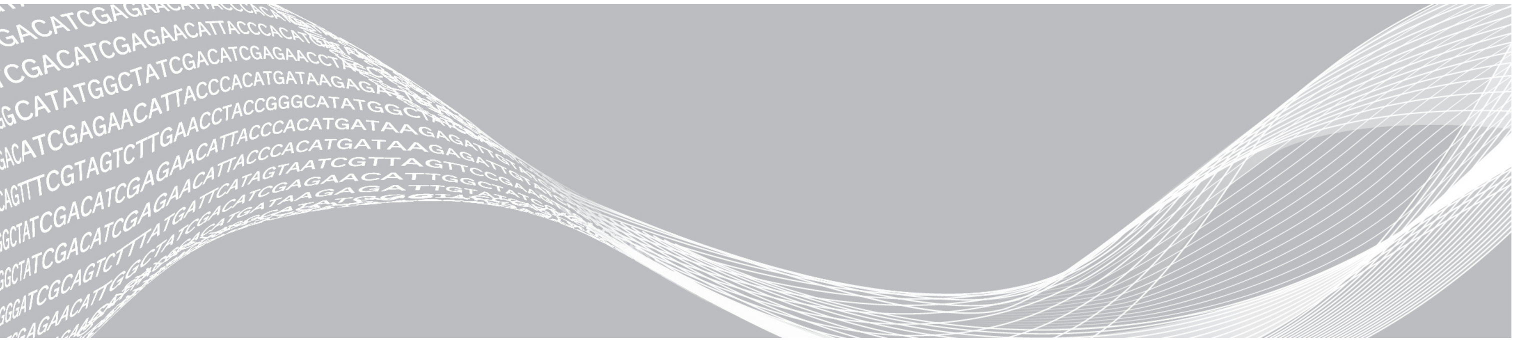
Local Run Manager DNA Enrichment Analysis Module

Workflow Guide

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



Document # 1000000155124 v00
ILLUMINA PROPRIETARY
February 2021

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

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

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

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

Overview

The Local Run Manager DNA Enrichment analysis module aligns reads against the whole genome reference, and then performs variant analysis for regions of interest specified in the manifest file.

The Local Run Manager DNA Enrichment 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 Enrichment 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 Enrichment analysis module requires the following files.

- **Manifest file**—The DNA Enrichment analysis module requires at least one manifest file. The manifest files are available for download from the Illumina website. The manifest file is a list of targeted regions and the chromosome start and end positions.
- **Reference genome**—The DNA Enrichment analysis module uses the hg19 as the reference genome. The reference genome provides the chromosome and start coordinate in the BAM file output.

Uploading Manifests

To import a manifest file for all runs using the DNA Enrichment 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 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 Enrichment 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. If needed, log in to Local Run Manager.
2. Select **Create Run**, and select **DNA Enrichment**.
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.

Specify Module-Specific Settings

1 Select an alignment method from the Aligner drop-down list.
- **BWA-MEM**—(Default) Optimized for Illumina sequencing data and reads ≥ 70 bp.
- **BWA-Backtrack Legacy**—Use with legacy data or reads < 70 bp.

2 Select a variant calling method from the Variant Caller drop-down list.
- **Starling**—(Default) Calls SNPs and small indels, and summarizes depth and probabilities for every site in the genome.
- **Somatic**—Identifies variants at low frequency and minimizes false positives. Recommended for analysis of tumor samples.
- **GATK**—Calls raw variants for each sample, analyzes variants against known variants, and then calculates a false discovery rate for each variant.

3 If using the Somatic Variant Caller, specify the following settings.
- **Variant Frequency**—Set to a threshold of 0.05 by default. Variants with a frequency below the specified threshold are not reported in VCF files.
- **Indel Repeat Filter Cutoff**—On by default. When enabled, indels are filtered when the reference has a 1-base or 2-base motif over 8 times next to the variant.

4 Select a threshold from the Manifest Padding drop-down list.
Set to 150 by default, this setting specifies the number of bases immediately upstream and downstream of the targeted regions used to calculate enrichment statistics. Options are 0–250 in increments of 50.

5 Select the **On/Off** toggle to enable or disable the following settings.
- **Flag PCR Duplicates**—On by default. When enabled, PCR duplicates are flagged in the BAM files and not used for variant calling. PCR duplicates are defined as 2 clusters from a paired-end run where both clusters have the exact same alignment position for each read.
- **Indel Realignment**—On by default. When enabled, regions containing indels are locally realigned to minimize the number of mismatches.

6 Select **Show Advanced Settings**, and then select the **On/Off** toggle to enable or disable Picard HS metrics. To upload a custom probe manifest file, select **Import** and navigate to the location of the file.

