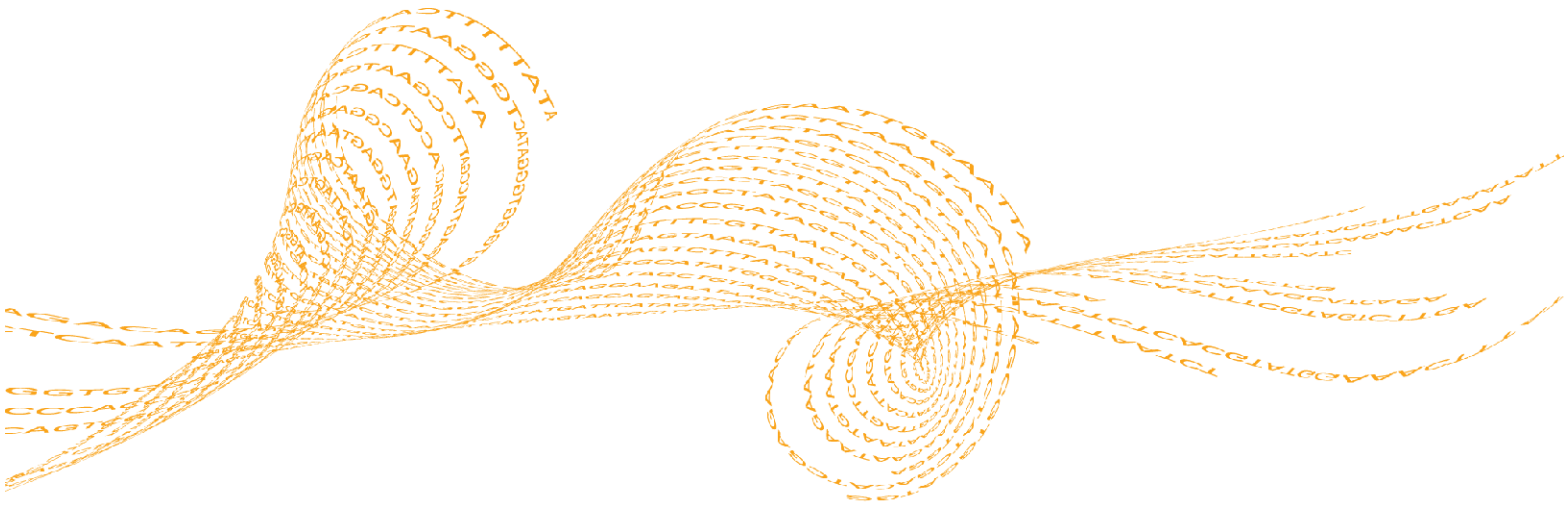
CASAVA 1.8.2

Quick Reference Guide

FOR RESEARCH USE ONLY



ILLUMINA PROPRIETARY
Part # 15011197 Rev C
October 2011

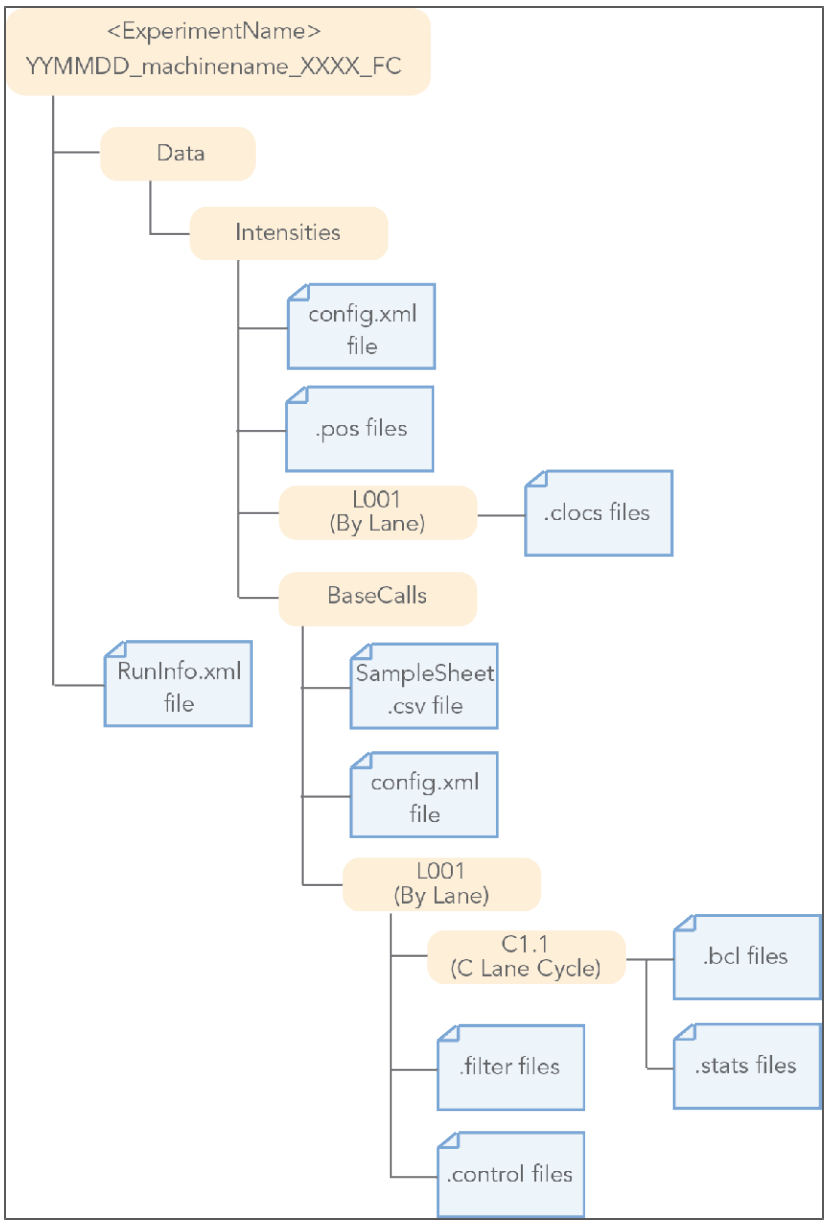
illumina®

Bcl Conversion and Demultiplexing

As of CASAVA 1.8, configureAlignment uses FASTQ files as input. Since Illumina sequencing instruments generate *.bcl files as primary sequencing output, CASAVA contains a BCL to FASTQ converter that combines these per-cycle *.bcl files from a run and translates them into FASTQ files. CASAVA 1.8 can start with bcl conversion and alignment as soon as the first read has been sequenced completely.

Bcl Conversion Input Files

Demultiplexing needs a BaseCalls directory and a sample sheet to start a run. These files are described below. See also image below.



BaseCalls Directory

Demultiplexing requires a BaseCalls directory as generated by RTA or OLB (Off-Line Basecaller), which contains the binary base call files (*.bcl files).

NOTE
As of 1.8, CASAVA does not use *_qseq.txt files as input anymore.

The BCL to FASTQ converter needs the following input files from the BaseCalls directory:

- *.bcl files.
- *.stats files.
- *.filter files.
- *.control files
- *.clocs, *.locs, or *_pos.txt files. The BCL to FASTQ converter determines which type of position file it looks for based on the RTA version that was used to generate them.
- RunInfo.xml file. The RunInfo.xml is at the top level of the run folder.
- config.xml file

RTA is configured to copy these files off the instrument computer machine to the BaseCalls directory on the analysis server. The files are described below.

Generating the Sample Sheet

The user generated sample sheet (SampleSheet.csv file) describes the samples and projects in each lane, including the indexes used. The sample sheet should be located in the BaseCalls directory of the run folder. You can create, open, and edit the sample sheet in Excel.

