MiSeq Reporter Enrichment
Workflow Reference Guide

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

ILLUMINA PROPRIETARY

illumina®

Document # 15042315 v01
September 2015

This document and its contents are proprietary to Illumina, Inc. and its affiliates ("Illumina"), and are intended solely for the contractual use of its customer in connection with the use of the product(s) described herein and for no other purpose. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way whatsoever without the prior written consent of Illumina. Illumina does not convey any license under its patent, trademark, copyright, or common-law rights nor similar rights of any third parties by this document.

The instructions in this document must be strictly and explicitly followed by qualified and properly trained personnel in order to ensure the proper and safe use of the product(s) described herein. All of the contents of this document must be fully read and understood prior to using such product(s).

FAILURE TO COMPLETELY READ AND EXPLICITLY FOLLOW ALL OF THE INSTRUCTIONS CONTAINED HEREIN MAY RESULT IN DAMAGE TO THE PRODUCT(S), INJURY TO PERSONS, INCLUDING TO USERS OR OTHERS, AND DAMAGE TO OTHER PROPERTY.

ILLUMINA DOES NOT ASSUME ANY LIABILITY ARISING OUT OF THE IMPROPER USE OF THE PRODUCT(S) DESCRIBED HEREIN (INCLUDING PARTS THEREOF OR SOFTWARE).

© 2015 Illumina, Inc. All rights reserved.

Illumina, **24sure**, **BaseSpace**, **BeadArray**, **BlueFish**, **BlueFuse**, **BlueGnome**, **cBot**, **CSPro**, **CytoChip**, **DesignStudio**, **Epicentre**, **ForenSeq**, **Genetic Energy**, **GenomeStudio**, **GoldenGate**, **HiScan**, **HiSeq**, **HiSeq X**, **Infinium**, **iScan**, **iSelect**, **MiSeq**, **MiSeqDx**, **MiSeq FGx**, **NeoPrep**, **NextBio**, **Nextera**, **NextSeq**, **Powered by Illumina**, **SureMDA**, **TruGenome**, **TruSeq**, **TruSight**, **Understand Your Genome**, **UYG**, **VeraCode**, **verifi**, **VeriSeq**, the pumpkin orange color, and the streaming bases design are trademarks of Illumina, Inc. and/or its affiliate(s) in the U.S. and/or other countries. All other names, logos, and other trademarks are the property of their respective owners.

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

Document #	Date	Description of Change
Document # 15042315 v01	September 2015	Changed the name of the guide from the MiSeq Reporter Enrichment Workflow Reference Guide to the MiSeq Reporter Enrichment Workflow Guide. In the BAM File Format section, revised the description of the alignment information in the file header, and updated the link for SAM format specifications. Added RunBwaAln parameter for setting the alignment method from BWA-MEM to BWA-backtrack.
Part # 15042315 Rev. E	December 2014	Added a note in the Demultiplexing section about the default index recognition for index pairs that differ by < 3 bases.
Part # 15042315 Rev. D	September 2014	Updated the description of the variants table with directions when there are > 2000 variants.
Part # 15042315 Rev. C	February 2014	Updated to changes introduced in MiSeq Reporter v2.4: • Added alignment method to the description of the BAM file header. • Added the command line and annotation algorithm to the description of VCF file header. Updated sample sheet OutputGenomeVCF parameter default setting information.
Part # 15042315 Rev. B	August 2013	• Added description of sample sheet settings OutputGenomeVCF and ManifestPaddingSize introduced in MiSeq Reporter v2.3. • Updated description of the Enrichment Summary output file. • Added description of the genome VCF (gVCF) file format. • Added descriptions of sample sheet settings applicable to the Enrichment workflow: Adapter, BaitManifestFileName, ExcludeRegionsManifestA, FlagPCRDuplicates, PicardHSMetrics, VariantCaller, and VariantMinimumGQCutoff (also known as VariantFilterQualityCutoff). • Added Baits file to input requirements.
Part # 15042315 Rev. A	June 2013	Initial release. The information provided within was previously included in the *MiSeq Reporter User Guide*. With this release, the *MiSeq Reporter User Guide* contains information about the interface, how to view run results, how to requeue a run, and how to install and configure the software. Information specific to the Enrichment workflow is provided in this guide.

