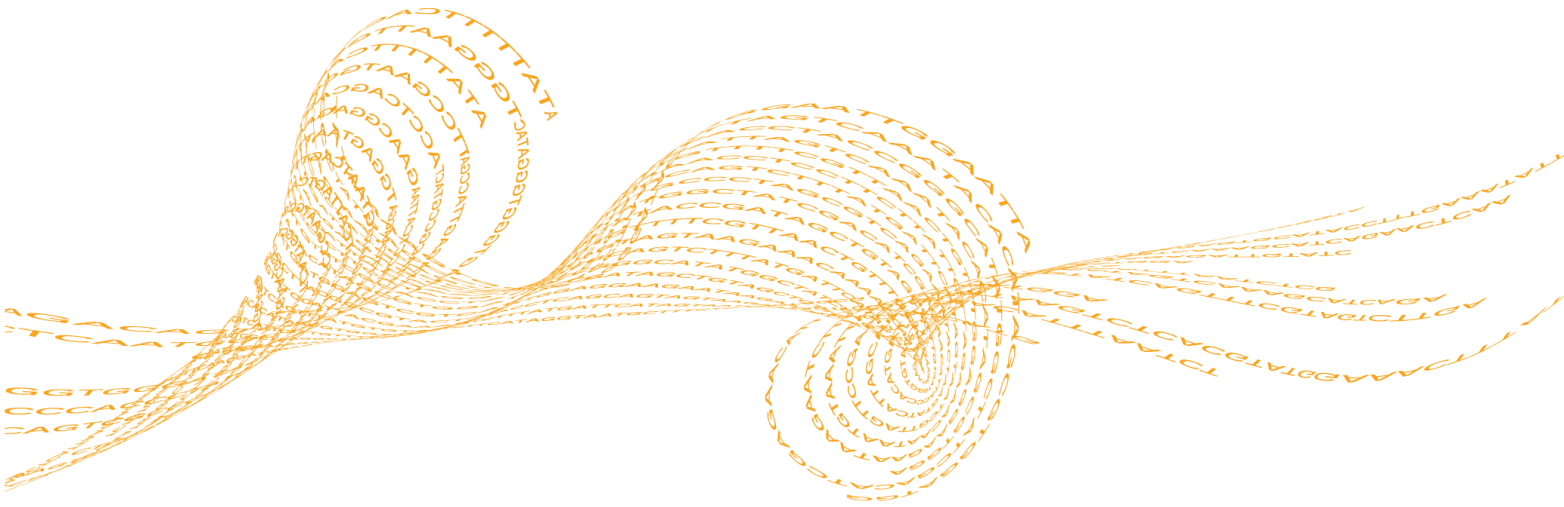
HiSeq Analysis Software v0.9

User Guide

FOR RESEARCH USE ONLY



ILLUMINA PROPRIETARY
Part # 15041353 Rev. B
May 2013

illumina®

Part #	Revision	Date	Description of Change
15041353	B	April 2013	Added adapter settings for Nextera Mate Pair Libraries. Corrected hg19 reference genome download URL.
15041353	A	March 2013	Initial release

Running HiSeq Analysis Software under Linux

1 Identify a run folder for your analysis. This can be but does not need to be the original run folder for the HiSeq flow cell.

2 Create a sample sheet named SampleSheet.csv in the top level of the run folder. See *Sample Sheet Settings* on page 20 for more information.

3 Start the analysis:

```
/path/to/illumina/HiSeqAnalysisSoftware/RunLatest -r
   /path/to/your/RunFolder
```

Where:

- `-r /path/to/your/RunFolder/` is the path to the run folder.

The analysis results will be written (by default) to the `Data/Intensities/BaseCalls/Alignment` subfolder of the run folder. This can be overridden using the `-a` command-line option.

Please note that if you log out of your terminal session, the HiSeq Analysis Software command may be prematurely terminated. You can retain terminal by adding the following to your command line:

```
nohup /path/to/illumina/HiSeqAnalysisSoftware/RunLatest -r
   /path/to/your/RunFolder 2>&1 > logfile.txt &
```

Running HiSeq Analysis Software on a Cluster

When running HiSeq Analysis Software on a cluster, you need to generate a shell script that will submit the analysis job to your queue manager. A simple shell script looks like this:

```
#!/usr/bin/env bash
/path/to/illumina/HiSeqAnalysisSoftware/RunLatest -r
   /path/to/your/RunFolder
```

The analysis software assumes that it has an entire cluster node available. In order to queue up several analyses without overloading a compute node, it is important to specify that each analysis job requires multiple slots on the node. SGE can be configured to support the threaded parallel environment, so that jobs that consume an entire node can be submitted.

For example, if you are using an SGE and want to submit a job with the shell script `Analysis.sh` that reserves X slots on a compute node from the queue `<queue name>`, use a command like this:

```
qsub -pe threaded X -q <queue name> Analysis.sh
```

Introduction

HiSeq Analysis Software (HAS) is a software package for processing sequencing results generated by Illumina HiSeq sequencers. HiSeq Analysis Software performs secondary analysis on the base calls and quality scores in bcl files generated on-instrument by the RTA software. HiSeq Analysis Software produces information about alignment, variants, or other analysis results for each sample in an analysis. The analysis is run with a simple command line. Each workflow includes recommended default settings for the available options, which are specified in the sample sheet.

The analysis workflows supported are:

- Enrichment analysis workflow
- Whole Genome Sequencing analysis workflow for the human hg19 genome.

NOTE

HiSeq Analysis Software takes a run folder with bcl files as input, along with a sample sheet, the hg19 reference genome, and—for the Enrichment analysis—a target manifest file and optional probe manifest file. HiSeq Analysis Software automatically performs bcl to fastq conversion and demultiplexing according to the sample sheet.

This user guide provides instructions for running HiSeq Analysis Software v0.9 through the Linux command-line interface for secondary analysis.

NOTE

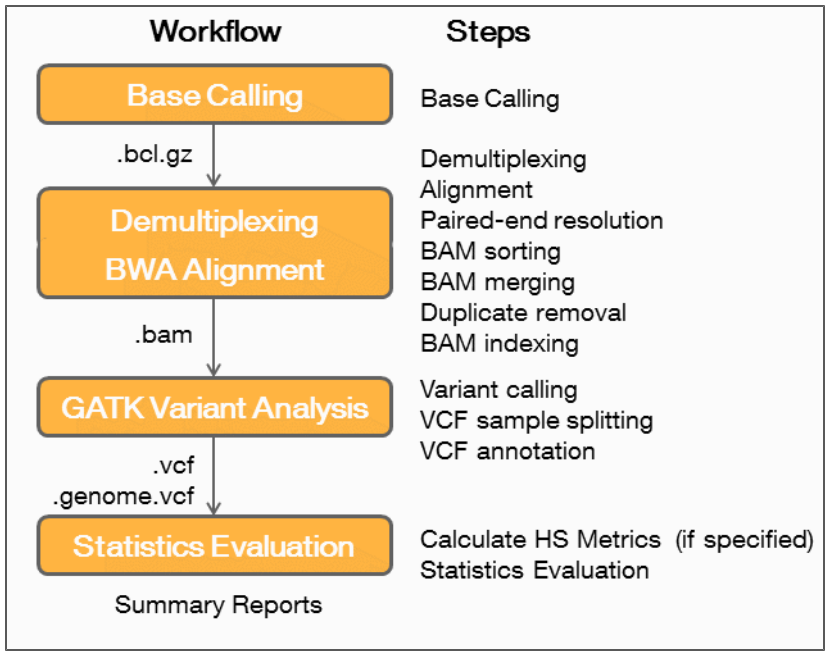
If you want to run HiSeq Analysis Software using a Graphical User Interface, install the Analysis Visual Controller (AVC). See *Additional Software* on page 12.

Enrichment Analysis Workflow

The Enrichment analysis workflow analyzes DNA that has been enriched for particular target sequences using Nextera Rapid Capture. Alignment is performed with BWA and variant calling with GATK. Variant analysis is performed for just the target regions. Statistics reporting accumulates coverage and enrichment specific statistics for each target as well as overall metrics.

The main output files generated by the Enrichment workflow are .bam files (containing the reads after alignment), .vcf files (containing the variant calls), and Genome VCF (.genome.vcf) files (describing the calls for all variant and non-variant sites in the genome).

Figure 1 Enrichment Workflow Diagram



BWA

The Enrichment workflow uses 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.. For more information, see http://bio-bwa.sourceforge.net/.

After BWA alignment, variant calling is done by GATK.