The sample sheet contains the following columns:

Column Header	Description
FCID	Flow cell ID
Lane	Positive integer, indicating the lane number (1-8)
SampleID	ID of the sample
SampleRef	The reference used for alignment for the sample
Index	Index sequences. Multiple index reads are separated by a hyphen (for example, `ACCAGTAA-GGACATGA`).
Description	Description of the sample
Control	Y indicates this lane is a control lane, N means sample
Recipe	Recipe used during sequencing
Operator	Name or ID of the operator
SampleProject	The project the sample belongs to

You can generate it using Excel or other text editing tool that allows .csv files to be saved. Enter the columns specified above for each sample, and save the Excel file in the .csv format. If the sample you want to specify does not have an index sequence, leave the Index field empty.

Illegal Characters

Project and sample names in the sample sheet cannot contain illegal characters not allowed by some file systems. The characters not allowed are the space character and

the following:

```
? ( ) [ ] / \ = + < > : ; " ' , * ^ | & .
```

Multiple Index Reads

If multiple index reads were used, each sample must be associated with an index sequence for each index read. All index sequences are specified in the **Index** field. The individual index read sequences are separated with a hyphen character (-). For example, if a particular sample was associated with the sequence `ACCAGTAA` in the first index read, and the sequence `GGACATGA` in the second index read, the index entry would be `ACCAGTAA-GGACATGA`.

Samples Without Index

As of CASAVA 1.8, you can assign samples without index to projects, sampleIDs, or other identifiers by leaving the Index field empty.

Running Bcl Conversion and Demultiplexing

Bcl conversion and demultiplexing is performed by one script, configureBclToFastq.pl. This section describes how to perform Bcl conversion and demultiplexing in CASAVA 1.8.

Usage of configureBclToFastq.pl

The standard way to run bcl conversion and demultiplexing is to first create the necessary Makefiles, which configure the run. Then you run `make` on the generated files, which executes the calculations.

1 Enter the following command to create a makefile for demultiplexing:
`/path-to-CASAVA/bin/configureBclToFastq.pl[options]`

NOTE
The options have changed significantly between CASAVA 1.7 and 1.8. See *Options for Bcl Conversion and Demultiplexing* on page 5.

2 Move into the newly created Unaligned folder specified by `--output-dir`.

3 Type the "make" command. Suggestions for "make" usage, depending on your workflow, are listed below.

Make Usage	Workflow
`nohup make -j N`	Bcl conversion and demultiplexing (default).
`nohup make -j N r1`	Bcl conversion and demultiplexing for read 1.

See *Makefile Options for Bcl Conversion and Demultiplexing* on page 7 for explanation of the options.

NOTE
The ALIGN option, which kicked off configureAlignment after demultiplexing was done in CASAVA 1.7, is no longer available.

4 After the analysis is done, review the analysis for each sample.

Options for Bcl Conversion and Demultiplexing

The options for demultiplexing are described below.

Option	Description	Examples
`--fastq-cluster-count`	Maximum number of clusters per output FASTQ file. Do not go over 16000000, since this is the maximum number of reads we recommend for one ELAND process. Specify 0 to ensure creation of a single FASTQ file. Defaults to 4000000.	`--fastq-cluster-count 6000000`
`-i, --input-dir`	Path to a BaseCalls directory.\ Defaults to current dir	`--input-dir <BaseCalls_dir>`
`-o, --output-dir`	Path to demultiplexed output. Defaults to <run_folder>/Unaligned Note that there can be only one Unaligned directory by default. If you want multiple Unaligned directories, you will have to use this option to generate a different output directory.	`--output-dir <run_folder>/Unaligned`
`--positions-dir`	Path to a directory containing positions files. Defaults depends on the RTA version that is detected.	`--positions-dir <positions_dir>`
`--positions-format`	Format of the input cluster positions information. Options: • .locs • .clocs • _pos.txt Defaults to .clocs.	`--positions-format .locs`
`--filter-dir`	Path to a directory containing filter files. Defaults depends on RTA version that is detected.	`--filter-dir <filter_dir>`
`--intensities-dir`	Path to a valid Intensities directory. Defaults to parent of base_calls_dir.	`--intensities-dir <intensities_dir>`
`-s,--sample-sheet`	Path to sample sheet file. Defaults to <input_dir>/SampleSheet.csv	`--sample-sheet <input_dir>/SampleSheet.csv`
`--tiles`	--tiles option takes a comma-separated list of regular expressions to match against the expected "s_<lane>_<tile>" pattern, where <lane> is the lane number (1-8) and <tile> is the 4 digit tile number (left-padded with 0s).	`--tiles=s_[2468]_[0-9][0-9][02468]5,s_1_0001`

`--use-bases-mask`	The `--use-bases-mask` string specifies how to use each cycle.
	• An “n” means ignore the cycle.
	• A “Y” (or "y") means use the cycle.
	• An “I” means use the cycle for the index read.
	• A number means that the previous character is repeated that many times.
	• The read masks are separated by commas ","
	The format for dual indexing is as follows:`--use-bases-mask Y*,I*,I*,Y*` or variations thereof as specified above.
	If this option is not specified, the mask will be determined from the 'RunInfo.xml' file in the run directory. If it cannot do this, you will have to supply the `--use-bases-mask`.
		`--use-bases-mask y50n,I6n,Y50n`
		`This means:`
		• Use first 50 bases for first read (Y50)
		• Ignore the next (n)
		• Use 6 bases for index (I6)
		• Ignore next (n)
		• Use 50 bases for second read (Y50)
		• Ignore next (n)

Option	Description	Examples
`--no-eamss`	Disable the masking of the quality values with the Read Segment Quality control metric filter.	`--no-eamss`
`--mismatches`	Comma-delimited list of number of mismatches allowed for each read (for example: 1,1). If a single value is provide, all index reads will allow the same number mismatches. Default is 0.	`--mismatches 1`
`--flowcell-id`	Use the specified string as the flowcell id. (default value is parsed from the config-file)	`--flowcell-id flow_cell_id`
`--ignore-missing-stats`	Fill in with zeros when *.stats files are missing	`--ignore-missing-stats`
`--ignore-missing-bcl`	Interpret missing *.bcl files as no call	`--ignore-missing-bcl`
`--ignore-missing-control`	Interpret missing control files as not-set control bits	`--ignore-missing-control`
`--with-failed-reads`	Include failed reads into the FASTQ files (by default, only reads passing filter are included).	`--with-failed-reads`
`--adapter-sequence`	Path to a FASTA adapter sequence file. If there are two adapters sequences specified in the FASTA file, the second adapter will be used to mask read 2. Else, the same adapter will be used for all reads. Default: None (no masking)	`--adapter-sequence <adapter dir>/adapter.fa`
`--man`	Print a manual page for this command	`--man`
`-h, --help`	Produce help message and exit	`-h`

Makefile Options for Bcl Conversion and Demultiplexing

The options for `make` usage in demultiplexing/analysis are described below.

Parameter	Description
nohup	Use the Unix nohup command to redirect the standard output and keep the "make" process running even if your terminal is interrupted or if you log out. The standard output will be saved in a nohup.out file and stored in the location where you are executing the makefile. nohup make -j n & The optional "&" tells the system to run the analysis in the background, leaving you free to enter more commands. We suggest always running nohup to help troubleshooting if issues arise.
-j N	The -j option specifies the extent of parallelization, with the options depending on the setup of your computer or computing cluster.
r1	Runs Bcl conversion for read 1. Can be started once the last read has started sequencing.
POST_RUN_ COMMAND_R1	A Makefile variable that can be specified either on the make command line or as an environment variable to specify the post-run commands after completion of read one, if needed. Typical use would be triggering the alignment of read 1.
POST_RUN_ COMMAND	A Makefile variable that can be specified on the make command line to specify the post-run commands after completion of the run.
KEEP_ INTERMEDIARY	The option KEEP_INTERMEDIARY tells CASAVA not to delete the intermediary files in the Temp dir after Bcl conversion is complete. Usage: `KEEP_INTERMEDIARY:=yes`

NOTE
If you specify one of the more specific workflows and then run a more

general one, only the difference will get processed. For instance:

```
make -j N r1
```

followed by:

```
make -j N
```

will do read 1 in the first step, and read 2 the second one.

