CASAVA 1.8.2 Release Notes

New Features:

- CASAVA 1.8.2 is the first version of the software package that supports the Nextera dual-indexing protocol. Single-index data continue to be supported in CASAVA 1.8.2 as before. The application of dual-indexing in CASAVA is described below:

    o In the sample sheet, a dash sign (-) should be used to separate the two indexes in the Index column.
    o In the demultiplexing workflow (configureBcltoFastq.pl), the --use-bases-mask parameter will need to be included if dual-indexing is indicated in the sample sheet:

        ▪ Example: --use-bases-mask=Y*,I*,I* (for single-end runs) or --use-bases-mask=Y*,I*,I*,Y* (for paired-end runs)

    o The parameter **--mismatches** has also been modified to support dual-indexing.

        ▪ Example: **--mismatches**=1,1 (allows 1 base mismatch in each of the index reads)
        ▪ By default, CASAVA 1.8.2 will not allow mismatches in either index read (**--mismatches**=0,0).

- CASAVA 1.8.2 is also the first version of CASAVA to perform adapter masking. The application of adapter masking in CASAVA is described below:

    o In the demultiplexing workflow (configureBcltoFastq.pl), the **--adapter-sequence** parameter is used to indicate the location of the FASTA file containing the adapter sequence(s) to be masked.
    o If only one sequence is specified within the FASTA file, the same sequence is used for all sequencing reads. If multiple sequences are specified, the first sequence is used for the first sequencing read, the second sequence for the second sequencing read, and so on.
    o Index reads are not affected by the adapter masking.

Improvements

- By default, fastq.gz files generated through the demultiplexing workflow (BcltoFastq.pl) now only contain pass-filter (PF) reads. To include all reads, the **--with-failed-reads** parameter needs to be included in the demultiplexing workflow (configureBcltoFastq.pl).
- CASAVA 1.8.2 now includes the index sequence in the header of undetermined fastq files to help troubleshoot potential indexing issues.
- CASAVA 1.8.2 can now generate one fastq file per read if the parameter **--fastq-cluster-count** is set to 0 in the demultiplexing workflow (configureBcltoFastq.pl).

CASAVA 1.8.1 Release Notes

- Minor bug fixes

    o Issue that caused the "Align-as-you-go" workflow to fail has been addressed.

        ▪ Please see CASAVA 1.8.0 Release Notes for the "Align-as-you-go" workflow.

    o Archival BAM now includes both mapped and unmapped read pairs.

    o QseqtoFastq now properly designates PF/non-PF reads.

CASAVA 1.8.0 Release Notes

(pg. ##) – User Guide page number

The primary scripts for CASAVA secondary analysis have changed in CASAVA 1.8.0. The new required workflows are as follows:

- (pg.22) configureBclToFastq.pl– Script which converts a directory of BCL files (from RTA/OLB) into a directory of compressed Fastq files, separated by Project/Sample. An example command line is shown below:

  {CASAVA-1.8.0 Install Directory}/bin/configureBclToFastq.pl

  o (pg. 36) CASAVA 1.8.0 expects this script to be run in the BaseCalls folder by default. To specify another input directory, the parameter **--input-dir** should be used.
  o (pg. 36) The results will be placed into the folder BaseCalls/../../../Unaligned by default.
    To specify another output directory, the parameter --output-dir should be used.
  o (pg. 36) A sample sheet is required for Indexed Runs or Runs where only a subset of lanes will be analyzed. To specify the location of a sample sheet, the parameter -- sample-sheet should be used. The sample sheet is optional for all other Runs.
  o (pg. 165) To convert a directory of qseq files, please use the configureQseqToFastq.pl script.
  o (pg. 36) By default, the number of clusters per compressed fastq file is set at 4,000,000.
    To adjust this number, use the parameter **--fastq-cluster-count**

- (pg. 48) configureAlignment.pl (formerly GERALD.pl) – Script which configures alignment on a directory of compressed Fastq files, and creates alignment directories separated by Project/Sample. An example command line is shown below:

  {CASAVA-1.8.0 Install Directory}/bin/configureAlignment.pl {Path to config file} --make

  o NOTE: CASAVA 1.8.0 expects this script to be run in the Unaligned folder by default. To specify another input directory, the parameter EXPT_DIR should be used.
  o NOTE: The results will be placed into the folder {Run Folder}/Aligned by default. To specify a different output directory, the parameter OUT_DIR should be used.
  o NOTE: By default, ELAND_FASTQ_FILES_PER_PROCESS is set at 3. If you wish to adjust this value, please ensure that the **--fastq-cluster-count** value (from configureBclToFastq.pl) multiplied by the ELAND_FASTQ_FILES_PER_PROCESS value equals between 10-13 million reads to ensure optimal ELAND performance.