GATK

Developed by the Broad Institute, the Genome Analysis Toolkit (GATK) calls raw variants for each sample read, analyzes the variants against known variants, and then applies a calibration procedure to compute a false discovery rate for each variant. Variants are flagged as homozygous (1/1) or heterozygous (0/1) in the VCF file sample column.

GATK is the standard variant caller after BWA alignment.

For more information, see http://www.broadinstitute.org/gatk.

Picard Metrics

Picard is a suite of tools in Java that work with next generation sequencing data in BAM format. HiSeq Analysis Software uses the CalculateHsMetrics tool in Picard to compute a set of Hybrid Selection specific metrics from an aligned SAM or BAM file. If a reference sequence is provided, AT/GC dropout metrics will be calculated. GC and mean coverage information for every target can also be computed.

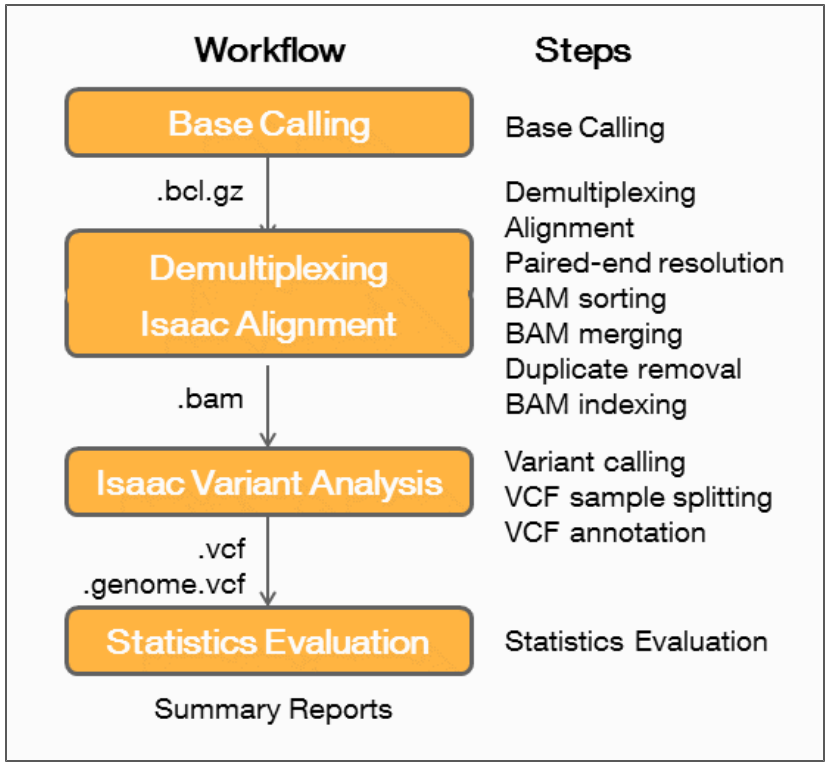
For more information, see: http://picard.sourceforge.net/command-line-overview.shtml

Whole Genome Sequencing Analysis Workflow

The Whole Genome Sequencing analysis workflow uses the Isaac Aligner and Isaac Variant Caller to compare the DNA sequence in the sample(s) against the human reference genome hg19. It identifies any variants (SNPs or indels) relative to the reference sequence.

The main output files generated by the Whole Genome Sequencing workflow are .bam files (containing the reads after alignment), .vcf files (containing the variant calls), and Genome VCF (.genome.vcf) files (describing the calls for all variant and non-variant sites in the genome).

Figure 2 Whole Genome Sequencing Workflow Diagram



Isaac

Alignment and variant calling in the Whole Genome Sequencing analysis workflow are performed with Illumina's Isaac Alignment Software and the Isaac Variant Caller (also referred to as Starling). The Isaac workflow generates output that consists of the realigned and duplicate marked reads in a BAM file format, variants in a VCF file format, a Genome VCF (gVCF) file that has an entry for every base in the reference, and a summary of the run quality.

Isaac Aligner

The Isaac aligner aligns DNA sequencing data, single or paired end, with read lengths and low error rates using the following steps:

- **Candidate mapping positions**—Identifies the complete set of relevant candidate mapping positions using a 32-mer seed-based search.

- **Mapping selection**—Selects the best mapping among all candidates.
- **Alignment score**—Determines alignment scores for the selected candidates based on a Bayesian model.
- **Alignment output**—Generates final output in a sorted duplicate-marked BAM file and summary file.

Candidate Mapping

To align reads, the Isaac aligner first identifies a small but complete set of relevant candidate mapping positions. The Isaac aligner begins with a seed-based search using 32-mers as seeds. The initial single-seed search is followed by a multi-seed only for the reads that were not mapped unambiguously with a single seed.

Mapping Selection

Following a seed-based search, the Isaac aligner selects the best mapping among all the candidates. For paired-end data sets, all mappings where only one end is aligned (called orphan mappings) trigger a local search to find additional mapping candidates defined by the expected minimum and maximum insert size (called shadow mappings). After optional trimming of low quality 3' ends and adapter sequences, the possible mapping positions of each fragment are compared, taking into account pair-end information when available, possible gaps using a banded Smith-Waterman gap aligner, and possible shadows. The selection is based on the Smith-Waterman score and on the log-probability of each mapping.

Alignment Scores

The alignment scores of each read pair are based on a Bayesian model, where the probability of each mapping is inferred from the base qualities and the positions of the mismatches. The final mapping quality is the alignment score, truncated to 60 if above 60, and possibly corrected to known ambiguities in the reference as flagged in the seeds. Following alignment, reads are sorted. Further analysis is performed to identify duplicates and optionally to realign indels.

Alignment Output

After sorting the reads, the Isaac aligner generates compressed binary alignment output files, called BAM (*.bam) files, using the following process:

- **Marking duplicates**—Detection of duplicates is based on the location and observed length of each fragment. The Isaac aligner identifies and marks duplicates even when they appear on oversized fragments or chimeric fragments.
- **Realigning indels**—The Isaac aligner keeps track of previously detected indels, over a window large enough for the current read length, and applies the known indels to all reads with mismatches.
- **Generating BAM files**—The first step in BAM file generation is creation of the BAM record, which contains all required information except the name of the read. The Isaac aligner reads data from base call (BCL) files that were written during primary analysis on the sequencer to generate the read names. The BAM file is then created by compressing data into blocks of 64 Kb or less.

Isaac Variant Caller

The Isaac Variant Caller (the algorithm is also referred to as Starling2) identifies single nucleotide polymorphisms (SNPs) and small indels using the following steps:

- **Read filtering**—Filters out reads failing quality checks.
- **Indel calling**—Identifies a set of possible indel candidates and realigns all reads overlapping the candidates using a multiple sequence aligner.
- **SNP calling**—Computes the probability of each possible genotype given the aligned read data and a prior distribution of variation in the genome.
- **Indel genotypes**—Calls indel genotypes and assigns probabilities.
- **Variant call output**—Generates output in a compressed genome variant call (gVCF) file. See *gVCF Files* on page 39 for details.

Indel Candidates

Input reads are filtered by removing any of the following:

- Reads that failed primary analysis quality checks.
- Reads marked as PCR duplicates.
- Paired-end reads not marked as a proper pair.
- Reads with a mapping quality less than 20.

Indel Calling

The variant caller proceeds with candidate indel discovery and generates alternate read alignments based on the candidate indels. As part of the realignment process, the variant caller selects a representative alignment to be used for site genotype calling and depth summarization by the SNP caller. This alignment is selected to be within a certain threshold of the most-likely of all alignments for a read.

SNP Calling

The variant caller runs a series of filters on the set of filtered and realigned reads for SNP calling without affecting indel calls. First, any contiguous trailing sequence of N base calls are trimmed from the ends of reads. Using a mismatch density filter, reads having an unexpectedly high number of disagreements with the reference are masked, as follows:

- The variant caller 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 applied to a 41-base window at the corresponding end of the read.
- The mismatch limit is applied to the entire read when the read length is 41 or shorter.

Indel Genotypes

All bases marked by the mismatch density filter and any N base calls that remain after the end-trimming step are filtered out by the variant caller. These filtered base calls are not used for site-genotyping but appear in the filtered base call counts in the variant caller output for each site.

All remaining base calls are used for site-genotyping. To account for the possibility of error dependencies, the genotyping method heuristically adjusts the joint error probability that is calculated from multiple observations of the same allele on each strand of the genome. This method treats the highest quality base call from each allele and strand as an independent observation and leaves the associated base call quality scores unmodified. However, quality scores for subsequent base calls for each allele and strand are adjusted to increase the joint error probability of the given allele above the error expected from independent base call observations.

