MiSeq Reporter TruSeq Amplicon Workflow Guide

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

ILLUMINA PROPRIETARY
Document # 15042314 v01
September 2015

illumina®

Revision History

Document #	Date	Description of Change
Document # 15042314 v01	September 2015	Changed the name of the guide from the MiSeq Reporter TruSeq Amplicon Workflow Reference Guide to the MiSeq Reporter TruSeq Amplicon 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. Updated the read stitching description to include information on what occurs when the Q-score is the same in an overlap region, and information on alignment in the BAM file for stitched reads.
Part # 15042314 Rev. D	December 2014	Added a note in the Demultiplexing section about the default index recognition for index pairs that differ by < 3 bases.
Part # 15042314 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. Changed sample sheet setting for analysis parameter CustomAmpliconAlignerMaxIndelSize to TruSeqAmpliconAlignerMaxIndelSize. Updated sample sheet OutputGenomeVCF parameter default setting information.
Part # 15042314 Rev. B	August 2013	• Updated to MiSeq Reporter v2.3: added sample sheet settings OutputGenomeVCF and StitchReads. • Added description of genome VCF (gVCF) files and read stitching. • Added descriptions of optional settings: Adapter, VariantCaller, and VariantMinimumGQCutoff (also known as VariantFilterQualityCutoff).
Part # 15042314 Rev. A	June 2013	Initial release. The TruSeq Amplicon workflow was formerly named the Custom Amplicon workflow. 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 TruSeq Amplicon workflow is provided in this guide.

Introduction

The TruSeq Amplicon workflow aligns TruSeq Amplicon reads against manifest files specified in the sample sheet, and then variants are identified.

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

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

Workflow Requirements

- **Manifest file**—The TruSeq Amplicon workflow requires at least 1 manifest file. The manifest file is provided with either your custom assay (CAT) when using the TruSeq Custom Amplicon kit or from the Illumina website when using the TruSeq Amplicon - Cancer Panel.
- **Reference genome**—The TruSeq Amplicon workflow requires the reference genome specified in the manifest file. The reference genome provides variant annotations and sets the chromosome sizes 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 (part # 15028392)*.

TruSeq Amplicon Workflow Overview

The TruSeq Amplicon workflow evaluates short regions of amplified DNA, or amplicons, for variants. Focused sequencing of amplicons enables high coverage of particular regions across many samples.

The TruSeq Amplicon 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

Clusters from each sample are aligned against amplicon sequences specified in the manifest file.

Each paired-end read is initially evaluated in terms of its alignment to the relevant probe sequences for that read. Read 1 is evaluated against the reverse complement 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 matches a probe sequence with no more than 1 mismatch, the full length of the read is aligned against the amplicon target for that sequence. This alignment is performed along the length of the amplicon target using a banded Smith-Waterman alignment algorithm.

The banded Smith-Waterman algorithm performs local sequence alignments to determine similar regions between 2 sequences. Instead of looking at the total sequence, the Smith-Waterman algorithm compares segments of all possible lengths, given the restriction that the maximum indel size is 25 bp.

Any alignments that include more than 3 indels are filtered from alignment results. Filtered alignments are written in alignment (BAM) files as unaligned and are not used in variant calling. Indels within the DLSO and ULSO are not observed given the assay chemistry.

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.

Variant Calling

SNPs and short indels are identified 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.

Variant Annotation

If the 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 occur within known genes are annotated.

Statistics Reporting

Statistics are summarized and reported, and written to the Alignment folder.

TruSeq Amplicon Summary Tab

The Summary tab for the TruSeq Amplicon workflow includes a low percentages graph, high percentages graph, clusters graph, and 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	This value is not applicable to the TruSeq Amplicon workflow and is always 0.

Mismatch Graph

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

TruSeq Amplicon Details Tab

The Details tab for the TruSeq Amplicon workflow includes a samples table, targets table, coverage graph, Q-score graph, variant score graph, consensus reads, 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%. For more information, see the *MiSeq Reporter Software Guide (document # 15042295)*.
- **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.