- (pg. 88) configureBuild.pl (formerly run.pl) – Script which configures variant calling on one or more Sample Alignment directories.

  {CASAVA-1.8.0 Install Directory}/bin/configureBuild.pl –id {Path to Sample Alignment Directory} and either -- refSequences={Path to Single-Entry Fasta Directory } or --samtoolsRefFile={ Path to Multi- Reference Fasta File}

  o NOTE: The results will be placed into the folder {Run Folder}/Build by default. To specify a different output directory, the parameter **-od** should be used.

- o NOTE: The **--refSequences** parameter should be used if **ELAND_GENOME** was used for alignment. The **--samtoolsRefFile** parameter should be used if **SAMTOOLS_GENOME** was used for alignment.
- o NOTE: To configure RNA variant calling analysis, please use the script **configureRnaBuild.pl** (formerly runRNA.pl)

General Updates

- (pg. 116) CASAVA 1.8.0 requires Cmake version 2.8 and Boost version 1.44. Both packages will be installed automatically if not detected in your environment during installation.
- Full paths should be used for all command line options for all workflow scripts.

BCL to Fastq Converter (configureBclToFastq.pl)

- (pg. 39) The output results for Fastq conversion and alignment will be organized by project.
  - o The project folders allow users to naturally group samples processed on a flow cell. Within the projects folder, there will be a folder for each sample.   Within a sample folder, the FASTQ files for the sample will be written such that each file is of a convenient size for downstream processing.
  - o The name of each FASTQ file provides the sample, the index used, the lane, and whether the file contains the first or second read. The names are based on information obtained from the sample sheet. If no sample sheet is provided, a default sample sheet is assumed. The same folder structure will accommodate single sample and indexed runs.
- The Converter expects the following RTA outputs (depending on version):
  - o RTA 1.8 and previous - CASAVA expects pos.txt files in the Intensities directory and filter files in the BaseCalls directory.
  - o RTA 1.9 - CASAVA expects .locs in the Intensities/L00* directories and filter/control files in the BaseCalls/L00* directories.
  - o RTA 1.10+ - CASAVA expects .clocs in the Intensities/L00* directories and filter/control files in the BaseCalls/L00* directories.
- NOTE: All reads are included in the FASTQ files, not just filtered reads. There is no option to parse out the non-passing filter reads from the generated FASTQ files.

Alignment (configureAlignment.pl)

- (pg. 57) The **ELAND_GENOME** parameter needs to be pointed to a folder of regular (non-squashed) fasta files. You can also use the **SAMTOOLS_GENOME** to point to a single multi-entry fasta file.
- (pg. 58) Semi-repeat resolution and the orphan-aligner are new features in ELAND, and are turned on by default in CASAVA 1.8. To use full-repeat resolution (which gives increased alignment % but takes almost twice as long to run as semi-repeat resolution), use the following parameter:

  INCREASED_SENSITIVITY --sensitive.

- (pg. 53) If you use the SAMTOOLS_GENOME parameter and your fasta files contain characters that CASAVA 1.8 restricts, you can choose to ignore this restriction to perform alignment by adding the following parameter to the GERALD config file:

  CHROM_NAME_VALIDATION off

  o NOTE: ELAND v2e (specifically the orphan aligner) cannot process chromosome names with a colon (:) or backslash (\). These are restricted characters by default, but can be used if the CHROM_NAME_VALIDATION is turned off. It is not recommended to use fasta files that contain colons or backslashes in the header entries.

- (pg. 54) For eland_rna, the new parameter **ELAND_RNA_GENOME_ANNOTATION** can be pointed at either a refFlat.txt.gz or seq.gene.md file.

- (pg. 67) The parameter **DATASET_POST_RUN_COMMAND** can be used in the config file to specify commands to be run upon the completion of alignment on each FASTQ files. To specify a command to be run at the completion of all analysis, please use the parameter **POST_RUN_COMMAND** on the make command line, as shown in the example below:

  make –j 8 POST_RUN_COMMAND:='echo Alignment completed'

Post-Alignment (configureBuild.pl/configureRnaBuild.pl)

- (pg. 103) Read alignments are now stored in the post-alignment build as BAM files. These files can be found in the Parsed directory as {chr. Name}/bam/sorted.bam.

  o The BAM output for RNA-Seq builds now represents each read which crosses a splice- junction as a single record, using the CIGAR 'N' character to indicate an alignment spanning an intron. This should allow the alignments to be more easily viewed in the Broad IGV and used in third-party analysis tools.

