DRAGEN v3.2.8 Software Release Notes

January 22, 2019

Introduction

This release note details the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform since the package containing DRAGEN v3.0.2.

If you are upgrading from a version prior to DRAGEN v3.0.2, please review the release notes for DRAGEN v3.0.2 for a list of features and bug fixes introduced in that version.

The software package includes:
- DRAGEN SW Intel Centos 6 - dragen-3.2.8.el6.x86_64
- DRAGEN SW Intel Centos 7 - dragen-3.2.8.el7.x86_64
- DRAGEN SW IBM Centos 7 - dragen-3.2.8.el7.ppc64le

Contents

New Features

- Accuracy improvements to Small Variant Pedigree Calling.
- Added support for continuous allele-frequency for Mitochondrial Small Variant calling.
- Added Manta 1.4.0 Structural Variant Caller to the workflow.
- Added support for Force Genotyping in Germline Small Variant calling.
- Added support for CNV self-normalization and generation of Kmer hashtable, with major improvements in FP.
- Added orientation bias filtering for FFPE artifacts in Somatic Small Variant calling
- Added new coverage metrics.
- Ability to run all the supported callers in a single workflow (map/align + VC + SV + CNV + RE).
- New BCL conversion interface (see below).

Fixes and Improvements

- Improved run time of gVCF output in BP_RESOLUTION mode.
- Improved compression of gVCF files.
- .
- Changes to the VCF and gVCF output formats (see below).

DeNovo Pedigree Calling

The SNP and Indel pedigree calling algorithm has been improved for accuracy. A brief summary of the changes is listed below. Please refer to the *User Guide section for DeNovo Joint Calling* for full details.

- When using a pedigree file for joint calling, DeNovo variants are marked in the FORMAT/DN field along with an associated quality score in the FORMAT/DQ field. Both new fields will be on the proband sample column

Example DeNovo variant

```
chr10     10370725          .          T          C
45.02     .
AC=1;AF=0.167;AN=6;DP=118;FS=1.170;MQ=244.54;MQRankSum=5.528;QD=1.07;ReadPosRan
kSum=-0.466;SOR=0.914          GT:AD:AF:DP:GQ:FT:PL:GL:GP:PP:DQ:DN
0/1:22,20:0.476:42:48:PASS:85,0,50:-8.456,0,-5:4.979e+01,6.730e-
05,5.300e+01:5,0,130:6.1924e+00:DeNovo
0/0:55,0:0.000:37:94:PASS:0,94,1485:.:.:0,51,217
0/0:54,0:0.000:36:85:PASS:0,85,1440:.:.:0,42,217
```

- Added support for sex chromosomes.
- Added support for duos and quads in the pedigree file.
  - For duos, there will be no DeNovo calls identified.
  - For quads, the proband is first child listed in the pedigree file and only DeNovo variants in the proband are identified. The second child (or other family member) is not used as part of the pedigree genotyping calculations.

Changes to the VCF and gVCF Output

- Added format fields ICNT, SPL in the gVCF file used for DeNovo pedigree calling.
- By default, the prefiltered VCF/gVCF file is no longer output. Added a new command line option, --vc-enable-prefilter-output, to control output of the prefiltered VCF/gVCF file.
- Added F1R2 and F2R1 annotations for somatic calls.
- Phasing changes:
  - The FORMAT/GT field now contains the phased genotype where applicable. The phased genotype used to be written in a separate FORMAT/PGT field that is now removed.
  - Renamed the FORMAT/PID to FORMAT/PS.

Changes to the BCL Pipeline

- The interface for BCL conversion has changed from DRAGEN 3.0 and is **not** backwards compatible. Please refer to the *User Guide section for Illumina BCL Data Conversion* for full details.
- The new interface is summarized below.