Variant Call Output

After the site and indel genotyping methods are complete, the variant caller applies a final set of heuristic filters to produce the final set of non-filtered calls in the output.

The output in the genome variant call (gVCF) file captures the genotype at each position and the probability that the consensus call differs from reference, which is expressed as a phred-scaled quality score.

More Information

Algorithm Descriptions

A detailed description of Isaac and its algorithms has been submitted for publication:

Come Raczy, Roman Petrovski, Chris Saunders, Ilya Chorny, Semyon Kruglyak, Elliott Margulies, Han-Yu Chuang, Morten Kallberg, Swathi A. Kumar, Arnold Liao, Kristina M. Little, Michael Stromberg, Stephen Tanner (submitted). Isaac: Ultra-fast whole genome secondary analysis on Illumina sequencing platforms.

Parameter Settings

The output file IsisLog.txt contains the parameter settings for the workflow used. See *Output Files* on page 32 for a description.

AVC

The Analysis Visual Controller software (AVC) can be used to control HiSeq Analysis Software with a Graphical User Interface (GUI) and is run on a Windows operating system. AVC can be downloaded from http://support.illumina.com/sequencing/sequencing_software/analysis_visual_controller_avc.ilmn. For more information on installing and using AVC software, see the *Analysis Vidsual Controller User Guide*.

AVC has the following benefits:

- Easy to use GUI for running the WGS and Enrichment workflows without using the Linux command line
- Easy management of the manifest file and the reference genome
- A mechanism for parallelizing analysis across samples using projects

NOTE
AVC uses a different sample sheet format than the one used for direct input to the HiSeq Analysis Software.

IEM

The easiest way to set up a workflow sample sheet is by using the Illumina Experiment Manager (IEM), version 1.5 or later. IEM is a tool that helps create the required sample sheet for use with HiSeq, HiScanSQ, GAIIx or MiSeq® system, and is available for download from http://support.illumina.com/.

Installing HiSeq Analysis Software

Before you install HiSeq Analysis Software, you need to check whether your system meets the requirements. If so, you can install HiSeq Analysis Software using the RPM package.

HiSeq Analysis Software uses the hg19 reference genome. Instructions for obtaining and unpacking it are below (*Unpacking Reference Genome* on page 17). Use of other genomes is not an option at this time.

You need root privileges to perform this installation, and to unpack the reference genome. we recommend that the same person performs both actions.

NOTE

You need root privileges to perform the installation and to unpack the reference genome. Illumina recommends that the same person with root privileges performs both actions. Installation may involve setting access privileges for other users.

Requirements

The requirements for running HiSeq Analysis Software on Linux are listed below.

- At least 48 GB of RAM.
- A Linux distribution supporting RPM installation (CentOS or similar).
- Hard drive (local drive, or network file system such as Isilon) with >=1 terabyte of disk space. Scratch storage should support speeds of at least 200 MB/s for favorable performance.
- Root (or sudo root) permissions required for software installation.
- 75 GB space in the opt directory, the default install directory. If that is not available you need to specify a different install location with the `--prefix` option.

Installing HiSeq Analysis Software RPM

NOTE
You need root privileges to perform this installation.

1 Copy the HiSeqAnalysisSoftware-x.x.x-x.x86_64.rpm file to the new system.

2 Install the rpm:

- For installation to the default `/opt/illumina` directory issue the `yum` install command:
  `sudo yum install HiSeqAnalysisSoftware-x.x.x-x.x86_64.rpm`
- If you want to install to a different location you need to use the `rpm` command with a `--prefix` option:
  `sudo rpm -i --prefix /install/path HiSeqAnalysisSoftware-x.x.x-x.x86_64.rpm`
  When using the `--prefix` option, the package will be placed inside the illumina directory within your custom directory.

NOTES

- The RPM will also install mono and python which are used by the HiSeq Analysis Software package.
- If using the `rpm -i --prefix` command to install the HiSeq Analysis Software and there are errors about missing dependencies, install the missing dependencies using `yum` and then try the installation again.
- The RPM is unsigned.
- You need 75 GB space in the opt directory, the default install directory. If that is not available you need to specify a different install location with the `--prefix` option.
- The owner of the installation directory is root but the permissions are 755 so everyone should be able to read and execute the program. However, if you do run into problems with permissions, change the ownership of the installation directory to a group containing all the users.

Symlink

The installation creates the following symlink for the Genomes root directory:

- `/opt/illumina/HiSeqAnalysisSoftware/Genomes -> /opt/illumina/scratch/iGenomes`
  To use a different Genomes root directory update the symlink.

Removal of HiSeq Analysis Software

To uninstall the RPM:

`sudo yum remove HiSeqAnalysisSoftware.x86_64`

To validate the installation you run the following script:
`/opt/illumina/scratch/InstallValidationData/ValidateInstall`

Update this path if you installed to a different location using the `--prefix` option.

A successful validation should dsiplay this output from the script on the screen:

```
Starting Isaac validation run...
Isaac run passed validation
Starting BWA/GATK validation run...
BWA/GATK run passed validation
Log file written to
/opt/illumina/scratch/InstallValidationData/ValidateInstall.log
```

An unsuccessful validation will print one of the following lines instead:

```
Isaac run failed
BWA/GATK run failed
```

Unpacking Reference Genome

NOTE
hg19 is the only large genome available with HiSeq Analysis Software. PhiX is automatically unpacked (to /opt/illumina/scratch/iGenomes) as part of the installation of the HiSeq Analysis Software.

1 Make sure you have python installed and the python unpacking script located in /opt/illumina/HiSeqAnalysisSoftware/unpackIsisReference.py, or, if you specified a custom install directory with the `--prefix` option, in /<Custom_Dir>/illumina/HiSeqAnalysisSoftware/unpackIsisReference.py).

2 Copy the packed reference genome to you computer (e.g. HiSeqAnalysisSoftware_UCSC_hg19.tar.gz).

NOTE
The reference genome is available from http://support.illumina.com/sequencing/sequencing_software/hiseq-analysis-software/downloads.ilmn

3 Issue the unpacking command:

```
python unpackIsisReference.py
--input-file HiSeqAnalysisSoftware_UCSC_hg19.tar.gz
--jobs 2
```

NOTE

- The genome size is 67 GB. If that does not fit in the default genome directory, you must specify a different location.
- The owner of the genome directory is root but the permissions are 755 so everyone should be able to read and execute. However, if you do run into problems with permissions, change the ownership of the genome directory to a group containing all the users.

The command line options are:

Option	Description
`-h, --help`	Show help message and exit
`--input-file=INPUTFILE`	Location of HiSeq Analysis Software reference to unpack
`--tmp=TMP`	Temp directory to use for unpacking (default /tmp)
`--genomes-root=GENOMESROOT`	Location to unpack reference (default /opt/illumina/HiSeqAnalysisSoftware/Genomes)
`--isaac-path=ISAACPATH`	Location of Isaac installation (/opt/illumina/Isaac/stable)
`--keep-tmp`	Don't delete temporary files
`--jobs=NUMJOBS`	Maximum number of parallel operations during Isaac reference unpacking

Setting up SGE for HiSeq Analysis Software

When running on a cluster like Sun Grid Engine (SGE), HiSeq Analysis Software assumes that it has an entire cluster node available. You should not attempt to run other jobs on a node where HiSeq Analysis Software is executing. In order to queue up several analyses without overloading a compute node, it is important to specify the following:

- Set your SGE to `fill_up` mode instead of the default round robin.

When running HiSeq Analysis Software on the cluster (see *Running HiSeq Analysis Software on a Cluster* on page 30), make sure to specify the following:

- Specify the slots that HiSeq Analysis Software will request from a node using the `-pe threaded X` in the `<qsub>` command line.

WARNING

The HiSeq Analysis Softwarewill use up all RAM on the node to which it is assigned. If you do not set up your SGE properly, you can get conflicts with other jobs.

Running HiSeq Analysis Software on Non-Dedicated Servers

Illumina does not recommend running HiSeq Analysis Software on a non-dedicated server without reserving an entire cluster node for HiSeq Analysis Software using the `-pe threaded X` option. HiSeq Analysis Software will consume all available compute resources in a cluster node.

User-Provided Files

Before you start an analysis, you have to modify the files listed below, and copy them to the run folder:

- **SampleSheet.csv**: contains user-specified analysis options for the run. You should modify the settings in this file and save it in the run folder prior to executing an analysis run. See *Sample Sheet Settings* on page 20 for details. You can use Illumina Experiment Manager version 1.5 or greater to create the sample sheet.

