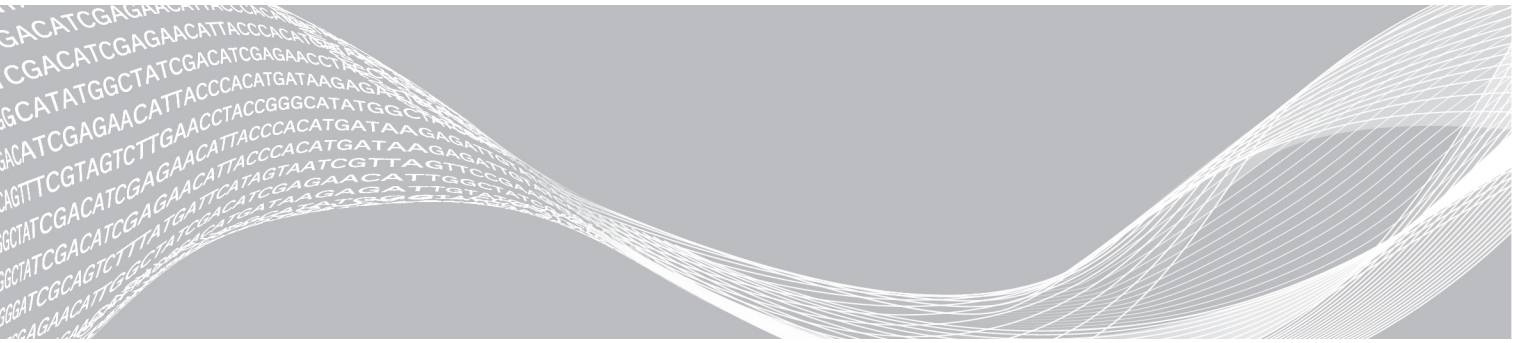
illumina®

BCL Convert

Software Guide



Document # 1000000094725 v00

ILLUMINA PROPRIETARY

June 2020

Introduction

The Illumina BCL Convert™ is a standalone local software app that converts the Binary Base Call (BCL) files produced by Illumina™ sequencing systems to FASTQ files. Additionally, BCL Convert provides adapter handling (through masking and trimming) and UMI trimming and produces metric outputs.

Additional Resources

The BCL Convert support pages on the Illumina support site provide additional resources. These resources include training, compatible products, and other considerations. Always check support pages for the latest versions.

Installation Requirements

The following are the minimum operating requirements for BCL Convert.

Hardware Requirements

Exact storage requirements vary based on throughput and sample count. See the BCL Convert pages on the Illumina support site for more information.

- Single multiprocessor or multicore computer.
- Minimum 32 GB RAM.

Software Requirements

- Linux CentOS 6 or higher
- Root access

Installation

BCL Convert is installed from a RPM package downloaded from the Illumina support site. The starting point is the binary executable `/usr/local/bin/bcl-convert`.

1 Install the RPM package using one of the following commands:
   - To install the software in the default location, enter:

     ```
     rpm --install <rpm package-name>
     ```

   - To specify a custom install location, enter:

     ```
     RPM --INSTALL --PREFIX <USER-SPECIFIED DIRECTORY>
        <rpm package-name>
     ```

Requirements to Run

BCL Convert requires the following files to be present in the run folder to complete analysis:

- BCL files (*.bcl, *.cbcl)
- Filter files (*.filter)
- Position files (*.locs, *.clocs, *s.locs)
- Aggregated files (*.bci)
- Run info file (*.xml)

- Config.xml— The config.xml file is only required for data produced by some systems. See the BCL Convert pages on the Illumina support site for more information.
- SampleSheet.csv

Sequencing Data

BCL Convert requires the following files to run. Different inputs might be required based on the sequencing system used to produce the data.

Input	Description
Base call files (*.bcl.bgzf)	Base call (BCL) files are compressed with the gzip (*.gz) or blocked GNU zip (*.bgzf) format.
Base call index files (*.bci)	Base call index files (BCI) files contain one record per lane in binary format. BCI files are acceptable as input by BCL Convert but are not used for analysis.
Concatenated base call files (*cbcl)	Concatenated base call (CBCL) files contain aggregated BCL data. Tiles from the same lane and surface are aggregated into one CBCL file for each lane and surface.
Filter files (*.filter)	Filter files are binary files that specify whether a given cluster passed filter.
Location files (*.locs, s.locs)	Location files (LOCS) are binary files that contain the cluster positions on the flow cell. CLOCS files are compressed versions of LOCS files.
Run information file (RunInfo.xml)	The run information file resides at the root level of the output folder. The file contains the run name, number of cycles, whether a read is an Index Read, and the number of lanes, swaths and tiles. If this file does not exist in the output folder, the software produces an error.
Configuration file (*.xml)	The configuration file resides in the BaseCalls folder and contains metadata on the sequencing run. The file is in XML format.