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 TruSeq 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.
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 TruSeq Amplicon workflow.

Q-Score Graph

Y Axis	X Axis	Description
Q-Score	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 TruSeq 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 are represented in the International Union of Pure and Applied Chemistry (IUPAC) convention.

Figure 1 Consensus Reads on Details Tab

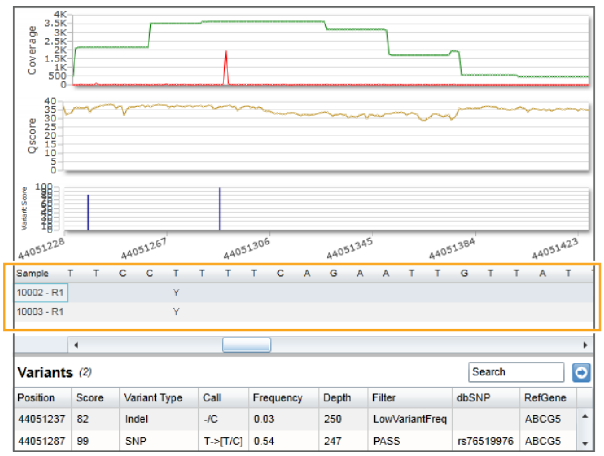

Table 1 IUPAC Nucleotide Codes

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

Nucleotide Code	Base
V	A, C, or G
N	any base
. or -	gap

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.
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 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.

Optional Settings for the TruSeq Amplicon 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 TruSeq Amplicon workflow.

Sample Sheet Settings for Analysis

Parameter	Description
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.
StitchReads	Settings are 0 or 1. Default is 0, paired-end reads are not stitched. If set to true (1), paired-end reads that overlap are stitched to form a single read. To be stitched, a minimum of 10 bases must overlap between Read 1 and Read 2. Paired-end reads that cannot be stitched are converted to 2 single reads. This setting requires MiSeq Reporter v2.3, or later.
TruSeqAmpliconAlignerMax IndelSize	By default, the maximum detectable indel size is 25. A larger value increases sensitivity to larger indels, but requires more time to complete alignment.
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.

Read Stitching

MiSeq Reporter v2.3, or later, is required to use the optional StitchReads setting.

When set to true (1), paired-end reads that overlap are stitched to form a single read in the FASTQ file. At each overlap position, the consensus stitched read has the base call and quality score of the read with higher Q-score.

For each paired read, a minimum of 10 bases must overlap between Read 1 and Read 2 to be a candidate for read stitching. The minimum threshold of 10 bases minimizes the

number of reads that are stitched incorrectly due to a chance match. Candidates for read stitching are scored as follows:

- For each possible overlap of 10 base pairs or more, a score of 1 – MismatchRate is calculated.
- Perfectly matched overlaps have a MismatchRate of 0, resulting in a score of 1.
- Random sequences have an expected score of 0.25.
- If the best overlap has a score of ≥ 0.9 ***and*** the score is ≥ 0.1 higher than any other candidate, then the reads are stitched together at this overlap.

Although the stitched reads are aligned as one, in the BAM file the stitched alignment is split into individual alignments.

During variant calling, stitched reads are processed together. A consensus read is generated by taking the base call and quality score of the read with the higher Q-score in the overlap region. When the Q-score is the same, but the base call differs, a "no call" is used at that position. Sometimes read stitching can improve the accuracy of variant calling.

Paired-end reads that cannot be stitched are converted to 2 single reads in the FASTQ file.

Sample Sheet Settings for Variant Calling

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. (*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 TruSeq Amplicon workflow and provide analysis results for alignment, variant calling, and coverage.

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.
AmpliconCoverage_M#.tsv	Contains details about the resulting coverage per amplicon per sample. M# represents the manifest number. 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.