The following user-provided files are only needed for the Enrichment analysis workflow:

- **Target manifest file (.txt)**: a tab-delimited values file that specifies targeted regions for the aligner and variant caller, which results in faster analysis times and visualization of results specific to these targeted regions. For Illumina-defined targets when using Nextera Rapid Capture, you can obtain a target manifest file from Illumina. You need to ensure that this file is present in the run folder. The format for this file is specified in *Target Manifest File* on page 26.
- **Probe manifest file (.txt)**: contains a list of all the baits (probes) regions and their chromosome start and end positions. The format for the baits file is specified in *Probe Manifest File Format* on page 26.

Sequencing Software Generated Files

The files listed below are generated by Illumina's sequencing software, and should be present in the Run Folder after a sequencing run, before analysis starts:

- **RunInfo.xml**: Produced by RTA, this file contains information about cycles per read, etc. See *RunInfo.xml File* on page 27.
- **s_x_z.bcl**: Base calls for lane x, cycle y, tile z needs to be present in Data\Intensities\BaseCalls\L00x\Cy.1\. See *BCL Files* on page 27.
- **s_x_yyyy.filter**: Filter file for lane x, tile yyyy needs to be present in Data\Intensities\BaseCalls\L00x\. See *Filter Files* on page 27.
- **Position files**, one of these types:
  - s_x_y.locs or s_x_y.clocs: Locs or compressed locs for lane x, tile y needs to be present in Data\Intensities\L00x\.
  - *_pos: Legacy text files (Data\Intensities\s_x_yyyy_pos.txt) are also supported.

  See *Position Files* on page 27.

Sample Sheet Settings

The sample sheet identifies the samples included in an analysis, and contains the index information used in sample demultiplexing. The Settings section of the Sample Sheet specifies the various analysis parameter settings to be used in the analysis. Each line in the Settings section contains a setting name in the first column and a value in the second column. Settings are not case sensitive.

Sample sheet settings are different for the Enrichment and WGS workflows.

NOTE
The sample sheet format used by the HiSeq Analysis Software is not the same as the one used in CASAVA or in AVC. Do not use CASAVA-style sample sheets when running the HiSeq Analysis Software from the command line.

The HiSeq Analysis Software is packaged with the human hg19 build reference genome provided by Illumina through iGenomes. Currently, HiSeq Analysis Software does not support large custom genomes.

NOTE
Isis does not use the [Reads] section of the sample sheet, so if you want to analyze shorter reads than you actually sequenced, you need to modify the RunInfo.xml file.

Sample Sheet Settings for Enrichment Workflow

The table below describes how to set up the sample sheets for the Enrichment workflow. You must create a sample sheet using IEM or manually using a text editor.

IEM

The easiest way to set up an Enrichment workflow sample sheet is by using the Illumina Experiment Manager (IEM), version 1.5 or later. IEM creates your sample sheet using a wizard-based application. IEM provides a feature for recording parameters for your sample plate, such as sample ID, dual indices, and other parameters applicable to your samples.

WARNING
Do not use earlier versions of IEM (below v1.5). This will generate a sample sheet that was formatted for CASAVA. Do not use CASAVA-style sample sheets when running HiSeq Analysis Software.

Use the following settings:

- At *Sample Prep Kit* selection in IEM, choose:
  - *Instrument Selection*: **HiSeq 2500/2000/1000** (etc)
  - *Application*: **HiSeq Enrichment**
- Select **Use Adapter Trimming** if your Illumina assay uses adapters.
  Shorter inserts can lead to sequencing into the adapter, and this feature helps filter out adapter sequence from the final sequence data.

You can download IEM from the Illumina website at from http://support.illumina.com/; for instructions on how to use the IEM application, see the *Illumina Experiment Manager User Guide* and *Quick Reference Card*. IEM can be run on any Windows platform.

Parameters

If you want to further customize the sample sheet for the Enrichment workflow beyond default settings or settings specified by IEM, you can use the following parameters in the [Settings] area of the sample sheets:

Parameter	Description
Adapter	Specify the 5' portion of the adapter sequence to prevent reporting sequence beyond the sample DNA. Illumina recommends adapter trimming for Nextera libraries (adapter sequence **CTGTCTCTTATACACATCT**).
AdapterRead2	Specify the 5' portion of the Read 2 adapter sequence to prevent reporting sequence beyond the sample DNA. Use this setting to specify a different adapter other than the one specified in the **Adapter** setting. If not specified, the **Adapter** setting is used for read 2.
Aligner	Specify the method for aligning reads against the reference genome. BWA is default and must be used for the Enrichment workflow.
BaitManifestFileName	Specifies the full path to the probe (bait) manifest file, or relative path if the probe manifest is present in the runfolder.
EnrichmentMaxRegionStatisticsCount	The number of output rows in the file EnrichmentStatistics.xml is limited to 40,000 by default to prevent problems with displaying exome sized data sets. This can be overridden with the sample sheet setting EnrichmentMaxRegionStatisticsCount Default: 40000
FlagPCRDuplicates	Settings are 0 or 1. Default is 0, do not flag. If set to 1 or TRUE, PCR duplicates are flagged in the BAM files and not used for variant calling. PCR duplicates are defined as two clusters from a paired-end run where both clusters have the exact same alignment positions for each read.
IndelRealignment	Enables indel realignment with GATK. Default is GATK; GATK must be used for the Enrichment workflow.
IndelRepeatFilterCutoff	This setting filters indels if the reference has a 1-base or 2-base motif repeated over eight times (by default) next to the variant. Default value for GATK: 8
MinimumCoverageDepth	The GATK variant caller does not report variants if the coverage depth at that location is less than the specified threshold. Decreasing this value will increase variant calling sensitivity, but raise the risk of false positives. Default value for GATK: 20
MinQScore	This setting specifies the minimum base Q-score to use as input to variant calling.
PicardHsMetrics	Default is 0. If set to 1 or TRUE, this setting allows the user to generate Picard HS metrics for the given probe (bait) manifest and target manifest file. If the bait file is not explicitly identified, the target manifest file is also used as the bait file.

Parameter	Description
QualityScoreTrim	If set to a value > 0, then the 3' ends of non-indexed reads with low quality scores are trimmed. Trimming is automatically applied by default for a value of 15 when using BWA for alignment.
StrandBiasFilter	This setting filters variants with a significant bias in read-direction. Variants filtered out in this way will have **sb** in the filter column of the VCF file, instead of **PASS**. Default value for GATK: -10
VariantCaller	Specify the variant caller. Default is GATK; this must be used in the Enrichment workflow.
VariantFilterQualityCutoff	This setting filters variants if the variant quality score is less than the specified threshold. Default value for GATK: 30
VariantFrequencyFilterCutoff	This setting filters variants with a frequency less than the specified threshold. Default value for GATK: 0.20
VariantMinimumGQCutoff	This setting filters sites if the genotype quality (GQ) is less than the threshold. Default value for GATK: 30

IEM

The easiest way to set up an Enrichment workflow sample sheet is by using the Illumina Experiment Manager (IEM), version 1.5 or later. IEM creates your sample sheet using a wizard-based application. IEM provides a feature for recording parameters for your sample plate, such as sample ID, dual indices, and other parameters applicable to your samples.

WARNING

Do not use earlier versions of IEM (below v1.5). This will generate a sample sheet that was formatted for CASAVA. Do not use CASAVA-style sample sheets when running HiSeq Analysis Software.

Use the following settings:

- When prompted to select a Sample Prep Kit in IEM, choose these options:
  - Instrument Selection: HiSeq 2000/1000 (etc)
  - Application: HiSeq Enrichment
- When using IEM for sample sheet generation, make sure you select Use Adapter Trimming when you create your sample sheet if your Illumina assay uses adapters. Shorter inserts can lead to sequencing into the adapter, and this feature helps filter out adapter sequence from the final sequence data.
- The manifest specifies targeted regions for the aligner and variant caller.
- When a run completes for HiSeq, HiScan SQ, or GA instruments, the product manifest file must be copied to the run folder before running the analysis. Also, a sample sheet must be created using IEM or manually using a text editor..

You can download IEM from the Illumina website at from http://support.illumina.com/; for instructions on how to use the IEM application, see the *Illumina Experiment Manager User Guide* and *Quick Reference Card*. IEM can be run on any Windows platform.

