	January 2013	Added statement regarding biomarker patents to copyright page. Noted that the source of the manifest file for the Enrichment workflow differs depending on the sample prep kit used to prepare libraries.
15028784	G	January 2013	Changed the following sample sheet setting definitions: • CustomAmpliconAlignerMaxIndelSize—Changed default to 25 • FilterPCRDuplicates—Changed to FlagPCRDuplicates • VariantCaller—Added Resequencing workflow Added sample sheet setting ExcludeRegionsManifestA for the Enrichment workflow and related Group column to the Enrichment manifest description. Added config setting GATKDownsampleDepth. Removed the output file DemultiplexComplete.txt.

Part #	Revision	Date	Description of Change
15028784	F	November 2012	Corrected the following information: • Output file name EnrichmentStatistics.xml • Description of variant analysis for Enrichment workflow • Added Enrichment workflow where applicable to variant calling introduction and manifest overview sections
15028784	E	November 2012	Added the following information introduced in MiSeq Reporter v2.1: • Description of the Enrichment workflow, analysis reports, and output files, including coverage files and gaps files • Instructions for sample sheet editing on sample sheet tab, including adding and deleting rows • Instructions for saving images of graphs in JPG format from the web interface • Description of consensus reads feature for the Custom Amplicon workflow • Information about adapter settings, descriptions of sample sheet setting AdapterRead2, and config settings AdapterTrimming Stringency and NMaskShortAdapterReads • Descriptions of manifest formats and file contents Updated the following information: • Changed computing requirements to 16 GB RAM, recommended • Removed column labeled Coverage from targets table in Library QC workflow • Updated VCF annotations to add LowDP, LowGQ, and LowQual • Updated sample sheet settings to add CustomAmpliconAlignerMaxIndelSize and VariantMinimumGQCutoff, and to update StandBiasFilter and MinumumCoverageDepth for the Enrichment workflow • Corrected config settings to remove FilterAmpliconNonVariantCallingReads and MinimumTilesPerAlignmentRead

Part #	Revision	Date	Description of Change
15028784	D	September 2012	Reorganized existing content into chapters and added an index to improve usability. Added the following information: • Overview of analysis metrics, including clusters passing filter, base call quality scores, and phasing/prephasing values • Overview of analysis procedures, including demultiplexing, FASTQ file generation, alignment, and variant calling, along with descriptions of alignment methods and variant callers • Overview of analysis folder structure and contents • Descriptions of output file formats, including demultiplexing, FASTQ, BAM, and VCF • Descriptions of output files generated for each workflow • Sample sheet settings for controlling analysis parameters • Description of a Windows service application • Description of adapter trimming compared to adapter masking • Instructions for off-instrument software installation • Basic troubleshooting steps and instructions for viewing log files
15028784	C	July 2012	This revision documents the web-based version of MiSeq Reporter v2.0.
15028784	B	April 2012	This revision includes new information covering the new release in MiSeq Reporter.
15028784	A	October 2011	Initial release.

Table of Contents

Chapter 1

Getting Started

Introduction

The MiSeq® System provides on-instrument secondary analysis using the MiSeq Reporter software, which performs secondary analysis on the base calls and quality scores generated by Real Time Analysis (RTA) during the sequencing run.

MiSeq Reporter performs analysis according to the analysis workflow specified in the sample sheet, and generates various types of information specific to the workflow upon completion of analysis. Results appear on the MiSeq Reporter web interface in the form of graphs and tables for each run. For more information, see *Analysis Workflows* on page 5.

MiSeq Reporter runs as a Windows service and is viewed through a web browser.

About Windows Service Applications

Windows service applications are continuously running applications that perform specific functions without user intervention, and will continue to run in the background as long as Windows is running. Because MiSeq Reporter runs as a Windows service, it automatically begins secondary analysis when primary analysis is complete. This is triggered by the presence of a file produced at the end of a sequencing run called RTAComplete.txt. For more information, see *Required Input Files* on page 15.

Sequencing During Analysis

If a new sequencing run is started on the MiSeq before secondary analysis of an earlier run is complete, secondary analysis is stopped automatically. MiSeq computing resources are dedicated to either sequencing or analysis and the system is designed in such a way that the sequencing run setup command overrides the analysis command.

To restart secondary analysis, use the Requeue feature on the MiSeq Reporter interface after the new sequencing run is complete. At that point, secondary analysis starts from the beginning.

Viewing MiSeq Reporter

The MiSeq Reporter interface can only be viewed through a web browser. To view the MiSeq Reporter interface during analysis, open any web browser on a computer with access to the same network as the MiSeq, and then connect to the HTTP service on port **8042** using one of the following methods:

- Connect using the instrument IP address followed by :8042.

IP Address	HTTP Service Port	HTTP Address
10.10.10.10, for example	8042	http://10.10.10.10:8042

- Connect using the network name for the MiSeq followed by :8042

Network Name	HTTP Service Port	HTTP Address
MiSeq01, for example	8042	http://MiSeq01:8042

For off-instrument installations of MiSeq Reporter, connect using the method for locally-installed service applications, **localhost** followed by :8042.

Off-Instrument	HTTP Service Port	HTTP Address
localhost	8042	http://localhost:8042

For more information, see *Installing MiSeq Reporter Off-Instrument* on page 121.

MiSeq Reporter Concepts

The following concepts and terms are common to MiSeq Reporter.

Concept	Description
Analysis Workflow	A secondary analysis procedure performed by MiSeq Reporter. The workflow for each run is specified in the sample sheet.
Manifest	The file that specifies a reference genome and targeted reference regions to be used in the alignment step. Manifests are required for the following workflows: Custom Amplicon, Enrichment, PCR Amplicon, and Targeted RNA.
Reference Genome	A FASTA format file that contains the genome sequences used during analysis. • In the Resequencing, LibraryQC, Custom Amplicon, and PCR Amplicon workflows, these sequences are used for alignment. • In the Assembly workflow, the genome is used as a reference to generate an assembly dot-plot. The FASTA files can use the extension *.fa or *.fasta. They are contained in subfolders of the Genome Repository, which is specified in the MiSeq Reporter.config file. For more information, see *MiSeq Reporter Configurable Settings* on page 116 and *Pre-Installed Databases and Genomes* on page 16.
Repository	A folder that holds the data generated during sequencing runs. Each run folder is a subfolder in the repository.
Run Folder	The folder structure populated by RTA primary analysis software (MiSeqOutput folder) or the folder populated by MiSeq Reporter (MiSeqAnalysis). For more information, see *MiSeqAnalysis Folder* on page 104.
Sample Sheet	A comma-separated values file (*.csv) that contains information required to set up and analyze a sequencing run, including a list of samples and their index sequences. The sample sheet must be provided during the run setup steps on the MiSeq. After the run begins, the sample sheet is renamed to SampleSheet.csv and copied to the run folders: MiSeqTemp, MiSeqOutput, and MiSeqAnalysis.

Analysis Workflows

The analysis workflow is a procedure performed by MiSeq Reporter. One analysis workflow must be specified in the sample sheet for each sequencing run. When the run is complete, MiSeq Reporter performs secondary analysis according to that workflow.

Analysis Workflow	Applications	Output
Assembly	Assembles small genomes from reads without the use of a genomic reference. If a genomic reference is specified, a dot-plot is generated with respect to the reference of genome position vs. assembled contigs.	Contigs in FASTA format.
Custom Amplicon	Sequencing of TruSeq Custom amplicons from probes targeting particular genome positions (up to approximately 1536 amplicons from up to 96 samples). Aligns reads against a manifest file specified in the sample sheet. Multiple manifests can be specified: one for each sample and one for control, for example.	Aligned reads in BAM format. Variant calls in VCF format.
Enrichment	Sequencing of DNA that has been enriched for particular target sequences using a pulldown assay. Aligns reads against the whole genome reference, and performs variant analysis for the regions of interest specified in the manifest file. Reporting accumulates coverage and other statistics for each amplicon.	Aligned reads in BAM format. Variant calls in VCF format.
Generate FASTQ	Generates intermediate analysis files in FASTQ format and then exits the workflow. Enables the use of third-party tools to analyze sequencing data.	Sequence files in FASTQ format.
Library QC	Aligns reads against reference genomes specified in the sample sheet, and then generates a sample report in LibraryQC.html.	Per-sample summary statistics.
Metagenomics	Classifies bacteria from a metagenomic sample by amplifying specific regions in 16S ribosomal RNA. No genomic reference is required for the metagenomics workflow. Reads are classified using a database of 16S rRNA data. For paired-end runs, each cluster is classified using base calls from both reads.	Read classifications by taxonomic group from kingdom through genus.

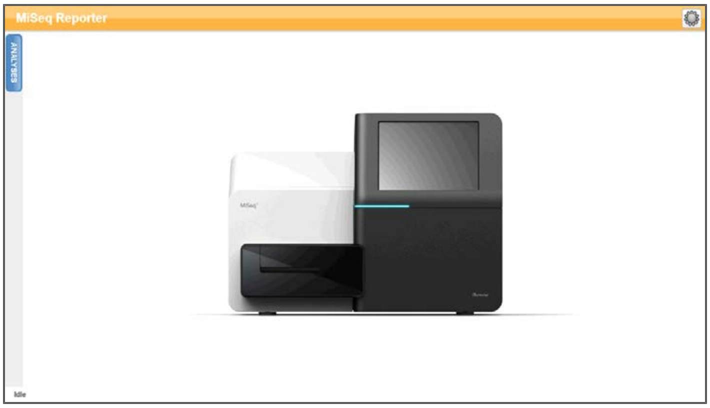
Analysis Workflow	Applications	Output
PCR Amplicon	Sequences any number of PCR amplicons that have been fragmented using Nextera tagmentation. Aligns reads against the reference genomes specified in the sample sheet. Performs variant analysis for the regions of interest specified in the manifest file.	Aligned reads in BAM format. Variant calls in VCF format.
Resequencing	Sequencing of a small genome, such as *E. coli*. Aligns reads against the reference genomes specified in the sample sheet and performs variant analysis.	Aligned reads in BAM format. Variant calls in VCF format.
Small RNA	Sequencing of miRNA. Aligns reads against databases for contaminants, mature miRNA, small RNA, and a genomic reference, in that order.	Reports on the relative abundance of each record.
Targeted RNA	Sequencing of TruSeq Targeted RNA libraries to perform multiplexed gene expression profiling for 12–1000 targets per sample from up to 384 samples. Aligns reads against a manifest file specified in the sample sheet. Quantifies the relative expression of genes and isoforms between several samples, and compare abundance across samples.	Reports on the relative abundance of each target in each sample.

MiSeq Reporter Interface

When MiSeq Reporter opens in the browser, the main screen appears with an image of the instrument in the center, the Settings and Help icons in the upper-right corner, and the Analyses tab in the upper-left corner.

- **MiSeq Reporter Help**—Select the Help icon to open MiSeq Reporter documentation in the browser window.
- **Settings**—Select the Settings icon ⚙ to change the server URL and Repository path.
- **Analyses Tab**—Select the Analyses tab to expand the tab and view a list of analysis jobs that are either completed, queued for analysis, or currently processing.

Figure 1 MiSeq Reporter Main Screen

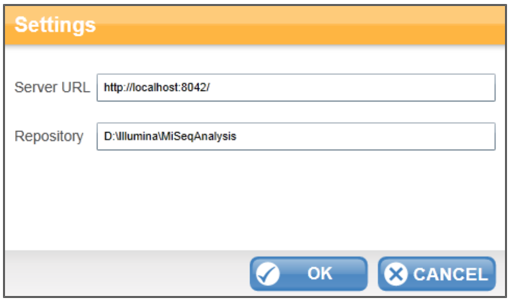


Server URL or Repository Settings

Use the Settings ⚙ feature to change the server URL and the Repository path:

- **Server URL**—The server on which MiSeq Reporter is running.
- **Repository path**—Location of the analysis folder where output files are written.

Figure 2 Settings for Server URL and Repository



Settings

Server URL: http://localhost:8042/

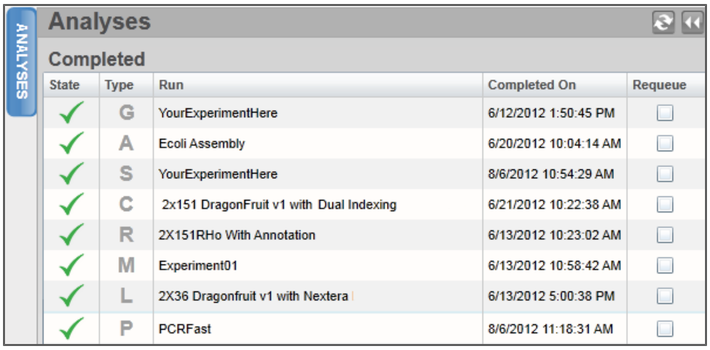
Repository: D:\Illumina\MiSeqAnalysis

Typically, it is not necessary to change these settings unless MiSeq Reporter is running off-instrument. In this case, set the repository path to the network location of the MiSeqOutput folder. For more information, see *Using MiSeq Reporter Off-Instrument* on page 123.

Analyses Tab

The Analyses tab lists all the sequencing runs located in the specified repository. From this tab, you can open the results from any of the runs listed, or requeue a selected run for analysis.

Select the **Refresh Analysis List** icon ↻ in the upper-right corner to refresh the list at any time.



Analyses

Completed

State	Type	Run	Completed On	Requeue
✓	G	YourExperimentHere	6/12/2012 1:50:45 PM	□
✓	A	Ecoli Assembly	6/20/2012 10:04:14 AM	□
✓	S	YourExperimentHere	8/6/2012 10:54:29 AM	□
✓	C	2x151 DragonFruit v1 with Dual Indexing	6/21/2012 10:22:38 AM	□
✓	R	2X151RHo With Annotation	6/13/2012 10:23:02 AM	□
✓	M	Experiment01	6/13/2012 10:58:42 AM	□
✓	L	2X36 Dragonfruit v1 with Nextera	6/13/2012 5:00:38 PM	□
✓	P	PCRFast	8/6/2012 11:18:31 AM	□

Figure 3 Analyses Tab Expanded

The Analyses tab columns are State, Type, Run, Completed On, and Requeue:

- **State**—Shows the current state of the analysis using one of three status icons.

Table 1 State of Analysis Icons

Icon	Description
✓	Indicates that secondary analysis completed successfully.
↻	Indicates that secondary analysis is in progress.
⚠	Indicates that secondary analysis was not completed successfully.

- **Type**—Lists the analysis workflow associated with each run using a single letter designation. Letter designators for each workflow are standard in the MiSeq Reporter interface.

Table 2 Letter Designators for Analysis Workflows

Letter	Workflow
A	Assembly
C	Custom Amplicon
E	Enrichment
G	GenerateFASTQ
L	Library QC
M	Metagenomics
P	PCR Amplicon
R	Resequencing
S	Small RNA
T	Targeted RNA

- **Run**—The name of the run as it is listed in the Experiment Name field of the sample sheet. If an experiment name was not included in the sample sheet prior to performing the sequencing run, this field is populated by the name of the run folder in the MiSeqOutput and MiSeqAnalysis folders.

Alternatively, you can specify a different name for the run by editing the Experiment Name field in the sample sheet. For more information, see *Editing the Sample Sheet in MiSeq Reporter* on page 12.

- **Completed On**—The date that secondary analysis completed.
- **Requeue**—Select the checkbox to requeue a specific job for analysis. The **Requeue** button appears. You might need to requeue analysis if the analysis was interrupted or if you want to reanalyze the data from that run.

