DRAGEN v3.5.7 Software Release Notes

February 17, 2020

Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform v3.5.7.

Changes are relative to DRAGEN™ v3.4.12. If you are upgrading from a version prior to DRAGEN™ v3.4.12, please review the release notes for a list of features and bug fixes introduced in subsequent versions.

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

The 3.5.7 software package includes:
- DRAGEN™ SW Intel Centos 6 - dragen-3.5.7.el6.x86_64
- DRAGEN™ SW Intel Centos 7 - dragen-3.5.7.el7.x86_64
- DRAGEN™ SW IBM PPC Centos 7 - dragen-3.5.7.el7.ppc64le.run

The following configurations are also available on request:
- Amazon Machine Image (AMI)
- RPM packages for Centos 7 and Ubuntu 14.04 for Amazon Web Services (AWS)

Contents

Highlights

Below is a summary of the changes included in v3.5.7. For full extensive details, please consult the Illumina DRAGEN Bio-IT Platform User Guide available on the support website at https://support.illumina.com/downloads/illumina-dragen-bio-it-platform-user-guide.html

Improved Mapper Accuracy with Hash Table Version 8

- New mapper improves DRAGEN™'s mapping rate on reads with mostly 0 MAPQ
    - HTv7 had a mapping rate lower by ~1.7% on average compared to BWA
    - HTv8 mapping rate is now on par or better than BWA
- Improved mapping rates
    - The DRAGEN™ read mapper is built upon a k-mer (seed) hash table: given a particular seed sequence the table provides a lookup of the genomic positions where the seed occurs. Seeds (possibly extended) that match more than N locations (N=16 by default) are not populated in the hash table. Prior to the v8 mapper, these locations are unavailable at run time, so mapping rate can suffer. With v8, high count seed location sets can be sampled at run time, so reads from highly repetitive regions now have a chance to map. For frequently occurring seeds, the hash table includes seed extension tree construction to constrain the number of reference positions tested for alignment.
- DRAGEN™ v3.5 requires reference hash tables to be re-built

Somatic SNV Caller

- Significant improvement in detection performance, both in terms of Sensitivity and Precision
- Improved robustness across:
    - Sequencing Instruments
    - Different Library Preps and mixed Library Preps
    - Datasets (mutation rich and not)
    - Tumor Purity
- Liquid Tumor Support
    - Support for Tumor in Normal contamination that is usually present in Liquid tumor T/N workflows
    - Command line option to enable a mode for liquid tumor analysis `--vc-enable-liquid-tumor-mode=true`
        - Flag is recommended to be enabled for liquid tumors with TIN present
        - Flag is recommended to be disabled for solid tumors. If left enabled, there is some degradation in precision
        - Non-default parameter settings can be used for cases of low tumor purity with high TIN contamination
- Run times preserved at <2 hours for T/N 100x/40x

- Separate mapping metrics generated for tumor and normal samples
- Tumor callability bed report is output by default. Callability is output for regions where the coverage is above a certain threshold

Somatic CNV Caller

- Addition of Somatic WGS CNV Calling
- Supported Modes
    - Tumor w/ Matched Normal
    - Tumor Only
- Supported Features
    - Tumor Purity and Ploidy Modeling
    - Both Total and Minor Allele Specific Copy Number Calling are output
        - CN, MCN

Structural Variant Caller

- Improved Germline WGS Accuracy
    - Insertion recall improved by 14.4% (GIAB v0.6)
        - MEI calling: Recall for insertions in the 200-500 base range is improved by more than 31%
    - Deletion recall improved by 2% (GIAB v0.6)
- Improved ease of use
    - The `--sv-reference` option is deprecated and can be removed. The `--sv-reference` option is no longer required when running the SV caller. Any argument given to `--sv-reference` will be ignored, reference genome is specified by the standard DRAGEN™ `--ref-dir` argument.
    - BED files can be specified to the `--sv-call-regions` argument in either uncompressed or (b)gzip-compressed format. A tabix index file is no longer required.
    - The SV caller automatically skips the same decoy chromosomes skipped by the small variant caller, as specified by the `vc-decoy-contigs` parameter.