Starting Bcl Conversion for Read 1

If you want to start Bcl to FASTQ conversion before completion of the run, use the makefile target `r1` at any time after the last read has started (for multiplexed runs, this is after completion of the indexing read).

1. Enter the following command to create a makefile for Bcl conversion:
   `/path-to-CASAVA/bin/configureBclToFastq.pl [options]`
2. Move into the newly created Unaligned folder specified by `--output-dir`.
3. Type the "make r1" command:
   `make -j 8 r1`

NOTE
the `-j <n>` command line option is supported to indicate up to <n> processes in parallel. However, for Bcl conversion the maximum level of parallelization is 8.

Starting Alignment

You can also start alignment before completion of the run using the target `r1` when running `make` for configureAlignment.

Alternatively, you can use the `POST_RUN_COMMAND_R1` variable to automatically start the alignment of read 1 at the end of the Bcl conversion. For example:

```
make -j 8 r1 POST_RUN_COMMAND_R1="cd ../Aligned ; make -j 16
   r1"
```

Starting the Second Read

To start Bcl conversion of the second read, use the regular make command in the Unaligned folder. Perform the following:

1. Move into the Unaligned folder specified by `--output-dir`.
2. Type the regular "make" command:
   `make -j 8`
3. After the analysis is done, review the analysis for each sample.

Bcl Conversion Output Folder

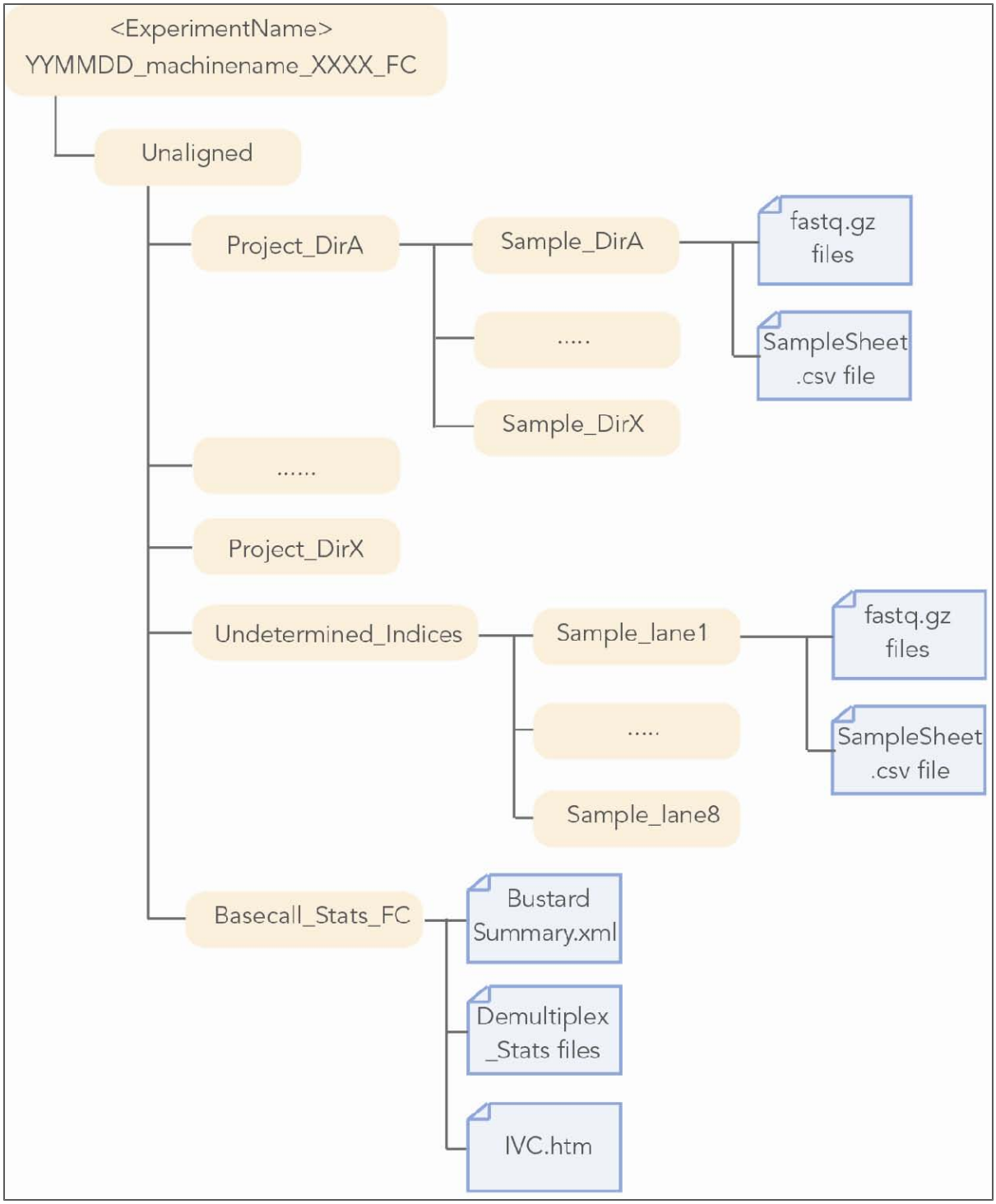
The Bcl Conversion output directory has the following characteristics:

- The project and sample directory names are derived from the sample sheet.
- The Demultiplex_Stats file shows where the sample data are saved in the directory structure.
- The Undetermined_indices directory contains the reads with an unresolved or erroneous index.
- If no sample sheet exists, CASAVA generates a project directory named after the flow cell, and sample directories for each lane.

- Each directory is a valid base calls directory that can be used for subsequent alignment analysis in CASAVA.

NOTE

If the majority of reads end up in the 'Undetermined_indices' folder, check the --use-bases-mask parameter syntax and the length of the index in the sample sheet. It may be that you need to set the --use-bases-mask option to the length of the index in the sample sheet + the character 'n' to account for phasing. Note that you will not be able to see which indices have been placed in the 'Undetermined_indices' folder