  When analysis is queued, the run appears at the bottom of the Analyses tab and indicated as in-progress with the icon ⟳.

Analysis Information and Results Tabs

After selecting a run from the Analyses tab, information and results for that run appear in a series of tabs on the MiSeq Reporter interface.

Analysis results that appear on the Summary and Details tabs vary by workflow. Information on the Analysis, Sample Sheet, Logs, and Errors tabs are similar for each workflow. All tabs are populated when analysis is complete.

Tab Name	Description
Summary Tab	Contains a summary of analysis results in graphs for mismatches, phasing and prephasing, alignment, and clusters passing filter, for example. For more information, see *Summary Tab* on page 9.
Details Tab	Contains details of analysis results in tables and graphs for samples, coverage, Qscores, variants and targets, for example. For more information, see *Details Tab* on page 10.
Analysis Tab	Contains logistical information about the run. For more information, see *Analysis Info Tab* on page 11.
Sample Sheet Tab	Contains run parameters specified in the sample sheet, and provides tools to edit the sample sheet and requeue the run. For more information, see *Sample Sheet Tab* on page 11.
Logs Tab	Lists every step performed during analysis. These steps are recorded in log files located in the Logs folder. A summary is written to AnalysisLog.txt, which is an important file for troubleshooting purposes.
Errors Tab	Lists any errors that occurred during analysis. A summary is written to AnalysisError.txt, which is an important file for troubleshooting purposes.

Summary Tab

The types of information that appear on the Summary tab are specific to the workflow used for the sequencing run and subsequent analysis.

The following table lists the reports for each workflow, which are indicated using a single letter designator as described in *Analyses Tab* on page 7.

Table 3 Summary Tab Information by Workflow

Summary Tab	A	C	E	L	M	P	R	S	T
Low Percentages Graph	■	■	■	■	--	■	■	--	■
High Percentages Graph	■	■	■	■	--	■	■	--	■
Clusters Graph	■	■	■	■	■	■	■	■	■
Mismatch Graph	--	■	■	■	--	■	■	--	■
Trimmed Lengths 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.
- **Trimmed Lengths Graph**—Specific only to the Small RNA workflow, shows a histogram of reads that were trimmed.

Details Tab

The types of information that appear on the Details tab are specific to the workflow used for the sequencing run and subsequent analysis.

The following table lists the reports for each workflow. Workflows are marked using single letter designators. For more information, see *Analyses Tab* on page 7.

Table 4 Details Tab Information by Workflow

Details Tab	A	C	E	L	M	P	R	S	T
Samples Graph	■	--	--	--	--	--	--	--	--
Samples Table	■	■	■	■	■	■	■	■	■
Targets Table	--	■	■	■	--	■	■	--	--
Coverage Graph	--	■	■	■	--	■	■	--	--
QScore Graph	--	■	■	■	--	■	■	--	--
Variant Score Graph	--	■	■	--	--	■	■	--	--
Variants Table	--	■	■	--	--	■	■	--	--
Comparison Graph	--	--	--	--	--	--	--	--	■
Comparison Table	--	--	--	--	--	--	--	--	■

- **Samples Graph**—Specific to the Assembly workflow, 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.
- **Targets Table**—Shows statistics for a particular sample and chromosome.
- **Coverage Graph**—Shows read depth at a given position in the reference.
- **Qscore 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%. For more information, see *Quality Scores* on page 19.
- **Variant Score Graph**—Shows the location of SNPs and indels.
- **Variants Table**—Summarizes differences between sample DNA and the reference. Both SNPs and indels are reported.
- **Comparison Graph**—Specific to the Targeted RNA workflow, compares the relative abundance of each RNA transcript between two selected samples.
- **Comparison Table**—Specific to the Targeted RNA workflow, provides a view of the relative abundance of each transcript in two selected samples.

Analysis Info Tab

Row	Description
Investigator	(Optional) The name of the investigator.
Read Cycles	A representation of the number of cycles in each read, including notation for any index reads. For example, 151, 8 (I), 8 (I), 151, indicates a first read of 151 cycles, followed by two eight-cycle index reads as noted by (I), and then a final read of 151 cycles.
Start Time	The clock time that secondary analysis was started.
Completion Time	The clock time that secondary analysis was completed.
Data Folder	The root level of the output folder produced by RTA primary analysis software (MiSeqOutput), which contains all primary and secondary analysis output for the run.
Analysis Folder	The full path to the Alignment folder in the MiSeqAnalysis folder (Data\Intensities\BaseCalls\Alignment).
Copy Folder	The full path to the Queued subfolder in the MiSeqAnalysis folder.

Sample Sheet Tab

Row	Description
Investigator Name	(Optional) The name of the investigator.
Project Name	(Optional) A descriptive name of the run.
Experiment Name	(Optional) A descriptive name of the experiment.
Date	The date the sequencing run was performed.
Workflow	The analysis workflow for the run.
Assay	The name of the assay used to prepare your samples.

Row	Description
Chemistry	The chemistry name identifies recipe fragments used to build the run-specific recipe. For runs using the Custom Amplicon or PCR Amplicon workflows, the name is amplicon. For all other workflows, the name is default or the field can be blank.
Manifests	The name of the manifest file that specifies alignments to a reference and targeted reference regions. This section is used with the Custom Amplicon, Enrichment, and PCR Amplicon workflows.
Reads	The number of cycles performed in Read 1 and Read 2. Index reads are not included in this section.
Settings	Optional run parameters. For more information, see *Sample Sheet Settings* on page 111.
Data	The sample ID, sample name, index sequences, and path to the genome folder. Requirements vary by workflow.

For information about sample sheet requirements, see the *MiSeq Sample Sheet Quick Reference Guide*, Part # 15028392.

Editing the Sample Sheet in MiSeq Reporter

You can edit the sample sheet for a specific run from the Sample Sheet tab on the MiSeq Reporter web interface. You might need to edit the sample sheet if you want to add settings to the Settings section, or edit the path to reference genomes if you are using the software off-instrument. A mouse and keyboard are required to edit the sample sheet.

- To edit a row in the sample sheet, click any field in the row and make required changes.
- To add a row to the sample sheet, click the row above the intended location of the new row and select **Add Row**.

- To delete a row from the sample sheet, click anywhere in the row and select **Delete Row**.

- After making changes to the sample sheet, select **Save and Requeue**. This saves any changes and initiates secondary analysis using the edited sample sheet.

- If a change to the sample sheet was made in error, click an adjacent tab before saving any changes. A warning appears that states changes were not saved. Click **Discard** to undo any changes or **Save** to save and requeue analysis.

Saving Graphs as Images

MiSeq Reporter provides the option to save an image of graphs shown on the Summary or Details tabs. Right-click any location on the Summary tab or the graphs location on the

Details tab, and then left-click **Save Image As**. When prompted, name the file and browse to a location to save the file.

All images are saved in a JPG (*.jpg) format. Graphs are exported as a single graphic for all graphs shown on the tab. A mouse is required to use this option.

Requeue Analysis

It is possible to requeue analysis from the MiSeq Reporter web interface. Before proceeding, check that a sequencing run is not in progress.

Each time analysis is requeued, the following folders and files are created:

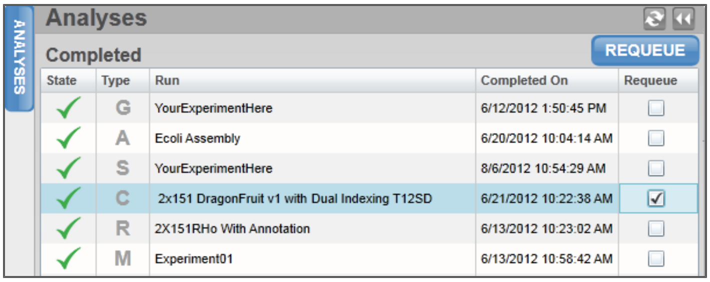
- A new Alignment folder is created in the MiSeqAnalysis folder with a sequential number appended to the folder name, such as Alignment2, for example.
  MiSeqAnalysis\<RunFolderName>\Data\Intensities\BaseCalls\Alignment2
- Existing intermediate analysis files written in FASTQ file format are overwritten with new analysis files. FASTQ files are written to the BaseCalls folder.
  MiSeqAnalysis\<RunFolderName>\Data\Intensities\BaseCalls.

NOTE
If changes were made to the sample sheet, make sure that the modified file is saved to the root level of the analysis folder and the file is named SampleSheet.csv.

1 From the MiSeq Reporter web interface, click **Analyses**.

2 Locate the run from the list of available runs on the Analyses tab, and click the Requeue checkbox adjacent to the run name.
If the run is not listed, confirm that the correct repository is specified using the Settings icon. For more information, see *Server URL or Repository Settings* on page 7.

Figure 4 Requeue Button



State	Type	Run	Completed On	Requeue
✓	G	YourExperimentHere	6/12/2012 1:50:45 PM	□
✓	A	Ecoli Assembly	6/20/2012 10:04:14 AM	□
✓	S	YourExperimentHere	8/6/2012 10:54:29 AM	□
✓	C	2x151 DragonFruit v1 with Dual Indexing T12SD	6/21/2012 10:22:38 AM	☑
✓	R	2X151RHo With Annotation	6/13/2012 10:23:02 AM	□
✓	M	Experiment01	6/13/2012 10:58:42 AM	□

3 Click **Requeue**. The State icon to the left of the run name changes to show that analysis is in progress.

- If analysis does not start, make sure that the following input files are present in the analysis run folder: SampleSheet.csv, RTAComplete.txt, and RunInfo.xml.
- During analysis, a status bar and elapsed time appear on the Analysis Info tab. To stop analysis, select the stop analysis icon next to the status bar on the Analysis Info tab.

Required Input Files

MiSeq Reporter requires the following primary analysis files generated during the sequencing run to perform secondary analysis or to re-queue analysis. Additional primary analysis files (*.bcl, *.filter, and *.locs) must be present to execute analysis correctly.

There is no need to move or copy files to another location before analysis begins. Required files are copied automatically to the MiSeqAnalysis folder during the sequencing process.

File Name	Description
RTAComplete.txt	A marker file that indicates RTA processing is complete. The presence of this file triggers MiSeq Reporter to queue analysis.
SampleSheet.csv	Provides parameters for the run and subsequent analysis. At the start of the run, the sample sheet is copied to the root level of the run folder and renamed SampleSheet.csv.
RunInfo.xml	Contains high-level run information, such as the number of reads and cycles in the sequencing run, and whether or not a read is indexed.

Primary Analysis Files

MiSeq Reporter requires the following primary analysis files to perform secondary analysis.

File Type	Path and File Name Example	Description
***.bcl files**	Data\Intensities\BaseCalls\L001\C1.1\s_1_3.bcl	Base calls for lane 1, cycle 1, tile 3
***.filter files**	Data\Intensities\BaseCalls\L001\s_1_0003.filter	Filter results file for lane 1, tile 3
***.locs files**	Data\Intensities\L001\s_1_3.locs	Location file for lane 1, tile 3

Pre-Installed Databases and Genomes

For most workflows, a reference is required to perform alignment. The MiSeq includes several pre-installed databases and genomes.

Pre-Installed	Description	
Databases	• miRbase for human • dbSNP for human • refGene for human	
Genomes	• *Arabidopsis thaliana* • cow (*Bos taurus*) • *E.coli* strain DH10b • human (*Homo sapiens*) build hg19	• mouse (*Mus musculus*) • rat (*Rattus norvegicus*) • yeast (*Saccharomyces cerevisiae*) • *Staphylococcus aureus*

The reference genome used for analysis by MiSeq Reporter is specified for each sample in the sample sheet (SampleSheet.csv). The full path to the folder containing the whole genome FASTA file must be specified in the sample sheet.

NOTE
You ***must*** enter the full path (UNC path) to the GenomeFolder in the sample sheet. Do not enter the path using a mapped drive.

NOTE
Introduced in MiSeq Reporter v2.1, you can specify genome references for multiple species in the same sample sheet for all workflows ***except*** the Small RNA workflow.

You can upload your own reference in FASTA format to the MiSeq computer. The reference must have a *.fa or *.fasta extension and contained in a single folder. You can upload several single *.fa or *.fasta files ***or*** a single multi-fasta file (recommended), but not a combination of both. Use the Manage Files feature in MCS to upload files.

NOTE
A limitation of custom genomes is that the chromosome name, which is the section of the > line up to any white space, must not contain the hash mark (#) or colon (:) characters. For best results, use only alpha-numeric characters as chromosome names.

Analysis Metrics and Procedures

Introduction

During the sequencing run, Real Time Analysis (RTA) generates data files that include analysis metrics used by MiSeq Reporter for secondary analysis. The following primary analysis metrics appear in secondary analysis reports:

- Clusters passing filter
- Base call quality scores
- Phasing and prephasing values

MiSeq Reporter performs secondary analysis using a series of analysis procedures, which include demultiplexing, FASTQ file generation, alignment, and variant calling.

Table 5 Analysis Procedures

Analysis Procedure	Description
Demultiplexing	Performed for all workflows if the run has index reads and the sample sheet lists multiple samples. For indexed libraries containing either one or two indices, this procedure separates data from pooled samples based on short index sequences that tag samples from different libraries.
FASTQ File Generation	Performed for all workflows. This procedure generates intermediate files in the FASTQ format that contain the non-index reads for each sample, excluding any reads identified as in-line controls and reads from any clusters that did not pass filter.
Alignment	Performed for all workflows, except Assembly and Metagenomics. Alignment compares sequences against the reference specified in the sample sheet to identify a relationship between the sequences and assigns a score based on regions of similarity. MiSeq Reporter uses alignment methods best-suited for the workflow. Aligned reads are written to files in BAM format.
Variant Calling	Performed for the Custom Amplicon, Enrichment, PCR Amplicon, and Resequencing workflows. Variant calling records SNPs and other structural variants in a standardized and parsable text file. MiSeq Reporter uses variant calling algorithms best-suited for the workflow. Variant calls are written to files in VCF format.

Analysis Metrics

During primary analysis, filters and statistical estimates measure data quality and later include these metrics with secondary analysis results. Metrics that appear in secondary analysis reports are clusters passing filter, base call quality scores, and phasing and prephasing values.

Clusters Passing Filter

This filter removes he least reliable data, often derived from overlapping clusters, by filtering raw data to remove any reads that do not meet the overall quality as measured by the Illumina chastity filter. The chastity of a base call is calculated as the ratio of the brightest intensity divided by the sum of the brightest and second brightest intensities.

Clusters passing filter are represented by PF in analysis reports. Clusters pass filter if no more than one base call in the first 25 cycles has a chastity of < 0.6.

Quality Scores

A quality score, or Q-score, is a prediction of the probability of an incorrect base call. A higher Q-score implies that a base call is more reliable and less likely to be incorrect.

Based on the Phred scale, the Q-score serves as a compact way to communicate very small error probabilities. Given a base call, X, the probability that X is not true, P(~X), is expressed by a quality score, Q(X), according to the relationship:

$$Q(X) = -10 \log_{10}(P(\sim X))$$

where P(~X) is the estimated probability of the base call being wrong.

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

Quality Score Q(X)	Error Probability P(~X)
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)

For more information on the Phred quality score, see http://en.wikipedia.org/wiki/Phred_quality_score.

During the sequencing run, base call quality scores are calculated after cycle 25 and results are recorded in base call (*.bcl) files, which contain the base call and quality score per cycle.

ASCII Format for Quality Scores

During analysis, base call quality scores are written to FASTQ files in an encoded compact form, which uses one byte per quality value and represents the quality score in an ASCII format (the value + 33) as illustrated in the following table.

