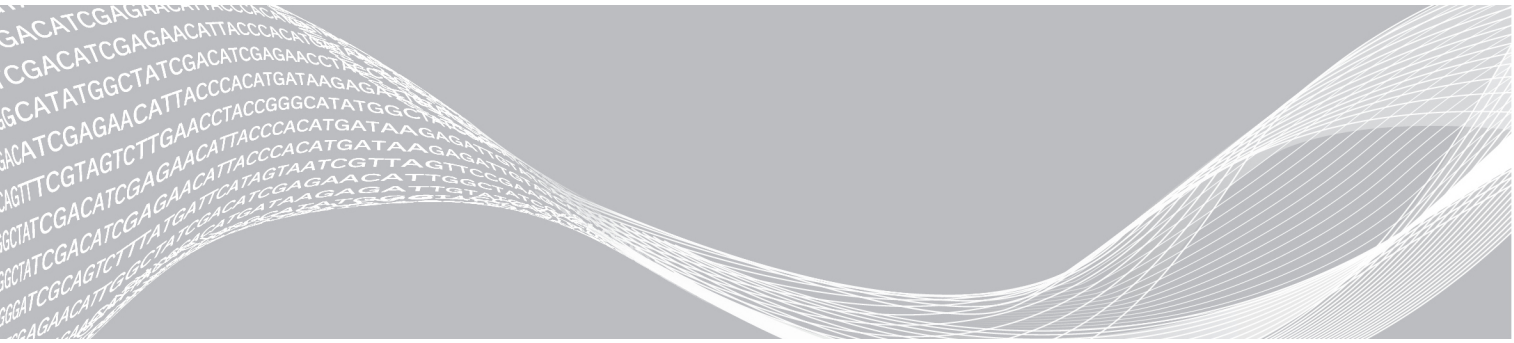
illumina®

Local Run Manager
DNA Amplicon Analysis Module

Workflow Guide

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



Document # 1000000155123 v00

ILLUMINA PROPRIETARY

February 2021

Overview

The Local Run Manager DNA Amplicon analysis module aligns amplicon reads against the reference specified in the manifest file. Variants are called for the targeted regions.

The Local Run Manager DNA Amplicon v3.0.0 (or later) analysis module can only be run on Local Run Manager v3.0 (or later). The analysis module requires the library prep and index kits that are used in the run to be decoupled.

Compatible Library Types

The DNA Amplicon analysis module is compatible with specific library types represented by library kit categories on the Create Run screen. For a current list of compatible library kits, see the Local Run Manager support page on the Illumina website.

Input Requirements

In addition to sequencing data files generated during the sequencing run, such as base call files, the DNA Amplicon analysis module requires the following files.

- Manifest file—The DNA Amplicon analysis module requires at least one manifest file. The manifest file is provided with either your custom assay (CAT) or from the Illumina website depending on the library kit used for the run.
- Reference genome—The DNA Amplicon analysis module requires the reference genome specified in the manifest file. The reference genome provides variant annotations and sets the chromosome sizes in the BAM output files.
- Annotation files—The DNA Amplicon analysis module requires the latest Illumina-provided annotation files. Using older annotations or annotations from any other source will result in failure during the annotation workflow. For the latest annotation files, see the Local Run Manager page on the Illumina website.

NOTE

The manifest file and reference genome selected for the run must be compatible or analysis will fail.

Optional Inputs

In addition to the manifest and reference genome information, the DNA Amplicon analysis module uses the following to provide additional information.

- Genotypes of interest VCF files—During variant calling, the positions in the VCF files are used to call genotypes for each sample, even if the position would not have been called. These forced calls are output in separate files (<sample>.genotype.vcf). Up to five VCF files can be used in the analysis.

Uploading Manifests and VCF Files

To import a manifest or genotypes of interest VCF file for all runs using the DNA Amplicon analysis module, use the Modules & Manifests option from the Tools menu. For more information, see the *Local Run Manager v3 Software Guide (document #1000000111492)*.

Alternatively, you can import a manifest or VCF file for the current run using **Import Manifests** on the Create Run screen.

About This Guide

This guide provides instructions for setting up run parameters for sequencing and analysis parameters for the DNA Amplicon analysis module. For information about the Local Run Manager dashboard and system settings, see the *Local Run Manager v3 Software Guide (document #1000000111492)*.