PopGen

- First release of DRAGEN™ ultrafast and scalable PopGen Software
- New workflow for population scale cohort analysis, using Gvcf Genotyper
- Improved run time compared to DRAGEN™ v3.4 using Combine GVCFs
    - 50x faster: From gVCFs to normalized & genotyped multi-sample VCF output
    - 13x faster: From gVCFs to joint genotyped multi-sample VCF output
    - Enables processing of larger cohorts on single nodes
- Improved software robustness when running larger cohorts

- The pipeline is backward compatible with gVCFs produced by prior versions of DRAGEN™
- Gvcf Genotyper usage:
    - Use new command line option `--enable-gvcf-genotyper=true`
    - Input is a list of single sample gVCF files
    - Output is a multi-sample VCF, where
        - All variants in the cohort are genotyped across all samples
        - Representation of variants is normalized across samples
    - Output may be passed to existing Joint Genotyper module for population-based genotype refinement
- Combine GVCFs is still available as-is for Pedigree based workflows. Gvcf Genotyper is now recommended for large cohort analysis.

RNA

- Improved Gene Fusion Detection accuracy
    - 38% fewer missed fusions than DRAGEN™ v3.4
    - 20% fewer additional calls than DRAGEN™ v3.4
    - Default gene-homology filter for human genome
    - New fusion output file, including details in CSV format
- Improved run time
    - 1.9x faster than DRAGEN™ v3.4
    - 4.5x faster than STAR-Fusion + Salmon

UMI

- DRAGEN™ v3.5 now offers highly accelerated workflow for mapping and UMI read collapsing
    - Both random (degenerate) and non-random (from specific set) UMIs are supported
    - DRAGEN™ UMI allows for collapsing of raw reads based on alignment position and UMI sequence.
    - UMI read collapsing drastically reduces the error, allowing accurate variant calling at very low allele frequencies.
    - Duplex collapsing for non-random UMIs further reduces the error rate and removes artifacts occurring on a single strand
    - Outputs mapped and UMI collapsed BAM/CRAM
- Exceptional run time performance
    - > 17x faster than 3rd party tools such as LocatIt and fgbio
    - DRAGEN™ UMI includes pre-collapsing and post-collapsing alignments

Germline Small Variant Caller

- Able to produce both VCF and gVCF in a single run
    - New option `--vc-enable-vcf-output` to output VCF in a GVCF run.
- Ploidy auto-detection is enabled by default and is used for handling X and Y chromosome automatically without an external tool to detect the sample sex
    - If `--sample-sex` is not specified from command line, the Ploidy Estimator module will run by default in order to estimate sex and the inferred sex will be passed to the small VC and Repeat Genotyper. The CNV caller uses its own sex inference feature
    - Each normalized per-contig median coverage is reported in a ploidy estimation metrics output
- Deprecated BQSR functionality
- Deprecated VQSR functionality

Metrics

- GC bias metrics calculation is available as an optional output, using `--gc-metrics-enable=true` command line option.

Methylation

- Improved settings for more accurate and precise mapping of methylation data
- Since DRAGEN™ v3.4 the following changes were made to default methylation settings

Param name	Old value	New value
`Aligner.aln-min-score`	22	0
`Aligner.min-score-coeff`	0	-0.2
`Aligner.match-score`	1	0

    - Limitation: if `--methylation-protocol` is set to something other than '`none`', then the `Aligner.match-score` and `Aligner.min-score-coeff` are hardcoded and command line options would not take effect. This limitation will be removed in the next version of DRAGEN™

BCL

- New Metrics added:
    - Adapter Trimming metrics. Output in `<output_directory>/Reports/Adapter_Metrics.csv`
    - Demultiplexed Read Counts Per Tile. Output 'IndexMetricsOut.bin' to `<output_directory>/Reports/` (similar to bcl2fastq2 output)
- Support for v2 sample sheet format used by NextSeq 1000/2000 sequencers

- Improvements:
    - Sample Sheet and RunInfo.xml files are copied to `<output_dir>/Reports`, for reference
    - Sample Sheet error handling is more extensive and uses a more consistent and accurate error message output format
    - Additional error checks for RunInfo.xml file: 0-cycle reads not supported; an index read after the 2 genomic reads not supported
    - Only one error message is output per missing cbcl input file

SW Installation Procedure

- Download the desired installer from the Illumina support website and unzip the package
- The archive integrity can be checked using: `./<DRAGEN 3.5.7 .run file> --check`
- Install the appropriate release based on your Linux OS with the command: `sudo sh <DRAGEN 3.5.7 .run file>`
- Cold boot (hard reset or power cycle) may be required after installation, depending on the currently installed version. Please follow the installer instructions. An updated FPGA shell image needs to load from flash, this is only achieved with cold boot.
- Installing prior releases after v3.4.5 was installed:
    - Installing a prior release, v3.3.7 or older, will require the following two 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