Example Enrichment Sample Sheet

An example of a sample sheet for the Enrichment workflow is shown below.

```
[Header]
IEMFileVersion,4
Investigator Name,John Doe
Project Name,Project A
Experiment Name,JD_0004
Date,12/5/2012
Workflow,Enrichment
Application,Enrichment
Assay,Nextera Enrichment
Description
Chemistry,Amplicon
[Manifests]
A,NexteraRapidCapture_Exome_TargetedRegions.txt

[Reads]
76
76

[Settings]
Adapter,CTGTCTCTTATACACATCT
Aligner,BWA
FlagPCRDuplicates,TRUE

[Data]
Sample_ID,Sample_
   Name,Lane,Index,GenomeFolder,Manifest,Runfolder
12plx_NA12144,,1+2,TAAGGCGA,Homo_
   sapiens/UCSC/hg19/Sequence/WholeGenomeFASTA,A,/Test01/Runs/T
   est/Enrich/130206_D0003_0508_AH01ABCDEF
```

Sample Sheet Settings for WGS

The table below describes the parameters that can be defined in the [Settings] area of sample sheets for the WGS workflow.

Parameter	Description
Adapter	Specify the 5' portion of the adapter sequence to prevent reporting sequence beyond the sample DNA. See the sample prep documentation for your sample. Nextera Mate Pair Libraries—Illumina recommends multiple adapter trimming for Nextera Mate Pair Libraries (adapter sequences CTGTCTCTTATACACATCT and AGATGTGTATAAGAGACAG) To trim two or more adapters, separate the sequences by a plus (+) sign. (i.e. CTGTCTCTTATACACATCT+ AGATGTGTATAAGAGACAG).

Parameter	Description
AdapterRead2	Specify the 5' portion of the Read 2 adapter sequence to prevent reporting sequence beyond the sample DNA. Use this setting to specify a different adapter other than the one specified in the **Adapter** setting. If not specified, the **Adapter** setting is used for read 2.
Aligner	Specify the method for aligning reads against the reference genome. For WGS workflow, use Isaac.
FlagPCRDuplicates	Settings are 0 or 1. Default is 1, filtering. If set to 1, PCR duplicates are flagged in the BAM files and not used for variant calling. PCR duplicates are defined as two clusters from a paired-end run where both clusters have the exact same alignment positions for each read.
QualityScoreTrim	If set to a value > 0, then the 3' ends of non-indexed reads with low quality scores are trimmed.
VariantCaller	Specify the variant caller Starling (this is the algorithm in Isaac Variant Caller)
IndelRepeatFilterCutoff	This setting filters indels if the reference has a 1-base or 2-base motif repeated over eight times (by default) next to the variant. Default value for Isaac variant caller: 8
MinimumCoverageDepth	This setting does not report variants if the coverage depth at that location is less than the specified threshold. Default value for Isaac variant caller: 0
MinQScore	This setting specifies the minimum base Q-score to use as input to variant calling. Default value for Isaac variant caller: 17
VariantFilterQualityCutoff	This setting filters variants if the variant quality score is less than the specified threshold. Default value for Isaac variant caller: 20
VariantFrequencyFilterCutoff	This setting filters variants with a frequency less than the specified threshold. Default value for Isaac variant caller: 0.20
VariantMinimumGQCutoff	This setting filters sites if the genotype quality (GQ) is less than the threshold. Default value for Isaac variant caller: 20

Example WGS Sample Sheet

An example of a sample sheet for the Whole Genome Sequencing workflow is shown below.

```
[Header]
Investigator Name,John Doe,
Experiment Name,Project A
Date,09/24/2012
Workflow,Resequencing

[Settings]

[Data]
SampleID,SampleName,GenomeFolder,RunFolder,Lanes
St1a,St1a,Homo_sapiens/UCSC/hg19/Sequence/WholeGenomeFASTA,
   /scratch/130114_HF1003_0083_BC0I1EACXX,1+2
St2a,St2a,Homo_
   sapiens/UCSC/hg19/Sequence/WholeGenomeFASTA,/scratch/130114_
   HF1009_0173_BC1341ACXX,1+2
```

Manifest Files

The Enrichment workflow uses two manifest files, a Target Manifest file and a Probe Manifest file. Fixed content manifest files are available in the Downloads section of the product support page. For Nextera Rapid Capture Exome and Expanded Exome target manifest and probes manifest files, go to www.illumina.com | Support | Kits and Reagents | Nextera Rapid Capture Exome | Downloads, or click http://support.illumina.com/sequencing/sequencing_kits/nextera_rapid_capture_exome_kit/downloads.ilmn.

Target Manifest File

The Enrichment workflow target manifest file is a tab-delimited file with two major sections. HiSeq Analysis Software ignores the version of the manifest but parses the `Reference Genome` and the `Regions` section. You need to copy the target manifest file into the run folder before analysis starts.

The following fields are used in the `Regions` section:

- **Name**—Unique user-specified name for the target.
- **Chromosome**—Chromosome from which the target originates.
- **Start**—1-based coordinate start position of the target.
- **End**—1-based and inclusive coordinate of the end position of the target.
- **Upstream Probe Length**—This field should be zero.
- **Downstream Probe Length**—This field should be zero.

Figure 3 Example Manifest File

```
[Header]
Manifest Version        1
ReferenceGenome Homo_sapiens\UCSC\hg19\Sequence\WholeGenomeFASTA

[Regions]
Name    Chromosome      Start   End     Upstream Probe Length   Downstream Probe Length
CEX-chr1-13403-13639    chr1    13403   13639   0       0
CEX-chr1-69089-70010    chr1    69089   70010   0       0
CEX-chr1-324439-325605  chr1    324439  325605  0       0
CEX-chr1-664485-665108  chr1    664485  665108  0       0
CEX-chr1-721405-721912  chr1    721405  721912  0       0
CEX-chr1-762080-762571  chr1    762080  762571  0       0
CEX-chr1-861320-861395  chr1    861320  861395  0       0
```

Probe Manifest File Format

The Enrichment workflow probe manifest file (baits file) is a tab-delimited file with a list of all the baits regions and their chromosome start and end positions. The probe manifest file is optional, and is only needed for Picard metrics. If you use the probe manifest file, copy it into the run folder before analysis starts, or else provide a full path.

The following fields are used in the `Regions` section:

- **Name**—Unique user-specified name for the bait.
- **Chromosome**—Chromosome from which the bait originates.
- **Start**—1-based coordinate start position of the bait.
- **End**—1-based and inclusive coordinate of the end position of the bait.
- **Upstream Probe Length**—This field should be zero.
- **Downstream Probe Length**—This field should be zero.

Sequencing Run Files

HiSeq Analysis Software needs several sequencing run files, which are described below.

RunInfo.xml File

The top level Run Folder contains a RunInfo.xml file. The file RunInfo.xml (normally generated by SCS/HCS) identifies the boundaries of the reads (including index reads).

The XML tags in the RunInfo.xml file are self-explanatory.

BCL Files

The BCL files can be found in the BaseCalls directory inside the run directory:

`Data/Intensities/BaseCalls/L<lane>/C<cycle>.1`

They are named as follows:

`s_<lane>_<tile>.bcl` or `s_<lane>_<tile>.bcl.gz`

The BCL files can be gzip compressed and the FASTQ generator accepts either format. The BCL files are binary base call files with the format described below.

Bytes	Description	Data type
Bytes 0–3	Number N of cluster	Unsigned 32bits little endian integer
Bytes 4–(N+3) Where N is the cluster index	Bits 0-1 are the bases, respectively [A, C, G, T] for [0, 1, 2, 3]: bits 2-7 are shifted by two bits and contain the quality score. All bits '0' in a byte is reserved for no-call.	Unsigned 8bits integer

Filter Files

The filter files can be found in the BaseCalls directory.

The *.filter files are binary files containing filter results; the format is described below.

Bytes	Description
Bytes 0–3	Zero value (for backwards compatibility)
Bytes 4–7	Filter format version number
Bytes 8–11	Number of clusters
Bytes 12–(N+11) Where N is the cluster number	unsigned 8-bits integer: • Bit 0 is pass or failed filter

Position Files

The BCL to FASTQ converter can use different types of position files and will expect a type based on the version of RTA used:

- *.locs: the locs files can be found in the Intensities/L<lane> directories.
- *.clocs: the clocs files are compressed versions of locs file and can be found in the Intensities/L<lane> directories.

Enrichment Test Data Set

A data set for testing the Enrichment workflow in HiSeq Analysis Software can be obtained from the Downloads section of the Nextera Rapid Capture Exome support webpage. Go to www.illumina.com | Support | Kits and Reagents | Nextera Rapid Capture Exome | Downloads, or click http://support.illumina.com/sequencing/sequencing_kits/nextera_rapid_capture_exome_kit/downloads.ilmn. Details on running the workflow, expected run time, input and output files can be found in the README.txt included in the download.

NOTE

The time needed for download of test data sets may vary depending on network performance and data set size.

Running HiSeq Analysis Software under Linux

The following section details running HiSeq Analysis Software under Linux.