New Command Line Options

The supported parameters for DRAGEN BCL conversion are listed below:

Usage

```
dragen --bcl-conversion-only true <options> --bcl-input-directory <path> --output-
directory <path> [OPTIONS]
```

Optional parameters

```
--sample-sheet     <path>
--strict-mode      {true,false}
--first-tile-only  {true,false}
```

- The input BCL root directory and output directory must be specified.
  - NOTE: The specified input path is not the BaseCalls directory but three levels higher, and should contain a file named 'RunInfo.xml' and a directory named 'Data'.
- All output FASTQ data will be placed into the directory specified by '--output-directory', along with report data.
- To prevent user overwriting of existing data, **BCL conversion will not overwrite existing folder. The output-directory specified must not exist.**
- If the location of the sample sheet is not specified with the '--sample-sheet' parameter, then the program will look for the sample sheet in the directory of '--bcl-input-directory', named as 'SampleSheet.csv', or quit with error if not found.
- If a flow cell is missing files used during conversion or a BCL file is corrupt, a warning will be output to the console and some data will be skipped. If '--strict-mode true' is specified, any missing files will result in termination of the program with a nonzero exit value.
- If the parameter '--first-tile-only true' is given, only a single tile is converted (tile '1101'), for each lane to be converted, instead of the entire lane. This is only useful for providing a fast run for automation testing and debugging purposes.

Sample Sheet Options

- Most run options are now specified in the SampleSheet.csv file under the [Settings] section, with setting name and value separated by a comma.
- This replaces many settings that were previously provided on the command line.
- The following settings are supported:

Supported Sample Sheet Options

```
AdapterBehavior,{Trim,Mask}
AdapterStringency,<0.5-1.0>
BarcodeMismatchesIndex1,<0-2>
BarcodeMismatchesIndex2,<0-2>
MinimumTrimmedReadLength,<#>
MaskShortReads,<#>
OverrideCycles,<use-bases-mask>
```

- These settings are similar to identical command line options in bcl2fastq.
- 'OverrideCycles' is similar to the 'use-bases-mask' parameter in bcl2fastq, except that semicolons separate read specifiers instead of commas.
- Each setting is shown below with its default value:

Settings example

```
[Settings]
AdapterBehavior,Trim
AdapterStringency,0.9
BarcodeMismatchesIndex1,1
BarcodeMismatchesIndex2,1
MinimumTrimmedReadLength,35
MaskShortReads,22
OverrideCycles,Y151;I8;I8;Y151
```

- None of these settings have to be provided if the default values are desired.
- The default *value* for 'OverrideCycles' varies with read configuration: the default behavior is to include all read bases for output and all index bases for barcode matching.

OverrideCycles

- In addition to masking out sections of reads and indexes, the OverrideCycles parameter is used to specify UMI sequences as well.
- Use a 'U' prefix to indicate bases to extract as UMI sequences. These bases will be trimmed from the read output and added to FASTQ sequence headers as UMI tags.

Override Cycles example

```
OverrideCycles,U6Y145;I8;I8;Y151
```

Sample Sheet [Data] Section

- The [Data] section of the sample sheet must contain a column labeled 'Sample_ID', and each of its entries must be unique within each lane.
- If demultiplexing is being performed, an 'index' column must be provided, along with an 'index2' column if using dual indices. Each of these must contain base sequences used for demultiplexing just as before.

    - NOTE: Every sample must have the same number of index bases within the column, and this number should match the length of the index shown in the RunInfo.xml file.
    - If a subset of the index is desired, then use the OverrideCycles setting to adjust the expect number of bases in each index accordingly.