Introduction

The Enrichment workflow aligns reads against the whole genome reference and performs variant analysis for regions of interest specified in the manifest file.

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

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

Workflow Requirements

- **Manifest file**—The Enrichment workflow requires at least 1 manifest file. Manifest files are available for download from the Illumina website. The manifest file is a list of targeted regions and the chromosome start and end positions.
- **Baits file**—Related to the manifest file, the Enrichment workflow requires a baits file, which lists the baits regions and the chromosome start and end positions.
- **Reference genome**—The Enrichment workflow requires the reference genome that is specified in the manifest file. The reference genome provides the chromosome and start coordinate in the BAM file output. Specify the path to the genome folder in the sample sheet. For more information, see the *MiSeq Sample Sheet Quick Reference Guide (document # 15028392)*.

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.

The Enrichment workflow demultiplexes indexed reads, generates FASTQ files, aligns reads to a reference, identifies variants, and writes output files to the Alignment folder.

Demultiplexing

Demultiplexing separates data from pooled samples based on short index sequences that tag samples from different libraries. Index reads are identified using the following steps:

- Samples are numbered starting from 1 based on the order they are listed in the sample sheet.
- Sample number 0 is reserved for clusters that were not successfully assigned to a sample.
- Clusters are assigned to a sample when the index sequence matches exactly or there is up to a single mismatch per Index Read.

NOTE
Illumina indexes are designed so that any index pair differs by ≥ 3 bases, allowing for a single mismatch in index recognition. Index sets that are not from Illumina can include pairs of indexes that differ by < 3 bases. In such cases, the software detects the insufficient difference and modifies the default index recognition (mismatch=1). Instead, the software performs demultiplexing using only perfect index matches (mismatch=0).

When demultiplexing is complete, 1 demultiplexing file named DemultiplexSummaryF1L1.txt is written to the Alignment folder, and summarizes the following information:

- In the file name, **F1** represents the flow cell number.
- In the file name, **L1** represents the lane number, which is always L1 for MiSeq.
- Reports demultiplexing results in a table with 1 row per tile and 1 column per sample, including sample 0.
- Reports the most commonly occurring sequences for the index reads.

FASTQ File Generation

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

FASTQ files are the primary input for alignment. The files are written to the BaseCalls folder (Data\Intensities\BaseCalls) in the MiSeqAnalysis folder, and then copied to the BaseCalls folder in the MiSeqOutput folder. Each FASTQ file contains reads for only 1 sample, and the name of that sample is included in the FASTQ file name. For more information about FASTQ files, see the *MiSeq Reporter Software Guide (document # 15042295)*.

Alignment

Reads are aligned against the entire reference genome using 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.

Paired-End Evaluation

For paired-end runs, the top-scoring alignment for each read is considered. Reads are flagged as an unresolved pair under the following conditions:

- If either read did not align, or the paired reads aligned to different chromosomes.
- If 2 alignments come from different amplicons or different rows in the Targets section of the manifest.

Bin/Sort

The bin/sort step groups reads by sample and chromosome, and then sorts by chromosome position. Results are written to 1 BAM file per sample.

Indel Realignment

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 using the Genome Analysis Toolkit (GATK), by default. GATK calls raw variants for each sample, analyzes variants against known variants, and then calculates 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 www.broadinstitute.org/gatk.

Alternatively, you can specify the somatic variant caller using the VariantCaller sample sheet setting. For more information, see *Optional Settings for the Enrichment Workflow* on page 12.

Variant 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.

Enrichment Summary Tab

The Summary tab for the Enrichment workflow includes a low percentage graph, high percentage graph, a clusters graph, and a mismatch graph.

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

Low Percentages Graph

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

High Percentages Graph

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

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

The Details tab for the Enrichment workflow includes a samples table, targets table, coverage graph, Q-score graph, variant score graph, and variants table.

- **Samples table**—Summarizes the sequencing results for each sample.
- **Targets table**—Shows statistics for a particular sample and chromosome.
- **Coverage graph**—Shows read depth at a given position in the reference.
- **Q-score graph**—Shows the average quality score, which is the estimated probability of an error measured in $10^{-(Q/10)}$. For example, a score of Q30 has an error rate of 1 in 1000, or 0.1%.
- **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. The variants table shows up to 2000 variants for the selected sample and chromosome. If there are > 2000 variants, open the .vcf file of the sample to view the complete list.

