MiSeq Reporter Amplicon DS
Workflow Guide

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

ILLUMINA PROPRIETARY
Document # 15042903 v02
February 2016

illumina®

Introduction

This guide describes the analysis steps performed in the Amplicon DS workflow, and the types of information and analysis files generated by the workflow.

Plug-In Workflow

The Amplicon DS workflow is provided as a software plug-in for MiSeq Reporter. A run folder associated with the Amplicon DS workflow is represented with the letter **U** in the MiSeq Reporter Analyses tab. Workflows marked as the letter **U** indicate a plug-in workflow.

Data do not appear in the form of graphs and tables on the MiSeq Reporter Summary tab and Details tab for plug-in workflows. However, the Analysis Info tab, Sample Sheet tab, Logs tab, and Errors tab are populated with information from the run and subsequent analysis. For more information about the MiSeq Reporter interface, see the *MiSeq Reporter Software Guide (document # 15042295)*.

Workflow Requirements

- **Two manifest files**—The Amplicon DS workflow requires 2 assay-specific manifest files, TruSightTumor-FPA-Manifest and TruSightTumor-FPB-Manifest. Download the manifest files from the Illumina website. If using DesignStudio with TruSeq Custom Amplicon and the design dual pools option, download the manifest files for pool A and pool B from MyIllumina.
- **Reference genome**—In addition to the manifest files, the Amplicon DS workflow requires the hg19 reference genome for coordinates and chromosome mapping. By default, this reference is included with the MiSeq Reporter software.
  Specify the path to the genome folder in the sample sheet. For more information, see the *MiSeq Sample Sheet Quick Reference Guide*.
- **MiSeq Reporter v2.2.29**, or later—Previous versions of MiSeq Reporter software are not compatible with the Amplicon DS software plug-in. MiSeq Reporter v2.3, or later does not require the software plug-in to perform the Amplicon DS workflow.
  MiSeq Reporter is available from the MiSeq Reporter support page on the Illumina website.

Amplicon DS Workflow Overview

The Amplicon DS workflow is uniquely suited for detection of somatic mutations in formalin-fixed paraffin-embedded (FFPE) samples.

This workflow independently processes variants from the forward and reverse strands of the sample material, and then algorithmically reconciles the calls.

The Amplicon DS 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 with 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.
- A table of demultiplexing results with 1 row per tile and 1 column per sample, including sample 0.
- 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

During the alignment step, the banded Smith-Waterman algorithm aligns clusters from each sample against amplicon sequences specified in the manifest file.

The banded Smith-Waterman algorithm performs local sequence alignments to determine similar regions between 2 sequences. Instead of comparing 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. This process allows alignment across small amplicon targets, often less than 10 bp.

Each paired-end read is 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).
- 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.

Alignments that include more than 3 indels are filtered from alignment results. Filtered alignments are written in alignment files as unaligned and are not used in variant calling.

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

The somatic variant caller identifies SNPs in 3 steps:

- Considers each position in the reference genome separately
- Counts bases at the given position for aligned reads that overlap the position
- Computes a variant score that measures the quality of the call.

Variant scores are computed using a Poisson model that excludes variants with a quality score below Q20. Additionally, the model only calls variants for bases that are covered at 300x or greater for a single amplicon.

Variants are first called for each pool separately. Then, variants from the 2 pools are compared and combined into a single output file.

If a variant meets the following criteria, the variant is marked as PASS in the variant file:

- Must be present in both pools
- Cumulatively have a depth of 1000 or an average depth of 500x per pool
- Have variant frequency of 3% or greater as reported in the merged variant call (VCF) files

For more information, see the *Amplicon DS Variant Caller Technical Note* on the TruSight Tumor 26 Kit support page.

Statistics Reporting

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

Optional Settings for the Amplicon DS 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: StitchReads,1

The following optional settings are compatible with the Amplicon DS workflow.

Sample Sheet Settings for Analysis

Parameter	Description
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.

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.

Analysis Output Files

The following analysis output files are generated for the Amplicon DS workflow and provide analysis results for alignment, variant calling, and coverage.