Off by default. When enabled, this setting generates Picard HS metrics. You have the option of uploading a custom probe manifest file.

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 Enrichment 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 2 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.

- **Quality Score Trim**—The BWA alignment algorithm automatically trims the 3' ends of non-indexed reads with low quality scores. By default, the value is set to 15.

Setting Name	Setting Value
QualityScoreTrim	Enter a value greater than 0.

- **Variant Frequency**—Filters variants with a frequency less than the specified threshold. If using the Somatic Variant Caller, adjust the value for this setting on the Create Run screen.

Setting Name	Setting Value
VariantFrequencyFilterCutoff	Enter a threshold value. With the Somatic Variant Caller, the default value is 0.05. With GATK or Starling, the default value is 0.20.

- **Indel Repeat Cutoff**—Filters insertions and deletions when the reference has a 1-base or 2-base motif over 8 times (by default) next to the variant. If using the Somatic Variant Caller, enable or disable this setting on the Create Run screen.

Setting Name	Setting Value
IndelRepeatFilterCutoff	Enter a threshold value. The default value is 8.

- **Variant Genotyping Quality**—Filters variants with a genotype quality (GQ) less than the specified threshold.

Setting Name	Setting Value
VariantMinimumGQCutoff	Enter a value less than 99. With GATK or Somatic Variant Caller, the default value is 30. With Starling, the default value is 20.

- **Variant Quality Cutoff**—Filters variants with a quality (QUAL) less than the specified threshold. QUAL indicates the confidence of the variant call.

Setting Name	Setting Value
VariantMiniumQualCutoff	Enter a threshold value. With GATK or Somatic Variant Caller, the default value is 30. With Starling, the default value is 20.

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 Enrichment 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 the following options:

- **Enter samples manually**—Use the blank table at the bottom of the Create Run screen.
- **Import sample sheet**—Navigate to an external file in a comma-separated values (*.csv) format.

After you have populated the samples table, you can export the sample information to an external file. You can use this file as a reference when preparing libraries or import the file when configuring 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 **[Optional]** Enter a project name in the Sample Project field.
Use alphanumeric characters, dashes, or underscores. Spaces are not allowed in this field.

7 **[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.

8 Select **Save Run**.

Import Sample Sheet

1 If you do not have a sample sheet to import, see *Enter Samples Manually* on page 6 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.

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 Enrichment 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.

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.

Adapter Trimming

The DNA Enrichment analysis module performs adapter trimming by default.

During longer runs, clusters can sequence beyond the sample DNA and read bases from a sequencing adapter. To prevent sequencing into the adapter, the adapter sequence is trimmed before the sequence is written to the FASTQ file. Trimming the adapter sequence avoids reporting false mismatches with the reference sequence and improves alignment accuracy and performance.

Alignment

During the alignment step, reads are aligned against the entire reference genome using the Burrows-Wheeler Aligner (BWA), which aligns relatively short nucleotide sequences against a long reference sequence. BWA automatically adjusts parameters based on read lengths and error rates, and then estimates insert size distribution.

The DNA Enrichment analysis module provides the option of using BWA-MEM or BWA-Backtrack Legacy for the alignment step.

BWA-MEM

BWA-MEM is the most recent version of the Burrows-Wheeler Alignment algorithm. Optimized for longer read lengths of ≥ 70 bp, BWA-MEM has a significant positive impact on detection of variants, especially insertions and deletions.

BWA-Backtrack

BWA-Backtrack is an earlier version of the Burrows-Wheeler Aligner algorithm that aligns sequencing read lengths in < 70 bp segments. Use this version for very short reads, or when consistency is required with previous study data.

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 Enrichment analysis module provides the option of using Starling, Somatic, or GATK for variant calling.

Starling

Starling calls both SNPs and small indels, and summarizes depth and probabilities for every site in the genome. Starling produces a VCF file for each sample that contains variants.

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

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.

GATK

The Genome Analysis Toolkit (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 (0/0 or 1/1) or heterozygous (0/1 or 1/0) in the VCF file sample column. For more information, see www.broadinstitute.org/gatk.

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

Analysis results are summarized on the Samples & 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	Description
Sample ID	The sample ID provided when the run was created.
Sample Name	The sample name provided when the run was created.
Run Folder	The name of the run folder.
Total PF Reads	The total number of reads passing filter.
Percent Q30 Bases	The percentage of bases called with a quality score ≥ Q30.
Median Read Length	The average read length in base pairs.
Adapters Trimmed	Indicator if adapter trimming was performed.

Enrichment Summary

Table 1 Enrichment Summary Table

Column Heading	Description
Target Manifest	The name of the file that specifies the reference and targeted reference regions.
Total Length of Targeted Reference	The total length in base pairs of sequenced bases in the target regions.
Padding Size	The length of sequence immediately upstream and downstream of the enriched targets.