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

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.
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.

Q-Score Graph

Y Axis	X Axis	Description
Q-Score	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 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 subsamples data in regions of high coverage. The GATK subsampling 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.

Optional Settings for the Enrichment Workflow

Sample sheet settings are optional commands that control various analysis parameters.

Settings are used in the Settings section of the sample sheet and require a setting name and a setting value.

If you are viewing or editing the sample sheet in Excel, the setting name resides in the first column and the setting value in the second column.

If you are viewing or editing the sample sheet in a text editor such as Notepad, follow the setting name is by a comma and a setting value. Do not include a space between the comma and the setting value.

Example: VariantCaller,Somatic

The following optional settings are compatible with the Enrichment workflow.

Sample Sheet Settings for Analysis

Parameter	Description
Adapter	Specify the 5' portion of the adapter sequence to prevent reporting sequence beyond the sample DNA. Illumina recommends adapter trimming for Nextera libraries and Nextera Mate Pair libraries. To specify 2 or more adapter sequences, separate the sequences by a plus (+) sign. For example: CTGTCTCTTATACACATCT+AGATGTGTATAAGAGACAG
AdapterRead2	Specify the 5' portion of the Read 2 adapter sequence to prevent reporting sequence beyond the sample DNA. Use this setting to specify a different adapter other than the one specified in the **Adapter** setting.
BaitManifestFileName	Specify the full path to the bait file. This setting is used only if the PicardHSmetrics setting is used and set to true.
EnrichmentMaxRegionStatisticsCount	Default is 40000. Sets the maximum number of rows shown in the Targets table and recorded EnrichmentStatistics.xml.
ExcludeRegionsManifestA	This setting excludes 1 or more region groups (separated by plus signs) from consideration. For example, if this setting specifies ABC+DEF, any region that has either ABC or DEF specified in the **Group** column of the manifest is ignored when parsing the manifest. No variant calling is performed for this region or reported in enrichment statistics. If the sample sheet contains more than 1 manifest, use multiple lines, such as ExcludeRegionsManifestB, ExcludeRegionsManifestC.

Parameter	Description
FlagPCRDuplicates	Settings are 0 or 1. Default is 1, filtering. If set to 1, PCR duplicates are flagged in the BAM files and not used for variant calling. PCR duplicates are defined as 2 clusters from a paired-end run where both clusters have the exact same alignment positions for each read. Duplicates are not flagged for single-read runs, including PCR duplicates. *(Formerly FilterPCRDuplicates. FilterPCRDuplicates is acceptable for backward compatibility.)*
ManifestPaddingSize	Settings are 0 or specified value. Default is 0. Specify the number of bases to extend the upstream and downstream ends of the targeted regions specified in the manifest. When padding is applied, read and base alignment statistics are reported for the targeted regions with and without the padding. Variants are reported for the targeted regions only. This setting requires MiSeq Reporter v2.3, or later.
OutputGenomeVCF	Settings are 0 or 1. Default is 1. If set to true (1), this setting turns on genome VCF (gVCF) output for single sample variant calling. If set to false (0), gVCF files are not generated. This setting requires MiSeq Reporter v2.3, or later.
PicardHSmetrics	Settings are 0 or 1. Default is 0. If set to true (1), this setting generates Picard HS metrics for the given bait and manifest file. If the bait file is not explicitly identified, the manifest file is used as the bait file. Use the BaitManifestFileName setting to specify the bait file.
RunBwaAln	Settings are 0 or 1. Default is 0, BWA-MEM alignment method. BWA-MEM is for ≥ 70 bp read lengths. If set to 1, BWA-backtrack is used for alignment. Formerly referred to as BWA, BWA-backtrack is an earlier version of the BWA. Use BWA-backtrack for < 70 bp read lengths, or when consistency is required with previous study data.
VariantCaller	Specify 1 of the following variant caller settings: • GATK (default) • Somatic (recommended for tumor samples) • None (no variant calling) When using the default variant caller for the workflow, it is not necessary to specify the variant calling method in the sample sheet.