File Name	Description
BAM files (*.bam)	Contains aligned reads for a given sample. Located in Data\Intensities\BaseCalls\Alignment.
VCF files (*.vcf)	Contains information about variants found at specific positions in a reference genome. Per-pool files are located in Data\Intensities\BaseCalls\Alignment\Variants. Consensus files are located in Data\Intensities\BaseCalls\Alignment.
gVCF files (*.genome.vcf)	Contains the genotype for each position, whether called as a variant or called as a reference. For more information, see *Genome VCF Files* on page 15. The genome VCF files generated for the Amplicon DS workflow are of a size that does not require block compression. Per-pool files are located in Data\Intensities\BaseCalls\Alignment\VariantCallingLogs. Consensus files are 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. For the Amplicon DS workflow, MiSeq Reporter generates variant call files as VCF files and genome VCF files:

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

Per-Pool and Consensus VCF Files

The Amplicon DS workflow generates 2 sets of variant call files:

- Per-pool VCF and gVCF files that are written to the VariantCallingLogs folder
- Consensus VCF and gVCF files that are written to the Alignments folder.

Base Calls

Alignment—Contains consensus VCF (*.vcf) and gVCF (*.genome.vcf) files.

VariantCallingLogs—Contains per-pool VCF (*.vcf) and gVCF (*.genome.vcf) files

Per-Pool VCF Files

Using the somatic variant caller, variants are called in the forward pool and the reverse pool to produce an independent set of VCF files for each pool. The set of per-pool VCF files include both VCF and gVCF files.

Per-pool VCF files are written to a subfolder of the Alignments folder named Variants. Per-pool VCF files use the following naming convention, where S# represents the order the sample is listed in the sample sheet:

- SampleName_S#.genome.vcf—Reports all sites
- SampleName_S#.vcf—Reports variants only

Merged VCF Files

MiSeq Reporter compares the per-pool VCF files generated for the forward and reverse pools, and combines the data at each position to create a final merged VCF file for the sample.

Merged VCF files are written to the Alignment folder. Merged VCF files use the following naming convention, where S# represents the order the sample is listed in the sample sheet:

- SampleName_S#.genome.vcf—Reports for all sites
- SampleName_S#.vcf—Reports variants only

The VCF file lists all called variants, including variants that were flagged as filtered, but not variants with a variant frequency of less than 3%. For variants that pass filters, PASS is written in the FILTER column of the VCF file. For variants that fail 1 or more filters, the filter name is written in the FILTER column. In particular, variants are filtered due to

probe bias (PB) when the variant frequency differs significantly between the 2 pools. For more information, see *VCF File Annotations* on page 14.

Variant calls from the 2 pools are merged using the following criteria.

Criteria	Result
A reference call in each pool	Reference call
A reference call in 1 pool and a variant call in the other pool	Filtered variant call
Two matching variant calls with similar frequencies in each pool	Variant call
Two matching variant calls with significantly different frequencies in each pool	Filtered variant call
Unmatched variant calls in each pool	Filtered variant call

Metrics from the 2 pools are merged using the following schema.

Metric	Schema
Depth	Addition of depths from both pools
Variant Frequency	Total variant counts/total coverage depth
Q-Score	Minimum value of both pools