Table 6 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

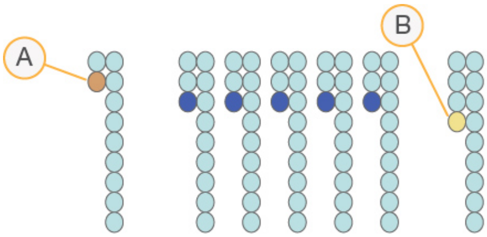
Table 6 ASCII Codes for Q-Scores 0–40

Symbol	ASCII Code	Q-score		Symbol	ASCII Code	Q-score
%	37	4		:	58	25
&	38	5		;	59	26
'	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 might become out of phase with the current incorporation cycle, either falling a base behind (phasing) or jumping a base ahead (prephasing). Phasing and prephasing rates indicate an estimate of the fraction of molecules that became phased or prephased in each cycle.

Figure 5 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 and Read 2 is required for prephasing calculations. Phasing and prephasing results are recorded in the file named phasing.xml, which is located in the folder Data\Intensities\BaseCalls\Phasing.

Phasing and prephasing calculations use statistical averaging over many clusters and sequences to estimate the correlation of signal between different cycles. Therefore, phasing estimates tend to be more accurate for tiles with larger numbers of clusters and a mixture of different sequences. Samples containing only a small number of different sequences do not produce reliable estimates. Sequencing into adapters or other highly homogeneous samples are expected to result in poor phasing estimates.

Demultiplexing

Demultiplexing is the first step in analysis if the sample sheet lists multiple samples and the run has index reads. Each index read sequence is compared 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 if they match the index sequence exactly, or if they have up to a single mismatch per index read.

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. For MiSeq, this is always L1.
- 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.

Additional demultiplexing files are generated for each tile of the flow cell. For more information, see *Demultiplexing File Format* on page 106.

FASTQ File Generation

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

FASTQ 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.

FASTQ files are the primary input for alignment. 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 Naming* on page 107.

Generate FASTQ Workflow

Typically, you do not need to access FASTQ files to obtain analysis results. However, you can use FASTQ files to perform secondary analysis with third-party analysis tools. To generate only FASTQ files, specify the GenerateFASTQ workflow in the sample sheet. This workflow generates FASTQ files and then exits analysis without proceeding to the alignment step.

FASTQ Config Settings

Some default settings for FASTQ file generation can be changed by editing the following settings in the MiSeq Reporter configuration file (C:\Illumina\MiSeq Reporter\MiSeq Reporter.exe.config):

- **ConvertMissingBclsToNoCalls**—By default, FASTQ files include all tiles. You can configure FASTQ file generation to treat *.bcl files that are missing or corrupt as no-calls (Ns) by changing the value to 1 (true).
- **CreateFastqForIndexReads**—By default, FASTQ files are not generated for index reads. You can override this setting by changing the value to 1 (true).
- **FilterNonPFReads**—By default, FASTQ files only include clusters passing filter. You can override this setting by changing the value to 0 (false).

For more information, see *MiSeq Reporter Configurable Settings* on page 116.

Quality Trimming

FASTQ file generation optionally performs quality trimming of the 3' portion of non-index reads with low quality scores, which is a step normally performed during alignment using BWA. For workflows that do not use BWA, you can use the **QualityScoreTrim** setting in the sample sheet to include trimming during FASTQ file generation. For more information, see *Sample Sheet Settings* on page 111.

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 107.

Alignment Methods

For workflows that include alignment, reads are aligned against the reference specified in the sample sheet or in a manifest using an alignment method best-suited for the workflow: Smith-Waterman or BWA, or Bowtie for the Small RNA workflow.

Smith-Waterman

For the Custom Amplicon workflow and Targeted RNA workflow, MiSeq Reporter uses a banded Smith-Waterman algorithm, which performs local sequence alignments to determine similar regions between two sequences. Instead of looking at the total sequence, the Smith-Waterman algorithm compares segments of all possible lengths. Local alignments are useful for dissimilar sequences that are suspected to contain regions of similarity within the larger sequence.

BWA

For Enrichment, Library QC, Resequencing, and PCR Amplicon workflows, MiSeq Reporter uses the Burrows-Wheeler Aligner (BWA), which 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.

When using BWA for alignment, GATK is used for variant calling, by default.

Bowtie

For the Small RNA workflow, MiSeq Reporter v2.2 uses the short-read aligner Bowtie to quickly align large sets of short sequences. For more information, see http://bowtie-bio.sourceforge.net.

Previous versions of MiSeq Reporter used the Illumina-implemented Eland algorithm for alignment.

Variant Calling

Variant calling records single nucleotide polymorphisms (SNPs), insertions and deletions (indels), and other structural variants in a standardized and parsable text file in the variant call format (VCF). For more information, see *VCF File Format* on page 108.

For each SNP or indel call, the probability of an error is provided as a variant quality score. Reads are re-aligned around candidate indels to improve the quality of the calls and site coverage summaries.

Variant Callers

For workflows that include variant calling, variants are detected using one of the following variant callers best-suited for the workflow: GATK, Somatic Variant Caller, or Starling. For more information, see *Aligners and Variant Callers by Workflow* on page 25.

GATK

Developed by the Broad Institute, the Genome Analysis Toolkit (GATK) calls raw variants for each sample read, analyzes the variants against known variants, and then applies a calibration procedure to compute a false discovery rate for each variant. Variants are flagged as homozygous (1/1) or heterozygous (0/1) in the VCF file sample column. For more information, see http://www.broadinstitute.org/gatk.

Somatic Variant Caller

Developed by Illumina, the somatic variant caller identifies variants present at low frequency in the DNA sample and minimizes false positives.

For SNP calling, the somatic variant caller considers each position in the reference genome separately, starting with the bases of aligned reads, and assigns a variant score measuring the accuracy of the call for the SNP. Variant scores are computed based on a Poisson model that excludes the SNP if the SNP has a quality score below Q20, which is a 1/100 chance of being a false positive.

For indels, the somatic variant caller analyzes how many alignments covering a given position include a particular indel compared to the overall coverage at that position. The somatic variant caller does not perform an indel re-alignment step included in other variant callers, such as GATK.

For more information, see the *Somatic Variant Caller Tech Note* available on the Illumina website.

Starling

Starling calls both SNPs and small indels, and summarizes depth and probabilities for every site in the genome. Upon completion, Starling produces html-formatted reports of SNPs and indels.

Starling treats each insertion or deletion as a single mismatch. Base calls with more than two mismatches to the reference sequence within 20 bases of the call are ignored. If the call occurs within the first or last 20 bases of a read, the mismatch limit is increased to 41 bases.

Starling can be used as an optional alternative variant caller with the Smith-Waterman aligner. However, Starling is used primarily with Eland, which has been deprecated in MiSeq Reporter v2.2. For more information, see *Sample Sheet Settings* on page 111.

Aligners and Variant Callers by Workflow

The following table lists aligners and variant callers that MiSeq Reporter uses for each workflow.

Workflow	Aligner	Variant Caller
Custom Amplicon	Smith-Waterman	GATK (default) or Somatic Variant Caller
Enrichment	BWA	GATK (default) or Somatic Variant Caller
Library QC	BWA	--
PCR Amplicon	BWA	GATK (default) or Somatic Variant Caller
Resequencing	BWA	GATK (default) or Somatic Variant Caller
Small RNA	Bowtie	--
Targeted RNA	Smith-Waterman	--

NOTE
Eland, the optional alternative alignment method for the Library QC and Resequencing workflows, has been deprecated in MiSeq Reporter v2.2. For more information, see *Aligner* in *Sample Sheet Settings* on page 111

Assembly Workflow

Chapter 3

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.

The Assembly workflow uses a default k-mer setting of 31. This can be changed to a value of up to 255 by adding the Kmer setting to the Settings section of the sample sheet. For more information, see *Sample Sheet Settings* on page 111.

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 setting MaximumMegabasesAssembly. For more information, see *Available Configurable Settings* on page 116.

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 Assemble_N_Rx.fastq.gz files contain the reads actually used in the assembly process. This selection process is random but not stochastic, meaning the same subset of reads will be selected each time the Assembly workflow is run. This subsampling of reads is done to prevent overloading of the RAM built into MiSeq instrument computer.

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

- Contigs are compared against the reference genome.
- Contigs are reordered to match the order of the reference genome, as closely as possible.
- The samples graph (dot-plot) is generated to summarize the match between contigs and the reference genome. For more information, see *Assembly Samples Graph* on page 30.

Using the Assembly workflow, small genomes (< 20 Mb) can be assembled from a MiSeq sequencing run. Because assembly relies upon significant coverage of the reference genome, this workflow is best suited for the assembly of bacterial genomes (such as *E. coli*).

The assembly process is performed by 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

Assembly summary information includes a low percentages graph, high percentages graph, and clusters graph.

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

Assembly details include a samples graph and a samples table.

Assembly 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 ideal 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.
Mean.Contig.Length	The average contig length for this sample.
Med.Contig.Length	The median contig length for this sample.
Min.Contig.Length	The minimum contig length for this sample.
Max.Contig.Length	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 such that the sum of contigs of equal length or longer is at least 50% of the total length of all contigs.

Assembly Analysis Files

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.
AssemblyRunStatistics.xml	Contains summary statistics specific to the run. Located at the root level of the run folder.
Contigs.fa	Contains the contigs for each assembly. Located in Data\Intensities\BaseCalls\Alignment.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one column per sample. Located in Data\Intensities\BaseCalls\Alignment.
DotPlot.png	Summarizes the match between contigs and the reference genome. 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.

Chapter 4

Custom Amplicon Workflow

Custom Amplicon Workflow Overview

The Custom Amplicon workflow evaluates short regions of amplified DNA, or amplicons, for variants. Focused sequencing of amplicons enables high coverage of particular regions across a large number of samples.

After demultiplexing and FASTQ file generation, the workflow performs the following steps:

- **Alignment**—Clusters from each sample are aligned against amplicon sequences specified in the manifest file.
  - For paired end data, each read is initially evaluated in terms of its alignment to the relevant probe sequences for that read. Read 1 is evaluated against the reverse compliment of the Downstream Locus-Specific Oligos (DLSO) and Read 2 is evaluated against the Upstream Locus-Specific Oligos (ULSO). If the start of a read sequence matches a probe sequence with no more than one mismatch, the full length of the read is then aligned against the amplicon target sequence for that probe sequence. This alignment is performed along the length of the amplicon target sequences using a banded Smith-Waterman alignment.
  - Indels within the DLSO and ULSO are not observed given the assay chemistry.
  - Any alignments that include more than three indels are filtered from alignment results and are not used in variant calling.
- **Paired-end evaluation**—For paired-end runs, the top-scoring alignment for each read is considered. If either read did not align or aligned to different chromosomes, the reads are flagged as an unresolved pair. Additionally, if the two alignments come from different amplicons (i.e., different rows in the Targets section of the manifest), the reads are flagged as an unresolved pair.
- **Bin/Sort**—Reads are grouped by sample and chromosome, and then sorted by chromosome position. Results are written to one BAM file per sample.
- **Variant calling**—SNPs and short indels are identified by the variant caller. The default variant caller is GATK. Optionally, you can specify the Somatic Variant Caller. For more information, see *Variant Calling* on page 24.
- **Variant analysis and annotation**—If a SNP database (dbsnp.txt) is available in the Annotation subfolder of the reference genome folder, any known SNPs or indels are flagged in the VCF output file. If a reference gene database (refGene.txt) is available in the Annotation subfolder of the reference genome folder, any SNPs or indels that fall within known genes are annotated.
- **Statistics reporting**—Statistics are summarized and reported, and written to the Alignment folder.

Custom Amplicon Summary Tab

Custom Amplicon summary information includes a low percentages graph, high percentages graph, clusters graph, and mismatch graph.

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.

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	This value is not applicable to the Custom Amplicon workflow and will always be zero.

Mismatch Graph

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

Custom Amplicon Details Tab

Custom Amplicon details include a samples table, targets table, coverage graph, QScore graph, variant score graph, consensus reads, and variants table.

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.
Cluster PF	The number of clusters passing filter for the sample.
Cluster Align	The total count of PF clusters aligning for the sample (Read 1/Read 2).
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).
Coverage	Median coverage (number of bases aligned to a given reference position) averaged over all positions.
Het SNPs	The number of heterozygous SNPs detected for the sample.
Hom SNPs	The number of homozygous SNPs detected for the sample.
Insertions	The number of insertions detected for the sample.
Deletions	The number of deletions detected for the sample.
Manifest	The name of the file that specifies the alignments to a reference and the targeted reference regions used in the Custom Amplicon workflow.
Genome	The name of the reference genome.

Targets Table

Column	Description
#	An ordinal identification number in the table.
Target ID	The name of the target in the manifest.
Chr	The reference target or chromosome name.
Start Position	The start position of the target region.
End Position	The end position of the target region.

Column	Description
Cluster PF	Number of clusters passing filter for the target displayed per read (Read 1/Read 2).
Mismatch	The percentage of mismatched bases to target averaged over all cycles, displayed per read. Mismatch = [mean(errors count in cycles) / cluster PF] * 100.
No Call	The percentage of no-call bases for the target averaged over cycles, displayed per read.
Het SNPs	The number of heterozygous SNPs detected for the target across all samples.
Hom SNPs	The number of homozygous SNPs detected for the target across all samples.
Insertions	The number of insertions detected for the target across all samples.
Deletions	The number of deletions detected for the target across all samples.
Manifest	The name of the file that specifies the alignments to a reference and the targeted reference regions used in the Custom Amplicon workflow.

Qscore Graph

Y Axis	X Axis	Description
Qscore	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.

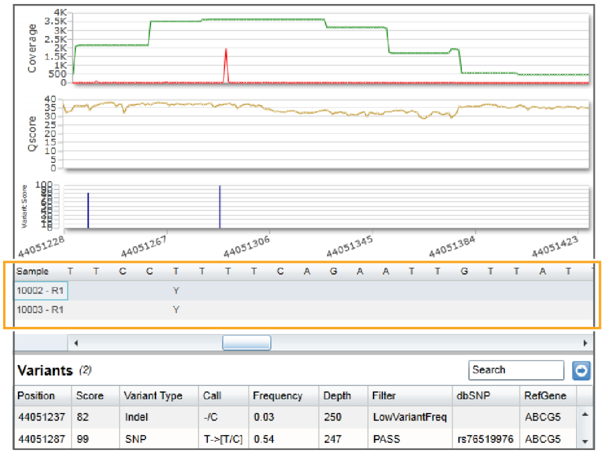
Variant Score Graph

Y Axis	X Axis	Description
Score	Position	Graphically depicts quality score and the position of SNPs and indels.

Consensus Reads

In the Custom Amplicon workflow, data are aligned to produce a consensus read, which reduces stochastic errors in a given sequence. Consensus reads are shown on the Details tab directly below the graphs and represented in the International Union of Pure and Applied Chemistry (IUPAC) convention.

Figure 6 Consensus Reads on Details Tab



Sample	T	T	C	C	T	T	T	T	C	A	G	A	A	T	T	G	T	T	A	T
10002 - R1					Y															
10003 - R1					Y															

Variants (2)

Search

Position	Score	Variant Type	Call	Frequency	Depth	Filter	dbSNP	RefGene
44051237	82	Indel	-/C	0.03	250	LowVariantFreq		ABCG5
44051287	99	SNP	T->[T/C]	0.54	247	PASS	rs76519976	ABCG5

Table 7 IUPAC Nucleotide Codes

Nucleotide Code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	Any purine: A or G
Y	Any pyrimidine: C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C, G, or T
D	A, G, or T
H	A, C, or T
V	A, C, or G
N	any base
. or -	gap

Variants Table

Column	Description
#	An ordinal identification number in the table.