Sample Sheet

A sample sheet (SampleSheet.csv) records information about samples, the corresponding indexes, and other information that dictates the behavior of the software. The default location of the sample sheet is the root output folder. To specify any CSV file in any location use the command `--sample-sheet`. When a sample sheet does not exist in the default location and no sample sheet is specified in the command line, the software produces an error.

Sample Sheet Versions

Both sample sheets V1 and V2 are supported by BCL Convert. The following table displays the different supported options for V1 and V2.

Sample Sheet V1	Sample Sheet V2
Supports both [Settings] and [settings]. Neither are required.	Supports only [BCLConvert_Settings]. Required.
Unrecognized settings triggers a warning.	Unrecognized settings produces an error and analysis aborts.

Settings Section

The software uses the settings section of the sample sheet to specify adapter trimming, cycle, UMI, and index options.

Table 1 Adapter Trimming Specifications

Setting	Description	Default
`AdapterRead1`	The sequence of the Read 1 adapter to be masked or trimmed. To trim multiple adapters, separate the sequences with a plus sign (+) to indicate independent adapters that must be independently assessed for masking or trimming for each read. Allowed characters: A, T, C, G.	Not applicable
`AdapterRead2`	The sequence of the Read 2 adapter to be masked or trimmed. To trim multiple adapters, separate the sequences with a plus sign (+) to indicate independent adapters that must be independently assessed for masking or trimming for each read. Allowed characters: A, T, C, G.	Not applicable
`AdapterBehavior`	Defines whether the software masks or trims Read 1 and/or Read 2 adapter sequence(s). When AdapterRead1 or AdapterRead2 is not specified, this setting cannot be specified. • mask—The software masks the identified Read 1 and/or Read 2 sequence(s) with N. • trim—The software trims the identified Read 1 and/or Read 2 sequence(s)	trim
`AdapterStringency`	The minimum match rate that triggers masking or trimming. This value is calculated as MatchCount / (MatchCount+MismatchCount). Accepted values are 0.5–1. The default value of 0.9 indicates that only reads with ≥ 90% sequence identity with the adapter are trimmed.	0.9
`MinimumAdapterOverlap`	Do not trim any bases unless the adapter matches are greater than or equal to the user specified number of bases. At least one `AdapterRead1` or `AdapterRead2` must be specified to use `MinimumAdapterOverlap`. Allowed characters: 1, 2, 3.	1

Table 2 Cycle, UMI, and Tile Specifications

Setting	Description	Default
BarcodeMismatchesIndex1	The number of mismatches allowed for index1. Accepted values are 0, 1, or 2.	1
BarcodeMismatchesIndex2	The number of mismatches allowed for index2. Accepted values are 0, 1, or 2.	1
MinimumTrimmedReadLength	The minimum read length after adapter trimming. The software trims adapter sequences from reads to the value of this parameter. Bases below the specified value are masked with N.	35
MaskShortReads	The minimum read length containing A, T, C, G values after adapter trimming. Reads with less than this number of bases become completely masked. If this value is less than 22, the default becomes the MinimumTrimmedReadLength.	22
OverrideCycles	Specifies the sequencing and indexing cycles that should be used when processing the data. The following format must be used: • Must be same number of semicolon delimited fields in string as sequencing and indexing reads specified in RunInfo.xml • Indexing reads are specified with an I. • Sequencing reads are specified with a Y. UMI cycles are specified with an U. • Trimmed reads are specified with N. • The number of cycles specified for each read must sum to the number of cycles specified for that read in the RunInfo.xml. • Only one Y or I sequence can be specified per read. Example: Y151;I8;I8;Y151	Use reads as specified in the RunInfo.xml
SoftwareVersion	**[Optional]** Records the version of BCL Convert intended to be used for conversion. Only supported by sample sheet V2.	Not applicable

Data Section

The data section is required. Headers for the data section should be [Data] or [data] for sample sheet V1 and [BCLConvert_Data] for sample sheet V2. BCL Convert uses columns in the Data section to sort samples and index adapters.