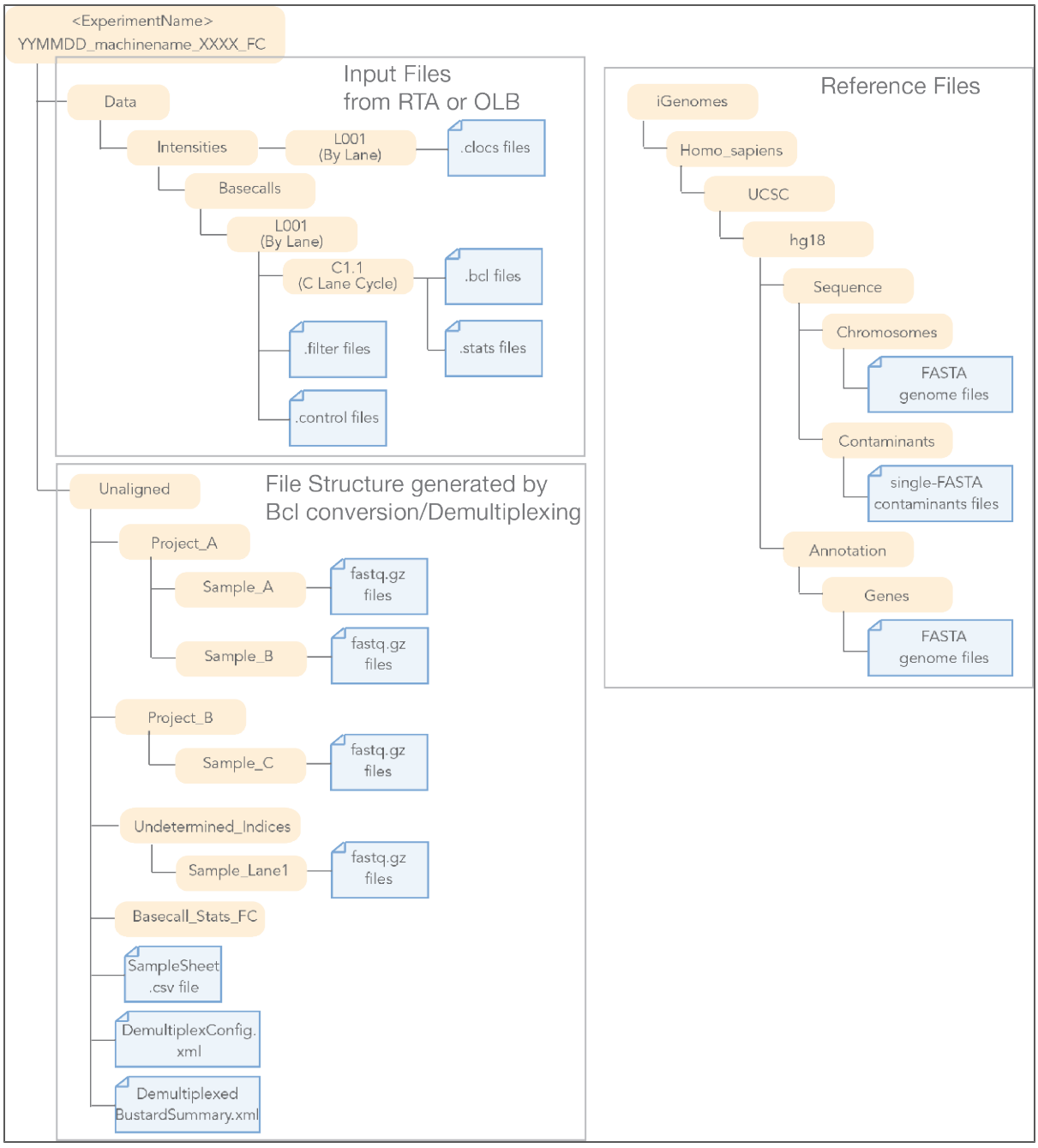


NOTE
There can be only one Unaligned directory by default. If you want multiple Unaligned directories, you will have to use the option `--output-dir` to generate a different output directory.

Sequence Alignment

configureAlignment is a CASAVA module that performs sequence alignments. This section describes running configureAlignment, parameters, analysis variables, configuration file options, and ELANDv2e alignments.

configureAlignment Input Files



Running configureAlignment

Standard configureAlignment Analysis

The standard way to run configureAlignment is to set the parameters in a configuration file, create a makefile, and start the analysis with the "make" command.

1 Edit the configureAlignment configuration file as described in *configureAlignment Configuration File* on page 13.

2 Check the analysis by running the configureAlignment.pl command without `--make`.

```
/path-to-CASAVA/bin/configureAlignment.pl config.txt
   --EXPT_DIR path_to_Unaligned_folder
```

3 Enter the configureAlignment.pl command, but now with `--make`. This creates the makefile for sequence alignment.

```
/path-to-CASAVA/bin/configureAlignment.pl config.txt
   --EXPT_DIR path_to_Unaligned_folder --make
```

4 Move into the newly created Aligned folder under the Run folder (see *configureAlignment Output Files* on page 16). Type the "make" command for basic analysis:

```
make
```

NOTE

You may prefer to use the parallelization option as follows:

```
make -j 3 all
```

The extent of the parallelization depends on the setup of your computer or computing cluster.

5 After the analysis is done, review the analysis.

ELAND_FASTQ_FILES_PER_PROCESS

CASAVA requires a minimum of 2 GB RAM per core. The parameter ELAND_FASTQ_FILES_PER_PROCESS (optional) in the configureAlignment config.txt specifies the maximum number of FASTQ files aligned by each ELAND process, to limit the per-core memory consumption.

The optimal value leads to approximately 10 to 13 million clusters in one set. Since the FASTQ file size (in reads) is determined by the Bcl conversion option --fastq-cluster-count , while the maximum number of files per process is determined by ELAND_FASTQ_FILES_PER_PROCESS, the product of these options should not exceed 16 million:

(`ELAND_FASTQ_FILES_PER_PROCESS` value) × (`--fastq-cluster-count` value) ≤ 16 million

CAUTION

Setting the right value for the ELAND_FASTQ_FILES_PER_PROCESS is very important. Too high may result in silent crashes due to too high memory utilization, and should be avoided. Too low may result in a decreased performance. Use is optional, and we generally recommend using default values.

The --fastq-cluster-count used during Bcl conversion can be found in Unaligned/Makefile.

configureAlignment Configuration File

This section describes the features and parameters of the configureAlignment configuration text file.

Config File Parameter List

The following tables list the parameters that can be specified in a configureAlignment configuration file.

Table 1 GERALD Configuration File Core Parameters

Parameter	Definition
`EXPT_DIR data/110113_ILMN-1_0217_FC1234/Unaligned`	Provide the path to the experiment (demultiplexed) directory in the run folder, if not specified on the command line. Usually the output folder from the BCL to FASTQ converter. The path should always be to the Unaligned directory, even when the run only contains one project
`USE_BASES nY*n`	Ignore the first and last base of the read. The USE_BASES string contains a character for each cycle. • If the character is "Y", the cycle is used for alignment. • If the character is "n", the cycle is ignored. • Wild cards (*) are expanded to the full length of the read.
`ELAND_GENOME /home/user/Genomes/Eland/BAC_plus_vector/`	Specify the single FASTA files that you want to use as genome reference for alignment with ELANDv2e.
`SAMTOOLS_GENOME`	Direct CASAVA to a multi-sequence FASTA reference file.
`ANALYSIS eland_extended`	Specify the type of alignment that should be performed. Available options are: • `ANALYSIS eland_extended` • `ANALYSIS eland_pair` • `ANALYSIS eland_rna` • `ANALYSIS none` The default is `ANALYSIS none`
`ELAND_FASTQ_FILES_PER_PROCESS N`	The maximum number of files analyzed by each ELAND process, needed to ensure that the memory usage stays below 2 GB. The optimal value is such that there are approximately 10 to 13 million lines (reads) in one set. Only available for ANALYSIS eland_extended, ANALYSIS eland_pair, and ANALYSIS eland_rna.

WARNING
Default for `USE_BASES` is `Y*n`, which means perform a single-read alignment and ignore the last base. If running ANALYSIS eland_pair, make sure to specify the USE_BASES option for two reads (for example USE_BASES Y*n,Y*n).

Optional Parameters

Table 2 configureAlignment Configuration File Optional Parameters

Parameter	Definition
`OUT_DIR`	Path to configureAlignment output. The path must be to a directory not already present.

Parameter	Definition
	Defaults to <run_folder>/Aligned Note that there can be only one Aligned directory by default. If you want multiple Aligned directories, you will have to use this option to generate a different output directory.
ELAND_RNA_GENOME_CONTAM	Points to the folder containing a set of **contaminant sequences** for the genome—typically the mitochondrial and ribosomal sequences. The files must be in single FASTA format.
ELAND_RNA_GENOME_ANNOTATION	Path to transcripts mapping to the genome (refFlat.txt.gz or seq_gene.md.gz). See also Using ANALYSIS eland_rna .
ELAND_RNA_GENE_MD_GROUP_LABEL	The group label above specifies which assembly to use in the seq_gene file, and is found in column 13 of the file. seq_gene files can hold entries for multiple assemblies. Example: `ELAND_RNA_GENE_MD_GROUP_LABEL GRCh37.p2-Primary Assembly`.