Set Parameters

1 Log in to Local Run Manager.

2 Select **Create Run**, and then select **DNA Amplicon**.

3 Enter a run name that identifies the run from sequencing through analysis.
The run name can contain alphanumeric characters, spaces, and the following special characters: `.~!@#$%-_{}.

4 **[Optional]** Enter a run description to further identify the run.
The run description can contain alphanumeric characters, spaces, and the following special characters: `.~!@#$%-_{}.

Specify Run Settings

1 Select the library prep kit from the Library Prep Kit drop-down list.

2 Select the index kit from the Index Kit drop-down list.

3 Specify the number of index reads.
- **0** for a run with no indexing
- **1** for a single-indexed run
- **2** for a dual-indexed run

If your index kit supports only one option, the index read is automatically selected.

4 Select the read type for the run.
If your index kit supports only one option, the read type is automatically selected.

5 Specify the number of cycles for the run.

6 **[Optional]** For Custom Primers, specify any custom primer information to be used for the run by selecting the appropriate checkboxes.
Custom primer options vary based on your instrument or Local Run Manager implementation.

NOTE

By default, the DNA Amplicon analysis module is set to two index reads of eight cycles each and the read type Paired End.

Specify Module-Specific Settings

1 Select a variant calling method from the Variant Caller drop-down list.
- **Somatic**—Identifies variants at low frequency and minimizes false positives. Recommended for analysis of tumor samples.
- **Germline**—Identifies small nonsomatic variants and is better used for variants found at higher frequencies. Recommended for all other sample types.

2 Enter a number, **10–10,000**, to define the Variant Caller Depth Filter level.

The default value is 10. Variants with a caller depth below the specified value are marked as filtered. Lower filter values may result in more false positive variants passing filter.

3 If using the Somatic Variant Caller, enter a number, **0.05–30**, to define the Variant Frequency.
The default is 5, which sets the threshold to 5%. Variants with a frequency below the specified threshold are not reported in VCF files.

4 Select an alignment method.
- **BWA**—Each read can be aligned to any position in the reference genome.
- **TruSeq Amplicon Aligner**—Each read is only aligned to the probe sequences defined in the manifest.

5 For Indel Realignment, select from the following options:
- **On**—Gemini performs indel realignment, which might improve medium-sized indel detection. However, overall accuracy can vary in different panels and total analysis time can increase. This option is the default setting.
- **Off**—Indel realignment is not performed.

6 **[Optional]** Select **+ Add a Genotype file** to use a genotype of interest file for the run. If you have not added any genotype of interest VCF files, import a VCF file as follows.

a Select **Import VCF**.
b Browse to the location of the genotype of interest file, select the file, and then select **Open**.
c Select **+ Add a Genotype file** and from the drop-down menu, select the VCF files to add to the run.

Up to five VCF files can be used in the analysis.

7 For Annotation, select from the following options.
- **RefSeq**—Variants are annotated using RefSeq transcripts.
- **Ensembl**—Variants are annotated using Ensembl transcripts.
- **None**—Variants are not annotated.

Variant annotation is supported for human genomes only, and not supported for custom genomes.

8 Set the Sample Identification Analysis option.
Generate sample fingerprints based on SNPs in the Sample ID panel spike-in to detect sample swaps. Sample identification is automatically disabled for nonhuman genomes.

NOTE

By default, the DNA Amplicon analysis module uses BWA Whole-Genome for alignment.

Specify Advanced Module Settings

1 Select **Show Advanced Settings** to view the available module settings.

2 **[Optional]** Enter a value, **2–1,000**, to define the Variant Quality Filter.
The default value is 30. Variants with a variant quality score below the specified threshold are flagged as filtered in VCF files.

Custom Analysis Settings

Custom analysis settings are intended for technically advanced users. It is recommended that you use this feature at your own risk.

Add a Custom Analysis Setting

1 From the Advanced Module Settings section of the Create Run screen, select **Show advanced module settings**.

2 Select **+ Add custom setting**.

3 In the custom setting field, enter the setting name as listed in the Available Analysis Settings section.

4 In the setting value field, enter the setting value.

5 To remove a setting, select ✖.

Available Analysis Settings

- **Adapter Trimming**—By default, adapter trimming is enabled in the DNA Amplicon analysis module. To specify a different adapter, use the Adapter setting. The same adapter sequence is trimmed for Read 1 and Read 2.
  - To specify two adapter sequences, separate the sequences with a plus (+) sign.
  - To specify a different adapter sequence for Read 2, use the AdapterRead2 setting.

