DRAGEN v4.0.5
Software Release Notes

Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform v4.0.5.

Changes are relative to DRAGEN™ v4.0.3. If you are upgrading from a version prior to DRAGEN™ v4.0.3, please review the release notes for a list of features and bug fixes introduced in subsequent versions.

DRAGEN™ Installers, User Guide and Release Notes are available here:
https://support.illumina.com/sequencing/sequencing_software/dragen-bio-it-platform.html

The v4.0.5 software package includes installers for the on-site server:

- DRAGEN™ SW Intel Centos 7 - dragen-4.0.5-8.el7.x86_64.run
- DRAGEN™ SW Intel Oracle 8 - dragen-4.0.5-8.el8.x86_64.run

The following configurations are also available on request:

- Amazon Machine Image (AMI)
- Microsoft Azure Image (VM)
- RPM packages for Centos 7 for Amazon Web Services (AWS)

Deprecated platforms:

- Support for DRAGEN™ Server v1 FPGA cards have been deprecated since DRAGEN™ v3.10
- Support for Ubuntu has been deprecated since DRAGEN™ v3.9
- Support for Intel CentOS 6 has been deprecated since DRAGEN™ v3.8

Contents

Overview

Below is a summary of the changes included in v4.0.5. This is a minor update to DRAGEN™ v4.0 targeting important bug fixes across features, including key accuracy improvements, as detailed below.

DRAGEN™ v4.0.5 also includes a beta version of a CYP21A2 Caller for early evaluation.

Updates and Fixes

Germline Small Variant Caller

- **Machine Learning (ML) improvements and bug fixes**
    - ML accuracy improvements for hs37d5/hg19. ML models for hs37d5/hg19 are updated, which fixes ML accuracy regressions of up to 16% with ML vs non-ML for hs37d5, observed with v4.0.3.
    - hg38/hg19/hs37d5 references are now validated with ML.
    - Fixes for several issues with ML:
        - Fix an issue where the computation of PL from GP and PRI is missing, for hethom calls where ML prediction does not match the VC call.
        - Fix some accuracy discrepancies between runs in VCF vs gVCF output mode when ML is enabled.
        - Fix handling of `PL` and `GP` in `0/0` calls, which lead to an accuracy regression on Joint Calling.
        - Fix a segmentation fault in overlapping mate handling.
        - Correct handling of `N` bases on some `mapq=0` reads, which gave a different result between BAM and FASTQ input.
        - Fix a run-run output variation, due to non-deterministic allele bases on homref events.
        - Fix a potential crash during the handling of disqualified reads.
- Fix an issue where some variants are not emitted, when evidence BAM is enabled.
- Additional and improved checks for reads with cigar length of 0, which lead to false assertions.
- Fix an issue where all reads are disqualified in regions with ForceGT only events.
- **Joint Calling**
    - Improve denovo SNV INDEL performance.
    - Fixed an issue where joint calling output for Mito positions are always PASS.

Somatic Small Variant Caller

- Improve elevated INDEL FP+FN seen in v4.0.3 compared to v3.10. SNP FP is rescued by up to 3% and INDEL FP is reduced by up to 7%.
- Improve slow down on mapper due to NTD error estimation.
- Improve VC run time with NTD error estimation. Resolves the slow down up to 33% seen on certain types of somatic tumor only samples v4.0.3.
- Fix an Out-of-Memory error when evidence BAM is enabled on high depth samples.
- Reduce memory usage in somatic mode, to generally improve stability by avoiding potential Out-of-Memory errors.
- Refactor TMB and Germline Filtering, to reduce peak memory usage and resolve Out-of-Memory issues.
- Fixed an issue where the MNV length overflows a variable, leading to a corrupted TAG and a downstream Germline Filtering that asserts.

Structural Variant Caller

- Fix an issue for FFPE samples, where the `sv.vcf.gz` file contains only 1 entry for a `MantaBND` instead of 2.
- Fix a crash caused by tiny candidate contig size generated in SV assembly stage.
- Fix a segmentation fault in Tumor Only mode due to long assembly size causing a 32bit integer overflow.
- After the introduction of the Tumor Only scoring model in v4.0.3, SV caller did not PASS all the variants in the hotspot region for somatic and Tumor Only analyses. Fix for the issue.