Paired-End Analysis Options

Table 3 configureAlignment Configuration File Paired-End Analysis Options

Parameter	Definition
ANALYSIS eland_pair	Use the paired-end alignment mode of ELANDv2e to align paired reads against a target.
USE_BASES Y*,nY*n	Use all bases of the first read and ignore the first and last base of the second read.
6:USE_BASES nY25	Ignore the first base on both the first and second read of lane 6; use 25 bases each and ignore any other bases for lane 6 only.
KAGU_PAIR_PARAMS	KAGU_PAIR_PARAMS passes options for paired-end runs to the alignmentResolver through the configureAlignment configuration file.

Specifying Analysis

Analysis can be specified by project, reference, sample, index, or lane, which is explained in this section.

Lane-Specific Analysis

By adding the lane number(s) followed by colon in front of an analysis option, you state that the analysis option is only for samples from that lane. The lane number is only valid for the configureAlignment settings on that same line.

For example, `567:ANALYSIS eland_extended` tells configureAlignment that eland_extended should be run on samples from lane 5, 6, and 7.

Sample-Specific Analysis

The config.txt file has some keywords that enable you to specify analysis for project, reference, sample, or index: PROJECT, REFERENCE, SAMPLE, and BARCODE. These keywords refer to the SampleProject, SampleRef, SampleID, and Index specified in the samplesheet.csv file located in the Unaligned directory of the run folder.

Lines starting with PROJECT, REFERENCE, SAMPLE, and BARCODE override any default settings specified in the config.txt file, but only for those samples for which the SampleProject, SampleRef, SampleID, or Index matches the PROJECT, REFERENCE, SAMPLE, or BARCODE. The override is only valid for the configureAlignment settings on that same line.

Combining Specificity

It is also possible to combine specific analyses, like in this example:

```
12: REFERENCE human ANALYSIS eland_pair
```

which tells configureAlignment to perform eland_pair analysis on the human reference samples from lanes 1 and 2.

Priority

If multiple specific settings conflict, configureAlignment uses the following order of priority:

1 PROJECT

2 REFERENCE

3 SAMPLE

4 BARCODE

5 Lane

6 Global settings

This means, PROJECT settings override any other settings, while REFERENCE settings can only be overruled by PROJECT settings, and so on.

WARNING

The attribute cannot be set for more than one scope at a time. In other words the following is not allowed:

```
PROJECT test BARCODE ACGT ANALYSIS eland_extended
```

Samples Without Index

Unless otherwise specified in the sample sheet, samples without index will end up in the project folder Undetermined_indices, and in a sample folder named after the lane (e.g. Sample_lane1).

If you want to specify analysis for these samples without index other than the global analysis, you can use identifiers PROJECT Undetermined_indices or SAMPLE lane1.

NOTE

Normally you would want to use:

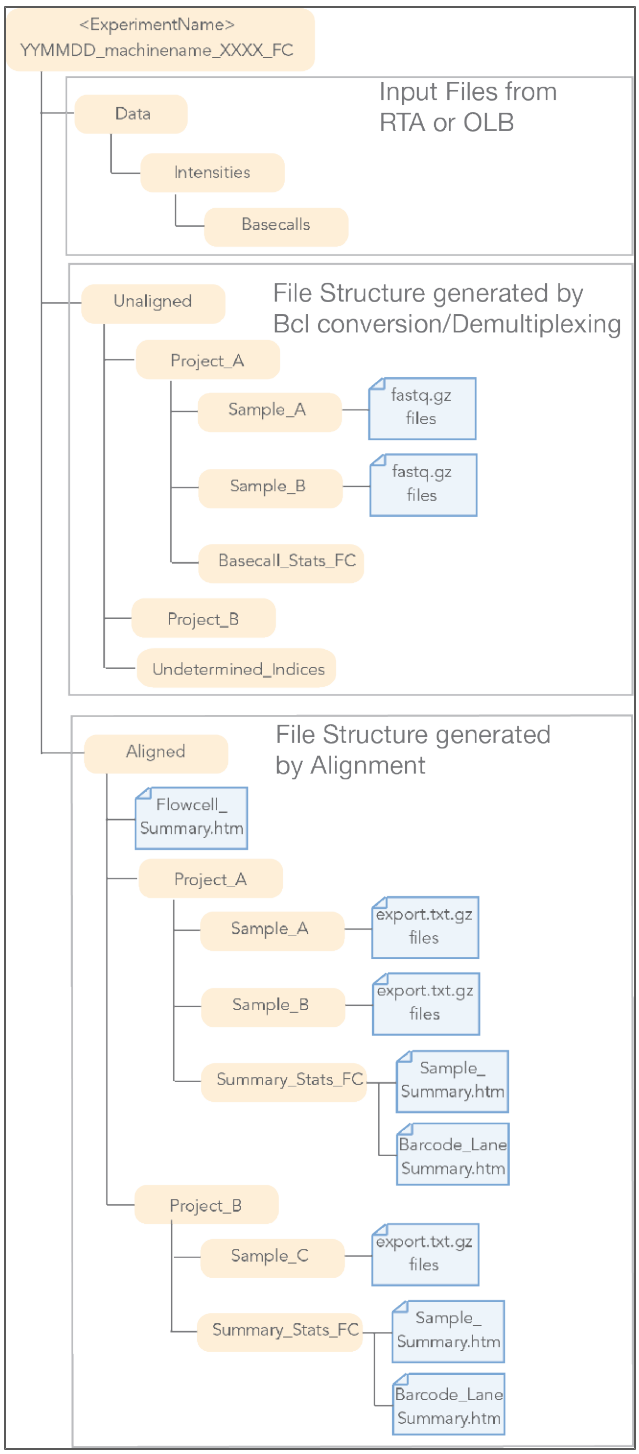
```
PROJECT Undetermined_indices ANALYSIS none
```

or

```
REFERENCE unknown ANALYSIS none
```

to avoid wasting CPU time on the Undetermined_indices data, which often is of poor quality.

configureAlignment Output Files



NOTE

There can be only one Aligned directory by default. If you want multiple Aligned directories, you will have to use the option `OUT_DIR` to generate a different output directory.

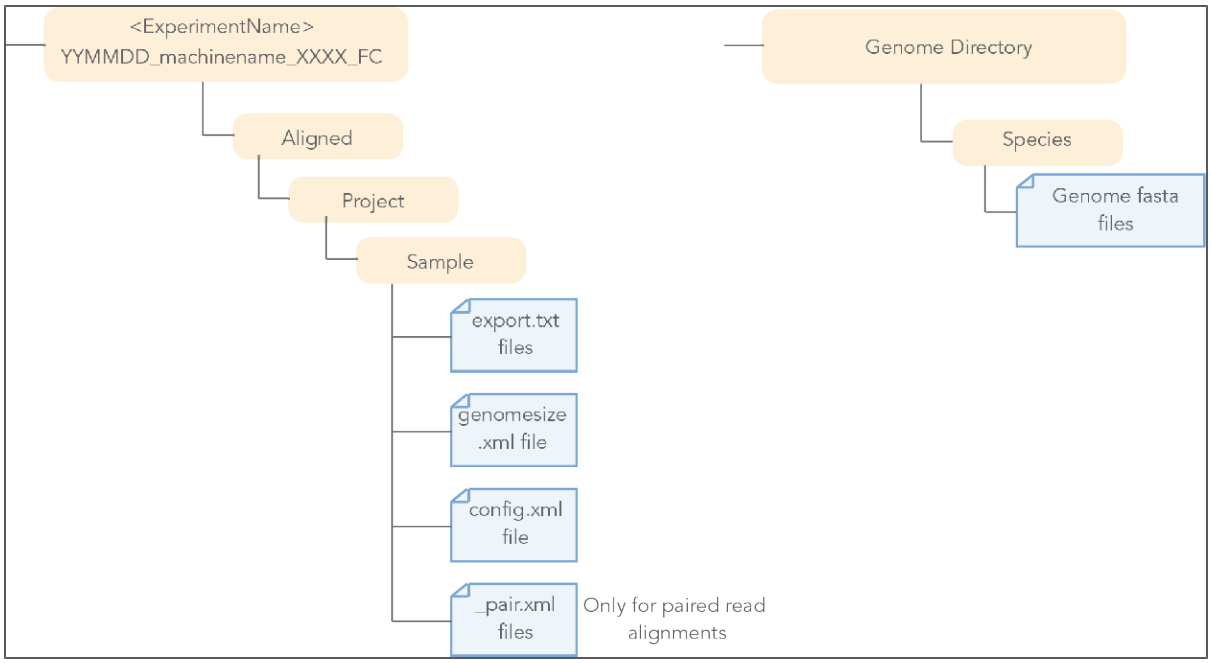
Variant Detection and Counting

This section explains how to use CASAVA1.8 to detect Single Nucleotide Polymorphisms (SNPs) and insertions/deletions (indels), and count hits on transcripts for RNA sequencing.