1 Identify a run folder for your analysis. This can be but does not need to be the original run folder for the HiSeq flow cell.

2 Create a sample sheet named SampleSheet.csv in the top level of the run folder. See *Sample Sheet Settings* on page 20 for more information.

3 Start the analysis:

```
/path/to/illumina/HiSeqAnalysisSoftware/RunLatest -r
   /path/to/your/RunFolder
```

Where:

- `-r /path/to/your/RunFolder/` is the path to the run folder.

The analysis results will be written (by default) to the `Data/Intensities/BaseCalls/Alignment` subfolder of the run folder. This can be overridden using the `-a` command-line option.

Please note that if you log out of your terminal session, the HiSeq Analysis Software command may be prematurely terminated. You can retain terminal by adding the following to your command line:

```
nohup /path/to/illumina/HiSeqAnalysisSoftware/RunLatest -r
   /path/to/your/RunFolder 2>&1 > logfile.txt &
```

TIP

- You can test the Enrichment workflow with an Illumina-provided data set; see *Enrichment Test Data Set* on page 28 for more information.
- If you have sample sheets with many samples, you can split them into smaller sample sheets and run each of these smaller analyses on separate computers to optimize your analysis time.

Options

The following command-line options can be used to customize the analysis run:

Option	Description
`-r`	Path to run folder (required).
`-a`	Path to analysis results folder. Default: the Data/Intensities/BaseCalls/Alignment subfolder of the run folder
`-c`	Resume analysis at indicated checkpoint (see *Resuming Analysis using Checkpoints* on page 31).

NOTE

- The analysis parameters are defined in the sample sheet.
- The workflow parameter settings used in the run are recorded in the output file IsisLog.txt. See *Output Files* on page 32 for a description.

Running HiSeq Analysis Software on a Cluster

When running HiSeq Analysis Software on a cluster, you need to generate a shell script that will submit the analysis job to your queue manager. A simple shell script looks like this:

```
#!/usr/bin/env bash
/path/to/illumina/HiSeqAnalysisSoftware/RunLatest -r
   /path/to/your/RunFolder
```

The analysis software assumes that it has an entire cluster node available. In order to queue up several analyses without overloading a compute node, it is important to specify that each analysis job requires multiple slots on the node. SGE can be configured to support the threaded parallel environment, so that jobs that consume an entire node can be submitted.

For example, if you are using an SGE and want to submit a job with the shell script `Analysis.sh` that reserves X slots on a compute node from the queue `<queue name>`, use a command like this:

```
qsub -pe threaded X -q <queue name> Analysis.sh
```

TIP

- If you have sample sheets with many samples, you can split them into smaller sample sheets and run each of these smaller analyses on separate nodes to optimize your analysis time.
- You can test the Enrichment workflow with an Illumina-provided data set; see *Enrichment Test Data Set* on page 28 for more information.

NOTE

See *Running HiSeq Analysis Software on Different System Types* on page 18 for the required SGE configuration

Options

The following command-line options can be used for customizing when running HiSeq Analysis Software on a cluster:

Option	Description
`-pe threaded X`	Requests an X-core node. The recommended value is the number of slots in a node.
`-q <queue name>`	Specifies the queue <queue name> for the node request defined in `-pe threaded X`. Optional.
`-M <email ID>`	Sends an email to <email ID> when the job is complete or when the job exits because of an error.
`-N <Job name>`	Associates a job name to the job. The name should begin with an alphabetical character.

NOTE

If your installation does not support SGE, these options will not work.

Resuming Analysis using Checkpoints

The Checkpoint.txt file, generated by the HiSeq Analysis Software, identifies the last successfully executed step in the workflow. These are useful if the analysis has unexpectedly stopped, or if you want to change an analysis parameter and don't want to start the analysis from the start.

For the different workflow configurations, the following checkpoints are generated for each analysis step:

Table 1 Checkpoints for Workflow Configurations

	WGS Workflow without Picard Metrics	Enrichment Workflow without Picard Metrics	Enrichment Workflow with Picard Metrics
Demultiplexing	N/A	1	1
Generating FastQ files	N/A	2	2
Alignment	1	3	3
Variant Analysis	2	4	4
Calculate Picard HS Metrics	N/A	N/A	5
Statistics Evaluation	3	5	6

The analysis can be resumed from any of these checkpoints by using the `-c` argument, as below:

```
/path/to/illumina/HiSeqAnalysisSoftware/RunLatest -r
   /path/to/your/RunFolder -c X -a /path/to/AlignmentFolder/
```

With `X` being the last successfully completed step or the step after which analysis should be resumed.

Common Data Files

File Name	Description
*.bam files	Contains aligned reads for a given sample. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment. For more information, see *BAM Files* on page 35.
*.vcf files	Contains information about variants found at specific positions in a reference genome. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment. For more information, see *VCF Files* on page 36.
*.genome.vcf.gz files	All variant and non-variant sites in the genome for SampleName. Block level compression is used (bgzip) to generate this file so that it is compatible with tabix indexing. Similarly, tabix generated index file corresponding to the genome.vcf.gz file are also generated. This information is represented using the Genome VCF (gVCF) conventions for VCF. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment. For more information, see *gVCF Files* on page 39 or https://sites.google.com/site/gvcftools/home/about-gvcf.

Common Metrics Files

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 <Run folder>\Data\Intensities\BaseCalls\Alignment.
*.CoverageHistogram.txt	Contains coverage depth information for every chromosome. This file can be used to generate a coverage histogram. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
DemultiplexSummaryF1L1.txt	Reports demultiplexing results in a table with one row per tile and one column per sample. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.

File Name	Description
ErrorsAndNoCallsByLaneTileReadCycle.csv	A comma-separated values file that contains the percentage of errors and no-calls for each tile, read, and cycle. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
Mismatch.htm	Contains histograms of mismatches per cycle and no-calls per cycle for each tile. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
Summary.htm	Contains a summary web page generated from Summary.xml. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
Summary.xml	Contains a summary of mismatch rates and other base calling results. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.

Enrichment Metrics Files

File Name	Description
*.coverage.csv	Contains a summary of the mean coverage and standard deviation for each region in the manifest. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
*.enrichment_summary.csv	A summary of performance metrics generated by the enrichment workflow. See *Summary of Enrichment Statistics* on page 43. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
EnrichmentStatistics.xml	Contains high-level summary statistics forabout the Enrichment run. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
*.gaps.csv	Contains a summary of gaps, their position and mean gap coverage in the alignment. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
*.HsMetrics.txt	Summary of hybrid selection metrics generated by the Picard suite of tools (CalculateHsMetrics.jar). More details can be found here http://picard.sourceforge.net/picard-metric-definitions.shtml Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.

File Name	Description
*_regions_Manifest_ Intervals.txt	A list of regions used to generate summary statistics. This is generated from the target manifest file provided in the sample sheet.

Common Run Progress Files

File Name	Description
QueuedForAnalysis.txt	Marker file that lists the HiSeq Analysis Software software version and indicates that analysis has begun. Located at the root level of the run folder.
IsisLog.txt	Processing log that describes every step that occurred during analysis of the current run folder. This file does not contain error messages. Located at the root level of the run folder.
IsisError.txt	Processing log that lists any errors that occurred during analysis. This file is present only if errors occurred. Located at 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 at the root level of the run folder.
Checkpoint.txt	Identifies the last successfully executed step in the workflow. This file is useful if a run failed and needs to be resumed. See *Resuming Analysis using Checkpoints* on page 31. Located in <Run folder>\Data\Intensities\BaseCalls\Alignment.
RunInfo.xml	Contains information about the run. Located at the root level of the run folder.

BAM Files

The Sequence Alignment/Map (SAM) format is a generic alignment format for storing read alignments against reference sequences, supporting short and long reads (up to 128 Mb) produced by different sequencing platforms. SAM is a text format file that is human-readable. The Binary Alignment/Map (BAM) keeps exactly the same information as SAM, but in a compressed, binary format that is only machine-readable.

Detailed Description

The file naming convention for aligned reads in BAM format is as follows: SampleName_S#.bam (where # is the sample number determined by ordering in the sample sheet).

Go to http://samtools.sourceforge.net/SAM1.pdf to see the exact SAM specification.

BWA adds some custom fields to the BAM output. See http://bio-bwa.sourceforge.net/bwa.shtml#4 for a description.

VCF is a text file format which contains information about variants found at specific positions in a reference genome. The file format consists of meta-information lines, a header line, and then data lines. Each data line contains information about a single variant.

More information is available here: http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41.

VCF Files from Isaac Variant Caller (WGS Workflow)

The file naming convention for VCF files is as follows: SampleName_S#.vcf (where # is the sample number determined by ordering in the sample sheet).