CNV Caller

- Remove unwanted assert during input file checking for Panel of Normals.
- In CNV VAF modeling, fix estimation of scale factor in presence of outliers.

Targeted Callers

- All targeted callers (CYP21A2, CYP2D6, CYP2B6, SMN, Star Allele) now have their outputs combined in JSON file `<prefix>.targeted.json` file.
- **New CYP21A2 Targeted Caller**
    - This is a Beta version for early evaluation. Upcoming DRAGEN™ v4.2 will contain the officially supported CYP21A2 Targeted Caller.
    - The CYP21A2 Caller can genotype the CYP21A2 gene from whole-genome sequencing (WGS) data. Recombinant and non-recombinant variants in CYP21A2 are reported in an output `<prefix>.targeted.json` file with the following fields:

Field name	Description	Type and possible values
recombinantHaplotypes	List of known variants from recombination that were detected in CYP21A2. An empty string indicates a haplotype without any recombination.	Array of recombinant variants with two entries. Each entry contain the IDs of recombinant variants separated by a '+'
variants	List of single site variants, non-CYP21A1P like variants (not a result of homology). An empty list if no variants are detected.	Array of non-CYP21A1P like variants
deletionBreakpointInGene	• null if total_CN > 3 • true if CN <= 3 and a deletion recombinant variant (more than one non-target site) is detected in CYP21A2 • false if CN <=3 and no deletion recombinant variant is detected in CYP21A2	true, false, null
totalCopyNumber	Total copy number of CYP21A2+CYP21A1P genes	non-negative integer

- **CYP2D6**

  - Improved detection of star allele genotypes when a rare deletion in the CYP2D6-downstream region is present.
  - Usage:
    - To enable the caller use `--enable-cyp21a2 true`. The caller can be enabled in parallel to all other callers or run from pre-aligned BAM/CRAM input.
- **SMN**
  - Support detection of `NM_000344.4:c.*3+80T>G`
- **Star Allele**
  - Fix a run time regression with Star Allele in v4.0.3, running for ~15 minutes instead of 5 minutes.
  - Fix a hang when CYP2D6 and EH are both enabled, when starting from BAM input.

Single-Cell

- Fix for a missing column for Feature/Peak ID in scRNA/scATAC output, causing compatibility issues for downstream tools.
- Fix for scATAC producing empty outputs (barcode list, matrix) when using combinatorial barcodes.

RNA Gene Fusion

- Add mitochondrial genes filter to fusion VCF header.
- Make Gene fusion VCF report PR:SR reads, like the SV caller.
- Fix for Gene fusion VCF qual score of "inf".

Mapper and Aligner

- Fix for incorrect CIGAR string produced by mapper, leading to crash in Variant Caller. The issue was only present when using specific mapper settings for PR overhang trimming.

BCL Conversion

- Fix a crash that can occur when using per-sample-settings with higher sample counts in a lane, due to a hash-table pre-size using a signed integer as input that is overflowed.
- Fix BCL FASTQ file paths in the `fastq_list.csv` when ORA-interleaved compression format is used. Previously the `fastq_list.csv` file contained two files (instead of one single interleaved file) under the "Read1File" and "Read2File" columns and the files were not named correctly.
- Some versions of RTA3 outputs CBCL with 0 qual + nonzero base for 0/"N#". In other words, masked nibbles (0 qual) do not zero out the base. bcl2fastq2 has a masking step, but DRAGEN™ and bcl-convert did not. This fix adds masking so that DRAGEN™ and bcl-convert matches bcl2fastq2 for those RTA3 outputs.
- Fix for no-sample-sheet setting that omits index sequences from fastq headers.
- Fix to remove BCL conversion thread settings limit of 64, a regression in v3.10, to allow runs on high core count systems.
- Fix a crash due to threading error on a 150k sample dataset.
- Fix for `Index_Hopping_Counts.csv` containing incorrect `Same_{Name,Project}`.