Variant Detection Input Files

The configureAlignment input files for CASAVA variant detection can be found in the Aligned directory of the run folder, and are described below.

In addition, CASAVA variant detection and counting uses annotation files (genome sequence files and refFlat.txt.gz or seq_gene.md.gz file) .



Running Variant Detection and Counting

Major Use Cases

- SNP and Indel Calling
  To run CASAVA with callSmallVariants and assembleIndels, enter:
  `/path-to-CASAVA/bin/configureBuild.pl [options]`
- SNP and Indel calling without large-indel assembly
  To run CASAVA with callSmallVariants, but without assembleIndels, enter:
  `/path-to-CASAVA/bin/configureBuild.pl --targets all noassembleIndels --variantsSkipContigs [options]`
- SNP and Indel calling, Single-end Build
  To run CASAVA with callSmallVariants for a single-end build, enter:
  `/path-to-CASAVA/bin/configureBuild.pl [options]`
- RNA Sequencing
  To run CASAVA for RNA Sequencing, enter:
  `/path-to-CASAVA/bin/configureRnaBuild.pl [options]`

Other Use Cases

- Help

  To get the CASAVA Help for callSmallVariants, enter:

  ```
  /path-to-CASAVA/bin/configureBuild.pl --help
     callSmallVariants
  ```

- Rerun callSmallVariants

  In any pre-existing build in which the sort module was previously completed (and the assembleIndels module for a paired end build), Small variant calling may be rerun using:

  ```
  /path-to-CASAVA/bin/configureBuild.pl -od $PROJECT_DIR --
     targets callSmallVariants
  ```

  NOTE
  We only support data sets originated from the same version of the software.

- Generate BAM File with Altered Alignments

  An advanced option useful for variant diagnosis is to create BAM files for those reads which had their alignments altered by the variant caller during local realignment. This may be done by adding the command `--variantsWriteRealigned` to any command-line which runs the variant caller.

The targets that define CASAVA analysis are listed in the tables below.

Option	Description
all	Run all pre-configured targets for the given analysis type (default), except for target bam.
sort	Bin reads and sort by position; Remove PCR duplicates for paired-end data.
assembleIndels	Search for candidate indels from paired-end reads via de-novo assembly of contigs which are aligned back to the reference.
callSmallVariants	Call SNPs and indels from locally re-aligned reads. Candidate indels from the assembleIndels target can be used to improve indel results. See also Target callSmallVariants.
rnaCounts	Calculate gene and exon counts in an RNA-Seq build.
bam	Aggregate all reads into a single BAM file with chromosome re-labeling. This target is not part of target all, and is therefore not done by default. Must be preceded by or combined with target sort. This BAM file is independent of the archival bam file, which can be produced using the option `--sortKeepAllReads`(see Archival Build).
gsIndex	Pre-compute Genome Studio linear index for all reads in the build.

If you run a target other than the default target (`all`), make sure to read the help written for the target. This will help you identify any dependencies for the target you want to run.

Target help can be accessed by typing:

```
Path/to/CASAVA/bin/configureBuild.pl --help <target>
```

NOTE
Prefixing any target name with `no` will exclude it from the targets list.
Example:

```
path-to-CASAVA/bin/configureBuild.pl --targets all
noassembleIndels --variantsSkipContigs [options]
```

Options

NOTE

The option --outDir is mandatory for all analysis types. CASAVA will not run if this option is missing.

CASAVA will only run without --inSampleDir if the build has been already configured with --inSampleDir before.

Global Options

The options described below are global options used to specify analysis across different targets.

Table 4 Major File Options for Variant Detection and Counting

Option	Application	Description
-id, --inSampleDir=PATH	SE, PE	PATH to the aligned sample input directory.
-od, --outDir=PATH	SE, PE	PATH to the build sample output directory.
-ref, --refSequences=PATH	SE, PE	PATH of the reference genome sequences. Default is buildDir/genomes/. The FASTA files should not be squashed for CASAVA.
--samtoolsRefFile=FILE	SE, PE	PATH to a single samtools-style reference file

Table 5 Behavioral Options for Variant Detection and Counting

Option	Application	Description
-a, --applicationType=TYPE	SE, PE	Type of analysis [DNA, RNA]; default is DNA.
-f, --force	SE, PE	Ignore errors from previous CASAVA execution.
-h, --help [TARGET]	SE, PE	Prints on screen usage guide. If TARGET is specified, prints usage guide for the corresponding plugin target
-j, --jobsLimit	SE, PE	Limit number of parallel jobs. Defaults: -1 (unlimited) for --sgeAuto. 1 for --workflowAuto. Do not set it to the maximum number of processors as this might cause the terminal to become unresponsive
--postRunCmd=CMDLINE	SE, PE	Post Run Commands can be launched after CASAVA completes by including the --postRunCmd option, followed by the commands to be launched
-sa, --sgeAuto	SE, PE	Generates the workflow definition file and runs it on SGE (use with --sgeQueue)
--sgeQsubFlags	SE, PE	Extra parameters to be passed to SGE qsub by the taskServer.pl
--sgeQueue	SE, PE	SGE queue name, used with --sgeAuto or --workflow (e.g: all.q)
--targets=LIST	SE, PE	Space-separated list of targets to run (see *The targets that define CASAVA analysis are listed in the tables below.* on page 18). Default is `all`.
--tempDir	SE, PE	Overrides default path for local temporary files
--verbose=NUMBER	SE, PE	Sets the verbose level (default is 0, which is the minimum).
--version	SE, PE	Prints version information.
-w, --workflow	SE, PE	Instead of running CASAVA , generates the workflow definition file tasks-DATA.txt

Option	Application	Description
-wa, --workflowAuto	SE, PE	Generates the workflow definition file and runs it. See --jobsLimit.
--workflowFile=FILE	SE, PE	Overrides workflow file name. Default is tasks.<date>.txt

Table 6 Global Analysis Options for Variant Detection and Counting

Option	Application	Description
--QVCutoff=NUMBER	PE	Sets the paired-end alignment score threshold to NUMBER (default 90).
--QVCutoffSingle=NUMBER	SE, PE	Sets the single-read alignment score threshold to NUMBER (default 10).
--read=NUMBER	PE	Limit input to the specified read only. Forces single-ended analysis on one read of a double-ended data set.
--singleScoreForPE=VALUE	PE	Sets the variant caller to filter reads with single score below QVCutoffSingle in PE mode YES|NO. Default NO.
--sortKeepAllReads	SE, PE	Generate an archive BAM file. Keep all purity filtered, duplicate and unmapped reads in the build. These reads will be ignored during variant calling.
--toNMScore=NUMBER	SE, PE	Minimum SE alignment score to put a read to NM. Default=-1 (-1 means option is turned off)
--ignoreUnanchored	PE	Ignore unanchored read pairs in indel assembly and variant calling. Unanchored read pairs have a single-read alignment score of 0 for both reads.