Setting Name	Setting Value
Adapter	Enter the sequence of the adapter to be trimmed.
AdapterRead2	Enter the sequence of the adapter to be trimmed.

- **Indel Repeat Filter**—Filters variants containing repeats when the reference has a 1-base or 2-base motif over 8 times (by default) next to the variant.

Setting Name	Setting Value
VariantFilterRMxN	Enter three values separated by a space, M N F, where • M is the maximum length of the repeat. • N is the minimum number of repetitions of the repeat. • F is the frequency below which a variant is filtered. The default is "5 9 0.35" for Germline variant calling and "3 6 0.20" for Somatic variant calling

Import Manifest Files for the Run

1 Make sure that the manifests you want to import are available locally or in an accessible network location.

2 Select **Import Manifests**.

3 Navigate to the manifest file and select the manifest that you want to add.

NOTE

To import manifests for any run using the DNA Amplicon analysis module, use the Modules & Manifests option from the Tools drop-down menu on the navigation bar.

Specify Samples for the Run

Specify samples for the run using one of the following options:

- **Enter samples manually**—Use the blank table at the bottom of the Create Run screen.
- **Import sample sheet**—Use an external CSV file to specify samples for the run.

After the samples table is populated, you can export the sample information to an external file. Use this file as a reference when preparing libraries or importing the file for another run.

Enter Samples Manually

To enter sample information manually, you must first select a Library Prep and Index Kit in the Run Settings section.

1 Adjust the samples table to an appropriate number of rows.
   - In the Rows field, use the up/down arrows or enter a number to specify the number of rows to add to the table. Select  to add the rows to the table.
   - Select  to delete a row.
   - Right-click on a row in the table and use the commands in the contextual menu.
2 Enter a unique sample ID in the Sample ID field.
   Use alphanumeric characters, dashes, or underscores. Spaces are not allowed in this field.
3 **[Optional]** Enter a sample description in the Description field.
   Use alphanumeric characters, dashes, or underscores. Spaces are not allowed in this field.
4 If you have a plated kit, select an index plate well from the Index well drop-down list.
5 Select a manifest file from the Manifest drop-down list.
6 Select a reference genome from the Genome Folder drop-down list.
7 **[Optional]** Enter a project name in the Sample Project field.
   Use alphanumeric characters, dashes, or underscores. Spaces are not allowed in this field.
8 **[Optional]** Select **Export Sample Sheet** to export the sample information in *.csv format.
   The exported sample sheet can be used as a template, or imported when creating new runs.
9 Select **Save Run**.

Import Sample Sheet

1 If you do not have a sample sheet to import, see *Enter Samples Manually* on page 7 for instructions on how to create and export a sample sheet. Edit the file as follows.
   a Open the sample sheet in a text editor.
   b Enter the sample information in the [Data] section of the file.
   c Save the file. Make sure that the sample IDs are unique.
2 Select **Import Sample Sheet** at the top of the Create Run screen and browse to the location of the sample sheet.
   Make sure that the information in the manifest and sample sheet is correct. Incorrect information can impact the sequencing run.

   NOTE

   During analysis, the iSeq™ 100, MiniSeq™, and NextSeq™ Systems automatically reverse complement the i5 indexes in custom library prep kits. If you are importing a sample sheet for a custom library prep kit, make sure that the i5 indexes are in the forward orientation.
3 When finished, select **Save Run**.

Sample Sheet Fields

Manual editing of the sample sheet is intended for technically advanced users. If settings are applied incorrectly, serious problems can occur.

Visit the Local Run Manager support page for available sample sheet settings. Settings must be entered as specified to avoid analysis failure.

Analysis Methods

The DNA Amplicon analysis module performs the following analysis steps and then writes analysis output files to the Alignment folder.

- Demultiplexes index reads
- Generates FASTQ files
- Aligns to a reference
- Identifies variants

Demultiplexing

Demultiplexing compares each Index Read sequence to the index sequences specified for the run. No quality values are considered in this step.

Index reads are identified using the following steps:

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

FASTQ File Generation

After demultiplexing, the software 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 the associated quality scores. Any controls used for the run and clusters that did not pass filter are excluded.

Each FASTQ file contains reads for only one sample, and the name of that sample is included in the FASTQ file name. FASTQ files are the primary input for alignment.

Read Stitching

The DNA Amplicon analysis module performs read stitching when using the TruSeq Amplicon aligner. Read stitching is not performed when BWA Whole-Genome is the selected aligner.

