BCL Convert Standalone v4.2.4 Software Release Notes

INTRODUCTION

These Release Notes detail the latest release of BCL Convert, including known issues.

BCL Convert converts per cycle binary data output by Illumina sequencers containing basecall files and quality scores to per read FASTQ files.

NEW FEATURES

- Output legacy stats for parity with *bcl2fastq2*
  - Using the option `--output-legacy-stats=true` (disabled by default) enables all the legacy stats outputs from *bcl2fastq2* in addition to existing report files:
    - DemultiplexingStats.xml
    - ConversionStats.xml
    - AdapterTrimming.txt
    - FastqSummaryF#L#.txt
    - DemuxSummaryF#L#.txt
    - Stats.json
    - Index.html, Lane.html, laneBarcode.html, tree.html
  - Will be output to the `<output directory>/Reports/` legacy folder.
- Support both 1 / 0 and True / False for command-line options and sample-sheet settings, for the `TrimUMI` and `CreateFastqForIndexReads` settings to improve backwards compatibility with *bcl2fastq2*.
- Make the combined index collision checking default to enabled for all lanes. Implement a new `IndependentIndexCollisionCheck` option to replace `CombinedIndexCollisionCheck`.
  - This important change reverts a strict check on dual index collisions added to BCL based on customer feedback. With this change, the default behavior matches *bcl2fastq2* and adds an option to change the behavior.
  - `IndependentIndexCollisionChecks` enables a stricter barcode collision check that requires only a single index read to collide rather than both, to abort.
    - Disabled by default.
    - Can provide lane numbers to specify which lanes to change the collision check behavior.
  - Deprecation the `CombinedIndexCollisionCheck` option that was introduced in v4.1.5, which enabled a more relaxed type of collision check that requires both index reads to have a collision to abort.

Version	Index collision check behavior
3.9.x	Relaxed by default. No option to change. Matches *bcl2fastq2*
3.10.x and 4.0.x	Strict by default. No option to change.
4.1.5	Strict by default. New option `CombinedIndexCollisionCheck` introduced to optionally relax the strictness
4.1.7 and 4.2.x	Relaxed by default. Remove `CombinedIndexCollisionCheck` option, add new `IndependentIndexCollisionCheck` option to allow optional strict checking. Default matches *bcl2fastq2*

- Support per sample settings, introduced with the NovaSeq X series instrument.
  - DRAGEN and bcl-convert v4.1 and later support the following settings as columns in the `[BCLConvert_Data]` section, allowing them to be specified differently for each sample:

`OverrideCycles, BarcodeMismatchesIndex1, BarcodeMismatchesIndex2, AdapterRead1, AdapterRead2, AdapterBehavior, AdapterStringency.`

- o These per-sample settings can be specified by omitting the setting from the `[BCLConvert_Settings]` section and instead adding a column to the `[BCLConvert_Data]` section with that setting name. Settings that do not apply to a sample (e.g., 'index2' if i5 is masked out for that sample) must be blank or 'na' in the entry for that sample.
- o This feature is only supported on version two (v2) sample sheets, and no setting can be specified both globally and per-sample. Specifying `OverrideCycles` differently per-sample allows mixing of different pools into the same lane but must still obey barcode mismatch constraints for all cycles that are used for demultiplexing by any sample in that lane.
- o The software will detect all conflicts between samples at the beginning of the conversion run, even between different pools.
- o Different strategies such as UMI indexes and dual-index inputs can be combined, provided `IndependentIndexCollisionCheck` is not enabled.

RESOLVED ISSUES

- Fix for index sequences missing from fastq headers when using `--no-sample-sheet` setting.
- Fix for BCL behavior being different than *bcl2fastq2* with respect to "`Sample_Name`" and "`Sample_Project`". In the special case of "`Sample_Name`" == "`Sample_ID`", *bcl2fastq2* does not create a "`Sample_ID`" subdirectory. This change makes bcl-convert behavior the same.
- Fix for false barcode collision reports when one sample's index is entirely trimmed out and another sample's index exists.
- Fix for BCL not aborting when single-index datasets have barcode collisions.
- Fix for incorrect yieldQ30/qscoresum stats when there is UMI in the first part of a read and TrimUMI is enabled (true by default).
- Fix for a false error when using global `BarcodeMismatchesIndex2` and a sample does not use index `BarcodeMismatchesIndex1`, when the sample sheet contains both single & dual-index samples.
- Fix for BCL failing with a "vector::reserve" message for mixed index strategies.
- Fix for BCL outputting many duplicate error messages for missing CBCL files.
- Fix for BCL Convert abort with an empty `[BCLConvert_Settings]` section in v2 sample sheets.
- Fix an issue where Undetermined FASTQ files are still created even after setting `--bcl-only-matched-reads` to true.
- Fix for failing validation check "No more than 27 total bases can be used as index bases".

KNOWN ISSUES

- If a directory is specified as input to `--sample-sheet`, BCL Convert will hang at the beginning of a run while trying to copy that path as a file to `<outdir>/Reports/SampleSheet.csv`
- BCL Convert does not validate when "`Logs`" or "`Reports`" is provided for a `Sample_Project`, and the software will be unable to create the subdirectories if these strings are provided.