Gvcf Genotyper

- Fix a memory leak during input VCF reading.
- Fix for incorrect `LPL` and `LAA` values in msVCF.

  - When `--gg-discard-ac-zero` is true, `--gg-remove-nonref` is false, and variant quality is lower than min_qual, the variant genotype is converted to 0/0 (or 0 on chrY) but the `AD` and `PL` fields are not changed. As a result, the `LPL` and `LAA` contained more original ALT alleles which are not supported by GT, so discard AC=0 option became invalid
  - When `--gg-remove-nonref` is true, for homref records, the `LAA` value should not be "1" but rather ".", since "1" refers to first ALT allele in msVCF.
- Fix for unnormalized variants in msVCF output of Lettuce samples.
- Fix to trim remaining alleles after AC=0 removal, if necessary. Enable trimming on lone REF allele after AC=0 discard
- When writing to allele counts and frequencies to the output msVCF file in some circumstances non-ref values were not correctly processed.
  - This occurred when the global ref allele is different from the batch ref allele. The non-ref allele is represented by the symbolic base sequence 'X' which does not change under right renormalization of the base sequences when the ref allele is lengthened to match the global equivalent. As such, no-ref must have been treated separately. Fix for this issue.

Other Bug Fixes

- Fix a race condition in XRT driver causing incorrect 64-bit reads, leading to incorrect FASTQC metrics on Azure cloud platforms.
- Fix a crash in HLA Typer when down sampling is enabled.
- Fix for EL8 driver DKMS 3.0 breaking networking on boot up.
- Fix for excessive watchdog logs filling up `/var/` partition.
- Fix a potential crash when processing multiple read group IDs, caused by a timing race condition.
- An invalid check for 10 required columns for the `--qc-cross-cont-vcf` file header leads to an exception. Fixed the check to require 8 columns. Also improved error handling for invalid file inputs, with clearer messages.
- Fix for mapper metrics being double counted when HLA is enabled.
- Fix for Out-of-Memory issue on the cloud when building systematic noise files. This function should never exceed 140GB memory consumption after fix.
- Two minor updates to systematic noise default settings for the VAF threshold and decimal precision. Based on new studies these settings slightly improve accuracy.
- New option `umi-parse-only` to enable the UMI parser for regular map/align without UMI (`enable-umi=false`). If user specifies the `umi-source`, it is saved to the output bam with `RX` tag.

Known Issues

Known issues of the DRAGEN™ v4.0.5 release

Component/s	Issue Description	Remedy/Workaround
BCL	When UMI is in the first part of a read and TrimUMI is enabled (true by default), then the quality score sum metrics (yieldQ30/qscoresum) are not accurate.	Minor inaccuracy, no workaround
BCL	bcl-convert hang but does not time out with a crash	None. A watchdog mechanism does not exist for sw-only bcl-convert