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 mismatch score is calculated. Perfectly matched overlaps have a MismatchRate of 0, resulting in a score of 1.
- If the best overlap has a score of ≥ 0.9 ***and*** the score is ≥ 0.1 higher than other candidates, the reads are stitched together at this overlap.

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

Although the stitched reads are aligned as a single sequence, the stitched read is split into individual alignments in the BAM file.

Alignment

Clusters from each sample are aligned against amplicon sequences from that sample's manifest file. The workflow supports the Truseq Amplicon and BWA Whole-Genome aligners.

BWA

The BWA aligner is used to align reads to the whole genome. If an aligned read comes from a target in the manifest, the probe bases are soft-clipped and the XN is set to the target name.

Alignments with a high number of mismatches are filtered from alignment results. Filtered alignments are written in alignment files as unaligned and are not used in variant calling.

TruSeq Amplicon Aligner

The probe matching step identifies potential reference amplicon sequences. Probe matching is performed over the first 15–20 bps of the read. A match is considered within three mismatches and two shifts, the total not exceeding three differences.

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, the full length of the read is aligned against the amplicon target for that sequence.

Alignments with high number of mismatches, or that include more than three indels are filtered from alignment results. Filtered alignments are written in alignment files, flagged as unaligned, and are not used in variant calling.

The aligner attempts to stitch Read 1 and Read 2 into one sequence. This ensures that any indels in the overlapping region are placed consistently in Read 1 and Read 2. The alignment for each read is derived from the alignment of the stitched sequence. If the reads cannot be stitched, or if the stitched sequence does not align, the aligner falls back to aligning Read 1 and Read 2 separately.

Variant Calling

Variant calling records single nucleotide polymorphisms (SNPs), insertions/deletions (indels), and other structural variants in a standardized variant call format (VCF).

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

The DNA Amplicon analysis module provides the option of using Germline or Somatic for variant calling.

Germline Variant Caller

Developed by Illumina, the germline variant caller identifies variants in DNA samples.

The germline variant caller genotypes candidate variants according to a set of configurable thresholds:

- Variants below 20% variant frequency are discarded.
- Variants with frequencies between 20% and 70% are classified as heterozygous (0/1 genotype).

- Variants with frequencies above 70% are classified as homozygous alternative variants (1/1 genotype).
- When multiple candidate variants are identified at the same locus, the caller attempts to merge these variants into a diploid-conform variant. The merging is not performed when genotypes of interest VCF files are provided.

A candidate variant is filtered under the following conditions:

- The frequency is below 20%.
- The depth is below the user-specified threshold (default: 20).
- The variant quality is below Q20.
- Significant strand bias is detected.
- The variant is an indel occurring in a homopolymer region.

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 three 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 using a Poisson model. Variants with a quality score below Q20 are excluded.

The somatic variant caller analyzes how many alignments covering a given position include a particular indel compared to the overall coverage at that position.

Sample Identification Analysis

Sample identification analysis (optional) is available for human DNA samples through the spike-in Sample ID panel. The analysis assigns a unique fingerprint to each sample, which enables sample tracking. The fingerprint contains nine letters. The first letter represents sample sex (F for female, M for male, or N for no call); the remaining eight letters represent the genotypes at eight predetermined loci covered by the sample ID panel, following the IUPAC coding convention. The genotyping is performed by calling Pisces in germline mode with the same thresholds (Germline mode) or default germline thresholds (Somatic mode).

Homologous genes on chrX (AMELX) and chrY (AMELY) are used for gender determination. Gender determination is decided by the equation:

$r = \frac{\mathrm{AMEY}}{\mathrm{AMEX}+\mathrm{AMEY}}$, where AMEX = read counts of AMEX amplicon and AMEY = read counts of AMEY amplicon.

If $r \geq 0.2$, then the sample is male. Otherwise, the sample is female. Total coverage of AMEX and AMEY (AMEX +AMEY) must be ≥ 30 to have confidence in the gender decision, otherwise, gender cannot be determined.

View Analysis Results

1 From the Local Run Manager dashboard, select the run name.

2 From the Run Overview tab, review the sequencing run metrics.

3 To change the analysis data file location for future requeues of the selected run, select the Edit icon,

and edit the output run folder file path.
The file path leading up to the output run folder is editable. The output run folder name cannot be changed.

4 [Optional] Select the Copy to Clipboard icon to copy the output run folder file path.