- If no 'Lane' column is provided in the [Data] section of the sample sheet, then every sample is extracted from every lane as per the following sample sheet example:

Sample Sheet [Data] Section example

```
[Settings]
AdapterRead1,AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

[Data]
Sample_ID,Sample_Plate,Sample_Well,I7_Index_ID,index,I5_Index_ID,index2,Sample_Project,Description
21599,,,D701,ATTACTCG,D501,TATAGCCT,,
21600,,,D701,ATTACTCG,D502,ATAGAGGC,,
21601,,,D701,ATTACTCG,D503,CCTATCCT,,
21602,,,D701,ATTACTCG,D504,GGCTCTGA,,
21607,,,D702,TCCGGAGA,D501,TATAGCCT,,
21608,,,D702,TCCGGAGA,D502,ATAGAGGC,,
21609,,,D702,TCCGGAGA,D503,CCTATCCT,,
21610,,,D702,TCCGGAGA,D504,GGCTCTGA,,
```

- It is occasionally useful to limit the lanes that will be converted during a single run, eg, if the files are written to a small local disk.
- To limit or specify which lanes must be converted, it is important to add a "Lane" column to the sample sheet.
- In the example below, two samples ("21599" and "21600") are multiplexed over 4 lanes. To convert only the first two lanes, simply remove information matching lanes "3" and "4".

Limiting which lanes are converted

```
[Settings]
AdapterRead1,AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

[Data]
Lane,Sample_ID,Sample_Plate,Sample_Well,I7_Index_ID,index,I5_Index_ID,index2,Sample_Project,Description
1,21599,,,D701,ATTACTCG,D501,TATAGCCT,,
2,21599,,,D701,ATTACTCG,D501,TATAGCCT,,
3,21599,,,D701,ATTACTCG,D501,TATAGCCT,, // remove these lines
4,21599,,,D701,ATTACTCG,D501,TATAGCCT,,
1,21600,,,D701,ATTACTCG,D502,ATAGAGGC,,
2,21600,,,D701,ATTACTCG,D502,ATAGAGGC,,
3,21600,,,D701,ATTACTCG,D502,ATAGAGGC,, // remove these lines
4,21600,,,D701,ATTACTCG,D502,ATAGAGGC,,
```

- It is important to keep a backup of the original sample sheet. This makes it possible to subsequently only convert lanes "3" and "4" by removing information pertaining to lanes "1" and "2".

Example

```
[Settings]
AdapterRead1,AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT

[Data]
Lane,Sample_ID,Sample_Plate,Sample_Well,I7_Index_ID,index,I5_Index_ID,index2,Sample_Project,Description
1,21599,,,D701,ATTACTCG,D501,TATAGCCT,, // during the second run remove these lines to only convert lanes 3
and 4
2,21599,,,D701,ATTACTCG,D501,TATAGCCT,,
3,21599,,,D701,ATTACTCG,D501,TATAGCCT,,
4,21599,,,D701,ATTACTCG,D501,TATAGCCT,,
1,21600,,,D701,ATTACTCG,D502,ATAGAGGC,,
2,21600,,,D701,ATTACTCG,D502,ATAGAGGC,,
3,21600,,,D701,ATTACTCG,D502,ATAGAGGC,,
4,21600,,,D701,ATTACTCG,D502,ATAGAGGC,,
```

Concurrent BCL and DRAGEN Analysis

The BCL conversion now makes use of the DRAGEN hardware. Previously the BCL conversion was a software only pipeline. To run the BCL conversion in parallel with DRAGEN analysis, the command line option "--bcl-use-hw false" needs to be used for the BCL execution.

Orientation Bias Filtering

Added support for orientation bias filtering on Somatic pipelines. See the *User Guide section for Orientation bias filter* for details.

Changes to the Small Variant Caller

It is now possible to specify the sex of the sample to the small variant caller using the "--vc-sex" option. Setting this option to MALE or FEMALE will allow the variant caller to use the correct ploidy for calling variants on the sex chromosomes.

- Specifically:
    - In male samples: Calls on chromosome Y are treated as haploid and calls on chromosome X will switch between diploid and haploid depending on whether the variant is in the PAR region or not.
    - In female samples: Calls on chromosome Y are treated as diploid and they will be filtered with a "PloidyConflict" filter tag, and calls on chromosome X are always treated as diploid.
    - If not set, the default behavior is to call all variants as diploid.