The header of the VCF file describes the tags used in the remainder of the file and has the column header. It looks like this:

```
##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=DPI,Number=1,Type=Integer,Description="Read depth
   associated with indel, taken from the site preceding the indel.">
##INFO=<ID=TI,Number=.,Type=String,Description="Transcript ID">
...
##INFO=<ID=IDREP,Number=A,Type=Integer,Description="Number of times RU
   is repeated in indel allele.">
##FILTER=<ID=IndelConflict,Description="Locus is in region with
   conflicting indel calls">
...
##FILTER=<ID=HighDepth,Description="Locus depth is greater than 3x the
   mean chromosome depth">
##fileDate=20130226
##source=starling
##source_version=2.0.2
##startTime=Tue Feb 26 04:18:06 2013
##cmdline=/illumina/development/Isis/2.3.20/ ... -genome-size 2861343702
##reference=file:///illumina/.../Homo_
   sapiens/UCSC/hg19/Sequence/WholeGenomeFasta/genome.fa
##contig=<ID=chrM,length=16571>
...
##contig=<ID=chrY,length=59373566>
##content=starling small-variant calls
##SnvTheta=0.001
##IndelTheta=0.0001
##MaxDepth_chr1=170.97
...
##MaxDepth_chrY=15.24
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE
```

A sample line of the VCF file is shown below. The data that is used to populate each column is also described:

```
chr22 16285888 rs76548004 T C 17 d15;q20 DP=11;TI=NM_
   001136213;GI=POTEH;CD GT:GQ 1/0:17
```

Setting	Description
CHROM	The chromosome of the reference genome. Chromosomes appear in the same order as the reference FASTA file (generally karyotype order)
POS	The 1-based position of this variant in the reference chromosome. The convention for .vcf files is that, for SNPs, this is the reference base with the variant; for indels or deletions, this is the reference base **immediately before** the variant. Variants are ordered by position.
ID	The rs number for the snp obtained from dbSNP. If there are multiple rs numbers at this location, the list is semi-colon delimited. If no dbSNP entry exists at this position, the missing value ('.') is used.
REF	The reference genotype. For example, a deletion of a single T could be represented by reference TT and alternate T.
ALT	The allele(s) that differ from the reference read. For example, an insertion of a single T could be represented by reference A and alternate AT.
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 should have a 0.1% error rate. Note that many variant callers assign quality scores (based on their statistical models) which are high relative to the error rate observed in practice.
FILTER	If all filters are passed, the' PASS' is written. The possible filters are as follows: • q20 – The variant score is less than 20. (Configurable using the VariantFilterQualityCutoff setting in the config file) • r8 – For an Indel, the number of adjacent repeats in the reference (of a 1- or 2-base repeat) is greater than 8. (Configurable using the IndelRepeatFilterCutoff setting in the config file)
INFO	These are the possible entries in the INFO column: • AD – Entry of the form X,Y where X is the number of reference calls, Y the number of alternate calls. • CD – A flag indicating that the snp occurs within the coding region of at least one refGene entry • DP – The depth (number of base calls aligned to a this position) • GI – A comma separated list of gene IDs read from refGene • NL – Noise level; an estimate of base calling noise at this position. • TI – A comma separated list of transcript IDs read from refGene • SB – Strand bias at this position. • VF – Variant frequency. The number of reads supporting the alternate allele.

Setting	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. The available fields are as follow: AD – Entry of the form X,Y where X is the number of reference calls, Y the number of alternate calls GQ – Genotype quality GT – Genotype. 0 corresponds to the reference base, 1 corresponds to the first entry in the ALT column, 2 corresponds to the second entry in the ALT column, etc. The '/' indicates that there is no phasing information. NL – Noise level; an estimate of base calling noise at this position SB – Strand bias at this position. Larger negative values indicate more bias; values near zero indicate little strand bias. VF – Variant frequency. The percentage of reads supporting the alternate allele.
SAMPLE	The sample column gives the values specified in the FORMAT column. One MAXGT sample column is provided for the normal genotyping (assuming the reference). For reference, a second column is provided for genotyping assuming the site is polymorphic.

VCF Files from GATK (Enrichment Workflow)

The GATK variant caller generates VCF files that are slightly different from those generated by Starling (Isaac Variant Caller). The core fields `CHROM`, `POS`, `ID`, `REF`, `ALT`, and `QUAL` are the same; see description above. The customizable fields `FILTER`, `INFO`, and `FORMAT` are different, and are defined in the header section; see http://www.broadinstitute.org/gatk/guide/article?id=1268 for more information.

gVCF Files

What is it?

HiSeq Analysis Software also produces the genome Variant Call Format file (gVCF). gVCF was developed to store sequencing information for both variant and non-variant positions, which is required for human clinical applications. gVCF is a set of conventions applied to the standard variant call format (VCF) 4.1 as documented by the 1000 Genomes Project. These conventions allow representation of genotype, annotation, and other information across all sites in the genome in a compact format. Typical human whole genome sequencing results expressed in gVCF with annotation are less than 1 Gbyte, or about 1/100 the size of the BAM file used for variant calling. If you are performing targeted sequencing, gVCF is also an appropriate choice to represent and compress the results.

gVCF is a text file format, stored as a gzip compressed file (*.genome.vcf.gz). Compression is further achieved by joining contiguous non-variant regions with similar properties into single 'block' VCF records. To maximize the utility of gVCF, especially for high stringency applications, the properties of the compressed blocks are conservative -- thus block properties like depth and genotype quality reflect the minimum of any site in the block. The gVCF file can be indexed (creating a .tbi file) and used with existing VCF tools such as tabix and IGV, making it convenient both for direct interpretation and as a starting point for tertiary analysis.

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

The following conventions are used in the Isaac Variant Caller gVCF files.

Samples per File

There is only one sample per gVCF file.

Non-Variant Blocks Using END Key

Continuous non-variant segments of the genome can be represented as single records in gVCF. These records use the standard 'END" INFO key to indicate the extent of the record. Even though the record can span multiple bases, only the first base is provided in the REF field to reduce file size.

The following is a simplified segment of a gVCF file, describing a segment of non-variant calls (starting with an A) on chromosome 1 from position 51845 to 51862.

```
##INFO=<ID=END,Number=1,Type=Integer,Description="End position
   of the variant described in this record">#CHROM POS ID REF
   ALT QUAL FILTER INFO FORMAT NA19238chr1 51845 . A . . PASS
   END=51862
```

Any fields provided for a block of sites such as read depth (using the DP key), will show the minimum value observed among all sites encompassed by the block. Each sample value shown for the block, such as the depth (using the DP key), is restricted to a range where the maximum value is within 30% or 3 of the minimum. i.e. for sample value range [x,y], y <= x+max(3,x*0.3). This range restriction applies to each of the ample values printed out in the final block record.

Indel Regions

Note that sites which are "filled in" inside of deletions have additional changes:

All deletions:

- Sites inside of any deletion are marked with the deletion's filters, in addition to any filters which have already been applied to the site.
- Sites inside of deletions cannot have a genotype or alternate allele quality score higher than the corresponding value from the enclosing indel.

Heterozygous deletions:

- Sites inside of heterozygous deletions are altered to have haploid genotype entries (e.g. "0" instead of "0/0", "1" instead of "1/1").
- Heterozygous SNV calls inside of heterozygous deletions are marked with the "SiteConflict" filter and their genotype is unchanged.

Homozygous deletions:

- Homozygous reference and no-call sites inside of homozygous deletions have genotype "."
- Sites inside of homozygous deletions which have a non-reference genotype are marked with a “SiteConflict” filter, and their genotype is unchanged.
- Site and genotype quality are set to "."

The above modifications reflect the notion that the site confidence is bound by the enclosing indel confidence.

Also note that on occasion, the variant caller will produce multiple overlapping indel calls which cannot be resolved into two haplotypes. If this occurs all indels and sites in the region of the overlap will be marked with the “IndelConflict” filter (see below).

Genotype Quality for Variant and Non-variant Sites

The gVCF file uses an adapted version of genotype quality for variant and non-variant site filtration. This value is associated with the key GQX. The GQX value is intended to represent the minimum of {Phred genotype quality assuming the site is variant, Phred genotype quality assuming the site is non-variant}. The reason for using this is to allow a single value to be used as the primary quality filter for both variant and non-variant sites. Filtering on this value corresponds to a conservative assumption appropriate for applications where reference genotype calls must be determined at the same stringency as variant genotypes, ie:

- An assertion that a site is homozygous reference at GQX >= 30 is made assuming the site is variant.
- An assertion that a site is a non-reference genotype at GQX >= 30 is made assuming the site is non-variant.