Column	Description
Lane	**[Optional]** When specified, the software generates FASTQ files only for the samples with the specified lane number. Only one valid integer is allowed, as defined by the RunInfo.xml.
Sample_ID	The sample ID.
index	The Index 1 (i7) index adapter sequence.
index2	The Index 2 (i5) Index adapter sequence.
Sample_Project	**Optional** Can only contain alphanumeric characters, dashes, and underscores. Duplicate data strings with different cases (eg `sampleProject` and `SampleProject`) are not allowed. If these are used, analysis fails. This column is not used unless the command line option `--bcl-sampleproject-subdirectories` is used. See *Command Line Options* on page 7 for more information on command line options.

Running BCL Convert

Use the following instructions to run BCL Convert. Use *Command Line Options* on page 7 to modify the software operation as needed.

1 Open a command-line window.

2 Enter the following command and add options as needed:

```
nohup /usr/local/bin/bcl-convert
```

For example, the following command line populates the output directory with FASTQ files:

```
nohup /usr/local/bin/bcl-convert -bcl-input-directory
<RunFolder> --output-directory <OutputDirectory>
```

Command Line Options

Use the following options to control BCL Convert.

Option	Description	Default
`--bcl-input-directory`	A main command-line option that indicates the path to the run folder directory.	Not applicable
`--output-directory`	A required command-line option that indicates the path to demultuplexed fastq output. The directory must not exist, unless -f, force is specified.	Not applicable
`-f --force`	**[Optional]** Allow for the directory specified by the --output- directory option to already exist.	Not applicable
`--sample-sheet`	**[Optional]** Indicates the path to the sample sheet to specify the sample sheet location and name, if different from the default.	<--bcl-input-directory>/SampleSheet.csv
`--bcl-only-lane`	**[Optional]** Convert only the specified lane number. The value must be less than or equal to the number of lanes specified in the RunInfo.xml. Must be a single integer value.	All lanes as specified in the RunInfo.xml
`--first-tile-only`	**[Optional]** true—Only process the first tile of the first swath of the top surface of each lane specified in the sample sheet. false—Process all tiles in each lane, as specified in the sample sheet.	false

Option	Description	Default
`--strict-mode`	**[Optional]** true—Abort the program if any filter, locs, bcl, or bci lane files are missing or corrupt. false—Continue processing if any filter, locs, bcl, or bci lane files are missing. Return a warning message for each missing or corrupt file.	false
`--bcl-sampleproject-subdirectories`	**[Optional]** true—Allows creation of Sample_Project subdirectories as specified in the sample sheet. This option must be set to true for the `Sample_Project` column in the data section to be used.	false
`-h`, `--help`	Produce a help message and exit the application.	Not applicable
`-V`, `--version`	Produce a help message and exit the application.	Not applicable

Ulimit Settings

BCL Convert requires high ulimit settings for both the number of open files allowed and maximum user process. If a run fails due to maximum user processes being set too low, an error message stating "Resource temporarily unavailable" occurs. By default, BCL Convert attempts to set the ulimit soft limit for the number of open files (ulimit-n) to 65535 and the maximum user processes to 32768. If those values exceed the hard limits of the system, the soft limit is set to the hard limit.

Missing File Handling

If `--strict-mode` is set to false, BCL Convert executes certain behaviors when it finds missing or corrupt files. The following are the possible behaviors according to file type and status.

File Type	Status	Behavior
*.bcl	Missing or Corrupt	All base calls of the cycle in the corresponding lane and tile are replaced with `N` and a quality score of `#`.
*.cbcl	Missing or Corrupt	All base calls of the cycle in the corresponding lane and surface are replaced with `N` and a quality score of `#`.
*.cbcl	Corrupt	All base calls of the cycle in the corresponding lane and tiles are replaced with `N` and a quality score of `#`.
*.locs	Missing or Corrupt	Produce FASTQ files with an automatically generated unique header for all reads in the corresponding lane and tiles.
*.filter	Missing or Corrupt	No FASTQ entries produced for any reads in the corresponding lane and tiles.
*.bci lane	Missing or Corrupt	No FASTQ entries produced for any reads in the corresponding lane and tiles.