5 Select the Sequencing Information tab to review run parameters and consumables information.

6 Select the Samples & Results tab to view the analysis report.
- If analysis was requeued, select the appropriate analysis from the Select Analysis drop-down list.
- From the left navigation bar, select a sample ID to view the report for another sample.

7 [Optional] Select the Copy to Clipboard icon to copy the Analysis Folder file path.

Analysis Report

Individual analysis results for each sample, as well as an aggregated report of all the samples, are available on the Samples and Results tab. The report is also available in a PDF file format for each sample and as an aggregate report in the Analysis folder.

Sample Information

Column Heading	Description
Sample ID	The sample ID provided when the run was created.
Total PF Reads	The total number of reads passing filter.
Percent Q30 Bases	The percentage of bases called with a quality score ≥ Q30.
Percent On-target Aligned Reads	The percentage of reads passing filter that aligned to the reference genome.
Autosome Call Rate	The percentage of bases covered by PASS variants reported on autosomes in the genome VCF file.

Amplicon Summary

Column Heading	Description
Manifest	The name of the file that specifies the reference and targeted reference regions.
Number of Amplicon Regions	The number of amplicon regions sequenced.
Total Length of Target Regions	The total length in base pairs of sequenced amplicons in the target regions.

Read Level Statistics

Column Heading	Description
Total Aligned Reads	The total number of reads that aligned to the reference for each read (Read 1 and Read 2).
Percent Aligned Reads	The percentage of reads that aligned to the reference for each read (Read 1 and Read 2).

Base Level Statistics

Column Heading	Description
Percent Q30 Bases	The percentage of bases called with a quality score ≥ Q30.
Total Aligned Bases	The total number of bases that aligned to the reference for each read (Read 1 and Read 2).
Percent Aligned Bases	The percentage of aligned bases averaged over cycles per read (Read 1 and Read 2).
Mismatch Rate	The percentage of bases that did not align to the reference averaged over cycles per read (Read 1 and Read 2).
Percent Mismatches	The average percentage of mismatches across both reads 1 and 2 over all cycles. Numbers are calculated per read.

Variants Summary

Row Heading	Description
Total Passing	The total number of variants passing filter for single nucleotide variations (SNVs), insertions, and deletions.
Percent Found in dbSNP	The percentage of variants called by the variant caller that are also present in dbSNP.
Het/Hom Ratio	The ratio of the number of heterozygous SNPs and number of homozygous SNPs detected for the sample.
Ts/Tv Ratio	The ratio of transitions and transversions in SNPs. • Transitions are variants of the same nucleotide type (pyrimidine to pyrimidine, C and T; or purine to purine, A and G). • Transversions are variants of a different nucleotide type (pyrimidine to purine, or purine to pyrimidine).

Coverage Summary

Column Heading	Description
Amplicon Mean Coverage	The total number of aligned bases divided by the targeted region size.
Uniformity of Coverage	The percentage of amplicon regions with coverage values greater than the low coverage threshold of 0.2 * amplicon mean coverage.

Variants by Sequence Context

Row Heading	Description
In Genes	The number of variants present in genes.
In Exons	The number of variants present in exons. Includes coding and UTR regions.
In Coding Regions	The number of variants present in coding regions.
In UTR Regions	The number of variants present in untranslated regions (UTR). Includes 5' and 3' UTR regions.
In Splice Site Regions	The number of variants present in splice site regions. Includes regions annotated as splice acceptor, splice donor, splice site, or splice region.

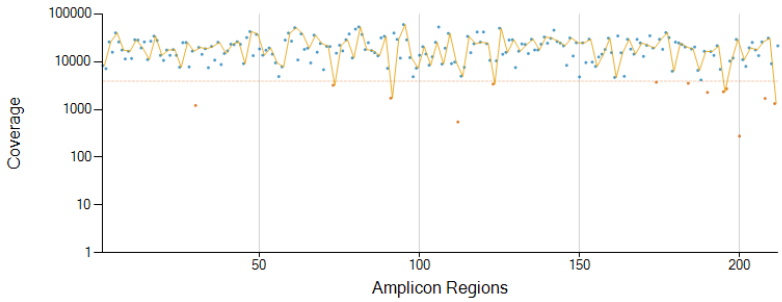
Variants by Consequence

Row Heading	Description
Frameshifts	The number of variants involving a number of base pairs that are not a multiple of 3, which disrupts the triple reading frame.
Nonsynonymous	The number of variants with an amino acid change in a coding region.
Synonymous	The number of variants within a coding region and without an amino acid change.
Stop Gained	The number of variants with the gain of a stop codon in a coding sequence.
Stop Lost	The number of variants with the loss of a stop codon in a coding sequence.