Column	Description
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.
Chr	The reference target or chromosome name.
Position	The position at which the variant was found.
Score	The quality score for this variant.
VariantType	The variant type, which can be either SNP or indel.
Call	A string representing how the base or bases changed at this location in the reference.
Frequency	The fraction of reads for the sample that includes the variant. For example, if the reference base at a particular position is A and sample 1 has 60 A reads and 40 T reads, then the SNP has a variant frequency of 0.4.
Depth	The number of reads for a sample covering a particular position. The GATK variant caller sub-samples data in regions of high coverage. The GATK sub-sampling limit is 5000 in MiSeq Reporter v2.2, raised from 250 in v2.1.
Filter	The criteria for a filtered variant.
dbSNP	The dbSNP name of the variant, if applicable.
RefGene	The gene according to RefGene in which this variant appears.

Custom Amplicon Analysis Files

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in Data\Intensities\BaseCalls\Alignment.
*.vcf files	Contains information about variants found at specific positions in a reference genome. Located in Data\Intensities\BaseCalls\Alignment.
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.
AmpliconCoverage_M#.tsv	Contains details about the resulting coverage per amplicon per sample. M# represents the manifest number. Located in Data\Intensities\BaseCalls\Alignment.
AmpliconRunStatistics.xml	Contains summary statistics specific to the run. Located at 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.
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.
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.

Custom Amplicon Manifest File Format

A manifest file is required input for the Custom Amplicon workflow. The manifest is provided by Illumina with your custom assay (CAT) and uses a ***.txt** file format. The manifest name for each sample is specified in the Data section of the sample sheet.

The Custom Amplicon manifest file contains a header section followed by two blocks of rows beginning with column headings, which are titled the Probes section and the Targets section:

- **Probes**—The Probes section has one entry for each pair of probes. The following columns for this section are required:
  - **Target ID**—A unique identifier consisting of numbers and letters, and used as the display name of the amplicon.
  - **ULSO Sequence**—Sequence of the upstream primer, or Upstream Locus-Specific Oligo, which is sequenced during Read 2 of a paired-end run. For more information, see *Custom Amplicon Workflow Overview* on page 34.
  - **DLSO Sequence**—Sequence of the downstream primer, or Downstream Locus-Specific Oligo. The reverse complement of this sequence forms the start of the first read. This sequence comes from the same strand as the ULSO sequence. For more information, see *Custom Amplicon Workflow Overview* on page 34.
- **Targets**—The Targets section has one entry for each amplicon that is amplified by a probe-pair. An expected off-target region should be included as well as the submitted genomic region. The following columns for this section are required:
  - **TargetA**—Matches a target ID in the Probes section that corresponds to the ULSO probe sequence in Read 1.
  - **TargetB**—Matches a target ID in the Probes section that corresponds to the DLSO probe sequence in Read 2.
  - **Target Number**—Number of the targeted genomic region. The target region for a probe pair has index of 1. Any off-target amplicons have an index of 2, 3, etc.
  - **Chromosome**—The chromosome of the amplicon (e.g., chr 1) that matches the reference chromosome.
  - **Start Position, End Position**—1-based chromosome endpoints of the entire amplicon including the sequence matching the probes. For example, if chromosome 1 started with **ACGTACACGT**, then a sequence with a Start Position of 2 and an End Position of 5 would be **CGTA**.
  - **Probe Strand**—The strand of the amplicon indicated as a plus (+) or minus (-).
  - **Sequence**—Sequence of the amplified region between the ULSO and DLSO. This sequence comes from the forward strand if Probe Strand is plus (+) or from the reverse strand if Probe Strand is minus (-).

Enrichment Workflow

Chapter 5

Enrichment Workflow Overview

The Enrichment workflow analyzes DNA that has been enriched for particular target sequences using a pulldown assay, and then fragmented using Nextera tagmentation.

After demultiplexing and FASTQ file generation, the workflow performs the following steps:

- **Alignment**—Reads are aligned against the entire reference genome using BWA. For more information, see *Alignment* on page 23.
- **Paired-end evaluation**—For paired-end runs, the top-scoring alignment for each read is considered. If either read did not align or aligned to different chromosomes, the reads are flagged as an unresolved pair.
- **Bin/Sort**—Reads are grouped by sample and chromosome, and then sorted by chromosome position. Results are written to one BAM file per sample.
- **Indel realignment**—When using BWA, reads near detected indels are realigned to remove alignment artifacts.
- **Variant calling**—Variant calling is performed only for the regions identified in the manifest file. The default variant caller is GATK. Optionally, you can specify the Somatic Variant Caller. For more information, see *Variant Calling* on page 24.
- **Variant analysis and annotation**—Variant analysis is performed only for the amplicon regions specified in the manifest file.
- **Statistics reporting**—Statistics are summarized and reported, and written to the Alignment folder.

Illumina recommends using the Adapter sample sheet setting for Nextera libraries, which includes adapter trimming during analysis to prevent reporting sequence beyond sample DNA. For more information, see *Sample Sheet Settings* on page 111.

Enrichment Summary Tab

Enrichment summary information includes a low percentage graph, high percentage graph, a clusters graph, and a mismatch graph.

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.

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.

Enrichment Details Tab

Enrichment details include a samples table, targets table, coverage graph, QScore graph, variant score graph, and variants table.

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.
Cluster PF	The number of clusters passing filter for the sample.
Cluster Align	The total count of PF clusters aligning for the sample (Read 1/Read 2).
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).
Coverage	Median coverage (number of bases aligned to a given reference position) averaged over all positions.
Het SNPs	The number of heterozygous SNPs detected for the sample.
Hom SNPs	The number of homozygous SNPs detected for the sample.
Insertions	The number of insertions detected for the sample.
Deletions	The number of deletions detected for the sample.
Median Len	The median fragment length for the sample.
Manifest	The name of the file that specifies the alignments to a reference and the targeted reference regions used in the Enrichment workflow.

Targets Table

By default, the maximum number of rows shown in the Targets table is 40000. This limit is intended to prevent display issues with exome-sized data sets. This limit can be changed using the sample sheet setting EnrichmentMaxRegionStatisticsCount. For more information, see *Sample Sheet Settings* on page 111.

Column	Description
#	An ordinal identification number in the table.
Name	The name of the target in the manifest.

Column	Description
Chr	The reference target or chromosome name.
Start Position	The start position of the target region.
End Position	The end position of the target region.
Cluster PF	Number of clusters passing filter for the target displayed per read (Read 1/Read 2).
Mismatch	The percentage of mismatched bases to target averaged over all cycles, displayed per read. Mismatch = [mean(errors count in cycles) / cluster PF] * 100.
No Call	The percentage of no-call bases for the target averaged over cycles, displayed per read.
Het SNPs	The number of heterozygous SNPs detected for the target across all samples.
Hom SNPs	The number of homozygous SNPs detected for the target across all samples.
Insertions	The number of insertions detected for the target across all samples.
Deletions	The number of deletions detected for the target across all samples.
Manifest	The name of the file that specifies the alignments to a reference and the targeted reference regions.

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.

Qscore Graph

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

Variant Score Graph

Y Axis	X Axis	Description
Score	Position	Graphically depicts quality score and the position of SNPs and indels.

Variants 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.
Chr	The reference target or chromosome name.
Position	The position at which the variant was found.
Score	The quality score for this variant.
Variant Type	The variant type, which can be either SNP or indel.
Call	A string representing how the base or bases changed at this location in the reference.
Frequency	The fraction of reads for the sample that includes the variant. For example, if the reference base at a particular position is A and sample 1 has 60 A reads and 40 T reads, then the SNP has a variant frequency of 0.4.
Depth	The number of reads for a sample covering a particular position. The GATK variant caller sub-samples data in regions of high coverage. The GATK sub-sampling limit is 5000 in MiSeq Reporter v2.2, raised from 250 in v2.1.
Filter	The criteria for a filtered variant.
dbSNP	The dbSNP name of the variant, if applicable.
RefGene	The gene according to RefGene in which this variant appears.
Genome	The name of the reference genome.

Enrichment Analysis Files

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in Data\Intensities\BaseCalls\Alignment.
*.coverage.csv	A comma-separated values file that contains information about mean coverage by target region. Located in Data\Intensities\BaseCalls\Alignment.
*.gaps.csv	A comma-separated values file that contains information about gaps in targeted regions. Located in Data\Intensities\BaseCalls\Alignment.
*.vcf files	Contains information about variants found at specific positions in a reference genome. Located in Data\Intensities\BaseCalls\Alignment.
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.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one 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.
EnrichmentStatistics.xml	Contains summary statistics specific to the run. Located at 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.
SampleName.enrichment_Summary.csv	Contains a summary of performance metrics generated by the Enrichment workflow. Located in Data\Intensities\BaseCalls\Alignment.
SampleName_regions_Manifest_intervals.txt	Contains a list of regions used to generate summary statistics. This file is generated from the manifest file specified in the sample sheet. Located in Data\Intensities\BaseCalls\Alignment.
Summary.htm	Contains a summary web page generated from Summary.xml. Located in Data\Intensities\BaseCalls\Alignment.

Enrichment Analysis File Formats

MiSeq Reporter generates two file formats that are unique to the Enrichment workflow: the coverage file (*.coverage.csv) and the gaps file (*.gaps.csv). Additionally, a summary metrics file named SampleName.enrichment_summary.csv is generated for each sample ID.

Coverage File Format

The coverage files generated by the Enrichment workflow contain information about mean coverage by targeted region, aligned reads in the sample, and the enrichment percentage. These files are in *.csv format, which can be loaded into a spreadsheet program such as Microsoft Excel for viewing, sorting, or graphing.

Coverage files contain a header section and a data section:

- **Header**—The header section contains one line per targeted region and each line begins with a # character. The first header line specifies the enrichment, which is defined as the fraction of aligned reads overlapping any of the targeted regions. The second header line specifies the number of reads aligning to targeted regions. The third header line specifies the column headings as shown in the following example:

```
#Enrichment: 55.3%
#Reads: 598713
#Chromosome,Start,Stop,RegionID,MeanCoverage
```

- **Data**—The data section includes columns described in the following table.

Column Heading	Description
Chromosome	Contains the chromosome of the targeted region.
Start	Contains the start position of the targeted region.
Stop	Contains the stop position of the targeted region.
RegionID	Contains the identity of the region as specified in the manifest.
MeanCoverage	Contains the mean coverage. Only reads mapped as proper pairs count toward the coverage calculation if the run is a paired-end run.

Gaps File Format

The gaps files generated by the Enrichment workflow contain information about targeted intervals where coverage fell below the threshold used to filter variants for low depth. This threshold is set using the **MinimumCoverageDepth** sample sheet setting. For more information, see *Sample Sheet Settings for Variant Calling* on page 112.

Given a depth threshold, a gap is defined as a consecutive run of bases in which all bases have coverage less than the threshold. It is in these regions that variants are filtered due to low depth. The gaps file lists all gaps identified in any targeted region.

Gaps files contain a header section and a data section:

- **Header**—The header section is a single line that specifies the following column headings:

```
#Chromosome,GapStart,GapStop,RegionID,MeanGapCoverage,
   RegionInterval,GapInterval
```

- **Data**—The data section includes columns described in the following table.

Column Heading	Description
Chromosome	Contains the chromosome of the targeted region.
GapStart	Contains the first coordinate of the gap.
GapStop	Contains the last coordinate of the gap.
RegionID	Contains the identity of the region as specified in the manifest.
MeanGapCoverage	Contains the mean coverage in the gap region. Only proper pairs are counted in a paired-end run.
RegionInterval	Contains a representation of the targeted interval in a format that can be easily copied and pasted into genome and read browsers.
GapInterval	Contains a representation of the gap interval in a format that can be easily copied and pasted into genome and read browsers.

Enrichment Summary File Format

Unique to the Enrichment workflow, MiSeq Reporter generates a summary of metrics for each sample ID written to files named SampleName.enrichment_summary.csv.

Heading	Description
Statistic	Definition.
Sample ID	Sample ID.
Run Folder	Path to the run folder.
Total Aligned Bases	Total aligned bases.
Targeted Aligned Bases	Total aligned bases in the target region.
Base Enrichment (not padded)	100*(Total aligned bases in targeted regions/total aligned bases).
Percent Duplicate Paired Reads	Percentage of paired reads that have duplicates.
Total Aligned Reads	Total number of aligned reads.
Targeted Aligned Reads	Number of reads that aligned to the target.
Read Enrichment	100*(Target aligned reads/Total aligned reads).
Total Length of Targeted Regions	Total length of sequenced bases in the target region.

Heading	Description
Mean Region Coverage Depth	The total number of targeted bases divided by the targeted region size. Roughly equivalent to the weighted mean of the region coverage in the <sample>.coverage.csv file.
Uniformity of Coverage (Pct > 0.2*mean)	The percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.
Target Coverage Above 1X	Percentage of targets with coverage greater than 1X.
Target Coverage Above 10X	Percentage of targets with coverage greater than 10X.
Target Coverage Above 20X	Percentage of targets with coverage greater than 20X.
Target Coverage Above 50X	Percentage of targets with coverage greater than 50X.
Insert Size Median	Median length of the sequenced fragment.
Insert Size Minimum	Minimum length of the sequenced fragment.
Insert Size Maximum	Maximum length of the sequenced fragment.
Insert Size SD	Standard deviation of the lengths of the sequenced fragment.
SNPs	Total number of SNPs present in the data set and pass the quality filters.
SNPs (Percent found in dbSNP)	100*(Number of SNPs in dbSNP/Number of SNPs).
SNP Het/Hom Ratio	Number of Heterozygous SNPs/Number of Homozygous SNPs.
SNP Ts/Tv Ratio	Transition rate of SNPs that pass the quality filters/Transversion rate of SNPs that pass the quality filter.
Indels	Total number of indels present in the data set that pass the quality filters.
Indels (Percent found in dbSNP)	100*(Number of Indels in dbSNP/Number of Indels).
Indel Het/Hom Ratio	Number of Heterozygous Indels/Number of Homozygous Indels.

Enrichment Manifest File Format

A manifest file is required input for the Enrichment workflow. The Enrichment manifest is provided for download from the Illumina website. The manifest name for each sample is specified in the Data section of the sample sheet.

The Enrichment manifest file begins with a header section comprising a header line followed by Manifest Version and ReferenceGenome.

The main section of the manifest file is the **Regions** section, which contains the following columns:

- **Name**—Unique user-specified name for the amplicon.
- **Chromosome**—Chromosome from which the amplicon originates.
- **Start**—1-based coordinate start position of the amplicon including the probe.
- **End**—1-based and inclusive coordinate of the end position of the amplicon including the probe.
- **Upstream Probe Length**—The length of the upstream (5') PCR probe. For the Enrichment workflow, this field should be set to zero.
- **Downstream Probe Length**—The length of the downstream (3') PCR probe. For the Enrichment workflow, this field should be set to zero.
- **Group**—(*For TruSight panels only*) If specified, this column can be used to group together regions (e.g., for a particular gene).

Chapter 6

Library QC Workflow

Library QC Workflow Overview

The Library QC workflow is intended for evaluating abundance, fragment length, and sample quality of DNA libraries. The analysis performed in the Library QC workflow is very similar to the Resequencing workflow with two exceptions:

- Alignment is performed using a faster, less sensitive setting that provides significantly faster processing time.
- No variant calling is performed after alignment. Instead a sample report is written to LibraryQC.html in the Alignment folder, which lists the characteristics of each DNA sample in terms of percentage of reads aligned. Results written to the sample report appear in the samples table for the run.