VCF File Annotations

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 (part # 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 TruSeq Amplicon 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 2 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.

Amplicon Coverage File

One amplicon coverage file is generated for each manifest. The M# in the file name represents the manifest number as it is listed in the sample sheet.

Each file begins with a header row that contains the sample IDs associated with the manifest. In the following example, sample ID 1 and sample ID 3 use one the first manifest in the sample sheet.

Figure 3 AmpliconCoverage_M1.tsv File

	1	3
AKT1ex2.chr14.105246425.105246553_tile_1.1	2022	5080
ALKex23.chr2.29443572.29443701_tile_2.1	8265	8794
APCex15_1.chr5.112173836.112173974_tile_1.1	25600	30728
APCex15_1.chr5.112173836.112173974_tile_2.1	7860	10106
APCex15_2.chr5.112174625.112174757_tile_1.1	10325	14223
APCex15_2.chr5.112174625.112174757_tile_2.1	26942	28129
APCex15_3.chr5.112174992.112176072_tile_2.1	5076	8121
APCex15_3.chr5.112174992.112176072_tile_3.1	18933	14776
APCex15_3.chr5.112174992.112176072_tile_4.1	25918	27048
APCex15_3.chr5.112174992.112176072_tile_5.1	13471	13235
APCex15_3.chr5.112174992.112176072_tile_7.1	30250	30355
APCex15_3.chr5.112174992.112176072_tile_8.1	6868	11041
APCex15_3.chr5.112174992.112176072_tile_9.1	15582	18799
APCex15_3.chr5.112174992.112176072_tile_10.1	27009	29830
APCex15_3.chr5.112174992.112176072_tile_11.1	19286	21314
APCex15_3.chr5.112174992.112176072_tile_12.1	16001	24434
BRAFex11.chr7.140481376.140481493_tile_1.1	69049	60563
BRAFex11.chr7.140481376.140481493_tile_2.1	32861	27975
BRAFex15.chr7.140453075.140453193_tile_1.1	29894	23146
CDH1ex8.chr16.68846038.68846166_tile_1.1	16844	15446
CDH1ex8.chr16.68846038.68846166_tile_2.1	15388	14331
CDH1ex9.chr16.68847216.68847398_tile_2.1	15969	15372
CDH1ex9.chr16.68847216.68847398_tile_3.1	17150	18529
CDH1ex12.chr16.68855904.68856128_tile_2.1	21237	21152
CDH1ex12.chr16.68855904.68856128_tile_3.1	24632	20282
CTNNB1ex2.chr3.41266017.41266151_tile_1.1	61125	43790
CTNNB1ex2.chr3.41266017.41266151_tile_2.1	15241	12005

Below the header rows are 3 columns:

- The first column is the Target ID as it is listed in the manifest.
- The second column is the coverage depth of reads passing filter.
- The third column is the total coverage depth.

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.
AmpliconRunStatistics.xml	Contains summary statistics specific to the run. Located in the root level of the run folder.

File Name	Description
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.
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.

TruSeq Amplicon Manifest File Format

A manifest file is required input for the TruSeq Amplicon workflow. The manifest is provided with your custom assay (CAT) when using the TruSeq Custom Amplicon kit or from the Illumina website when using the TruSeq Amplicon - Cancer Panel. The manifest uses a ***.txt** file format and the manifest name for each sample is specified in the Data section of the sample sheet.

The TruSeq Amplicon manifest file contains a header section followed by 2 blocks of rows beginning with column headings, which are titled the Probes section and the Targets section:

- **Probes**—The Probes section has 1 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 *TruSeq Amplicon Workflow Overview* on page 5.
- **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 *TruSeq Amplicon Workflow Overview* on page 5.
- **Targets**—The Targets section has an entry for each amplicon that a probe-pair amplifies. An expected off-target region is 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, and so forth.
- **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 (-).

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 2 Illumina General Contact Information

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

Table 3 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