- BCL Convert does not support the `--first-tile-only` option being specified for SP flow cells, but the new `--tiles` option can be used as a substitute.
- Does not error when no tile list exists in the RunInfo.xml file and `--tile` or `--exclude-tiles` is specified in the command line
- BCL will abort with an exit code of 1 if a .filter file is missing from a single lane dataset
- There will be many duplicate adapters error for all `sampleIDs`, and for any cbcl errors
- When an index collision exists in a lane that has been excluded via the `ExcludeTiles` setting, the software will still error as it is a sample sheet validation error
- Performance issues occur with very high sample counts

- BCL Convert has differences in legacy stats compared to *bcl2fastq2*
    - Formatting differences
    - JSON library to parse the data will work
    - AdapterTrimming.txt
        - Different number of cycles listed (*bcl2fastq2* lists too many cycles)
            - *bcl2fastq2* enumerates too many cycles for some reads, particularly when one read is shorter than the other, and sets the "PercentageOfBases" value of the extra cycles to value 0. The extra cycles are found at the beginning of the enumerated cycles. For example, when Read1 is 88 cycles and Read2 is 76 cycles, *bcl2fastq2* will enumerate cycles 0-88 for Read1 and 0-88 for Read2. On the other hand, bcl-convert will enumerate cycles 0-88 for Read1 and 0-76 for Read2. The cycles 0-11 for *bcl2fastq2* will have value 0.
        - Different values for "PercentageOfBases" (*bcl2fastq2* values are based on incorrect total number of cycles)
            - "PercentageOfBases" value differs between *bcl2fastq2* and bcl-convert, particularly when OverrideCycles is used. This is because *bcl2fastq2* counts the OverrideCycles-trimmed bases in the total bases, while bcl-convert omits those bases from the total count. For the case in which *bcl2fastq2* lists 98 cycles for a 94 cycle read, the *bcl2fastq2* "PercentageOfBases" value is equal to BCL-Convert value * (94 / 98), which is inaccurate.
        - Rounding differences (*bcl2fastq2* is less accurate)
    - ConversionStats.xml
        - Difference in "Raw" cluster metrics (*bcl2fastq2* uses statistical best guess, is not deterministic)
            - Even though "Pf" cluster metrics are identical between *bcl2fastq2* and bcl-convert for all samples, the "Raw" cluster metrics differ specifically for 'Sample name="Undetermined"' and 'Sample name="all"' (but are identical for all other samples).
        - Top Unknown Barcodes differ in sequence and value (*bcl2fastq2* uses statistical best guess, is not deterministic)
        - Difference in cluster count for "Raw" stats when CreateFastqForIndexReads is enabled
            - Having CreateFastqForIndexReads on causes *bcl2fastq2* to sometimes double the intended value, even for single-index inputs. This is likely a bug in *bcl2fastq2*.
    - DemuxSummaryF1L1.txt and DemuxSummaryF1L2.txt
        - Different "Most Popular Unknown Index Sequences" sequences and counts (*bcl2fastq2* uses statistical best guess, is not deterministic)
        - Different rounding of decimals (*bcl2fastq2* is less accurate)

- o **FastqSummaryF#L#.txt**
    - ▪ Difference in count for "NumberOfReadsRaw" stats when CreateFastqForIndexReads is enabled
        - • Having CreateFastqForIndexReads on causes *bcl2fastq2* to sometimes double the intended value, even for single-index inputs. This is a bug in *bcl2fastq2*.
- o Stats.json
    - ▪ *bcl2fastq2* has a space before the colon ':' but bcl-convert does not
        - • *bcl2fastq2* example: '<key>: <value>' OR '<key> : <value>'
        - • bcl-convert example: '<key>: <value>'
    - ▪ Top Unknown Barcodes differ in sequence and value (*bcl2fastq2* uses statistical best guess, is not deterministic)
- o Html reports
    - ▪ Differences in Top Unknown Barcodes sequences and counts (*bcl2fastq2* uses statistical best guess) *(same as ConversionStats.xml)*
    - ▪ Difference in values for "Clusters (Raw)" and "% PF Metrics" when CreateFastqForIndexReads is enabled
        - • Having CreateFastqForIndexReads on causes *bcl2fastq2* to sometimes double the intended value, even for single-index inputs. This is a bug in *bcl2fastq2*.
- o IndexMetricsOut.bin
    - ▪ Bcl Convert will use the value for Sample_Project from the Sample Sheet in the IndexMetrisOut.bin file even when the use of the Sample_Project column is disabled via the command line (no differences to *bcl2fastq2* as it is a new feature to disable the column)
        - • When Sample_Project is disabled, bcl-convert is expected to behave exactly as if the Sample_Project column did not exist in the sample sheet. In this case, the value for Sample_Project should be the following string: "default". However, bcl-convert uses the value of Sample_Project from the sample sheet.
        - • Note that the Sample_Name column behaves correctly - bcl-convert correctly uses Sample_ID in IndexMetricsOut.bin instead of Sample_Name when the Sample_Name column is disabled.

RELEASE HISTORY

Revision	Release Reference	Originator	Description of Change
00	1088629	Cobus De Beer	Initial release