Enrichment values are calculated without padding. If a targeted region overlaps another regions, positions are adjusted to remove the overlap.

Read Level Enrichment

Table 2 Read Level Enrichment Table

Column Heading	Description
Total Aligned Reads	The total number of reads that aligned to the reference.
Percent Aligned Reads	The percentage of reads that aligned to the reference.
Targeted Aligned Reads	The number of reads that aligned to the target.
Read Enrichment	The percentage of targeted aligned reads over total aligned reads.
Padded Target Aligned Reads	The number of reads that aligned to the padded target.
Padded Read Enrichment	The percentage of padded target aligned reads over total aligned reads.

Base Level Enrichment

Table 3 Base Level Enrichment Table

Column Heading	Description
Total Aligned Bases	The total number of bases that aligned to the reference.
Percent Aligned Bases	The percentage
Targeted Aligned Bases	The total number of aligned bases in the target region.
Base Enrichment	The percentage of aligned bases in targeted regions over total aligned bases.

Column Heading	Description
Padded Target Aligned Bases	The total number of aligned bases in the padded target region.
Padded Base Enrichment	The percentage of total aligned bases in padded target regions over total aligned bases.

Small 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

Table 4 Coverage Summary Table

Column Heading	Description
Mean Region Coverage Depth	The total number of aligned bases divided by the targeted region size.
Uniformity of Coverage	The percentage of targeted base positions with coverage values greater than the low coverage threshold of 0.2 * mean region coverage depth.
Target Coverage at 1X	The percentage of coverage greater than 1X.
Target Coverage at 10X	The percentage of coverage greater than 10X.
Target Coverage at 20X	The percentage of coverage greater than 20X.
Target Coverage at 50X	The percentage of coverage greater than 50X.

The Mean Coverage by Targeted Region graph shows the mean coverage across target regions.

Depth of Coverage in Targeted Regions

Table 5 Depth of Coverage in Targeted Regions Table

Column Heading	Description
Depth of Sequencing Coverage	The coverage depth based on the number of sequence bases that align to the position.
Number of Targeted Bases Covered at Depth	The number of targeted bases that have at least the indicated depth of coverage.
Total Targeted Bases Covered	The total number of bases that align to the target regions that have at least the indicated depth of coverage.
Target Coverage	The percentage of targeted bases that meet the indicated depth of coverage.

Reads marked as duplicates are not included.

Fragment Length Summary

The fragment length summary section lists the average length of the sequenced fragment for the selected sample, the minimum fragment length, the maximum fragment length, and the range of variability listed as standard deviation. To account for potential outliers, the minimum and maximum are calculated from values within ~ 3 standard deviations, excluding the lower and upper 0.15% of the data.

Duplicate Information

The duplicate information section lists the percentage of clusters for a paired-end sequencing run that are considered to be PCR duplicates. PCR duplicates are defined as 2 clusters from a paired-end run where both clusters have the exact same alignment positions for each read.

Analysis Output Files

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

File Name	Description
Demultiplexing (*.demux)	Intermediate files containing demultiplexing results.
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.
Consensus variant call files in the VCF format (*.vcf)	Contains variants called at each position.
Coverage file (*.coverage.csv)	Contains information about mean coverage by target regions aligned reads in the sample, and enrichment percentage.
Gaps file (*.gaps.csv)	Contains information about gaps in targeted intervals where coverage fell below the threshold used to filter variants.

Demultiplexing File Format

The process of demultiplexing reads the index sequence attached to each cluster to determine from which sample the cluster originated. The mapping between clusters and sample number is written to a demultiplexing (*.demux) file for each tile of the flow cell.

The demultiplexing file naming format is s_1_X.demux, where X is the tile number.

Demultiplexing files start with a header:

- Version (4-byte integer), currently 1
- Cluster count (4-byte integer)

The remainder of the file consists of sample numbers for each cluster from the tile.

When the demultiplexing step is complete, the software generates a demultiplexing file named DemultiplexSummaryF1L1.txt.

- In the file name, **F1** represents the flow cell number.
- In the file name, **L1** represents the lane number.
- Demultiplexing results in a table with one row per tile and one column per sample, including sample 0.
- The most commonly occurring sequences in index reads.