- (pg. 90) Post-alignment can now be run in archival mode. In this mode all reads provided as input for post-alignment will be stored in their entirety in the post-alignment BAM files and the entire build can be optionally consolidated to a single archival BAM file. This mode can be activated by using the "--sortKeepAllReads" flag. Note the following changes to the BAM files in archival mode:

  o Purity filtered and PCR duplicate reads will be stored in the chromosome BAM files with the appropriate annotations. These reads will not be used for variant-calling or RNA-Counting.

  o A new directory will be created in the Parsed folder called 'notMapped'. This directory will contain all read pairs where both reads could not be mapped, or single reads which could not be mapped in a single-ended build. These reads are divided into the following categories: 1) whether no match was found, 2) the reads could not be uniquely placed, 3) a qc-failure occurred, or 4) the reads were repeat-masked by ELAND.

- (pg. 147) CASAVA 1.8 uses a new snp-caller which models the probability distribution of diploid genotypes at each site. Snp and site-genotype files now provide a most likely genotype together with snp and genotype Q-scores for each site.

- (pg. 141) The CASAVA 1.8 indel caller now provides calls for dense, overlapping and larger indels. By default indels of up to 300 bases will be genotyped for a paired-end run. In addition the indel caller now reports and genotypes open-ended breakpoints for all other large, complex or incomplete indels.

New Summary Documents

- (pg. 42) Demultiplex_Stats.htm

    o Summary File created at the end of the BCL to Fastq Converter (configureBclToFastq.pl).
    o Located in the Unaligned/Basecall_Stats_{Flow Cell ID} folder
    o Includes a summary of Barcode Lane Statistics and Sample Information for the entire Flow Cell

- (pg. 74) Sample_Summary.htm

    o Project Summary File created at the end of the project Alignment (configureAlignment.pl)
    o Located in the Aligned/{Project Name}/ Summary_Stats_{Flow Cell ID}
    o Barcode_Lane_Summary.htm – more detailed summary file, per barcode lane.

        ▪ Located in the Aligned/{Project Name}/ Summary_Stats_{Flow Cell ID}

    o **Mismatch Rate** is the new term for Error Rate, to account for the fact that mismatches include sequencing errors as well as SNV's.

- (pg. 79) Flowcell_Summary_{Flow Cell ID}.htm

    o Summary File created at the end of the Alignment (configureAlignment.pl)
    o Located in the Aligned folder
    o A summary of all Sample_Summary files (for all Projects on the Flow Cell).

Align-as-you-go

- (pg. 38) CASAVA-1.8 provides a mechanism to align the first read of a paired-end run before completion of the run. This can be done by using the target `r1` available in the Makefiles for both the BCL-to-FASTQ converter and the alignment. These targets can be invoked at any time after the last read has started (after completion of the indexing read for multiplexed runs). They are invoked by typing the following command while in the appropriate directory:

    make r1

    o As usual, the "-j <n>" command line option is supported to indicate up to <n> processes in parallel. However, for the BCL conversion, the maximum level of parallelization is 8.
    o For instance, assuming that the BCL conversion to FASTQ and the alignment have been configured at the default locations (Unaligned and Aligned subdirectories of the run folder) and that the current directory is the run folder, the conversion and the alignment of the first read can be done using the following commands:

```
cd Unaligned
make -j 8 r1
cd ../Aligned
make -j 16 r1
```

- o For the situation where the jobs are submitted to a queuing system, or any situation where the conversion of the BCLs has to be asynchronous, it is possible to use the `Makefile` variable `POST_RUN_COMMAND_R1` to automatically start the alignment of read 1 at the end of the BCL conversion. That variable can be set either as an environment variable, or on the make command line, in which case the command for both the conversion and the alignment would be:

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

Validation Datasets

- Two validation datasets have been included with CASAVA 1.8.0 specifically to ensure your version of CASAVA 1.8.0 has been successfully installed. These datasets start with BCL files and run all the way through Variant Calling.

  - o After the completion of each validation run, CASAVA 1.8.0 will print out whether each main part of CASAVA—Demux (BCLtoFastq), Alignment, and Build (Variant Calling)-- has Passed or Failed.
  - o NOTE: These datasets do not represent the quality of actual sequencing data generated using Illumina technology. The results from the validation datasets were not designed to be analyzed further.

- To run end-to-end validation on the DNA (small) test data set included with CASAVA, you can use the following commands:

```
{CASAVA-1.8.0 Install Directory}/bin/configureValidation.pl --output-dir ./ValidationDefault
make -C ValidationDefault
```

> use make -j <parallel jobs> to speed up the conversion

- To run end-to-end validation on the indexed RNA (large) test data set included with CASAVA, you can use the following commands:

```
{CASAVA-1.8.0 Install Directory}/bin/configureValidation.pl
--source-dir {CASAVA-1.8.0 Install Directory}/share/CASAVA-1.8.0/examples/Validation/110120_P20_0993_A805CKABXX
--output-dir ./ValidationMultiplexed

make -C ValidationMultiplexed
```

> use make -j <parallel jobs> to speed up the conversion