- New command line options:

v3.0 options	v3.2 options	Description
n/a	vc-tumor-sample-name	Variant caller sample name for tumor sample
n/a	vc-enable-triallelic-filter	Enable triallelic site filter for somatic mode.

Changes to the Mitochondrial Calling

Significant changes have been made to the processing of the mitochondrial chromosome compared to prior versions. In older versions, chrM was either handled as diploid (if no sex was provided on the command line) or haploid (if sex was provided on the command line).

Due to the nature of chromosome M, neither a haploid or diploid model is adequate because a given cell has many copies of the haploid mitochondrial chromosome and these mitochondria copies don't share the exact same DNA sequence. Typically, there are approximately 100 mitochondria in each mammalian cell, and each mitochondrion harbors 2–10 copies of mitochondrial DNA (mtDNA). For example, if 20% of the chrM copies have a variant, then the AF will be 20%. This is also referred to as "continuous allele frequency (AF)", and the expectation is that the AF of variants on the chrM is anywhere between 0% and 100%.

In DRAGEN 3.2, chrM is now processed through a continuous AF pipeline. In this case, a single ALT allele is considered, and the AF is estimated, and can be anywhere between 0%–100%.

As a result, the mitochondrial chromosome processing is now more accurate than in previous versions because low AF calls would not have been output previously. In the current version, you will be able to see all the single ALT allele variants on the mitochondrial chromosome, across the whole range of AF.

- Changes to mitochondrial VCF records
  - The FORMAT/GT is hard-coded to 0/1 even if the AF=100%. FORMAT/GT of 0/1 in this context should be interpreted as FORMAT/GT 1 with unconstrained AF.
  - The confidence score is output in INFO/LOD and the QUAL is marked as '.'.
- Known issues
  - The INFO/DP and FORMAT/DP are currently artificially low due to an improper default setting of the downsampling parameters. This will be addressed in a future release.

Changes to the CNV Caller

- The CNV user guide has been merged into the main user guide.
- Added support for additional options when using BAM/CRAM input.
  - The CNV caller can now be run as an additional caller alongside mapping/aligning and small variant calling starting from BAM/CRAM input, without the need to invoke it as a separate caller.
  - In addition, CNV can still be run as a standalone caller.
  - To support this additional functionality, the standalone CNV caller now requires that "--enable-map-align false" be set when mapping and alignment of the input BAM/CRAM is not needed. In DRAGEN v3.0 and earlier, the CNV caller with BAM/CRAM input only ran as a standalone caller and ignored the "--enable-map-align true" argument.
- Added support for self-normalization mode and generation of a Kmer hashtable.
  - Self-normalization is a new algorithm that allows for single sample CNV calling without a panel of normals.
  - To run self-normalization mode, rebuild the hashtable with option "--enable-cnv true". Then run CNV with "--cnv-enable-self-normalization true".
  - NOTE: The new Kmer file generated during hashtable generation is used to determine regions of low complexity. When running in self-normalization mode, these regions are

filtered out during CNV calling, reducing the total number of false positive calls compared to the prior DRAGEN versions.
    - Please refer to the *User Guide section on CNV* for details.

- Added support for calling on sex chromosomes in self-normalization mode

Changes to the Repeat Expansions Caller

- The command line options have been renamed for uniformity. The options from v3.0 are no longer supported.
- The Expansion Hunter version remains unchanged at v2.5.6.