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:SampleNumber

Example:

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAAA9#:<#<;<<<????#=
```

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 of SampleName_S#.bam, in which # 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.
- NM—Edit distance tag, which records the Levenshtein distance between the read and the reference.
- MD—Mismatching positions/bases (BWA only).
- MQ—Mapping quality (if applicable).
- AS—Paired-end alignment quality.
- XS—Suboptimal alignment score

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.

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. 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. • LowDP—Applied to sites with depth of coverage below a cutoff. • LowGQ—The genotyping quality (GQ) is below a cutoff. • LowQual—The variant quality (QUAL) is below a cutoff. • LowVariantFreq—The variant frequency is less than the given threshold. • R8—For an indel, the number of adjacent repeats (1-base or 2-base) in the reference is greater than 8. • SB—The strand bias is more than the given threshold. Used with the Somatic Variant Caller and GATK.

Heading	Description
INFO	Possible entries in the INFO column include: • AC—Allele count in genotypes for each ALT allele, in the same order as listed. • AF—Allele Frequency for each ALT allele, in the same order as listed. • AN—The total number of alleles in called genotypes. • CD—A flag indicating that the SNP occurs within the coding region of at least 1 RefGene entry. • DP—The depth (number of base calls aligned to a position and used in variant calling). • EXON—A comma-separated list of exon regions read from RefGene. • FC—Functional Consequence. • GI—A comma-separated list of gene IDs read from RefGene. • QD—Variant Confidence/Quality by Depth. • TI—A comma-separated list of transcript IDs read from RefGene.
FORMAT	The format column lists fields separated by colons. For example, GT:GQ. The list of fields provided depends on the variant caller used. Available fields include: • AD—Entry of the form X,Y, where X is the number of reference calls, and Y is the number of alternate calls. • DP—Approximate read depth; reads with MQ=255 or with bad mates are filtered. • GQ—Genotype quality. • GQX—Genotype quality. GQX is the minimum of the GQ value and the QUAL column. In general, these 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. Used with the Somatic Variant Caller and GATK. • 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 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.

Coverage File Format

Coverage files can be copied into a spreadsheet program such as Microsoft Excel for viewing, sorting, or graphing.

Coverage files contain a header section and a data section:

- Header—Contains 1 line per targeted region that begins with a # character.
  - The first header line specifies the enrichment, which is defined as the fraction of aligned reads overlapping any of the targeted regions.

  - The second header line specifies the number of reads aligning to targeted regions.
  - The third header line specifies the column headings for the data section.
- Data—The data section includes the following information.

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

Gaps File Format

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

Gaps files contain a header section and a data section:

- Header—The header section specifies the column headings for the data section.
- Data—The data section includes the following information.

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

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
AdapterCounts.txt	Contains a summary of the number of reads that had adapter trimming performed per sample.
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.

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

File Name	Description
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 will be empty if no errors occurred. Located in the root level of the run folder.
CompletedJobInfo.xml	Written after analysis is complete. Contains information about the run, such as date, flow cell ID, software version, and other parameters. Located in the root level of the run folder.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one column per sample.
ErrorsAndNoCallsByLaneTileReadCycle.csv	A comma-separated values file that contains the percentage of errors and no-calls for each tile, read, and cycle.
Mismatch.htm	Contains histograms of mismatches per cycle and no-calls per cycle for each tile.
EnrichmentStatistics.xml	Contains summary statistics specific to the run. Located in the root level of the run folder.
Summary.xml	Contains a summary of mismatch rates and other base calling results.
Summary.htm	Contains a summary web page generated from Summary.xml.

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
      - 📄 FastqSummaryF1L1.txt
      - 📄 Sample1_S1_L001_R1_001.fastq.gz
      - 📄 Sample2_S2_L001_R2_001.fastq.gz
      - 📄 Undetermined_S0_L001_R1_001.fastq.gz
      - 📄 Undetermined_S0_L001_R2_001.fastq.gz
- 📁 Alignment_## or Alignment_Imported_##
  - 📁 [Timestamp of Run]
    - 📁 DataAccessFiles
    - 📁 Logging
    - 📁 Plots
    - 📁 VariantCallingLogs
    - 📄 AdapterCounts.txt
    - 📄 AdapterTrimming.txt
    - 📄 Checkpoint.txt
    - 📄 CompletedJobInfo.xml
    - 📄 DemultiplexSummaryF1L1.txt
    - 📄 EnrichmentStatistics.xml
    - 📄 [SampleName]_S1.bam
    - 📄 [SampleName]_S1.bam.bai
    - 📄 [SampleName]_S1.coverage.csv
    - 📄 [SampleName]_S1.CoverageHistogram.txt
    - 📄 [SampleName]_S1.fragmentlength.csv
    - 📄 [SampleName]_S1.gaps.csv
    - 📄 [SampleName]_S1.genome.vcf.gz
    - 📄 [SampleName]_S1.readlength.csv
    - 📄 [SampleName]_S1.summary.csv
    - 📄 [SampleName]_S1.vcf
    - 📄 SampleSheetUsed.csv

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_#**.

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.