Library QC Summary Tab

Library QC summary information includes a low percentages graph, high percentages graph, clusters graph, and mismatch graph.

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

Library QC details include a samples table, targets table, coverage graph, and QScore graph.

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 three standard deviations from the median.
Max Len	The high percentile of fragment lengths for this sample as they correspond to three 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.

Column	Description
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.

Qscore Graph

Y Axis	X Axis	Description
Qscore	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.

Library QC Analysis Files

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in Data\Intensities\BaseCalls\Alignment.
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.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one 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 at 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.

Chapter 7

Metagenomics Workflow

Metagenomics Workflow Overview

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

After demultiplexing and FASTQ file generation, the workflow performs a classification of reads.

Classification of Reads

Classification of reads is performed using a Bayesian classifier. 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.

Summary statistics provide the total number of classified clusters for each sample at each taxonomic level. Statistics are written to the file Classification.txt. For more information, see *Metagenomics Analysis Files* on page 67.

Current Taxonomy

The source of the taxonomy stored in Taxonomy.dat is the GreenGenes database: http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/current_GREENGENES_gg16S_unaligned.fasta.gz

Currently, the following taxonomic counts are available for the Metagenomics workflow.

Taxonomy	Count
Kingdoms	2
Phyla	32
Classes	73
Orders	149
Families	379
Genera	1311

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, typicallly 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: http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/current_GREENGENES_gg16S_unaligned.fasta.gz

Metagenomics Summary Tab

Metagenomics summary information includes a clusters graph.

Clusters Graph

Table 9

Table 8 Cluster Graph Information

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

Metagenomics details include a samples table and clusters pie chart.

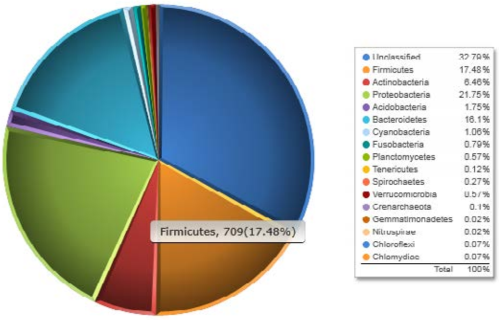
Samples Table

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.
Taxomic 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, Genus.
Clusters Classified	The number of clusters that were confidently classified at this taxonomic level.

Metagenomics Pie Chart

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

Figure 7 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.

Metagenomics Analysis Files

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.
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 at the root level of the run folder.

Chapter 8

PCR Amplicon Workflow

PCR Amplicon Workflow Overview

The PCR Amplicon workflow sequences any number of PCR amplicons that have been fragmented using Nextera tagmentation.

After demultiplexing and FASTQ file generation, the workflow performs the following steps:

- **Alignment**—Reads are aligned against the reference genomes specified in the sample sheet using BWA. For more information, see *Alignment* on page 23.
- **Paired-end evaluation**—For paired-end runs, the top-scoring alignment for each read is considered. If either read did not align or aligned to different chromosomes, the reads are flagged as an unresolved pair.
- **Bin/Sort**—Reads are grouped by sample and chromosome, and then sorted by chromosome position. Results are written to one BAM file per sample.
- **Indel realignment**—When using BWA, reads near detected indels are realigned to remove alignment artifacts.
- **Variant calling**—SNPs and short indels are identified by the variant caller. The default variant caller is GATK. Optionally, you can specify the Somatic Variant Caller. For more information, see *Variant Calling* on page 24.
- **Variant analysis and annotation**—Variant analysis is performed only for the amplicon regions specified in the manifest file. Variant calling is performed only for those bases between the PCR amplification probes to avoid including synthetic DNA in variant calls.
- **Statistics reporting**—Statistics are summarized and reported, and written to the Alignment folder.

Illumina recommends using the Adapter sample sheet setting for Nextera libraries, which includes adapter trimming during analysis to prevent reporting sequence beyond sample DNA. For more information, see *Sample Sheet Settings* on page 111.

PCR Amplicon Summary Tab

PCR Amplicon summary information includes a low percentage graph, high percentage graph, a clusters graph, and a mismatch graph.

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.

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.

PCR Amplicon Details Tab

PCR Amplicon details include a samples table, targets table, coverage graph, QScore graph, variant score graph, and variants table.

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.
Cluster PF	The number of clusters passing filter for the sample.
Cluster Align	The total count of PF clusters aligning for the sample (Read 1/Read 2).
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).
Coverage	Median coverage (number of bases aligned to a given reference position) averaged over all positions.
Het SNPs	The number of heterozygous SNPs detected for the sample.
Hom SNPs	The number of homozygous SNPs detected for the sample.
Insertions	The number of insertions detected for the sample.
Deletions	The number of deletions detected for the sample.
Median Len	The median fragment length for the sample.
Manifest	The name of the file that specifies the alignments to a reference and the targeted reference regions.

Targets Table

Column	Description
#	An ordinal identification number in the table.
Name	The name of the target in the manifest.
Chr	The reference target or chromosome name.
Start Position	The start position of the target region.
End Position	The end position of the target region.

Column	Description
Cluster PF	Number of clusters passing filter for the target displayed per read (Read 1/Read 2).
Mismatch	The percentage of mismatched bases to target averaged over all cycles, displayed per read. Mismatch = [mean(errors count in cycles) / cluster PF] * 100.
No Call	The percentage of no-call bases for the target averaged over cycles, displayed per read.
Het SNPs	The number of heterozygous SNPs detected for the target across all samples.
Hom SNPs	The number of homozygous SNPs detected for the target across all samples.
Insertions	The number of insertions detected for the target across all samples.
Deletions	The number of deletions detected for the target across all samples.
Manifest	The name of the file that specifies the alignments to a reference and the targeted reference regions.

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.

Qscore Graph

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

Variant Score Graph

Y Axis	X Axis	Description
Score	Position	Graphically depicts quality score and the position of SNPs and indels.

Variants 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.

Column	Description
Sample Name	The sample name from the sample sheet.
Chr	The reference target or chromosome name.
Position	The position at which the variant was found.
Score	The quality score for this variant.
Variant Type	The variant type, which can be either SNP or indel.
Call	A string representing how the base or bases changed at this location in the reference.
Frequency	The fraction of reads for the sample that includes the variant. For example, if the reference base at a particular position is A and sample 1 has 60 A reads and 40 T reads, then the SNP has a variant frequency of 0.4.
Depth	The number of reads for a sample covering a particular position. The GATK variant caller sub-samples data in regions of high coverage. The GATK sub-sampling limit is 5000 in MiSeq Reporter v2.2, raised from 250 in v2.1.
Filter	The criteria for a filtered variant.
dbSNP	The dbSNP name of the variant, if applicable.
RefGene	The gene according to RefGene in which this variant appears.
Genome	The name of the reference genome.

PCR Amplicon Analysis Files

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in Data\Intensities\BaseCalls\Alignment.
*.vcf files	Contains information about variants found at specific positions in a reference genome. Located in Data\Intensities\BaseCalls\Alignment.
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.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one 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.
PCRAmpliconRunStatistics.xml	Contains summary statistics specific to the run. Located at 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.

PCR Amplicon Manifest File Format

A manifest file is required input for the PCR Amplicon workflow. The manifest is generated using the Illumina Experiment Manager and uses a tab-delimited ***.AmpliconManifest** file format. The manifest name for each sample is specified in the Data section of the sample sheet.

The PCR Amplicon manifest file begins with a header section comprising a header line followed by two columns: ReferenceGenome and Reference Genome Folder. The main section of the file is the **Regions** section, which contains the following columns:

- **Name**—Unique user-specified name for the amplicon.
- **Chromosome**—Chromosome from which the amplicon originates.
- **Amplicon Start**—1-based coordinate start position of the amplicon including the probe.
- **Amplicon End**—1-based and inclusive coordinate of the end position of the amplicon including the probe.
- **Upstream Probe Length**—The length of the upstream (5') PCR probe. For the Enrichment workflow, this field should be set to zero.
- **Downstream Probe Length**—The length of the downstream (3') PCR probe. For the Enrichment workflow, this field should be set to zero.

Resequencing Workflow

Chapter 9

Resequencing Workflow Overview

The Resequencing workflow compares the DNA sequence in the samples against a reference genome and identifies any variants (SNPs or indels) relative to the reference sequence.

After demultiplexing and FASTQ file generation, the workflow performs the following steps:

- **Alignment**—Reads are aligned against the reference genomes specified in the sample sheet using BWA, by default. For more information, see *Alignment* on page 23.
- **Paired-end evaluation**—For paired-end runs, the top-scoring alignment for each read is considered. If either read did not align or aligned to different chromosomes, the reads are flagged as an unresolved pair.
- **Bin/Sort**—Reads are grouped by sample and chromosome, and then sorted by chromosome position. Results are written to one BAM file per sample.
- **Indel realignment**—When using BWA, reads near detected indels are realigned to remove alignment artifacts.
- **Variant calling**—SNPs and short indels are identified by the variant caller. The variant caller specified in the sample sheet is GATK by default when using BWA for alignment. Optionally, you can specify the Somatic Variant Caller. For more information, see *Variant Calling* on page 24.
- **Variant analysis and annotation**—If a SNP database (dbsnp.txt) is available in the Annotation subfolder of the reference genome folder, any known SNPs or indels are flagged in the VCF output file. If a reference gene database (refGene.txt) is available in the Annotation subfolder of the reference genome folder, any SNPs or indels that fall within known genes are annotated.
- **Statistics reporting**—Statistics are summarized and reported, and written to the Alignment folder.

Resequencing Summary Tab

Resequencing summary information includes a low percentage graph, high percentage graph, a clusters graph, and a mismatch graph.

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.

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.

Resequencing Details Tab

Resequencing details include a samples table, targets table, coverage graph, QScore graph, variant score graph, and variants table.

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.
Cluster PF	The number of clusters passing filter for the sample that aligned to the reference genome.
Cluster Align	The total count of PF clusters aligning for the sample (Read 1/Read 2).
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).
Coverage	Median coverage (number of bases aligned to a given reference position) averaged over all positions.
Het SNPs	The number of heterozygous SNPs detected for the sample.
Hom SNPs	The number of homozygous SNPs detected for the sample.
Insertions	The number of insertions detected for the sample.
Deletions	The number of deletions detected for the sample.
Median Len	The median fragment length for the sample.
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.
Cluster Align	The total count of PF clusters aligning for the sample (Read 1/Read 2).

Column	Description
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).
Coverage	Median coverage (number of bases aligned to a given reference position) averaged over all positions.
Het SNPs	The number of heterozygous SNPs detected for the sample.
Hom SNPs	The number of homozygous SNPs detected for the sample.
Insertions	The number of insertions detected for the sample.
Deletions	The number of deletions detected for the sample.
Genome	The name of the reference genome.

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.

Qscore Graph

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

Variant Score Graph

Y Axis	X Axis	Description
Score	Position	Graphically depicts quality score and the position of SNPs and indels.

Variants 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.

Column	Description
Chr	The reference target or chromosome name.
Position	The position at which the variant was found.
Score	The quality score for this variant.
Variant Type	The variant type, which can be either SNP or indel.
Call	A string representing how the base or bases changed at this location in the reference.
Frequency	The fraction of reads for the sample that includes the variant. For example, if the reference base at a particular position is A and sample 1 has 60 A reads and 40 T reads, then the SNP has a variant frequency of 0.4.
Depth	The number of reads for a sample covering a particular position. The GATK variant caller sub-samples data in regions of high coverage. The GATK sub-sampling limit is 5000 in MiSeq Reporter v2.2, raised from 250 in v2.1.
Filter	The criteria for a filtered variant.
dbSNP	The dbSNP name of the variant, if applicable.
RefGene	The gene according to RefGene in which this variant appears.
Genome	The name of the reference genome.

Resequencing Analysis Files

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in Data\Intensities\BaseCalls\Alignment.
*.vcf files	Contains information about variants found at specific positions in a reference genome. Located in Data\Intensities\BaseCalls\Alignment.
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.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one 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 at 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.

Small RNA Workflow

Chapter 10

Small RNA Workflow Overview

The Small RNA workflow measures the abundance of various types of short RNA sequences, particularly miRNA. It is suitable for identifying and quantifying miRNA expression and for comparing abundance across samples.

NOTE
The FASTQ (fastq.gz) files generated during the Small RNA workflow are intermediate files, and do not contain all of the reads generated during sequencing. These files only contain reads not removed by the duplicate filtering or alignment steps.

The workflow performs the following steps:

- **Mask adapter**—As specified by the Adapter setting in the sample sheet, mask the adapter sequence from all reads by replacing adapter bases with N (no-call) and set the quality score to a value of 2. Then, filter reads that do not pass filter and collapse duplicate identical reads to a single entry. A binary.count file is written to specify the number of times each read sequence was observed for each sample, and a histogram of read lengths is written to TrimmerHistogram.txt.

  This method of adapter masking is performed by default in the Small RNA workflow. For more information, see *Sample Sheet Settings* on page 111.
- **Alignment**—Using Bowtie v0.12.8, each cluster is aligned against reference databases specified in the sample sheet. All samples must use the same set of reference databases.

Database Name	Order of Precedence	Column in Sample Sheet
Contaminants	1	Specified in Contaminants column
Mature miRNA	2	Specified in miRNA column
RNA	3	Specified in RNA column
Reference genome	4	Specified in GenomeFolder column

- **Statistics reporting**—Clusters are aligned in order of precedence using the following criteria:
  - A cluster that aligns to the contaminants database is considered to be a contaminant even if it also aligns to the reference genome.
  - Clusters that match the miRNA database and not contaminants are counted as miRNA.
  - By default, only exact matches to mature miRNA records are counted.
  - Matches must align to the start of the reference sequence, on the same strand, with adapter-masked read lengths identical to the length of the reference sequence.
  - If the same sequence maps to multiple database records with the same number of mismatches, counts are split across records.

Small RNA Summary Tab

Small RNA summary information includes a clusters graph and trimmed lengths graph.

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.

Trimmed Lengths

Y Axis	X Axis	Description
Clusters	Trimmed Lengths	Histogram of reads that were trimmed.

Small RNA Details Tab

Small RNA details include a samples table, pie chart, and graph.

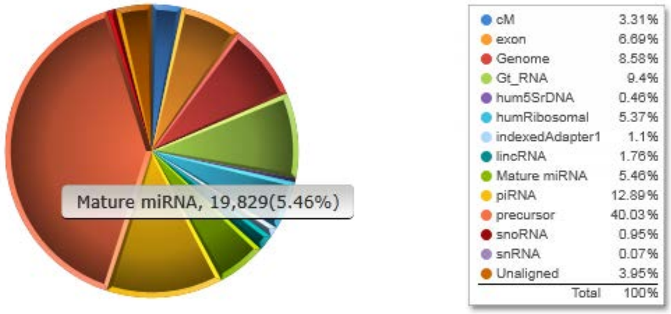
Samples Table

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.
Cluster Raw	The number of raw clusters detected for the sample.
Cluster PF	The number of clusters passing filter for the sample.
Cluster Align Contam	The number of clusters that match records in the Contaminants database.
Cluster Align miRNA	The number of clusters that exactly match records in the Mature miRNA database.
Cluster Align RNA	The number of clusters that match records in the RNA database.
Cluster Align Genome	The number of clusters that match records in the genomic database.
Cluster Unaligned	The number of clusters that did not align against any reference database.