Options for Target sort

Option	Application	Description
--rmDup=YES|NO	PE	Turn On/Off PCR duplicate marking/removal for paired-end reads (default YES).
--sortBufferSize=INTEGER	SE, PE	Buffer size used by the read sorting process, in megabytes (default: 1984).
--sortKeepAllReads	SE, PE	Run the sort module in archival mode instead of the default filtered mode.

Options for Target assembleIndels

Option	Application	Description
--indelsSpReadThresholdIndels=NUMBER	PE	Spanning read score threshold. The higher the single read alignment score before realignment, the more unlikely it is to see this pattern of mismatches given the read's quality values. Default threshold value is 25. Drop this value to add more reads into the indel finding process, at the possible expense of introducing noise. For an alignment with no mismatches this option should be set at zero.
--indelsPrasThreshold=NUMBER	PE	Paired read alignment score threshold. If a read has a paired read alignment score of at least this, then it is used to update the base quality stats for that sample prep. Default is calculated based off the data.

Option	Application	Description
--indelsAlignScoreThresh=NUMBER	PE	If an alignment score for a read exceeds this threshold after realignment then the output file is updated to incorporate this new alignment. Otherwise the read's entry remains as per the input file. Default value is 120. A low value will cause some reads to be wrongly placed (albeit within a small interval).
--indelsSdFlankWeight=NUMBER	PE	Number of standard deviations to use when defining the genomic interval to align the read to (default: 1).
--indelsMinGroupSize=NUMBER	PE	Only output clusters if they contain at least this many reads.
--indelsSpReadThresholdClusters=NUMBER	PE	Spanning read score threshold. This is calculated in exactly the same way as --indelsSpReadThresholdIndels. However it is used in the opposite way. Here the point to find reads with few or no mismatches, which are presumed to arise from repeats and not from indels, and exclude them from the clustering process.
--indelsMinCoverage=NUMBER	PE	Minimum coverage to extend contig (default 3).
--indelsMinContext=NUMBER	PE	Demand at least x exact matching bases either side of variant (default is 6). The idea here is to ensure that an indel has a minimum number of exactly matching bases on either side. Setting this to zero might be good for finding reads which align to breakpoints.
--indelsSaveTempFiles	PE	Add this flag to save intermediate output files from each stage of the indel assembly process.

Options for Target callSmallVariants

Table 7 Workflow Options for callSmallVariants

Option	Application	Description
--variantsSkipContigs	PE	By default information from the assembleIndels module is used (and required) in paired-end DNA Sequencing analysis. This option disables use of indel contigs during variant calling, and only uses gapped alignment to find indels.
--variantsNoSitesFiles	SE, PE	Do not write out the sites.txt.gz files.
--variantsNoReadTrim	SE, PE	By default, the ends of reads can be trimmed if the alignment path through an indel is ambiguous. This option disables read trimming and chooses the ungapped sequence alignment for any ambiguous read segment. Note that this can trigger spurious SNP calls near indels.
--variantsWriteRealigned	SE, PE	Write only those reads which have been realigned to bam file: "sorted.realigned.bam" for each reference sequence.

Table 8 Read Mapping Options for callSmallVariants

Option	Application	Description
--variantsSEMapScoreRescue	PE	Include reads if they have an SE mapping score equal to or above that set by the "--QVCutoffSingle" option, even if the read pair fails the PE mapping score threshold.

Option	Application	Description
`--variantsIncludeSingleton`	PE	Include paired-end reads which have unmapped mate reads. Note that "--variantsSEMapScoreRescue" must also be specified because ELAND gives singleton reads a PE mapping score of zero.
`--variantsSEMapScoreRescue`	PE	Include reads if they have an SE mapping score equal to or above that set by the "--QVCutoffSingle" option, even if the read pair fails the PE mapping score threshold.

Table 9 SNP and Indel Options for callSmallVariants

Option	Application	Description
`--variantsNoCovCutoff`	SE, PE	Disables the SNP and indel coverage filters detailed below for the options: `--variantsSnpCovCutoff` and `--variantsIndelCovCutoff`. This setting is recommended for targeted resequencing and RNA-Seq (Note it is already set by default for RNA-Seq).

Table 10 SNP Options for callSmallVariants

Option	Application	Description
`--variantsSnpTheta=FLOAT`	SE, PE	The frequency with which single base differences are expected between two unrelated haplotypes (default is 0.001).
`--variantsSnpCovCutoffAll`	SE, PE	By default the mean chromosomal depth filter is based on "used-depth" (the number of basecalls used by the snp-caller after filtration) calculated from all known sites (non-N) in the reference sequence. When this option is set, the threshold and the filtration use the full depth at all known sites in the reference sequence.
`--variantsSnpCovCutoff=FLOAT`	SE, PE	SNPs are filtered out of the final output if the depth of reads used for that site is greater than this value times the mean chromosomal used-depth. (default 3.0) The filter may be disabled for targeted resequencing or other applications by setting this value to -1 (or any negative number).
`--variantsMDFilterCount=INTEGER`	SE, PE	The mismatch density filter removes all basecalls from consideration during SNP calling where greater than 'variantsMDFilterCount' mismatches to the reference occur on a read within a window of 1+2*'variantsMDFilterFlank' positions encompassing the current position. The default value for 'variantsMDFilterCount' is 2 and for 'variantsMDFilterFlank is 20. Set either value to less than 0 to disable the filter.
`--variantsMDFilterFlank=INTEGER`	SE, PE	The mismatch density filter removes all basecalls from consideration during SNP calling where greater than 'variantsMDFilterCount' mismatches to the reference occur on a read within a window of 1+2*'variantsMDFilterFlank' positions encompassing the current position. The default value for 'variantsMDFilterCount' is 2 and for 'variantsMDFilterFlank is 20. Set either value to less than 0 to disable the filter.

Option	Application	Description
--variantsIndependentErrorModel	SE, PE	This switch turns off all error dependency terms in the SNP calling model, resulting in a simpler model where each basecall at a site is treated as an independent observation.
--variantsMinQbasecall=INTEGER	SE, PE	The minimum basecall quality used for SNP calling. (default is 0).
--variantsSummaryMinQsnp=INTEGER	SE, PE	The snps.txt files contain all positions where Q(snp) > 0, however it is expected that only a higher Q(snp) subset of these will be used dependent upon the false positive tolerance of a user's workflow. For this reason summary statistics about the called SNPs are created at a higher "average-application" threshold, which can be set using this option (default is 20).

Table 11 Indel Options for callSmallVariants

Option	Application	Description
--variantsIndelTheta=FLOAT	SE, PE	The frequency with which indels are expected between two unrelated haplotypes (default is 0.0001). See Theta for more explanation.
--variantsIndelCovCutoff=FLOAT	SE, PE	Indels are filtered out of the final output if the local sequence depth is greater than this value times the mean chromosomal depth. The sequence depth of the indel is approximated by the depth of the site 5' of the indel. (default 3.0) The filter may be disabled for targeted resequencing or other applications by setting this value to -1 (or any negative number).
--variantsCanIndelMin=INTEGER	SE, PE	Unless an indel is observed in at least this many gapped or assembleIndels reads, the indel cannot become a candidate for realignment and genotype calling. (default: 3)
--variantsCanIndelMinFrac=FLOAT	SE, PE	Unless an indel is observed in at least this fraction of intersecting reads, the indel cannot become a candidate for realignment and genotype calling. (default: 0.02)
--variantsSmallCanIndelMinFrac=FLOAT	SE, PE	In addition to the above filter for all indels, for indels of size 4 or less, unless the indel is observed in at least this fraction of intersecting reads, the indel cannot become a candidate for realignment and genotype calling. (default: 0.1)
--variantsIndelErrorRate=FLOAT	SE, PE	Set the indel error rate used in the indel genotype caller to a constant value of f (0<=f<=1). The default indel error rate is taken from an empirical function accounting for homopolymer length and indel type (i.e. insertion or deletion). This flag overrides the default behavior with a constant error rate for all indels.
--variantsSummaryMinQindel=INTEGER	SE, PE	The indels.txt files contain all positions where Q(indel) > 0, however it is expected that only a higher Q(indel) subset of these will be used dependent upon the false positive tolerance of a user's workflow. For this reason summary statistics about the called snps are created at a higher "averege-application" threshold, which can be set using this option (default is 20).