Coverage by Amplicon Region Plot

The Coverage by Amplicon Region plot shows the coverage across amplicon regions. Regions with coverage values lower than the coverage threshold are highlighted in red. The average of all values is indicated by an orange line.

Figure 1 Coverage by Amplicon Region Plot (Example)



Analysis Details – Settings

Column Heading	Description
Run Folder	The Run ID of the analyzed run.
Analysis Folder	The file location of the analysis information.
Reference Genome	The reference genome used for the analysis.
Annotation Source	The database used to annotate the variants.
Depth Threshold	The Variant Caller Depth Filter level setting.
Manifest	The manifest file used for the run.

Analysis Details – Software Versions

Column Heading	Description
DNA Amplicon Workflow	Software version of the DNA Amplicon app.
<Aligner name>	Software version of the aligner used in the run.
<Variant Caller name>	Software version of the variant caller used in the run.
Illumina Annotation Engine	Software version of the variant annotator used in the run.

Column Heading	Description
BWA Aligner	Software version of the Burrows-Wheeler Aligner.
bammetrics	Software version of the BamMetrics workflow.
SAMtools	Software version of the Sequence Alignment/Map (SAM) format.

Analysis Details – Data Collections

Column Heading	Description
Annotation Dataset	The software version of the annotation dataset.

Analysis Output Files

The following analysis output files are generated for the DNA Amplicon analysis module and provide analysis results for alignment and variant calling. Analysis output files are located in the Alignment folder.

File Name	Description
DemuxSummary*	Intermediate files containing demultiplexing summary results.
FASTQ (*.fastq.gz)	Intermediate files containing quality scored base calls. FASTQ files are the primary input for the alignment step.
Alignment files in the BAM format (*.bam)	Contains aligned reads for a given sample.
Variant call files in the genome VCF format (*.genome.vcf)	Contains the genotype for each position, whether called as a variant or called as a reference.
Variant call file in the VCF format (*.vcf)	Contains variants called at each position from both pools.
AmpliconCoverage_M1.tsv	Contains information about coverage per amplicon per sample for each manifest provided. M# represents the manifest number.

Sample ID Files

If the Sample Identification Analysis option is enabled, the DNA Amplicon module creates output files that contain summary information for each sample in the analysis. These files can be used to detect sample swap and determine gender.

- *.sampleID.csv—Summary of the sample ID panel variant calls
- *.sampleID.vcf—Variant call file with the SNPs used for sample ID analysis
- *.sampleID.fingerprint.txt—Contains the sample's fingerprint

Sample ID Genotypes of Interest

The DNA Amplicon app creates a VCF file (*.sampleID.vcf) and a CSV file (*.sampleID.csv). Each row in the output files represent a variant call of interest. The fields of the CSV file are defined in the following table.

Statistic	Definition
Chr	Name of the reference chromosome.
Position	Position within the reference chromosome.
ID	The identifier in dbSNP
RefAllele	The reference allele

Statistic	Definition
AltAllele	The alt allele
Filters	The filters that have been applied to the variant.
GT	The genotype of the variant following the standard vcf annotation (eg, 0/0, 0/1, 1/1, 2/2, ./.).
GenotypeText	Human readable genotype of the variant: REF (reference), HET (heterozygous), HOM (homozygous alternate), NO_CALL, FORCED_REPORT, or OTHER_CALL.
GenotypeForwardAlleles	The observed alleles on the forward strands, separated by "/". For example: Ref/Ref, Ref/Alt, Alt/Alt, or NA for no call.
GenotypeTopAlleles	The two single nucleotide alleles on the top strands, concatenated (eg, AT). When TopBotStrand is TOP, GenotypeTopAlleles contains bases in GenotypeForwardAlleles ordered alphabetically; When TopBotStrand is BOT, GenotypeTopAlleles contains complements of the base ordered alphabetically; When TopBotStrand is NA, GenotypeTopAlleles is NA.
TopBotStrand	Top or bottom strand (ie, TOP, BOT, or NA for indels). See the table below for the strand designation.
Allele1Top	The first (top) allele in the allele pair GenotypeTopAlleles (A/T/G/C or NA for indels).
Allele2Top	The second (top) allele in the allele pair GenotypeTopAlleles (A/T/G/C or NA for indels).
VariantQuality	Phred-scaled quality score indicating how confident we are in this asserted haplotype.
GQ	Phred-scaled quality score indicating how confident we are in this asserted genotype.
AD	The number of reads containing the variant allele.
DP	The number of reads aligned at this position.
VF	The proportion of the variant allele among all alleles being considered.
GoIFileName	The name of the input genotypes of interest (GoI) VCF files. If multiple GoI files contain the same variant, the first GoI file name is used.