v3.0 options	v3.2 options	Description
RepeatGenotyping.enable	repeat-genotype-enable	Enable calling of repeat-expansion variants
RepeatGenotyping.specs	repeat-genotype-specs	Directory with repeat-specification files
RepeatGenotyping.sex	repeat-genotype-sex	Sex of the sample; must be either 'male' or 'female'
RepeatGenotyping.ref-fasta	repeat-genotype-ref-fasta	FASTQ format reference file to use for repeat expansion
RepeatGenotyping.min-anchor-mapq	repeat-genotype-min-anchor-mapq	Minimum MAPQ of a read anchor
RepeatGenotyping.skip-unaligned	repeat-genotype-skip-unaligned	Whether to skip unaligned reads when searching for IRRs
RepeatGenotyping.min-baseq	repeat-genotype-min-baseq	Minimum quality of a high-confidence base call
RepeatGenotyping.region-extension-length	repeat-genotype-region-extension-length	How far from on/off-target regions to search for informative reads
RepeatGenotyping.min-score	repeat-genotype-min-score	Minimum weighted purity score required to flag a read as an in-repeat read; must be between 0 and 1
RepeatGenotyping.read-depth	repeat-genotype-read-depth	Read depth; calculated if not set
RepeatGenotyping.read-length	repeat-genotype-read-length	Read sequence length; calculated if not set

Changes to Metrics

Coverage Metrics

- Coverage metrics reports are now decoupled from variant calling.
- Default reports are always generated for the whole genome as well as for the vc-target-bed if specified.
- Up to three regions of interest can be specified by the user, along with additional report types requested for these regions.
- Default reports and user-requested reports are generated for each of the user-specified region of interest.
- Please refer to the *User Guide section on Metrics* for a detailed description of the coverage reports.

Updates to the command line

- The legacy command line options "--vc-enable-depth-of-coverage", "--vc-enable-history-of-coverage", "--vc-depth-intervals-bed" options are still supported but will soon be deprecated.
- Usage of the new command line options is recommended.

    - NOTE1: If the legacy options are specified the new options are deactivated and multiple coverage reports will not run.

- ERRATA to the User Guide:

    - NOTE2: The arguments to "--qc-coverage-report-N"must be either "full_res" or "cov_report"'. These are keywords to select the type of output report.
    - NOTE3: Specifying "--qc-coverage-report-N" is optional and will produce additional reports. By default, four mean coverage reports are generated.

Default reports generated

Report name	DRAGEN output file type
Hist	_hist.csv
Overall mean coverage	_overall_mean_cov.csv
Per contig mean coverage	_contig_mean_cov.csv
Predicted ploidy	_ploidy.csv

Optional reports generated

Report name	DRAGEN output file type
full_res	<coverage region>_full_res.bed
cov_report	<coverage region>_cov_report.bed

Example usage

```
$ dragen … \
--qc-coverage-region-1 <bed file 1> \
--qc-coverage-report-1 full_res \
--qc-coverage-region-2 <bed file 2> \
--qc-coverage-region-3 <bed file 3> \
--qc-coverage-report-3 cov_report full_res
```

Duplicate Reads

The Duplicate reads metric is split into two new metrics:

1. duplicates marked (map and aligned duplicates), and
2. duplicates removed (duplicates and mates - some of which may be unmapped).

DeNovo Metrics

- DRAGEN now produces DeNovo metrics when the Joint Caller is called with a pedigree file.
- The metrics report DeNovo SNP and Indel counts separately for DeNovo calls above a specified minimum DQ quality threshold.
- The pass thresholds can be set using new command line options.
    - --qc-snp-denovo-quality-threshold (default 0.05)
    - --qc-indel-denovo-quality-threshold (default 0.02)

Known Limitations

SW Installation

1. Install the appropriate release based on your Linux OS with command -> sudo sh <DRAGEN .run file>.
2. Cold boot the server so that the new SW if fully installed with the updated FPGA HW image. This will be not necessary if already on previous 3.0.2 release.

```
md5checksum:

1d2467560858de47854d8648dee749dc dragen-3.2.8.el6.x86_64.run

2cd0225f39b1215c573f30d7abb81ac6 dragen-3.2.8.el7.x86_64.run

7aaf569d21dffa2edc3742efcc6ca6e2 dragen-3.2.8.el7.ppc64le.run
```