BCL, Ora Compression	Wrong filenames in fastq_list.csv when converting BCL to ora with interleaved option	fastq_list.csv cannot be used to decompress, must be fixed manually.
BCL, Ora Compression	DRAGEN BCL silently disregards ORA Compression commands when --no-sample-sheet option is used	The --no-sample-sheet option is meant purely for one legacy use case and should not be used with BCL to Ora
BCL, Ora Compression	BCL ORA-interleaved Compression outputs FASTQs missing "_001" filename suffix. For original filenames ending in "R1_001.fastq", "R2_001.fastq" the decompressed file names are "R_1.fastq", "R_2.fastq"	File names to be improved in future version.
Biomarkers	TMB does not error out on invalid input file. The input is ignored and run completes with no output	No workaround
CNV VC	CNV can over-segment calls in some cases, which affects mutational signature calculations	No workaround
CNV VC	Some samples may fail ploidy detection	No workaround
Compression	When CRAM is decompressed with a different HashTable ref than the map/align, and CRAM is also output, and SV caller is enabled; then the SV caller crashes reading CRAM input due to mismatched reference in the cram interface.	No issue for BAM output, or when FASTA is used decompress the CRAM file. Issue is specifically affecting SV caller only. Workaround: use FASTA instead of HT when decompressing cram with different ref in this use case. A fix for this issue will be released
CYP21A2 Caller	CYP21A2 may mis-detect homozygous recombinant variants	No workaround
DNA Alignment	For some input values of RGPL, the software generates a RG line that is not SAM compliant and produce Bam that has compatibility issues with some 3rd party tools	Use uppercase RGPL name
DNA Alignment	When running different read trimmers back-to-back, crash during RecomputeTags	Workaround: Issue a dragen_reset in between the runs.
GBA Caller	GBA regression for LB-01223 with map/align enabled	No workaround
GVCF Genotyper	GVCF Genotyper regression for Mito calls, when running non-iterative GG + Joint Genotyping (in that order)	Use Iterative GVCF Genotyper
GVCF Genotyper	Missing genotypes coded as haploid instead of diploid	No workaround
HLA	HLA crash with when run with enable-map-align=false. HLA is not possible without map/align	Only enable HLA with map/align run
HW GRAPH, RNA VC	RNA VC hits ERROR: Invalid node flags	Rare hardware error. Re-run sample as it is expected to pass

Infrastructure	dragen_hugepagctl conflicts with other programs that allocate hugepages	Manual configuration of hugepages
Joint Genotyping	WGS Denovo trio run accuracy regression on v4.0, due to STR context change, resulting in threshold changes and shifts the relative FP and FN performance	None
Methyl-Seq	Methyl CX report file may contain trailing data from prior run, when same output folder is used	Run new analysis in new output folder
Paralog Caller	Callers using pair-by-name running from BAM input w/out map/align may hang when coverage is very deep.	Enable the callers (CYP2D6, EH) during the map/align step instead
Paralog Caller	GBA variant NM_000157.4:c.84dup reported as NM_001005741.2:c.84dupG	No workaround
SNV Somatic	Somatic T/N end-to-end runtime increases significantly when map/align output is enabled	No workaround
SNV VC	Germline run time is roughly 6.3% slower with graph aligner and graph reference is used, compared to non-graph. The increased run time is in both mapper and variant caller phases.	None
SNV VC	A gene panel accuracy test dropped one SNV in v3.10 and v4.0 that was called in v3.9	None. Missing variant is caused by an improvement to the read trimmer in HW that correctly trims a read but leads to a missing edge in the graph causing the variant to be lost.
Somatic	Elevated FPs for ICGC datasets in v3.10/v4.0 compared to v3.9: a 5-6% increase in the SNP FPs and a 25%-30% increase in the INDEL FPs.	None
Sort and Dupmark	Very large samples fail with dumpark=hash (default)	Run with dupmark=sort.
SV	Extra SV call in FLT3-ITD hotspot region for FLT3_C317.TCGA-AB-2830 T/N sample	None
UMI	Certain Methylation UMI analyses fail on BSSH App	No workaround
UMI	Some UMI samples can run into OOM condition.	Sample can succeed with specific settings of the fuzzy window

SW Installation Procedure

- Download the desired installer from the Illumina support website and unzip the package
- The archive integrity can be checked using: `./<DRAGEN 4.0.5 .run file> --check`
- Install the appropriate release based on your Linux OS with the command: `sudo sh <DRAGEN 4.0.5 .run file>`
- Please follow the installer instructions. Server power cycle may be required after installation, depending on the currently installed version. If an updated FPGA shell image needs to load from flash, this is only achieved with power cycle.
    - A power cycle is required when upgrading from v3.3.7 or older
    - A power cycle is required when downgrading to v3.3.7 or older
    - A power cycle is not required when upgrading from a release after v3.3.7
- Procedure to downgrade to v3.3.7 or older:
    - Requires the following three steps. The prior .mcs file needs to be flashed manually:
        - Install the prior release: `sudo sh <DRAGEN 3.3.7 .run file>`
        - `program_flash /opt/edico/bitstream/07*/*.mcs`
        - Power cycle