Sample ID Fingerprint

The *.sampleID.fingerprint.txt file contains a 9-letter IUPAC code that represents the sample's fingerprint.

For example in the IUPAC code, `FYGACRCRW`, the first letter, `F`, represents the gender call. The remaining letters represent the genotypes at the eight SNPs.

Gender call	IUPAC Code
Male	M
Female	F
No call	N

Genotype call	IUPAC Code
AC	M
AG	R
AT	W
CG	S
CT	Y

Genotype call	IUPAC Code
GT	K
AA	A
CC	C
GG	G
TT	T

FASTQ File Format

FASTQ is a text-based file format that contains base calls and quality values per read. Each record contains 4 lines:

- The identifier
- The sequence
- A plus sign (+)
- The Phred quality scores in an ASCII + 33 encoded format

The identifier is formatted as:

@Instrument:RunID:FlowCellID:Lane:Tile:X:Y ReadNum:FilterFlag:0:IndexSequences or Pairs

Example:

```
@M22:213:FC1234567-ABCDE:1:1101:16239:1359 1:N:0:AACCCCTC+TGTTCTCT
```

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 samtools.github.io/hts-specs/SAMv1.pdf.

BAM files use the file naming format, SampleName_S#.bam. The variable,#, is the sample number determined by the order that samples are listed for the run. In multinode mode, the S# is set to S1, regardless of the order of the sample.

BAM files contain a header section and an alignment 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.
- Alignments—Contains read name, read sequence, read quality, alignment information, and custom tags. The read name includes the chromosome, start coordinate, alignment quality, and match descriptor string.

The alignments section includes the following information for each read or read pair:

- RG—Read group, which indicates the number of reads for a specific sample.
- BC—Barcode tag, which indicates the demultiplexed sample ID associated with the read. (Applies only to runs aligned by the TruSeq Amplicon Aligner).
- XN—Identifies the target name if the read comes from a target.
- SM—Single-end alignment quality.
- AS—Paired-end alignment quality.

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

VCF File Format

Variant Call Format (VCF) is a common file format developed by the genomics scientific community. It contains information about variants found at specific positions in a reference genome. VCF files end with the .vcf or .vcf.gz suffixes.

The VCF file header includes the VCF file format version and the variant caller version and lists the annotations used in the remainder of the file. If Illumina Annotation Engine is listed, the Illumina internal algorithm annotated the VCF file. The last line in the header contains the column headings for the data lines. Each of the VCF file data lines contains information about one variant.

VCF File Headings

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 SNVs, this position is the reference base with the variant. For indels, this position is the reference base immediately preceding the variant.
ID	The rs number for the SNP obtained from dbSNP.txt, if applicable. If multiple rs numbers exist at this location, the list is delimited by semicolons. If a dbSNP entry does not exist 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 in relation to the error rate observed.