Small RNA Pie Chart

The Small RNA pie chart provides a visualization of clusters identified as mature miRNA, other forms of RNA, genomic sequence, or contaminants.

Figure 8 Small RNA Pie Chart



Common categories for the Small RNA pie chart are as follows:

- Unaligned clusters that did not align against any reference
- Genome clusters that aligned to the reference genome
- miRNA clusters that aligned to the mature miRNA database

Hits to the mature miRNA database are counted only if the cluster aligned to the correct strand and position for the mature miRNA.

The remaining category names in the Small RNA pie chart are taken from the FASTA file names in the databases. For example, if the RNA database contains a file named rRNA.fa, then matches to this file are reported as the category **rRNA**.

Small RNA Graph

The Small RNA graph provides a plot of the common mature miRNA sequences for a sample and their abundances. The most common miRNA sequences for the selected sample (up to ten records) are shown in proportion to the number of clusters matched.

Small RNA Analysis Files

File Name	Description
*.export.contam, *.contam	Lists the number of clusters that exactly match records in the Contaminants database. These files are used to populate the samples table and pie chart. Located in Data\Intensities\BaseCalls\Alignment.
*.export.genome, *.genome	Lists the number of clusters that exactly match records in the genomic database. These files are used to populate the samples table and pie chart. Located in Data\Intensities\BaseCalls\Alignment.
*.export.mirna, *.mirna	Lists the number of clusters that exactly match records in the Mature miRNA database. These files are used to populate the samples table and pie chart. Located in Data\Intensities\BaseCalls\Alignment.
*.export.rna, *.rna	Lists the number of clusters that exactly match records in the RNA database. These files are used to populate the samples table and pie chart. Located in Data\Intensities\BaseCalls\Alignment.
AdapterTrimming.txt	Lists the number of trimmed bases and percentage of bases for each tile. Located in Data\Intensities\BaseCalls\Alignment.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one column per sample. Located in Data\Intensities\BaseCalls\Alignment.
SmallRNARunStatistics.xml	Contains summary statistics specific to the run. Located at the root level of the run folder.
TrimmerHistogram.txt	Contains a histogram of masked read lengths. Located in Data\Intensities\BaseCalls\Alignment.

Targeted RNA Workflow

Chapter 11

Targeted RNA Workflow Overview

Designed for TruSeq Targeted RNA libraries, the Targeted RNA workflow analyzes RNA sequences for a set of predefined transcripts. Transcripts are targeted by identifying regions of interest which may contain splice junctions, exons, cSNPs, or other expressed sequences. MiSeq Reporter aligns reads, estimates the depth of sequencing for each sample, estimates variance, and calculates the significance of the observed differential expression.

After demultiplexing and FASTQ file generation, the workflow performs the following steps:

- **Alignment**—Reads from each sample are aligned against references specified in the manifest using a banded Smith-Waterman alignment. For the Targeted RNA workflow, this process allows alignment across very small amplicon targets, often less than 10 bp.

  In addition to creating BAM files during alignment, the Targeted RNA workflow produces target hits files that contain raw aligned replicate counts for each transcript. For more information, see *Target Hits File Format* on page 100.
- **Differential expression analysis**—Differential expression analysis performs depth of sequencing normalization, variance estimation, and p-value calculation:
  - **Depth of sequencing normalization**—This step estimates the depth of sequencing in order to compare two different samples. For example, if the average counts of non-differentially expressed genes in one sample are twice as high as in another sample, the depth of sequencing for the first sample should be twice that of the other sample.

    For each transcript, the geometric mean of aligned read counts is calculated for all sample IDs. The median of the ratio of raw counts to the geometric mean is used as the scaling factor for each replicate. When evaluating non-differentially expressed transcripts, dividing raw counts by the scaling factor yields equivalent scaled counts. By default, the expression levels for each replicate are normalized based on the total number of aligned reads.

    NOTE
    Optionally, you can normalize based on a specific gene or genes. To customize normalization, add a **Normalize** column to the Data section of the sample sheet, and specify the gene name. Separate two or more gene names with either a semicolon (;) or a plus sign (+). All samples for a given manifest ***must*** use the same normalization method.
  - **Variance estimation**—The squared coefficient of variation is calculated by dividing the raw variance by the square of the mean. The bias inherent in estimating variance is correlated with the number of biological replicates.

    Variance is calculated using data from all sample IDs rather than calculating for each sample name separately. In this workflow, biological replicates are denoted by using a common sample name.
  - **P-value calculation**—The normalized transcript abundance is modeled by a negative binomial distribution, and this model is used to derive a p-value for the differential expression (up- or down-regulations) of each transcript. Q-values are computed using the Benjamini-Hochberg procedure to control the false discovery rate (FDR) by correcting for multiple hypothesis testing.
- **Statistics reporting**—Statistics are summarized and reported, and written to the Alignment folder.

Targeted RNA Summary Tab

Targeted RNA summary information includes a low percentage graph, high percentage graph, a clusters graph, and a mismatch graph.

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.
	PrePhasing 1	The percentage of molecules in a cluster that run ahead of the current cycle within Read 1.
	Mismatch 1	The average percentage of mismatches for Read 1 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.
	I20 / I1 1	The ratio of intensities at cycle 20 to the intensities at cycle 1 for Read 1.

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.

Mismatch Graph

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

Targeted RNA Details Tab

Targeted RNA details include a samples table, comparison graph, and comparison table.

Samples Table

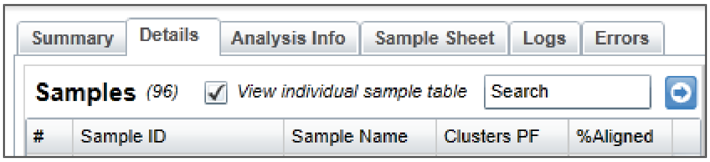
The sample table has two views, the default view, which is described in the following table, and a view that shows data for individual samples.

Column	Description
#	An ordinal identification number in the table.
Sample Name	The sample name from the sample sheet.
Manifest	The name of the manifest file that specifies alignments to a reference and targeted reference regions.
# Replicates	The number of biological replicate samples (grouped by sample name) processed in the sequencing run.
Clusters PF	The number of clusters passing filter for the sample.

Samples Table - View Individual Sample Table

Select the checkbox **View individual sample table** to change the sample table view.

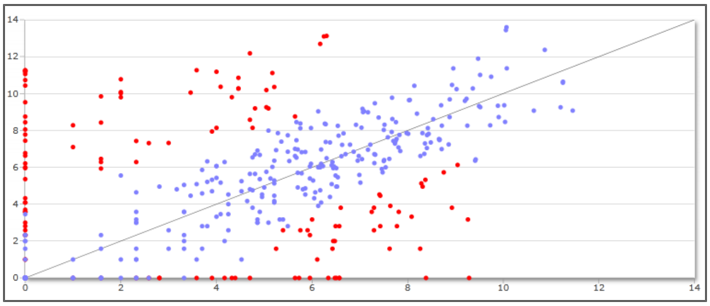
Figure 9 View Individual Sample Table Checkbox



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 PF	The number of clusters passing filter for the sample.
%Aligned	The percentage of clusters successfully aligned.
Manifest	The name of the file that specifies the alignments to a reference and the targeted reference regions.

Comparison Graph

Figure 10 Comparison Graph

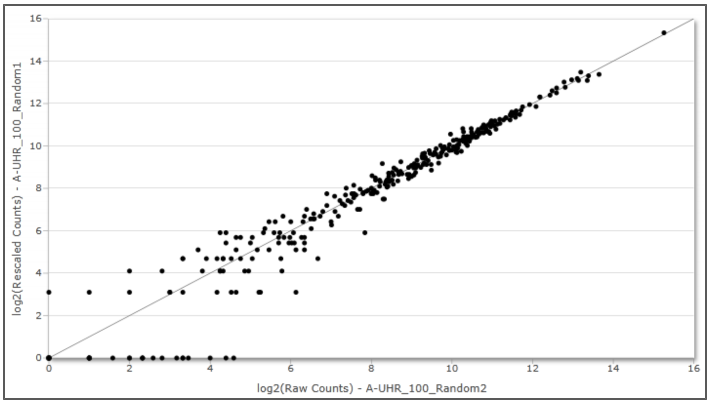


The comparison graph is a scatter plot that compares the relative abundance of each RNA transcript between two selected samples. Data points close to the diagonal line (y=x) denote transcripts with similar abundance, while points distant from this line represent transcripts expressed at different levels.

If you click on an individual point on the graph, the corresponding transcript in the comparison table is highlighted. Points that fall below the q-value threshold are highlighted in red. The **Q-value Threshold** slider can be adjusted for higher specificity or sensitivity.

Comparison Graph - View Individual Sample Table

Figure 11 Comparison Graph - View Individual Sample Table



Select the checkbox **View individual sample table** to compare counts for two replicates. No statistical tests are performed when comparing single replicates. However, the raw counts can be compared as a way to identify a mislabeled replicate.

Comparison Table

The comparison table provides a summary of the relative abundance of each transcript in the two selected samples. Raw and normalized counts are provided in the comparison table.

Column	Description
Gene	The name of the gene.

Column	Description
Amplicon ID	The unique identifier for the transcript or isoform from the manifest. The Amplicon ID is a combination of fields in the manifest, as follows: GeneName.Transcript.LeftExon.RightExon.AssayID.
Q-value	The p-value adjusted for false discovery rate (FDR) using the Benjamini-Hochberg method.
Mean Raw Counts 1	The mean of counts for sample 1 across replicates.
Mean Raw Counts 2	The mean of counts for sample 2 across replicates.
Mean Normalized Counts 1	The mean raw counts 1 after library size normalization.
Mean Normalized Counts 2	The mean raw counts 2 after library size normalization.
Fold Change	The ratio of mean normalized counts for sample 2 divided by mean normalized counts for sample 1 (mnc2/mnc1).
log2(Fold Change)	The $\log_2$(ratio of mnc2/mnc1).
P-value	The statistical significance of the differential expression.

Comparison Table - View Individual Sample Table

Select the checkbox **View individual sample table** to change the comparison table view.

Column	Description
Gene	The name of the gene.
Amplicon ID	The unique identifier for the transcript or isoform from the manifest. The Amplicon ID is a combination of fields in the manifest, as follows: GeneName.Transcript.LeftExon.RightExon.AssayID.
Raw Counts 1	The raw counts for sample 1 across replicates.
Raw Counts 2	The raw counts for sample 2 across replicates.
Rescaled Counts 2	The number of reads for sample 2, rescaled to the same overall coverage as sample 1.

Targeted RNA Analysis Files

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in Data\Intensities\BaseCalls\Alignment.
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.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one 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.
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.
TargetedRNARunStatistics.xml	Contains summary statistics specific to the run. Located at the root level of the run folder.
TargetedRNASeqGene-Expression.tsv	Contains the genes used for normalization and normalization results. Located in Data\Intensities\BaseCalls\Alignment.
TargetedRNASeqGene-Expression_M#.tsv	Contains sample correlation and differential expression results. Located in Data\Intensities\BaseCalls\Alignment.
TargetHitsPerSample_M#.tsv	Contains the raw aligned replicate counts for each transcript. Located in Data\Intensities\BaseCalls\Alignment.

Targeted RNA Analysis File Formats

MiSeq Reporter generates two tab-delimited file formats that are unique to the Targeted RNA workflow: TargetHitsPerSample_M#.tsv and (TargetedRNASeqGeneExpression.tsv.

Target Hits File Format

The target hits file, TargetHitsPerSample_M#.tsv, is a tab-delimited file that contains the raw aligned replicate counts for each transcript. One output file is created for each manifest using the file naming format of *_M1.tsv, *_M2.tsv, *_M3.tsv, etc.

The header of the target hits file contains one row of column headings including sample IDs, followed by a row of sample names. For differential analysis, there must be at least two distinct sample names in the sample sheet.

The target hits file contains the following fields.

Column Heading	Description
Gene Name	The name of the gene.
Amplicon ID	The amplicon identifier constructed from the gene name, transcript ID, left exon, right exon, and assay ID.
Assay ID	The unique identifier for the probe set.
Sample ID	Aligned count for all transcripts for this sample. There is one column for each sample using this manifest.

Gene Expression File Format

The gene expression file, TargetedRNASeqGeneExpression.tsv, is a tab-delimited text file that is split into two main sections: Sample Correlation and Differential Expression. This is the final result of the Targeted RNA workflow.

Sample Correlation Section

The Sample Correlation section contains the following fields.

Column Heading	Description
Sample Name 1	The first sample (one or more replicates) being compared.
Sample Name 2	The second sample (one or more replicates) being compared.
R^2	The square of the correlation coefficient.
R^2 (ignore zero counts)	The square of the correlation coefficient, except that all transcripts with zero expression in one or both samples are ignored.

Differential Expression Section

The Differential Expression section consists of one table for each pair of samples. Each table has the following fields.

Column Heading	Description
Gene	The name of the gene.
Amplicon ID	The amplicon identifier constructed from the gene name, transcript ID, left exon, right exon, and assay ID.
Assay ID	The unique identifier for the probe set.
Sample Name 1	The first sample being compared, which can be one or more replicates.
Sample Name 2	The second sample being compared, which can be one or more replicates.
Mean Raw Counts 1	The mean of counts for sample 1 across replicates.
Mean Raw Counts 2	The mean of counts for sample 2 across replicates.
Mean Normalized Counts 1	The mean raw counts 1 after library size normalization.
Mean Normalized Counts 2	The mean raw counts 2 after library size normalization.
Fold Change	The ratio of mean normalized counts for sample 2 divided by mean normalized counts for sample 1 (mnc2/mnc1).
log2(Fold Change)	The $\log_2$(ratio of mnc2/mnc1).
P-value	The statistical significance of the differential expression.
Q-value	The p-value adjusted for false discovery rate (FDR) using the Benjamini-Hochberg method.

Targeted RNA Manifest File Format

A manifest file is required input files for the Targeted RNA workflow. The manifest is provided by Illumina with your Targeted Oligo Pool (TOP) and uses a ***.txt** file format. The manifest name for each sample is specified in the Data section of the sample sheet.

The Targeted RNA manifest is a tab-delimited file that contains a header section followed by two blocks of rows beginning with column headings, which are titled the Probes section and the Targets section.

MiSeq Reporter ignores the Header section and uses only the following fields in the Probes and Targets sections.

- **Probes**—The following fields for this block are required:
  - **Gene Name**—The gene name.
  - **Transcript ID**—The identifier for the transcript isoform.
  - **Assay ID**—A unique identifier for the probe set.
  - **Left Exon**—The zero-based index of the 5' exon at the splice junction.
  - **Right Exon**—The zero-based index of the 3' exon at the splice junction.
  - **ULSO Sequence**—Sequence of the upstream primer, or Upstream Locus-Specific Oligo.
  - **DLSO Sequence**—Sequence of the downstream primer, or Downstream Locus-Specific Oligo. The reverse complement of this sequence forms the start of the first read. This sequence comes from the same strand as the ULSO sequence.
- **Targets**—The following fields for this section are required:
  - **Assay ID**—A unique identifier for the probe set. This field matches Assay ID in the Probes section.
  - **Target Sequence**—Full sequence of the amplicon body, including probes.

NOTE
Because the Target Sequence specifies the full amplicon sequence, there is no need to specify a genome in the sample sheet.

Gene Name and Assay ID are carried through to the end of the workflow along with the **Amplicon ID**, which is constructed from a combination of fields that ensure uniqueness, as follows:

{Gene Name}.{Transcript ID}.{Left Exon}.{Right Exon}.{Assay ID}

The Amplicon ID is stored in the gene expression file. For more information, see *Gene Expression File Format* on page 100.