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 14.

```
##fileformat=VCFv4.1
##FORMAT=<ID=GQX,Number=1,Type=Integer,Description="Minimum of
   {Genotype quality assuming variant position,Genotype quality
   assuming non-variant position}">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype
   Quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allele
   Depth">
##FORMAT=<ID=VF,Number=1,Type=Float,Description="Variant
   Frequency">
##FORMAT=<ID=NL,Number=1,Type=Integer,Description="Applied
   BaseCall Noise Level">
##FORMAT=<ID=SB,Number=1,Type=Float,Description="StrandBias
   Score">
##FORMAT=<ID=PB,Number=1,Type=Float,Description="Probe-pool
   bias. Values closer to 0 indicate more bias toward one probe
   pool (and less confidence in a variant call)">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
```

```
##INFO=<ID=TI,Number=.,Type=String,Description="Transcript ID">
##INFO=<ID=GI,Number=.,Type=String,Description="Gene ID">
##INFO=<ID=EXON,Number=0,Type=Flag,Description="Exon Region">
##INFO=<ID=FC,Number=.,Type=String,Description="Functional
   Consequence">
##FILTER=<ID=LowVariantFreq,Description="Low variant frequency
   < 0.01">
##FILTER=<ID=LowGQ,Description="GQ below < 30.00">
##FILTER=<ID=R8,Description="IndelRepeatLength is greater than
   8">
##FILTER=<ID=LowDP,Description="Low coverage (DP tag),
   therefore no genotype called">
##FILTER=<ID=SB,Description="Variant strand bias too high">
##FILTER=<ID=PB,Description="Probe pool bias - variant not
   found, or found with low frequency, in one of two probe
   pools">
##fileDate=20130320
##source=CallSomaticVariantsv3.1.1.0
##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=\\ussd-
   file\Depts\LifeSci\Bioinformatics\Genomes\Homo_
   sapiens\UCSC\hg19\Sequence\WholeGenomeFASTA
##phasing=none
##contig=<ID=chr1,length=249250621>
##contig=<ID=chr2,length=243199373>
##contig=<ID=chr3,length=198022430>
##contig=<ID=chr4,length=191154276>
##contig=<ID=chr5,length=180915260>
##contig=<ID=chr7,length=159138663>
##contig=<ID=chr9,length=141213431>
##contig=<ID=chr10,length=135534747>
##contig=<ID=chr12,length=133851895>
##contig=<ID=chr14,length=107349540>
##contig=<ID=chr15,length=102531392>
##contig=<ID=chr16,length=90354753>
##contig=<ID=chr17,length=81195210>
##contig=<ID=chr18,length=78077248>
##contig=<ID=chr19,length=59128983>
##contig=<ID=chr20,length=63025520>
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 3
```

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. • **PB**—The prevalence of the variant is significantly biased between the 2 forward and reverse probe pools. • **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*.
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; however, 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. • **PB**—The probe pool bias score assigned at a position. Larger negative values indicated less bias. • **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 Amplicon DS 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.

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 2 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.
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.

Manifest File Format

The Amplicon DS workflow requires 2 manifest files supplied by Illumina, 1 for the forward pool and 1 for the reverse pool. The manifest files use a ***.txt** file format.

NOTE
There is no need to modify manifests files. The following manifest file description is provided for reference only.

The Amplicon DS manifest file contains a header section followed by 3 blocks of rows beginning with column headings: Probes, Targets, and Intervals.

- **Probes**—The Probes section has 1 entry for each pair of probes.

Column Heading	Description
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.
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 Targets section includes 1 entry for each amplicon amplified by a probe-pair. An expected off-target region is included in addition to the submitted genomic region.

Column Heading	Description
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.
TargetNumber	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 on.
Chromosome	The chromosome of the amplicon 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 (–).

Column Heading	Description
Clip Read Direction	Indicates the read direction and position relative to the reference chromosome where soft-clipping occurs. A plus (+) sign indicates read direction for the forward strand. A minus (-) sign indicates read direction for the reverse strand. Position is indicated in the Clip Upstream and Clip Downstream columns. Position numbers are inclusive relative to the soft-clipping, 1-based.

- **Intervals**—The Intervals section has 1 entry for each interval of interest, and restricts variant calling to these intervals. The Intervals section is organized with the following headings: Species, Build ID, Chromosome, Target Start, and Target Stop.

Revision History

Document #	Date	Description of Change
Document # 15042903 v02	February 2016	Updated the Workflow Requirements section with information on how to access manifest files when using DesignStudio with TruSeq Custom Amplicon. Removed compatibility information stating that the Amplicon DS workflow can only be used with the TruSight Tumor assay.
Document # 15042903 v01	September 2015	Changed the name of the guide from MiSeq Reporter Amplicon-DS Workflow Reference Guide to the MiSeq Reporter Amplicon DS Workflow Guide. Removed the Installation section because MiSeq Reporter v2.3 and later does not require a separate plug-in installer. 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 # 15042903 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 # 15042903 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.
Part # 15042903 Rev. B	August 2013	• Updated to MiSeq Reporter v2.3: added sample sheet setting StitchReads; added description of read stitching. • Noted that MiSeq Reporter v2.3 does not require the Amplicon DS software plug-in to perform analysis of TruSight® Tumor libraries.
Part # 15042903 Rev. A	May 2013	Initial release.

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	Japan	0800.111.5011
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
Belgium	0800.81102	Norway	800.16836
China	400.635.9898	Singapore	1.800.579.2745
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	Taiwan	00806651752
Hong Kong	800960230	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000
Italy	800.874909		

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
5200 Illumina Way
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