Sample Sheet Settings for Variant Calling

Setting Name	Description
MinimumCoverageDepth	The variant caller filters variants if the coverage depth at that location is less than the specified threshold. Decreasing this value increases variant calling sensitivity, but raises the risk of false positives. **Default value:** • 20—GATK
StrandBiasFilter	This setting filters variants that have a significant bias in read-direction. Variants filtered in this way have **SB** in the filter column of the VCF file, instead of **PASS**. **Default value:** • -10—GATK • 0.5—Somatic variant caller
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. (*Formerly, VariantFilterQualityCutoff, which is acceptable for backward compatibility*.) **Default value:** • 30—GATK • 30—Somatic variant caller

Analysis Output Files

The following analysis output files are generated for the Enrichment workflow and provide analysis results for alignment, variant calling, and performance metrics.

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.
SampleName.enrichment_ Summary.csv	Contains a summary of performance metrics generated by the Enrichment workflow. Located in Data\Intensities\BaseCalls\Alignment.

Alignment Files

Alignment files contain the aligned read sequence and quality score. MiSeq Reporter generates alignment files in the BAM (*.bam) file format.

BAM File Format

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

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

BAM files contain a header section and an alignments section:

- **Header**—Contains information about the entire file, such as sample name, sample length, and alignment method. Alignments in the alignments section are associated with specific information in the header section.
  Alignment methods include banded Smith-Waterman, Burrows-Wheeler Aligner (BWA), and Bowtie. The term Isis indicates that an Illumina alignment method is in use, which is the banded Smith-Waterman method.
- **Alignments**—Contains read name, read sequence, read quality, alignment information, and custom tags.

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

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

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

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

Variant Call Files

Variant call files contain all called variants. MiSeq Reporter generates variant call files in the VCF (*.vcf) file format and genome VCF (*.gVCF), if configured to do so using the optional sample sheet setting, OutputGenomeVCF.

- VCF files contain information about variants found at specific positions.
- gVCF files contain information about all sites within the region of interest.

VCF File Format

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

VCF files use the file naming format 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 VCF file format version and the variant caller version. The header lists the annotations used in the remainder of the file. If MARS is listed as the annotator, the Illumina internal annotation algorithm is in use to annotate the VCF file. The VCF header also contains the command line call used by MiSeq Reporter to run the variant caller. The command-line call specifies all parameters used by the variant caller, including the reference genome file and .bam file. The last line in the header is column headings for the data lines. For more information, see *VCF File Annotations* on page 18.

```
##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>
##annotator=MARS
##CallSomaticVariants_cmdline=" -B D:\Amplicon_DS_Soma2\121017_
   M00948_0054_000000000-
```

```
A2676_Binf02\Data\Intensities\BaseCalls\Alignment3_Tamsen_
   SomaWorker -g [D:\Genomes\Homo_sapiens
\UCSC\hg19\Sequence\WholeGenomeFASTA,] -f 0.01 -fo False -b 20
   -q 100 -c 300 -s 0.5 -a 20 -F 20 -gVCF
True -i true -PhaseSNPs true -MaxPhaseSNPLength 100 -r D:
\Amplicon_DS_Soma2\121017_M00948_0054_000000000-A2676_Binf02"
##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. 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 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 position is the reference base with the variant; for indels or deletions, this position is the reference base immediately before the variant.
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 semicolon 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 is represented as reference TT and alternate T. An A to T single nucleotide variant is represented as reference A and alternate T.
ALT	The alleles that differ from the reference read. For example, an insertion of a single T is represented as reference A and alternate AT. An A to T single nucleotide variant is represented as reference A and alternate T.
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 has 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.

Heading	Description
FILTER	If all filters are passed, **PASS** is written in the filter column. • **LowDP**—Applied to sites with depth of coverage below a cutoff. Configure cutoff using the **MinimumCoverageDepth** sample sheet setting. • **LowGQ**—The genotyping quality (GQ) is below a cutoff. Configure cutoff using the **VariantMinimumGQCutoff** sample sheet setting. • **LowQual**—The variant quality (QUAL) is below a cutoff. Configure using the **VariantMinimumQualCutoff** sample sheet setting. • **LowVariantFreq**—The variant frequency is less than the given threshold. Configure using the **VariantFrequencyFilterCutoff** sample sheet setting. • **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** sample sheet setting; available only for somatic variant caller and GATK. For more information about sample sheet settings, see *MiSeq Sample Sheet Quick Reference Guide (document # 15028392)*.
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 1 RefGene entry. • **DP**—The depth (number of base calls aligned to a 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.

Heading	Description
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 values are similar; 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, and so on. 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 less bias; values near 0 indicate more bias. • **VF**—Variant frequency; the percentage of reads supporting the alternate allele.
SAMPLE	The sample column gives the values specified in the FORMAT column.

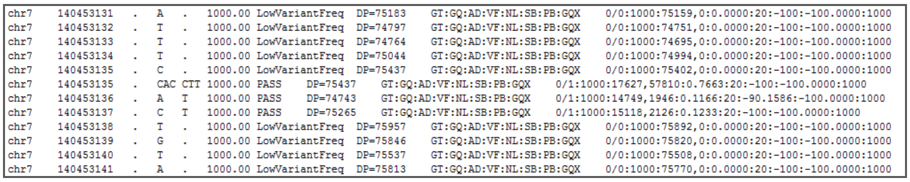
Genome VCF Files

Genome VCF (gVCF) files are VCF v4.1 files that follow a set of conventions for representing all sites within the genome in a reasonably compact format. The gVCF files generated in the Enrichment workflow include all sites within the region of interest specified in the manifest file.

For more information, see sites.google.com/site/gvcftools/home/about-gvcf.

The following example illustrates the convention for representing nonvariant and variant sites in a gVCF file.

Figure 1 Example gVCF File