Appendix A

Folders, File Formats, and Settings

MiSeqAnalysis Folder

The MiSeqAnalysis folder is the main run folder for MiSeq Reporter. The relationship between the MiSeqOutput and MiSeqAnalysis run folders can be summarized as follows:

- During sequencing, RTA populates the MiSeqOutput folder with files generated during primary analysis.
- With the exception of focus images and thumbnail images, RTA copies files to the MiSeqAnalysis folder in real time. When primary analysis is complete, RTA writes the file RTAComplete.xml to both run folders.
- MiSeq Reporter monitors the MiSeqAnalysis folder and begins secondary analysis when the file RTAComplete.xml appears.
- As secondary analysis progresses, MiSeq Reporter writes analysis output files generated during secondary analysis to the MiSeqAnalysis folder and then copies the output files to the MiSeqOutput folder.

Analysis Files Common to All Workflows

MiSeq Reporter generates analysis results in output files written to the MiSeqAnalysis folder. Some files are common to all workflows, such as marker files and log files. Other files are specific to workflows. See the associated analysis workflow chapter for a description of workflow-specific files.

File Name	Description
QueuedForAnalysis.txt	Marker file that lists the MiSeq Reporter software version and indicates that analysis has begun. Located at the root level of the run folder.
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 at 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 at 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 at the root level of the run folder.

Folder Structure

- 📁 **Data**
  - 📁 **Intensities**
    - 📁 **Basecalls**
      - 📁 **Alignment**—Contains *.bam and *.vcf files, if applicable.
      - 📁 **L001**—Contains one subfolder per cycle, each containing *.bcl files.
      - 📄 Sample1_S1_L001_R1_001.fastq.gz
      - 📄 Sample2_S2_L001_R1_001.fastq.gz
      - 📄 Undetermined_S0_L001_R1_001.fastq.gz
    - 📁 **L001**—Contains *.locs files, one for each tile.
  - 📁 **RTA Logs**—Contains log files from primary analysis.
- 📁 **InterOp**—Contains binary files used by Sequencing Analysis Viewer (SAV).
- 📁 **Logs**—Contains log files describing steps performed during sequencing.
- 📁 **Queued**—A working folder for MiSeq Reporter; also called the copy folder.
- 📄 AnalysisError.txt
- 📄 AnalysisLog.txt
- 📄 CompletedJobInfo.xml
- 📄 QueuedForAnalysis.txt
- 📄 [Workflow]RunStatistics
- 📄 RTAComplete.xml
- 📄 RunInfo.xml
- 📄 runParameters.xml
- 📄 SampleSheet.csv

For descriptions of files located at the root level of the analysis folder, see *Required Input Files* on page 15 and *Analysis Files Common to All Workflows* on page 104.

When using BaseSpace for secondary analysis without replicating analysis locally, the local MiSeqAnalysis folder is empty.

Alignment Folder Contents

Most secondary analysis files are written to the Alignment folder. If analysis is requeued, analysis results are written to additional Alignment folders named **AlignmentN**, where N indicates the number that analysis was requeued. The contents of the Alignment folder depend on the analysis workflow performed.

Log files from any sub-programs used in secondary analysis, such as BWA or GATK, are written to Data\BaseCalls\Alignment\Logging.

Analysis File Formats

Analysis results are written to file formats specific to their function and purpose.

Analysis Step	Format	Purpose
Demultiplexing	*.demux	Intermediate files containing demultiplexing results.
FASTQ	*.fastq.gz	Intermediate files containing quality scored base calls. FASTQ files are the primary input for the alignment step.
Alignment	*.bam	Compressed binary files containing sequence alignment data. BAM files are the primary input for the variant calling step.
Variant Calling	*.vcf	Text files containing SNPs, indels, and other structural variants.

Other file formats used in analysis results are *.txt, *.xml, *.htm, and *.png. Many of these files contain information that appears in tables, graphs, and charts on the MiSeq Reporter web interface.

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 one demultiplexing (*.demux) file for each tile of the flow cell.

Demultiplexing files are in a binary format and written to the L001 folder in Data\Intensities\BaseCalls\L001 and use the file naming format of 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.

FASTQ File Format

FASTQ file is a text-based file format that contains base calls and quality values per read. Each record is represented by four 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 Naming

FASTQ files are named with the sample name and the sample number, which 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, which is a required field in the sample sheet and must be unique.
- **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, there will be at least one file with R2 in the file name for Read 2.
- **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).

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 on the SAM Tools web site: http://samtools.sourceforge.net.

BAM files are written to the alignment folder in Data\Intensities\BaseCalls\Alignment in 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 and sample length. Alignments in the alignments section are associated with specific information in the header section.
- **Alignments**—Contains read name, read sequence, read quality, 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 includes the chromosome and start coordinate **chr22 17552189**, the alignment quality **8**, and the match descriptor **35M * 0 0**.

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

BAM index files (*.bam.bai) are the index file for the corresponding BAM file. The BAM index file is used by BAM tools to more quickly search through the BAM file.

VCF File Format

VCF is a widely used file format developed by the genomics scientific community, and contains information about variants found at specific positions in a reference genome.

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

VCF File Header—Includes the file format version, file date, name and version of the variant caller, the VCF annotations used in the remainder of the file, and column headings for the data lines. For descriptions, see *VCF File Annotations* on page 109.

```
##fileformat=VCFv4.1
##FORMAT=<ID=GQX,Number=1,Type=Integer>
##FORMAT=<ID=AD,Number=.,Type=Integer>
##FORMAT=<ID=DP,Number=1,Type=Integer>
##FORMAT=<ID=GQ,Number=1,Type=Float>
##FORMAT=<ID=GT,Number=1,Type=String>
##FORMAT=<ID=PL,Number=G,Type=Integer>
##FORMAT=<ID=VF,Number=1,Type=Float>
##INFO=<ID=TI,Number=.,Type=String>
##INFO=<ID=GI,Number=.,Type=String>
##INFO=<ID=EXON,Number=0,Type=Flag>
##INFO=<ID=FC,Number=.,Type=String>
##INFO=<ID=IndelRepeatLength,Number=1,Type=Integer>
##INFO=<ID=AC,Number=A,Type=Integer>
##INFO=<ID=AF,Number=A,Type=Float>
##INFO=<ID=AN,Number=1,Type=Integer>
##INFO=<ID=DP,Number=1,Type=Integer>
##INFO=<ID=QD,Number=1,Type=Float>
##FILTER=<ID=LowQual>
##FILTER=<ID=R8>
##reference=file://d:\Genomes\Homo_
   sapiens\UCSC\hg19\Sequence\WholeGenomeFASTA\genome.fa
##source=GATK 1.6
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 10002 - R1
```

VCF File Data Lines—Contains information about a single variant. Information in data lines are listed under the column headings included in the header.

VCF File Headings

The VCF file format is flexible and extensible, so not all VCF files contain exactly the same fields. The following tables describe VCF files generated by MiSeq Reporter.

Heading	Description
CHROM	The chromosome of the reference genome. Chromosomes appear in the same order as the reference FASTA file.
POS	The single-base position of the variant in the reference chromosome. For SNPs, this is the reference base with the variant; for indels or deletions, this is the reference base immediately before the variant. Variants are ordered by position.

Heading	Description
ID	The rs number for the SNP obtained from dbSNP.txt, if applicable. If there are multiple rs numbers at this location, the list is semi-colon delimited. If no dbSNP entry exists at this position, a missing value marker ('.') is used.
REF	The reference genotype. For example, a deletion of a single T might be represented by reference TT and alternate T.
ALT	The alleles that differ from the reference read. For example, an insertion of a single T could be represented by reference A and alternate AT.
QUAL	A Phred-scaled quality score assigned by the variant caller. Higher scores indicate higher confidence in the variant and lower probability of errors. For a quality score of Q, the estimated probability of an error is 10-$^{(Q/10)}$). For example, the set of Q30 calls should have a 0.1% error rate. Many variant callers assign quality scores based on their statistical models, which are high relative to the error rate observed.

VCF File Annotations

Heading	Description
FILTER	If all filters are passed, **PASS** is written in the filter column. • **R8**—For an indel, the number of adjacent repeats (1-base or 2-base) in the reference is greater than 8. This filter is configurable using the **IndelRepeatFilterCutoff** setting in the config file or the sample sheet. • **SB**—The strand bias is more than the given threshold. This filter is configurable using the **StrandBiasFilter** setting in the sample sheet; available only for somatic variant caller and GATK. • **LowDP**—Applied to sites with depth of coverage below a cutoff. Configure cutoff using **MinimumCoverageDepth** setting in the sample sheet. • **LowGQ**—The variant score (GQ) is below a cutoff. Configure cutoff using **VariantMinimumGQCutoff** in the sample sheet. • **LowQual**—The variant score (QUAL) is below a cutoff. Configure using the **VariantMinimumQualCutoff** setting in the sample sheet. • **LowVariantFreq**—The variant frequency is less than the given threshold. Configure using the **VariantFrequencyFilterCutoff** setting in the sample sheet. For more information, see *Sample Sheet Settings for Variant Calling* on page 112.

Heading	Description
INFO	Possible entries in the INFO column include: • **AC**—Allele count in genotypes for each ALT allele, in the same order as listed. • **AF**—Allele Frequency for each ALT allele, in the same order as listed. • **AN**—The total number of alleles in called genotypes. • **CD**—A flag indicating that the SNP occurs within the coding region of at least one refGene entry. • **DP**—The depth (number of base calls aligned to a this position and used in variant calling). In regions of high coverage, GATK down-samples the available reads. • **Exon**—A comma separated list of exon regions read from refGene. • **FC**—Functional Consequence. • **GI**—A comma separated list of gene IDs read from refGene. • **QD**—Variant Confidence/Quality by Depth. • **TI**—A comma separated list of transcript IDs read from refGene.
FORMAT	The format column lists fields separated by colons. For example, GT:GQ. The list of fields provided depends on the variant caller used. Available fields include: • **AD**—Entry of the form X,Y, where X is the number of reference calls, and Y is the number of alternate calls. • **DP**—Approximate read depth; reads with MQ=255 or with bad mates are filtered. • **GQ**—Genotype quality. • **GQX**—Genotype quality. GQX is the minimum of the GQ value and the QUAL column. In general, these are similar values; taking the minimum makes GQX the more conservative measure of genotype quality. • **GT**—Genotype. 0 corresponds to the reference base, 1 corresponds to the first entry in the ALT column, 2 corresponds to the second entry in the ALT column, etc. The forward slash (/) indicates that no phasing information is available. • **NL**—Noise level; an estimate of base calling noise at this position. • **PL**—Normalized, Phred-scaled likelihoods for genotypes. • **SB**—Strand bias at this position. Larger negative values indicate more bias; values near zero indicate little bias. • **VF**—Variant frequency; the percentage of reads supporting the alternate allele.
SAMPLE	The sample column gives the values specified in the FORMAT column.

Sample Sheet Settings

You can specify settings in the Settings section of the sample sheet to control various analysis parameters. See the *MiSeq Sample Sheet Quick Reference Guide*, Part # 15028392 for settings specific to sequencing.

Each line in the Settings section contains a setting name in the first column and a value in the second column. Settings are not case sensitive.

Parameter	Description
Adapter	Specify the 5' portion of the adapter sequence to prevent reporting sequence beyond the sample DNA. For more information, see *Adapter Settings* on page 115. • **Nextera libraries**—Illumina recommends adapter trimming for Nextera libraries (adapter sequence **CTGTCTCTTATACACATCT**). • **Small RNA libraries**—By default, adapter trimming is performed using the standard adapter sequence **TGGAATTCTCGGGTGCCAAGGC**. This adapter can be overridden using the sample sheet. For other libraries, see the associated sample prep documentation.
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. For more information, see *Read 1 and Read 2 Adapters* on page 115.
Aligner	**For Resequencing and Library QC workflows**— As of MiSeq Reporter v2.2, Eland has been deprecated, but not removed. You can use the Aligner setting to specify Eland, if needed for backward-compatibility. When using the default aligner for any workflow, you do not need to specify the alignment method in the sample sheet.
CustomAmpliconAlignerMaxIndelSize	**For Custom Amplicon workflow**—The maximum detectable indel size using the Custom Amplicon workflow. The default is 25. Setting this to a larger value increases the sensitivity to larger indels, but requires more time to complete alignment.
EnrichmentMaxRegionStatisticsCount	**For the Enrichment workflow**—Default is 40000. Sets the maximum number of rows shown in the Targets table and recorded EnrichmentStatistics.xml.

Parameter	Description
ExcludeRegionsManifestA	**For the Enrichment workflow**—Use this setting to exclude one of more region groups (separated by plus signs) from consideration. For example, if this is set to ABC+DEF, any region that has either ABC or DEF specified in the **Group** column of the manifest will be ignored when parsing the manifest. This means that no variant calling will be performed for this region and it will not be reported in enrichment statistics. If the sample sheet contains more than one manifest, use multiple lines, such as ExcludeRegionsManifestB, ExcludeRegionsManifestC, etc.
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 two clusters from a paired-end run where both clusters have the exact same alignment positions for each read. (Formerly FilterPCRDuplicates. FilterPCRDuplicates is acceptable for backward compatibility.)
Kmer	**For the Assembly workflow**—Use this setting to override the k-mer size used by Velvet. Default is 31; odd-numbered values up to 255 are supported.
QualityScoreTrim	If set to a value > 0, then the 3' ends of non-indexed reads with low quality scores are trimmed. Trimming is automatically applied by default at a value of 15 when using BWA for alignment.
TaxonomyFile	**For the Metagenomics workflow**—Use this setting to override the taxonomy database (default is taxonomy.dat).
VariantCaller	**For the Custom Amplicon and Resequencing workflows**—Specify one of the following variant caller settings: • GATK (default) • Somatic (for tumor samples) • Starling (legacy variant caller) • None (no variant calling) When using the default variant caller for any workflow, you do not need to specify the variant calling method in the sample sheet.

Sample Sheet Settings for Variant Calling

Some sample sheet settings specify parameters for variant calling. Most settings and default values are specific to a variant caller.

Setting Name	Description
FilterOutSingleStrandVariants	This setting filters variants if they are only found in one read-direction. This setting applies to the Resequencing and PCR Amplicon workflows only; it does not apply to the Custom Amplicon workflow. **Default value and variant caller:** • 1 (on)—Somatic Variant Caller (0 for Custom Amplicon workflow)
IndelRepeatFilterCutoff	This setting filters indels if the reference has a 1-base or 2-base motif over 8 times (by default) next to the variant. **Default value and variant caller:** • 8—Somatic Variant Caller • 8—GATK • 8—Starling
MinimumCoverageDepth	The variant caller filters variants if the coverage depth at that location is less than the specified threshold. Decreasing this value will increase variant calling sensitivity, but raise the risk of false positives. **Default value and variant caller:** • 20—GATK (Enrichment workflow only; 0 for any other workflow)
MinQScore	This setting specifies the minimum base call Q-score to use as input to variant calling. **Default value and variant caller:** • 20—Somatic Variant Caller • 0—Starling
StrandBiasFilter	This setting filters variants with a significant bias in read-direction. Variants filtered out in this way will have **sb** in the filter column of the VCF file, instead of **PASS**. **Default value and variant caller:** • 0.5—Somatic Variant Caller • -10—GATK (Enrichment workflow only; no filter for any other workflow)
VariantFrequencyEmitCutoff	This variant caller does not report variants with a frequency less than the specified threshold. **Default value and variant caller:** • 0.01—Somatic Variant Caller
VariantFrequencyFilterCutoff	This setting filters variants with a frequency less than the specified threshold. **Default value and variant caller:** • 0.01—Somatic Variant Caller • 0.20—GATK • 0.20—Starling

Setting Name	Description
VariantMinimumGQCutoff	This setting filters variants if the genotype quality (GQ) is less than the threshold. GQ is a measure of the quality of the genotype call and has a maximum value of 99. (VariantFilterQualityCutoff is acceptable for backward compatibility.) **Default value and variant caller:** • 30—Somatic Variant Caller • 30—GATK • 20—Starling
VariantMinimumQualCutoff	This setting filters variants if the quality (QUAL) is less than the threshold. QUAL indicates the confidence that the variant is genuine. **Default value and variant caller:** • 30—Somatic Variant Caller • 30—GATK • 20—Starling

Adapter Settings

It is possible that some clusters will sequence beyond the sample DNA and read bases from a sequencing adapter. This particularly applies to longer read lengths up to 250 cycles.