Analysis Methods

BCL Convert performs the following analysis:

- Demultiplexing
- Unique molecular identifier (UMI) trimming

- Adapter trimming and masking

Demultiplexing

BCL Convert produces one FASTQ file for each sample for each lane and read. Demultiplexing behaviors are as follows:

- When a sample sheet contains multiplexed samples, the software:
  - Places reads without a matching index adapter sequence in the Undetermined_S0 FASTQ file.
  - Places reads with valid index adapter sequences in the sample FASTQ file.
- When a sample sheet contains one unindexed sample, all reads are placed in the sample FASTQ files (one each for Read 1 and Read 2).
- All reads that do not demultiplex to the samples defined in the Data section of the sample sheet are placed in one FASTQ file named Undetermined_S0 per lane.
- When the Lane column in the Data section is not used, all lanes are converted. Otherwise, only populated lanes are converted.

UMI Trimming

The software is capable of trimming UMI sequences from the genomic or index sequences. The cycles of the sequencing read that correspond to the UMI are specified in the OverrideCycles parameter in the Settings section of the sample sheet. See the *Settings Section* on page 5 to set the OverrideCycles parameter.

The following are details of the behavior of reads specified as UMIs:

- UMI is always trimmed from the sequence.
- UMI sequence can be specified in the index and genomic reads. More than one UMI sequence can be specified per read.
- The specified UMI cycles are applied to all clusters. There is no mechanism to apply UMI based on lane or sample.
- UMI sequences can only be specified at the beginning and end of sequencing and index reads. UMIs cannot be located in the middle of a read.

Adapter Handling

The software can mask or trim user specified adapter sequences from read data so that those adapter sequences are not passed to any downstream analysis steps.

Additional details of the adapter handling capabilities are as follows:

- The software masks the identified adapter sequence with N so that the overall read length is constant across all clusters in the read.
- The software masks or trims the identified adapter sequence from the read. The length for each cluster may vary due to trimming.
- The software assumes that input adapter sequences can only contain A, C, G, or T.

Output Files

BCL Convert generates the following files as output:

- FASTQ files
- Demultiplex statistics file

- FASTQ list file
- Index hopping metrics file
- Top unknown barcodes file
- FASTQ complete file
- Log files
- Adapter metrics file

FASTQ Files

As converted versions of BCL files, FASTQ files are the primary output of the BCL Convert. Like BCL files, FASTQ files contain base calls with associated Q-scores. Unlike BCL files, which contain per-cycle data, FASTQ files contain the per-read data that most analysis applications require.

The software generates one FASTQ file for every sample, read, and lane. For example, for each sample in a paired-end run, the software generates two FASTQ files: one for Read 1 and one for Read 2. In addition to these sample FASTQ files, the software generates two FASTQ files per lane containing all unknown samples. FASTQ files for Index Read 1 and Index Read 2 are generated since the sequence is included in the header of each FASTQ entry.

FASTQ Files Directory

The software writes compressed, demultiplexed FASTQ files to the directory defined in the command line `<--output-directory`.

Reads with unidentified indexes are recorded in one file named Undetermined_S0_. If a sample sheet includes multiple samples per lane, the indices must be specified. If they are not, the software displays a missing barcode error and ends the analysis.

NOTE

The software allows one unindexed sample because identification is not necessary to sequence one sample. However, sequencing multiple samples requires multiplexing so the samples can be identified for analysis.

FASTQ List File

The FASTQ list file (*csv) provides an association between the sample indexes, lane, and the output FASTQ file names. Any BAM tags provided can be used in subsequent DRAGEN processing and are copied into the resulting BAM file.

The following columns are provided per unique sample_ID and lane combination:

- RGID: index1.index2.lane
- RGSM: Sample_ID
- RGLB: UnknownLibrary
- Lane
- Read1File: path to Read 1 FASTQ file
- Read2File: path to Read 2 FASTQ file

Metrics Output

BCL Convert produces the following metrics output files. All the metrics output files are located in the reports folder of the output directory.