```
chr7    140453131   .   A   .   1000.00 LowVariantFreq  DP=75183    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:75159,0:0.0000:20:-100:-100.0000:1000
chr7    140453132   .   T   .   1000.00 LowVariantFreq  DP=74797    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:74751,0:0.0000:20:-100:-100.0000:1000
chr7    140453133   .   T   .   1000.00 LowVariantFreq  DP=74764    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:74695,0:0.0000:20:-100:-100.0000:1000
chr7    140453134   .   T   .   1000.00 LowVariantFreq  DP=75044    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:74994,0:0.0000:20:-100:-100.0000:1000
chr7    140453135   .   C   .   1000.00 LowVariantFreq  DP=75437    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:75402,0:0.0000:20:-100:-100.0000:1000
chr7    140453135   .   CAC CTT 1000.00 PASS    DP=75437    GT:GQ:AD:VF:NL:SB:PB:GQX    0/1:1000:17627,57810:0.7663:20:-100:-100.0000:1000
chr7    140453136   .   A   T   1000.00 PASS    DP=74743    GT:GQ:AD:VF:NL:SB:PB:GQX    0/1:1000:14749,1946:0.1166:20:-90.1586:-100.0000:1000
chr7    140453137   .   C   T   1000.00 PASS    DP=75265    GT:GQ:AD:VF:NL:SB:PB:GQX    0/1:1000:15118,2126:0.1233:20:-100:-100.0000:1000
chr7    140453138   .   T   .   1000.00 LowVariantFreq  DP=75957    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:75892,0:0.0000:20:-100:-100.0000:1000
chr7    140453139   .   G   .   1000.00 LowVariantFreq  DP=75846    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:75820,0:0.0000:20:-100:-100.0000:1000
chr7    140453140   .   T   .   1000.00 LowVariantFreq  DP=75537    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:75508,0:0.0000:20:-100:-100.0000:1000
chr7    140453141   .   A   .   1000.00 LowVariantFreq  DP=75813    GT:GQ:AD:VF:NL:SB:PB:GQX    0/0:1000:75770,0:0.0000:20:-100:-100.0000:1000
```

NOTE

The gVCF file shows no-calls at positions with low coverage, or where a low-frequency variant (< 3%) occurs often enough (> 1%) that the position cannot be called to the reference. A genotype (GT) tag of **./.** indicates a no-call.

Enrichment Analysis File Formats

MiSeq Reporter generates 2 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 1 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 *Optional Settings for the Enrichment Workflow* on page 12.

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,RegionInt
     erval,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.

NOTE
Changes introduced in MiSeq Reporter v2.3 include the addition of padding size and the organization of rows by run, read, base, coverage, and variant statistics. If the sample sheet setting ManifestPaddingSize is applied, the summary file includes additional rows for statistics with padding. For more information, see *Sample Sheet Settings for Analysis* on page 12.

The following table describes the file contents as of MiSeq Reporter v2.3.

Statistic	Description
Sample ID	Sample ID.
Run folder	Path to the run folder.
Padding size	The number of bases specified for padding. Padding is 0, by default.
Total length of targeted reference	Total length of sequenced bases in the target reference.
Total aligned reads	Total aligned reads.
Targeted aligned reads	Number of reads that aligned to the target.
Padded targeted aligned reads	Number of reads that aligned to the padded targeted regions. (This statistic only appears if the ManifestPaddingSize setting was applied.)
Read enrichment	100*(Target aligned reads/Total aligned reads).

Statistic	Description
Padded read enrichment	100*(Padded targeted aligned reads/Total aligned reads). (This statistic only appears if the ManifestPaddingSize setting was applied.)
Total aligned bases	Total aligned bases.
Targeted aligned bases	Total aligned bases in the target region.
Padded targeted aligned bases	Total aligned bases in the padded targeted regions. (This statistic only appears if the ManifestPaddingSize setting was applied.)