Section Descriptions

The gVCF file contains the following sections:

- Meta-information lines start with ## and contain meta-data, config information, and define the values that the INFO, FILTER and FORMAT fields can have.
- The header line starts with # and names the fields that the data lines use. These are #CHROM, POS, ID,REF, ALT, QUAL, FILTER, INFO, FORMAT, followed by one or more sample columns.
- Data lines that contain information about one or more positions in the genome.

Note that if you extract the variant lines from a gVCF file, you produce a conventional variant VCF file.

Field Descriptions

The fixed fields #CHROM, POS, ID, REF, ALT, QUAL are defined in the VCF 4.1 standard provided by the 1000 Genomes Project, while the fields ID, INFO, FORMAT, and sample are described in the meta-information. Descriptions are provided below.

Field Description

- **CHROM**: Chromosome: an identifier from the reference genome or an angle-bracketed ID String ("<ID>") pointing to a contig.
- **POS**: Position: The reference position, with the 1st base having position 1. Positions are sorted numerically, in increasing order, within each reference sequence CHROM. There can be multiple records with the same POS. Telomeres are indicated by using positions 0 or N+1, where N is the length of the corresponding chromosome or contig.
- **ID**: Semi-colon separated list of unique identifiers where available. If this is a dbSNP variant it is encouraged to use the rs number(s). No identifier should be present in more than one data record. If there is no identifier available, then the missing value should be used.
- **REF**: Reference base(s): A,C,G,T,N; there can be multiple bases. The value in the POS field refers to the position of the first base in the string. For simple insertions and deletions in which either the REF or one of the ALT alleles would otherwise be null/empty, the REF and ALT strings include the base before the event (which is reflected in the POS field), unless the event occurs at position 1 on the contig in which case they include the base after the event. If any of the ALT alleles is a symbolic allele (an angle-bracketed ID String "<ID>") then the padding base is required and POS denotes the coordinate of the base preceding the polymorphism.
- **ALT**: Comma separated list of alternate non-reference alleles called on at least one of the samples. Options are:
  - Base strings made up of the bases A,C,G,T,N
  - angle-bracketed ID String (”<ID>”)
  - breakend replacement string as described in the section on breakends.

  If there are no alternative alleles, then the missing value should be used.
- **QUAL**: Phred-scaled quality score for the assertion made in ALT. i.e. -10log_10 prob (call in ALT is wrong). If ALT is ”.” (no variant) then this is -10log_10 p(variant), and if ALT is not ”.” this is -10log_10 p(no variant). High QUAL scores indicate high confidence calls. Although traditionally people use integer phred scores, this field is permitted to be a floating point to enable higher resolution for low confidence calls if desired. If unknown, the missing value should be specified. (Numeric)
- **FILTER**: PASS if this position has passed all filters, i.e. a call is made at this position. Otherwise, if the site has not passed all filters, a semicolon-separated list of codes for filters that fail. gVCF files use the following values:
  - **PASS**: position has passed all filters
  - IndelConflict: Locus is in region with conflicting indel calls.
  - SiteConflict: Site genotype conflicts with proximal indel call. This is typically a heterozygous SNV call made inside of a heterozygous deletion.
  - LowGQX: Locus GQX is less than 30 or not present.
  - HighDPFRatio: The fraction of basecalls filtered out at a site is greater than 0.3.
  - HighSNVSB: SNV strand bias value (SNVSB) exceeds 10.
  - HighSNVHPOL: SNV contextual homopolymer length (SNVHPOL) exceeds 6.
  - HighREFREP: Indel contains an allele which occurs in a homopolymer or dinucleotide track with a reference repeat greater than 8.
  - HighDepth: Locus depth is greater than 3x the mean chromosome depth.
- **INFO**: Additional information. INFO fields are encoded as a semicolon-separated series of short keys with optional values in the format: <key>=<data>[,data]. gVCF

files use the following values:

- **END**: End position of the region described in this record.
- BLOCKAVG_min30p3a: Non-variant site block. All sites in a block are constrained to be non-variant, have the same filter value, and have all sample values in range [x,y], y <= max(x+3,(x*1.3)). All printed site block sample values are the minimum observed in the region spanned by the block.
- SNVSB: SNV site strand bias.
- SNVHPOL: SNV contextual homopolymer length.
- CIGAR: CIGAR alignment for each alternate indel allele
- RU: Smallest repeating sequence unit extended or contracted in the indel allele relative to the reference. RUs are not reported if longer than 20 bases.
- REFREP: Number of times RU is repeated in reference.
- IDREP: Number of times RU is repeated in indel allele.

- **FORMAT**: Format of the sample field. FORMAT specifies the data types and order of the subfields. gVCF files use the following values:
  - GT: Genotype.
  - GQ: Genotype Quality.
  - GQX: Minimum of {Genotype quality assuming variant position,Genotype quality assuming non-variant position}.
  - DP: Filtered basecall depth used for site genotyping.
  - DPF: Basecalls filtered from input prior to site genotyping.
  - AD: Allelic depths for the ref and alt alleles in the order listed. For indels this value only includes reads which confidently support each allele (posterior prob 0.999 or higher that read contains indicated allele vs all other intersecting indel alleles).
  - DPI: Read depth associated with indel, taken from the site preceding the indel.
- **SAMPLE**: Sample fields as defined by the header.

Summary of Enrichment Statistics

The Enrichment workflow provides summary statistics for the target file provided. These statistics are present in the SampleName_S1.enrichment_summary.csv file in the Alignment folder. The Enrichment Summary Report is intended as a simple to use report on a wide range of metrics. A brief description of the results is below.

NOTE

- The metrics generated by the Enrichment workflow are comparable, but not identical, to those generated by Picard HS metrics. HiSeq Analysis Software generated metrics do not filter PCR duplicates.
- The metrics also don't incorporate padding around targeted regions, which will make the per cent enrichment metrics lower than tools that incorporate padding, such as Picard HsMetrics.
- Illumina cannot provide technical support on the interpretation of the outputs of 3rd party tools such as Picard. Creation of input files for and automatic execution of Picard are provided as a convenience for our customers. Please refer to the Picard user documentation for any questions about this tool.

Statistic	Definition
Sample ID	ID of the Sample
Run Folder	Path to the Run folder
Total aligned bases	Total aligned bases
Targeted aligned bases	Total aligned bases in the target region
Base enrichment (not padded)	100*(Total Aligned Bases in Targeted Regions/ Total Aligned Bases)
Percent duplicate paired reads	Percentage of paired reads which have duplicates
Total aligned reads	The total number of reads that aligned reads
Targeted aligned reads	Number of reads that aligned to the target
Read enrichment	100*(Target aligned reads/ Total aligned reads)
Total length of targeted reference	Total length of sequenced bases in the target region
Mean region coverage depth	The total number of targeted bases divided by the targeted region size. Roughly equivalent to the weighted mean of the region coverage's from the <sample>.coverage.csv file.
Uniformity of coverage (Pct > 0.2*mean):	The percentage of targeted base positions in which the read depth is greater than 0.2 times the mean region target coverage depth.

Statistic	Definition
Target coverage at 1X	Percentage targets with coverage greater than 1X
Target coverage at 10X	Percentage targets with coverage greater than 10X
Target coverage at 20X	Percentage targets with coverage greater than 20X
Target coverage at 50X	Percentage targets with coverage greater than 50X
Insert Size median	Median length of the sequenced fragment
Insert Size min	Minimum length of the sequenced fragment
Insert Size max	Maximum length of the sequenced fragment
Insert Size SD	Standard deviation of the sequenced fragment
SNPs	Total number of SNPs present in the dataset and pass the quality filters
SNPs (Percent found in dbSNP)	100*(Number of SNPs in dbSNP/ Number of SNPs)
SNP Ts/Tv Ratio	Transition rate of SNPs that pass the quality filters/Transversion rate of SNPs that pass the quality filters
SNP Het/Hom Ratio	Number of Heterozygous SNPs/ Number of Homozygous SNPs
Indels	Total number of Indels present in the dataset and pass the quality filters
Indels (Percent found in dbSNP)	100*(Number of Indels in dbSNP/ Number of Indels)
Indel Het/Hom ratio	Number of Heterozygous Indels/ Number of Homozygous Indels

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 2 Illumina General Contact Information

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

Table 3 Illumina Customer Support Telephone Numbers

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

MSDSs

Material safety data sheets (MSDSs) are available on the Illumina website at www.illumina.com/msds.

Product Documentation

Product documentation in PDF is available for download from the Illumina website. Go to www.illumina.com/support, select a product, then click **Documentation & Literature**.

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