- Demultiplex statistics file (*.csv)
- Index hopping metrics file (*.csv)
- Index metrics out file (*.bin)
- Adapter metrics file (*.csv)

Demultiplex Statistics File

The demultiplex statistics file (*.csv) provides the number of passing filter reads that are assigned to each sample in the sample sheet, as well as the set of undetermined reads treated as one sample. The file also contains information about the quality scores of bases in the passing filter reads assigned to each sample. For each sample ID in each lane, the following information is provided:

- Number of reads
- Number of perfect Index Reads
- Number of One Mismatch Index Reads
- Number of >= Q30 Bases (Passing Filter)
- Mean Quality Score (Passing Filter)

Index Hopping Metrics File

The index hopping metrics file (*.csv) contains the number of reads for each expected and hopped index for unique, dual index runs. The count is only reported for UDIs (Unique Dual Indices) per lane, where no barcode collision is detected in either index. Each pair of entries within each index must have a distance between bases of at least 2n+1, where n is the barcode mismatch tolerance specified for the index, in order for index hopping metrics to be output for the given lane.

For non index runs, single index runs, or lanes that do not contain UDIs, the file is output with only the header.

Index Metrics Out File

The index metrics out file is a binary file in BIN format that contains index statistics for each sample and index combination per lane provided to BCL Convert in the sample sheet. The content of the file is documented as follows:

- Byte 0: file version number (2)
- The remaining bytes represent records, which are composed of the following information:
  - 2 bytes: lane number (uint16)
  - 4 bytes: tile number (uint32)
  - 2 bytes: read number (uint16)
  - 2 bytes: indexLength, the length in bytes of index name (uint16)
  - indexLength bytes: string representing index name
  - 8 bytes: number of occurrences of index (uint64)
  - 2 bytes: sampleLength, the length in bytes of the sample name (uint16)
  - sampleLength bytes: string representing sample name
  - 2 bytes: projectLength, the length in bytes of the project name (uint16)
  - projectLength bytes: string representing project name

Adapter Metrics File

The adapter metrics file (*.csv) reports the number of bases detected to belong to adapters for each read per sample ID. This information allows for the detection of the genomic yield by subtracting the count of adapter sequence bases. Adapter sequences can be trimmed or masked from the output, depending upon the settings.

Each Read Group is reported, defined as the unique combination of lane, sample ID and index pair. The columns are left blank when they do not apply to the given sample. For a run without adapters, the file is output with only the header. The headings are as follows:

Heading	Description
`Lane`	Lane number.
`Sample_ID`	Sample ID.
`index`	The Index 1 (i7) index adapter sequence.
`index2`	The Index 2 (i5) index adapter sequence.
`R1_AdapterBases`	Number of bases trimmed or masked from the corresponding read for the corresponding and sample. This includes Ns that replace base pairs according to the MaskShortReads setting.
`R1_SampleBases`	Number of PF bases included in the corresponding read not belonging to an adapter for the corresponding and sample.
`R2_AdapterBases`	Number of bases trimmed or masked from the corresponding read for the corresponding sample, including Ns that replace base pairs according to the `MaskShortReads` setting.
`R2_SampleBases`	Number of PF bases included in the corresponding read not belonging to an adapter for the corresponding and sample.
`# Reads`	Number of reads.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

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

Illumina Customer Support Telephone Numbers

Region	Toll Free	Regional
North America	+1.800.809.4566	
Australia	+1.800.775.688	
Austria	+43 800006249	+43 19286540
Belgium	+32 80077160	+32 34002973
China	400.066.5835	
Denmark	+45 80820183	+45 89871156
Finland	+358 800918363	+358 974790110
France	+33 805102193	+33 170770446
Germany	+49 8001014940	+49 8938035677
Hong Kong, China	800960230	
Ireland	+353 1800936608	+353 016950506
Italy	+39 800985513	+39 236003759
Japan	0800.111.5011	
Netherlands	+31 8000222493	+31 207132960
New Zealand	0800.451.650	
Norway	+47 800 16836	+47 21939693
Singapore	+1.800.579.2745	
South Korea	+82 80 234 5300	
Spain	+34 911899417	+34 800300143
Sweden	+46 850619671	+46 200883979
Switzerland	+41 565800000	+41 800200442
Taiwan, China	00806651752	
United Kingdom	+44 8000126019	+44 2073057197
Other countries	+44.1799.534000	

Safety data sheets (SDSs)—Available on the Illumina website at support.illumina.com/sds.html.

Product documentation—Available for download from support.illumina.com.

Document # 1000000094725 v00

Illumina
5200 Illumina Way
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com

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

illumina®