Base enrichment	100*(Total aligned bases in targeted regions/total aligned bases).
Padded base enrichment	100*(Total aligned bases in padded targeted regions/total aligned bases). (This statistic only appears if the ManifestPaddingSize setting was applied.)
Percent duplicate paired reads	Percentage of paired reads that have duplicates.
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 at 1X	Percentage of targets with coverage greater than 1X.
Target coverage at 10X	Percentage of targets with coverage greater than 10X.
Target coverage at 20X	Percentage of targets with coverage greater than 20X.
Target coverage at 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 Ts/Tv ratio	Transition rate of SNPs that pass the quality filters/Transversion rate of SNPs that pass the quality filter.

Statistic	Description
SNP Het/Hom ratio	Number of Heterozygous SNPs/Number of Homozygous SNPs.
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.

Supplementary Output Files

The following output files provide supplementary information, or summarize run results and analysis errors. Although, these files are not required for assessing analysis results, they can be used for troubleshooting purposes.

File Name	Description
AdapterTrimming.txt	Lists the number of trimmed bases and percentage of bases for each tile. This file is present only if adapter trimming was specified for the run. Located in Data\Intensities\BaseCalls\Alignment.
AnalysisLog.txt	Processing log that describes every step that occurred during analysis of the current run folder. This file does not contain error messages. Located in the root level of the run folder.
AnalysisError.txt	Processing log that lists any errors that occurred during analysis. This file is present only if errors occurred. Located in the root level of the run folder.
CompletedJobInfo.xml	Written after analysis is complete, contains information about the run, such as date, flow cell ID, software version, and other parameters. Located in the root level of the run folder.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with 1 row per tile and 1 column per sample. Located in Data\Intensities\BaseCalls\Alignment.
EnrichmentStatistics.xml	Contains summary statistics specific to the run. Located in the root level of the run folder.
ErrorsAndNoCallsByLaneTileReadCycle.csv	A comma-separated values file that contains the percentage of errors and no-calls for each tile, read, and cycle. Located in Data\Intensities\BaseCalls\Alignment.
Mismatch.htm	Contains histograms of mismatches per cycle and no-calls per cycle for each tile. Located in Data\Intensities\BaseCalls\Alignment.

File Name	Description
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.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.

Enrichment Manifest File Format

A manifest file is required input for the Enrichment workflow. The Enrichment workflow 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 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 is set to 0.
- **Downstream Probe Length**—The length of the downstream (3') PCR probe. For the Enrichment workflow, this field is set to 0.
- **Group**—(*For TruSight panels only*) If specified, this column can be used to group regions (e.g., for a particular gene).

Notes

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 1 Illumina General Contact Information

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

Table 2 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Italy	800.874909
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
Belgium	0800.81102	Norway	800.16836
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000

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

Product documentation—Available for download in PDF from the Illumina website. Go to support.illumina.com, select a product, then select **Documentation & Literature**.

Illumina
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com