Option	Application	Description
--variantsMaxIndelSize= INTEGER	SE, PE	Sets the maximum indel size for realignment and indel genotype calling. Whenever an indel larger than this size is nominated by a de-novo assembly contig it is handled as two independent breakpoints. Note that increasing this value should lead to an approximately linear increase in variant caller memory consumption. The default value is 300 for paired-end builds and 50 for single-end builds.

Options for Target rnaCounts

Option	Application	Description
--refFlatFile	SE	Name and location of UCSC refFlat.txt.gz file. The file must be gz-compressed.
--seqGeneMdFile	SE	Name and location of NCBI seq_gene.md.gz file.
--seqGeneMdGroupLabel	SE	The group label specifies which assembly to use in the seq_gene file, and is found in column 13 of the file. seq_gene files can hold entries for multiple assemblies. **Required** for RNA counting when you use the annotation seqGeneMd file from NCBI.

Options for Target bam

Option	Application	Description
--bamChangeChromLabels= OFF/NOFA/UCSC	SE, PE	Change chromosome labels in the bam plugin output. The available behaviors are: OFF Use unmodified CASAVA chromosome labels (default behavior). NOFA Remove any ".fa" suffix found on each chromosome label. For example "c11.fa" is changed to "c11". UCSC Remove any ".fa" suffix found on each chromosome label and attempt to map the result to the corresponding UCSC human chromosome label. For example "c11.fa" is changed to "chr11".
--bamSkipRefSeq	SE, PE	Do not generate a reference sequence file with each bam file. The default behavior can be restored with --no-bamSkipRefSeq.

Targeted Resequencing

Since targeted resequencing only sequences part of a genome, we recommend using the option --variantsNoCovCutoff to turn off high-coverage filtration of SNPs and indels.

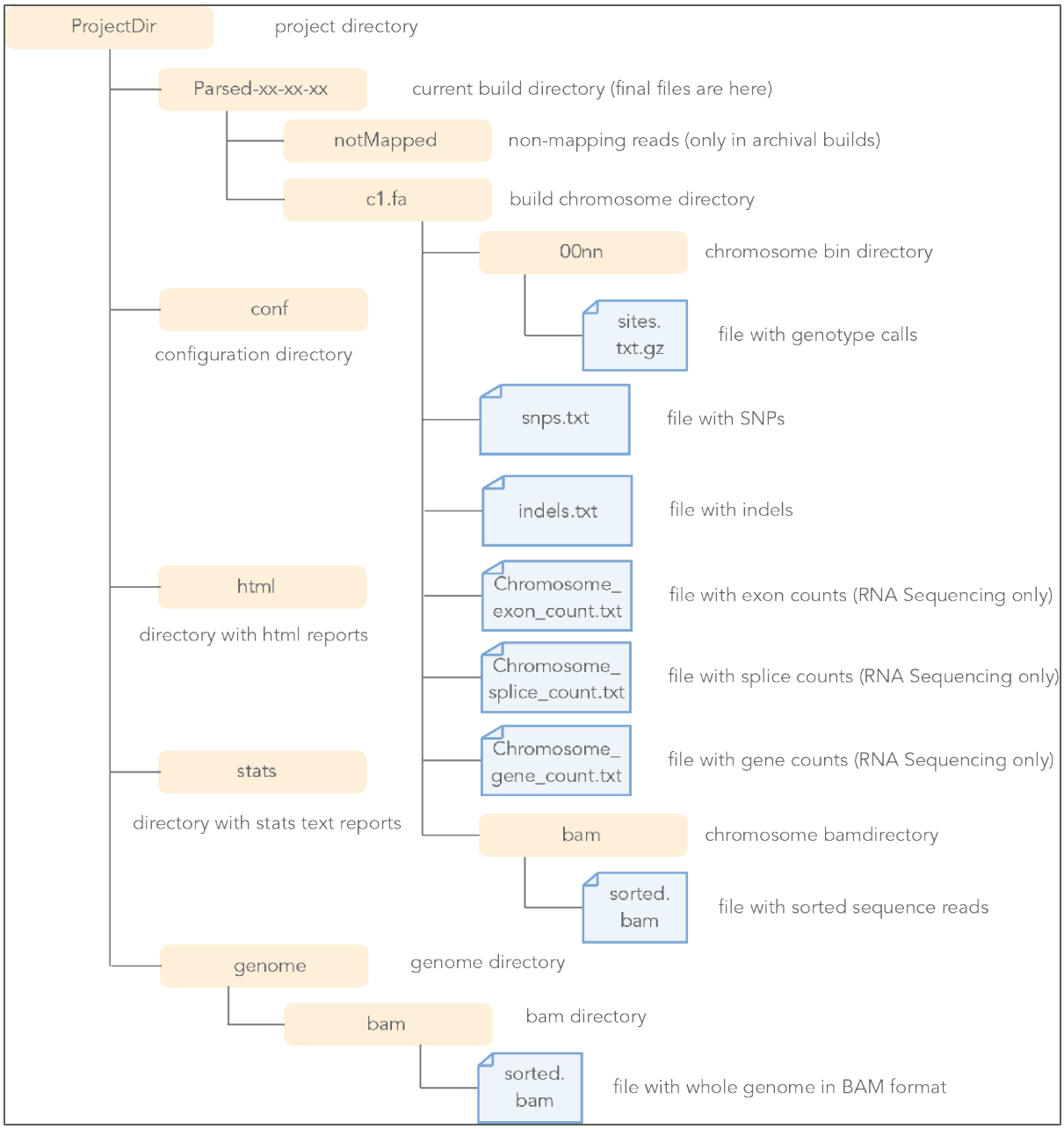
Variant Detection and Counting Output Files

Once the post-alignment build is complete, all relevant information is listed in the build directory, such as:

- Build summary html pages.
  The build summary html pages are located in the buildDir/html folder, and provides access to run information and graphs of important statistics.

- Variant calls and counts.
  The CASAVA build contains sequence, SNP, indels, and (for RNA Sequencing) counts information, and is located in buildDir/Parsed_DATE.
- Computer readable statistics.
  Computer readable statistics are located in buildDir/stats.
- Configuration files.
  CASAVA configuration files are located in buildDir/conf.

Build Directory



Build Html Page

The build html page is located in buildDir/html. When you open the file Home.html, you will find a list of all runs, and a link to statistics.

The **Report Menu** link on the build html page will lead you to graphs and tables for important statistics:

- Coverage
- Duplicates
- Indels statistics
- SNPs statistics

Technical Assistance

For technical assistance, contact Illumina Customer Support.

Table 12 Illumina General Contact Information

Illumina Website	http://www.illumina.com
Email	techsupport@illumina.com

Table 13 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 http://www.illumina.com/msds.

Product Documentation

If you require additional product documentation, you can obtain PDFs from the Illumina website if PDFs are available. Go to http://www.illumina.com/support/documentation.ilmn. When you click on a link, you will be asked to log in to iCom. After you log in, you can view or save the PDF. To register for an iCom account, please visit https://icom.illumina.com/Account/Register.

Illumina, Inc.
9885 Towne Centre Drive
San Diego, CA 92121-1975
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America) techsupport@illumina.com
www.illumina.com