DRAGEN v3.3.7
Software Release Notes

April 29, 2019

Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform since the package containing DRAGEN v3.2.8

If you are upgrading from a version prior to DRAGEN v3.2.8, please review the release notes for DRAGEN v3.2.8 for a list of features and bug fixes introduced in that version.

The software package includes:

- DRAGEN SW Intel Centos 6 - dragen-3.3.7.el6.x86_64
- DRAGEN SW Intel Centos 7 - dragen-3.3.7.el7.x86_64
- DRAGEN SW IBM Centos 7 - dragen-3.3.7.el7.ppc64le

Contents

DRAGEN v3.3.7 Fixes

- Methyl-Seq: Fix for crash when using --methylation-reports-only option
- Joint Caller: Fix handling of extremely large multi-sample VCF records
- Combine GVCF: Fix for crash in Tabix update during gVCF generation
- VC: Improve the heuristic filter based on TLEN, in the presence of structural variants on Novaseq
  - Prior to DRAGEN v3.3.7: Discard the event if TLEN < (read length + 6bp),
  - From DRAGEN v3.3.7 onward: Discard the event if TLEN > Mean TLEN +- (2.25*TLEN std dev)
- AWS: Fix to avoid reference HT re-loading between runs, to improve run times of back-to-back analysis for exomes and panels.
- CNV: Fix father/proband de novo calling on chrY
- CNV: Fix for issue that causes self-normalization to fail due to insufficient data

DRAGEN v3.3.5 Release Notes

Highlights

- Speed and Accuracy improvements to the Somatic T/N pipeline
  - Up to 6x speed improvement. Most 100x/40x T/N datasets now complete in under 2 hours.
  - Accuracy improvement compared to DRAGEN v3.2, on-par or better than GATK 4.1.
  - Accuracy has been validated against a range of tumor purities, library preps, and sequencing instruments.
- CNV DeNovo calling
  - Support for DeNovo calling and scoring with pedigree input.
- Speed improvement of BCL conversion
  - Up to 2x speed improvement on NovaSeq data when using DRAGEN Phase2 servers.
  - Up to 1.2-1.5x speed improvement on other servers.
- Structural Variants
  - DeNovo pedigree scoring.
  - Caller upgraded to Manta v1.5.1.
- Metrics
  - Added a model for detection of sample cross-contamination in human species.
  - Added MAPQ and BQ coverage filters for each selected coverage region.
- Repeat Expansion Calling
  - Updated repeat expansion caller to GraphExpansionHunter, which allows calling more complex repeat loci.
- Added RNA Quantification module to estimate transcript-level gene-expression results.
- Added support for quad and multi-generation pedigree calling in a single execution of the small variant joint caller.
- Added Beta support for read collapsing on the Illumina TSO-500 UMI design.

Summary of Changes

A summary of key changes is listed below. Please refer to the DRAGEN Bio-IT Platform User Guide for more information.

Somatic T/N Small Variant Calling

- Up to 6-fold speed improvement on datasets that were previously HMM-limited, with typical 100X/40X tumor-normal runs now finishing within 1h40m on a local server or 2h30m on AWS.

- Accuracy optimized for a broader range of datasets, with improvement for both snvs and indels on most datasets.

CNV Caller

- The CNV caller now supports de novo calling.
  - Multisample VCF support, starting from normalized signal files (*.tn.tsv) of single sample runs.
  - New *.tn.tsv files must be generated with this version of DRAGEN to be compatible with the de novo CNV caller.
  - De novo calling and scoring for valid trios defined in a pedigree file.
  - Multiple trios supported.
- Output VCF changes
  - The ID field in the output VCF now also encodes the contig of the event.
    - Now formatted as DRAGEN:<event>:<chr>:<start>-<stop>.
    - This is to comply with the VCF spec and ensure that the ID field is unique within the VCF.
    - Example: DRAGEN:LOSS:chr1:2841405-2847435
  - Added support for tabix of CNV VCFs.
- Filtering changes
  - Introduced a new option cnv-filter-bin-support-ratio to allow control of filtering events based on number of supporting bins.

Structural Variant Caller

- Structural Variant caller is updated to Manta 1.5.x
  - Improved accuracy, particularly improved precision for germline calling.
  - Improved runtime.
- Supports de novo calling
  - De novo scoring for a valid trio defined in a pedigree file.
  - Adds DQ and DN tags in the FORMAT field in the multi-sample VCF of germline calls.
- Output changes
  - VCF format changes
    - Change filters for easy interpretation of multi-sample germline variant vcf.
      - Add record-level filter 'SampleFT' when no sample passes all sample level filters.
      - Add sample-level filter 'HomRef' for homogyzous reference calls.
      - No more sample-level filters are applied at the record level even if it applies to all samples.
    - Change representation of inversions in the VCF output
      - Intrachromosomal translocations with inverted breakpoints are now reported as two breakend (BND) records.
      - Previously they were reported in the VCF using the inversion (INV) allele type.
  - The SV final output VCF is now available in the <output-directory>/<output-file-prefix>.sv.vcf.gz
  - The SV intermediate outputs moved to the <output-directory>/sv folder

Metrics

- The coverage report output file names have changed.
- All reports for qc-coverage-region-*i* are output in qc-coverage-region-*i*_*.bed and qc-coverage-region-*i*_*.csv files, where *i* can be 1, 2, or 3.

Repeat Expansion Calling

- Updated GraphExpansionHunter toallow calling more complex repeat loci, including loci with multiple flanking STRs and SNVs.

- The new version of repeat calling does not rely on unaligned (eg, fully in-repeat) reads. This makes calls more robust to sequencing bias, which means that repeat lengths will not be estimated beyond the library fragment size.
- This version uses a new repeat-spec (variant catalog) format; see the DRAGEN Bio-IT Platform User Guide and repeat-specs/hg19/variant_catalog.json for an example.
- Add SMA calling using the same graph alignment approach. This feature allows detecting absence of the fully functional allele at the duplicated SMN1/2 locus. Variant catalogs with SMN are found in repeat-specs/experimental.
- Realignments of repeat reads are now output in BAM format.

RNA Quantification

- Added support for quantification of gene expression from RNA-seq data.
- The module outputs the estimated expression of annotated transcripts and genes in Transcripts per Million (TPM) and read-counts units, using an EM method for deconvolution.
- Optionally includes GC-bias correction.
- Quantification can be enabled with map/align in RNA mode, by setting –enable-rna-quantification to true and supplying the transcript annotation file (GTF/GFF) with –annotation-file.
- Supports both stranded and un-stranded paired-end RNA-Seq protocol.
- RNA Quantification is still in Beta.

Known Limitations

- Structural Variant caller should be run with a BED file containing the set of regions to call, to avoid SV calls on alt and decoy contigs commonly found in hg38 references.
- CNV de novo calling not supported for Father/Proband calling on chrY in DRAGEN v3.3.5

SW Installation

1. Install the appropriate release based on your Linux OS with the following command:

   ```
   sudo sh <DRAGEN .run file>
   ```

2. Cold boot the server so that the new SW is fully installed with the updated FPGA HW image.

md5checksum:

```
f3c69c4552173c01a564453423771094 dragen-3.3.7.el6.x86_64.run
49902f5d64eb57d2dc8d0a78ccaeb25e dragen-3.3.7.el7.x86_64.run
14f45cf03395f07349e21f96b8aae0e8 dragen-3.3.7.el7.ppc64le.run
```