Using the **Adapter** setting in the Settings section of the sample sheet, you can specify the 5' portion of the adapter sequence to prevent reporting sequence beyond the sample DNA in FASTQ files. Trimming the adapter sequence avoids reporting spurious mismatches with the reference sequence, and improves performance in accuracy and speed of alignment.

The Adapter setting performs adapter trimming or adapter masking depending on the aligner used for the workflow. For workflows using BWA, reads are trimmed from the start of the adapter sequence. When using Eland (deprecated in v2.2), reads are N-masked, or replaced with Ns (no-call), from the start of the adapter.

Read 1 and Read 2 Adapters

If you specify an adapter sequence using the **Adapter** setting in the sample sheet, the same adapter is trimmed for Read 1 and Read 2. If you want to trim a different adapter sequence in Read 2, use the **AdapterRead2** setting in the sample sheet and specify the adapter sequence in the same way that you would using the **Adapter** setting.

To trim two or more adapters, separate the sequences by a plus (+) sign.

How Adapter Trimming Works

MiSeq Reporter considers each potential adapter start position (n) within the sequence starting at the first base (n=0). The process continues to count matches and mismatches between sequence (n) and adapter (0), sequence (n + 1), and adapter (1), and so on. This loop terminates if the following occurs:

```
MismatchCount > 1 and MismatchCount > MatchCount
```

Otherwise, the count continues until the end of the sequence or end of the adapter is reached, whichever comes first. The sequence is trimmed starting at position n, if:

```
MatchCount / (MatchCount + MismatchCount) > Cutoff
```

By default, the cutoff is 0.9 or < 10% mismatch rate. This default setting can be modified using the configurable setting **AdapterTrimmingStringency**.

Masking Short Reads

MiSeq Reporter includes a setting that prevents reads that have been almost entirely trimmed or masked from confounding downstream analysis, which is based on the following criteria:

- If the adapter is encountered within the first 32 bases of the read, the adapter sequence is N-masked.
- If the adapter is identified in the first 32 bases ***and*** the read includes ten or more bases from the start of the adapter, the entire read is N-masked.

This ten-base limit is controlled by the configuration setting **NMaskShortAdapterReads**.

MiSeq Reporter Configurable Settings

Typically, you do not need to change configurable settings. However, if you want to customize analysis results, you can edit settings in MiSeq Reporter.exe.config located in the MiSeq Reporter installation folder, C:\Illumina\MiSeqReporter, by default. Always restart the service after modifying the config file.

The editable portion of this file is contained between the <appSettings> tags, which shows key/value pairs for the parameter settings applied.

```
<appSettings>
  <add key="Repository" value="D:\Illumina\MiSeqAnalysis" />
  <add key="GenomePath" value="C:\Illumina\MiSeqReporter\Genomes" />
  <add key="TempFolder" value="D:\Illumina\MiSeqAnalysis\Temp" />
  <add key="EnableHTTPService" value="1"/>
  <add key="ClientSettingsProvider.ServiceUri" value="" />
  <add key="CopyToRTAOutputPath" value="1"/>
  <add key="DemuxMaxSequencesToReport" value="100"/>
  <add key="MaximumHoursPerProcess" value="24"/>
</appSettings>
```

Available Configurable Settings

The following configurable settings are used in MiSeq Reporter.exe.config.

Setting Name	Values and Description
AdapterTrimmingStringency	0.9 (default) The minimum match rate allowed in adapter trimming. The default value sets adapter trimming to only trim sequences with > 90% sequence identity with the adapter.
ConvertMissingBclsToNoCalls	1 (true; default) 0 (false) If set to true, any missing or invalid *.bcl files cause MiSeq Reporter to log an error and flag the tile as having no-calls (Ns) for the affected cycle. If set to false, any missing or truncated *.bcl files cause MiSeq Reporter to log an error and abort analysis.
CopyToRTAOutputPath	1 (true; default) 0 (false) If set to true, copy all alignment data to the <OutputDirectory> specified in the RTAConfiguration.xml file, which is located in Data\Intensities.
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.

Setting Name	Values and Description
EnableHTTPService	1 (true; default) 0 (false) Determines whether MiSeq Reporter provides the web interface.
FilterNonPFReads	1 (true; default) 0 (false) Determines whether those clusters that fail the chastity filter are filtered from all FASTQ files.
GATKDownsampleDepth	5000 (default) When using GATK for variant calling, reads in regions of high depth are (optionally) randomly down-sampled. • Set to a higher value to retain more reads. • Set to 0 to disable down-sampling. *This might lead to increased runtime and memory use on high-coverage runs.*
IndelRepeatFilterCutoff	8 (default) By default, indels are flagged as filtered if the reference has a 1- or 2-base motif repeated eight or more times next to the variant. This threshold (8) can be adjusted with this setting.
MaximumGigabytesPerProcess	Varies The maximum gigabytes of memory allowed for a child process. By default, this threshold is adjusted automatically based on the memory available on the system.
MaximumHoursPerProcess	1.5 (default) The maximum number of hours to allow a child process to run.
MaximumMegabasesAssembly	550 (default) The maximum number of megabases to assemble. Larger values require more RAM.
MinimumAlignReadLength	21 (max; default) 8 (min) The minimum length of a non-indexed read to align using BWA or Eland (deprecated in v2.2).
NMaskShortAdapterReads	10-base (default) The number of bases from the start of the adapter that triggers N-masking of the entire read.
RetainTempFiles	0 (false; default) 1 (true) If set to true, temporary files are retained. This requires large amounts of disk space and intended for use in troubleshooting only.
VariantFilterQualityCutoff	30 (default) for GATK and Somatic Variant Caller 20 (default) for Starling SNPs with variant quality scores below this threshold are flagged as filtered in the *.vcf files.

Restarting the Service

After updating MiSeq Reporter.exe.config, you must restart the service, as follows:

1. From the Control Panel, select **Administrative Tools** | **Services**.
2. Select **MiSeq Reporter service**, and then click the **Restart Service** icon ⏯.

Appendix B

Installation and Troubleshooting

MiSeq Reporter Off-Instrument Requirements

Installing another copy of the MiSeq Reporter software on a Windows computer other than the MiSeq computer allows secondary analysis of sequencing data while the MiSeq performs a subsequent sequencing run.

For more information, see *Installing MiSeq Reporter Off-Instrument* on page 121.

Computing Requirements

MiSeq Reporter software requires the following computing components:

- 64-bit Windows OS (Vista, Windows 7, Windows Server 2008 64-bit)
- ≥ 8 GB RAM minimum; ≥ 16 GB RAM recommended
- ≥ 1 TB disk space
- Quad core processor (2.8 Ghz or higher)
- Microsoft .NET 4

Supported Browsers

MiSeq Reporter can be viewed with the following web browsers:

- Safari 5.1.7 or later
- Chrome 20.0 or later
- Firefox 13.0.1 or later
- Internet Explorer 8 or later

Downloading and Licensing

1. Download a second copy of the MiSeq Reporter software from the Illumina website. A MyIllumina login is required.
2. Accept the end-user licensing agreement (EULA) when prompted during installation. No license key is required as this additional copy is free of charge.

Installing MiSeq Reporter Off-Instrument

To install MiSeq Reporter on a different Windows computer other than the MiSeq instrument computer, first set up **Log on as a service** permissions, and then run the installation wizard. After installation is complete, configure the software to point to the appropriate Repository and GenomePath.

Uninstall Previous Versions of MiSeq Reporter

If version 1.0.27 or earlier of MiSeq Reporter is installed on the computer, first uninstall it before running the installation wizard, as described here.

> NOTE
> If version 1.0.28 or later is installed or if you have never installed MiSeq Reporter on this computer, skip to the next section, *Set Up User or Group Accounts on Windows 7*.

1. [Optional] Save a copy of the folder where the FASTA files for the reference genomes are stored.
2. From the Windows **Start** menu, select **Control Panel**, and then click **Programs**.
3. Click **Programs and Features**.
4. Right-click **MiSeq Reporter**, and then click **Uninstall**.
5. Click **OK** through any prompts.

Set Up User or Group Accounts on Windows 7

To configure user or group accounts to enable **Log on as a service** permissions, you must administrator rights to the computer. If you do have administrator rights or need assistance setting up a user or group account, contact your local facility administrator.

1. From the Windows **Start** menu, select **Control Panel**, and then click **System and Security**.
2. Click **Administrative Tools**, and then double-click **Local Security Policy**.
3. From the Security Settings tree on the left, double-click **Local Policies** and then click **User Rights Assignments**.
4. In the details pane on the right, double-click **Log on as a service**.
5. In the Properties dialog box, click **Add User or Group**.
6. Type the name of the user or group account that will run MiSeq Reporter on this computer, and then click **Check Names** to validate the account.
7. Click **OK** through any open dialog boxes and then close the control panel.

For more information, see http://technet.microsoft.com/en-us/library/cc739424(WS.10).aspx on the Microsoft website.

Run the MiSeq Reporter Installation Wizard

1. Download and unzip the MiSeq Reporter installation package from the Illumina website.
2. Double-click the setup.exe file.
3. Click **Next** through the prompts in the installation wizard.

4 When prompted for a user name and password, specify the user name and password for an account with **Log on as a service** permissions, as set up in the previous step.

5 Continue through any remaining prompts.

Configure MiSeq Reporter

To configure MiSeq Reporter to locate the run folder and reference genome folder, edit the configuration file in a text editor, such as Notepad.

1 Navigate to the installation folder (C:\Illumina\MiSeq Reporter, by default) and open the file MiSeq Reporter.exe.config in a text editor.

2 Locate the **Repository** tag and change the **value** to the default data location on the off-instrument computer.

```
<add key="Repository" value="E:\Data\Repository" />
```

Alternatively, this location can be a network location accessible from the off-instrument computer.

3 Locate the **GenomePath** tag and change the **value** to the location of the folder containing reference genomes files in FASTA format.

```
<add key="GenomePath" value="E:\MyGenomes\FASTA" />
```

Start the MiSeq Reporter Service

After completing the installation, the MiSeq Reporter service should start automatically. If the service does not start, start it manually as described below or by rebooting the computer.

1 From the Windows **Start** menu, right-click **Computer** and select **Manage**.

2 From the Computer Management tree on the left, double-click **Services and Applications** and then click **Services**.

3 Right-click **MiSeq Reporter** and select **Properties**.

4 On the General tab, make sure the **Startup Type** is set to **Automatic**, and then click **Start**.

5 On the Log On tab, set the **user name** and **password** for a Services account that has permissions to write to the server. Illumina recommends the **Local System** account for most users. For assistance or site-specific network requirements, contact the local facility administrator.

6 Click **OK** through any open dialog boxes and then close the Computer Management window.

7 After starting the MiSeq Reporter service, connect to the software locally using http://localhost:8042 in a web browser.

Using MiSeq Reporter Off-Instrument

To use MiSeq Reporter off-instrument you first need to make sure that folders containing run data and reference genomes are accessible. You can do this by specifying a network location or a location on your local computer.

1 If you are not using a network location for sequencing data and reference genomes, copy the following folders to your local computer:
   - Copy run data from the MiSeq computer in D:\MiSeqOutput\<RunFolder>.
   - Copy reference genomes from the MiSeq computer in C:\Illumina\MiSeq Reporter\Genomes.

2 Open a web browser to http://localhost:8042, which opens the MiSeq Reporter web interface.

3 If the location of the run data differs from the location specified in MiSeq Reporter.exe.config, you can change the path using the **Settings** ⚙ icon in the top-right corner of the web interface.

> NOTE
> Specifying the repository path in Settings is temporary. The next time you restart your computer, the path defaults to the Repository location specified in MiSeq Reporter.exe.config.

4 Select **Analyses** on the left-side of the web interface to view the runs available in the specified Repository location.

5 Before you requeue analysis of a run using an off-instrument installation of MiSeq Reporter, you must update the path of the GenomeFolder in the sample sheet to the new location. You can do this from the Sample Sheet tab. After updating the GenomeFolder path, click **Save and Requeue**.
For more information, see *Editing the Sample Sheet in MiSeq Reporter* on page 12.

Troubleshooting MiSeq Reporter

MiSeq Reporter runs as Windows service application. User accounts must be configured to enable **Log on as a service** permissions before installing MiSeq Reporter. For more information, see *Set Up User or Group Accounts on Windows 7* on page 121.

For more information, see http://msdn.microsoft.com/en-us/library/ms189964.aspx.

Service Fails to Start

If the configuration file, MiSeq Reporter.exe.config, has an invalid file format, the MiSeq Reporter service will not start. To confirm that this is the cause, check the Window Event Log and view the details of the error message.

1 Open the **Control Panel** and select **Administrative Tools**.

2 Select **Event Viewer**.

3 In the Event Viewer window, select **Windows Logs | Application**. The error listed in the event log should list the syntax error in MiSeq Reporter.exe.config.

Files Failed to Copy

If files fail to copy to the intended location, check the following settings:

1 Check the path to the specified repository folder or MiSeqOutput folder:
   - If you are using MiSeq Reporter off-instrument, check the repository location using Settings ⚙ on the MiSeq Reporter web interface.
   - If you are using MiSeq Reporter on-instrument, check the MiSeqOutput folder location on the MCS Run Options screen, Folder Settings tab.

   You must use the full UNC path (e.g., \\server1\Runs). Because MiSeq Reporter runs as a Windows service, it does not recognize user-mapped drives (e.g., Z:\Runs).

2 Confirm that you have write-access to the output folder location. If you need assistance, contact your facility administrator.

3 Check that copying is not disabled in the MiSeq Reporter.exe.config. This setting is located in the <appSettings> section and the value should be set to **1**.

```
<add key="CopyToRTAOutputPath" value="1"/>
```

Viewing Log Files for a Failed Run

Viewing logs files can help identify specific errors for troubleshooting purposes.

1 To view the log files using the MiSeq Reporter web browser interface, select the run in the Analyses tab.

2 Select the Logs tab to view a list of every step that occurred during analysis. Log information is recorded in AnalysisLog.txt, which is located in the root level of the MiSeqAnalysis folder.

3 Select the Errors tab to view a list of errors that occurred during analysis. Error information is recorded in AnalysisError.txt, which is located in the root level of the MiSeqAnalysis folder.

Index

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 10 Illumina General Contact Information

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

Table 11 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Italy	800.874909
Austria	0800.296575	Netherlands	0800.0223859
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

MSDSs

Material safety data sheets (MSDSs) are available on the Illumina website at www.illumina.com/msds.

Product Documentation

Additional product documentation in PDF is available for download from the Illumina website. Go to www.illumina.com/support, 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