VCF File Annotations

Heading	Description
FILTER	If all filters are passed, **PASS** is written in the filter column. • **ForcedReport**—Filter if the variant would normally fail emit filters. Is printed to VCF because it is in one of the genotypes of interest VCF files. • **LowDP**—Applied to sites with depth of coverage below a cutoff. • **LowGQ**—The genotyping quality (GQ) is below a cutoff. • **LowVariantFreq**—The variant frequency is less than the given threshold. • **MultiAllelicSite**—Filter if the variant is in a multiallelic site that breaks ploidy assumptions. Only applicable to germline variant calling. • **q{threshold}**—Quality below {threshold}. • **R8**—For an indel, the number of adjacent repeats (1-base or 2-base) in the reference is greater than 8. • **R{thresholdM}x{thresholdN}**—Filter if the variant is in a repeat region, where a repeat is defined as any region where the reference has motif up to length thresholdM that repeats thresholdN or more times. Only applicable to indels that contain the repeat motif, and are under the cutoff frequency. • **SB**—The strand bias is more than the given threshold.
INFO	Possible entries in the INFO column include: • **AF1000G**—The allele frequency from all populations of 1000 genomes data. • **AA**—The inferred allele ancestral (if determined) to the chimpanzee/human lineage. • **clinvar**—Clinical significance. Format: GenotypeIndex|Significance. • **cosmic**—The numeric identifier for the variant in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Format: GenotypeIndex|Significance. • **CSQ**—Consequence type as predicted by IAE. Format: GenotypeIndex|HGNC|Transcript ID|Consequence. • **CSQR**—Predicted regulatory consequence type. Format: GenotypeIndex|RegulatoryID|Consequence • **DP**—The total depth (number of base calls aligned to a position and used in variant calling). • **EVS**—Allele frequency, coverage, and sample count taken from the Exome Variant Server (EVS). Format: AlleleFreqEVS|EVSCoverage|EVSSamples. • **GMAF**—Global minor allele frequency (GMAF); technically, the frequency of the second most frequent allele. Format: GlobalMinorAllele|AlleleFreqGlobalMinor. • **phyloP**—PhyloP conservation score. Denotes how conserved the reference sequence is between species throughout evolution. • **RefMinor**—Denotes positions where the reference base is a minor allele and is annotated as though it was a variant.
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**—Allele Depth; if the GT is 0/0, the AD is the reference count. If the GT is 0/1 or 1/1, the AD is of the form X,Y, where X is the reference allele count and Y is the alternative allele count. If the GT is 1/2, the AD is of the form Y,Z, where Y and Z are the alternative allele 1 and 2 counts. • **DP**—Total depth used for variant calling. • **GQ**—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. • **SB**—Strand bias at this position. Larger negative values indicate less bias; values near 0 indicate more bias. • **VF**—Variant frequency; if the GT is 0/0, the VF is the nonreference frequency. If the GT is 0/1 or 1/1, the VF is the frequency of the variant allele. If the GT is 1/2, the VF is the frequency of the two variant alleles, together.
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 include all sites within the region of interest in a single file for each sample.

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

If the genotypes of interest feature is turned on, the gVCF file may show variant calls of interest that are requested by the user. These calls may have a filter value of "ForcedReport", indicating that the calls were force written to the gVCF file.

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

Amplicon Coverage File

An amplicon coverage file is generated for each manifest file. The M# in the file name represents the manifest number as it is listed in the samples table for the run.

Each file includes a header row that contains the sample IDs associated with the manifest. Under the header row are three columns that list the following information:

- The Target ID as it is listed in the manifest.
- The coverage depth of reads passing filter.
- 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. All files are located in the Alignment folder unless otherwise specified.

File Name	Description
WorkflowLog.txt	Processing log that describes every step that occurred during analysis of the current run folder. This file does not contain error messages.
WorkflowError.txt	Processing log that lists messages or errors that occurred during analysis.
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.
[Workflow]RunStatistics.xml	Contains summary statistics specific to the run. 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.

Analysis Folder

The analysis folder holds the files generated by the Local Run Manager software.

The relationship between the output folder and analysis folder is summarized as follows:

- During sequencing, Real-Time Analysis (RTA) populates the output folder with files generated during image analysis, base calling, and quality scoring.

- RTA copies files to the analysis folder in real time. After RTA assigns a quality score to each base for each cycle, the software writes the file RTAComplete.txt to both folders.
- When the file RTAComplete.txt is present, analysis begins.
- As analysis continues, Local Run Manager writes output files to the analysis folder, and then copies the files back to the output folder.

Folder Structure

- **Data**
  - **Intensities**
    - **BaseCalls**
      - **L001**—Contains one subfolder per cycle, each containing *.bcl files.
    - **L001**—Contains *.locs files, 1 for each tile.
  - **RTA Logs**—Contains log files from RTA software analysis.
- **Alignment_Imported_#**
  - **Time Stamp**—Contains *.bam files and *.vcf files, and files specific to the analysis module.
    - **Fastq**
      - Sample1_S1_L001_R1_001.fastq.gz
      - Sample2_S1_L001_R1_001.fastq.gz
      - Undetermined_S0_L001_R1_001.fastq.gz
- **Logs**—Contains log files describing steps performed during sequencing.
- **Queued**—A working folder for software; also called the copy folder.
- CompletedJobInfo.xml
- AmpliconRunStatistics
- RunInfo.xml
- runParameters.xml

Alignment Folders

When analysis begins, the Local Run Manager creates an Alignment folder named **Alignment_#**, where # is a sequential number.

If you created the run by importing the information for a run that has already been analyzed, the Alignment folder is named **Alignment_Imported_#**.

Document # 1000000155123 v00

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

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

illumina®