DRAGEN v4.0.3 Software Release Notes

Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform v4.0.3.

Changes are relative to DRAGEN™ v3.10.8. If you are upgrading from a version prior to DRAGEN™ v3.10.8, 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 v4.0.3 software package includes installers for the on-site server:
- DRAGEN™ SW Intel Centos 7 - dragen-4.0.3-8.el7.x86_64.run
- DRAGEN™ SW Intel Oracle 8 - dragen-4.0.3-8.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 v4.0.3. DRAGEN™ v4.0 offers new callers, as well as speed and accuracy gains and new feature introductions across most callers. For full extensive details, please consult the latest 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

New Product Resource Files

New product resource files that accompany the v4.0 software is available for download on the Illumina support website https://support.illumina.com/sequencing/sequencing_software/dragen-bio-it-platform/product_files.html

The following product resource files are updated for v4.0:

- Alt-masked v2 graph genomes for hg19 and hg38, updated with HLA anchored hash tables
- Custom multigenome reference genome builder resource files for hg38, hg19, hs37d5
- Imputation reference panel package IRPv1 for hg38 containing genetic maps and reference panels
- Structural variant caller systematic noise baseline collection, containing BEDPE files for GRCh37 and hg38

Interface Changes

DRAGEN v4.0 introduces interface changes in some components, that may prevent plug-and-play

- Machine Learning (ML) is enabled in the germline small variant caller. QUAL, GT, GQ fields are re-calibrated in the VCF. NOTE: The distribution of QUAL values will now be different between ML enabled or disabled. Hard-filtering is also applied at QUAL 3 when ML is enabled
- DRAGEN aligner handling of split reads has been updated
- The systematic noise builder tool command line options have been updated
- The HLA biomarker component has been replaced with a new method and new command line interface, and requires an anchored_hla hash table to be built
- Single Cell processing command line options have been updated
- BCL conversion now requires each index in a dual setup to individually meet the hamming distance requirements for BarcodeMismatchesIndex#
- There is a new AWS FPGA shell, which prevents running DRAGEN v4.0 followed by an older version 3.x back-to-back on same instance. A dragen_reset must be performed with v4.0 first

New Callers and Major Features

DNA and RNA Alignment

- We implemented a much more rigorous method for determining split-read alignments and influencing primary alignments and MAPQs with split-read analysis
- This sophisticated new method can support up to a maximum of 4095 secondary alignments and 4095 supplementary alignments per read

Germline Small Variant Caller

- **Machine Learning (ML) is enabled as default**
  - `--vc-ml-enable-recalibration` True by default but auto-disabled for non-human samples / amplicon workflows
  - *`--vc-ml-dir`* Defaults to the packaged ML model, customers no longer need to download the model in advance. Files are present at `/opt/edico/resources/ml-model`. DRAGEN automatically uses the appropriate ML model
  - The ML model version is updated, prior ML model files are not compatible with v4.0
- **Force Genotyping with VCF**
  - Force Genotyping was originally designed to work with GVCF output mode. In v4.0 we made many improvements to FGT to ensure it runs smoothly with VCF output and other features such as MNV
- **Mitochondrial variant calling updates**
  - Default mitochondrial allele frequency thresholds have been updated. AF call threshold 1%, filter threshold 2%
  - Accuracy improvements have been achieved by
    - Tuning of mito-specific columnwise detection parameters
    - Implementation of mito-specific De Bruijn graph algorithm and parameter updates to reduce FP
  - The following command line options are available
    - `--vc-enable-af-filter-mito` Enable the allele frequency filter in mito vc calling (Default=true)
    - `--vc-af-call-threshold-mito` Mitochondrial AF threshold for emitting calls in the vcf (Default=1%)
    - `--vc-af-filter-threshold-mito` Mitochondrial AF threshold to mark emitted vcf call as filtered (Default=2%)

Somatic Small Variant Caller

- **High Coverage Analysis**
  - DRAGEN can now analyze data sets with higher coverage than ever before. Out of the box, 1000x for a tumor-only sample is supported. For tumor-normal datasets the supported coverage is 750/250x
  - To allow the analysis of deeper regions, the system must limit the size of a region. By default, a callable region is limited to 13GB. The threshold can be adjusted using `--vc-max-callable-region-memory-usage`
- **Using somatic BAF for CNV**
  - In v3.10 `--cnv-use-somatic-vc-baf` was introduced to enable running CNV without a separate germline run. In v4.0, considerable under-the-hood changes have been made to improve compatibility with other run modes. For instance, forcing germline variants at the same time now works better than before
- **NTD error bias estimation**
  - The NTD error bias estimation feature accounts for oxidation/deamination artifacts, as is common in FFPE samples
  - Up to v3.10 this feature was enabled by default only in UMI pipelines. In v4.0, changes were made to correct for the effect of basecall quality, and the feature is enabled by default in non-UMI somatic pipelines. The feature causes a 5-12% slowdown in run times

- An autodetect feature has been added (off by default), such that the feature can be activated only when DRAGEN detects an artifact to correct
- Usage:
  - `--vc-enable-unequal-ntd-errors` Now takes options `true`, `false`, and `auto`, with default changed from `false` to `true`. Set to `false` to disable, and `auto` for autodetect mode
  - Add `--vc-snp-error-cal-bed=<bedfile>` to override the default regions used for error estimation

- **N-base Indel variants**
  - DRAGEN does not call N-base indel variants by default (both N as ref and alt allele). Option `--vc-enable-n-indels=true` (default=false) will let DRAGEN call N-base indel variants
- **Mutation Annotation Format (MAF) Output Conversion**
  - New in v4.0, DRAGEN supports the output conversion of a Nirvana Annotation JSON file to a Mutation Annotation Format (MAF) file
  - Two modes are supported
    - *Integrated mode*: Runs the MAF conversion at the end of a standard somatic variant caller workflow in one command line (e.g., FASTQ/BAM input to VCF+MAF output)
    - *Standalone mode*: supports conversion of a user provided VCF or annotated JSON input
  - One of two supported annotation sources must be specified - `Refseq` or `Ensembl`
  - The output is a tab-separated MAF file in the form `<file_name>.<source>.maf`
  - When starting from VCF input, the Nirvana Variant Annotation must also be enabled
  - Usage:
    - *Integrated mode*

      ```
      dragen -r <ref_dir> --enable-variant-caller=true --output-
      directory <out_dir> --output-file_prefix <out_prefix> --enable-
      variant-annotation=true --variant-annotation-assembly
      <GRCh37/GrCh38> --variant-annotation-data <data_dir> --enable-
      maf-output=true --maf-transcript-source=<Refseq/Ensembl>
      ```

    - *Standalone mode*
      - VCF input (output directory and output file prefix are optional - Annotation and MAF files will be written to location of VCF by default)

        ```
        dragen -r <ref_dir> --output-directory <out_dir> --output-
        file_prefix <out_prefix> --enable-maf-output=true --maf-
        input-vcf=<VCF_path> --maf-transcript-
        source=<Refseq/Ensembl> --enable-variant-annotation=true --
        variant-annotation-assembly <GRCh37/GrCh38> --variant-
        annotation-data <data_dir>
        ```

      - JSON input (output directory and output file prefix are optional - Annotation and MAF files will be written to location of JSON by default)

        ```
        dragen -r <ref_dir> --output-directory <out_dir> --output-
        file_prefix <out_prefix> --enable-maf-output=true --maf-
        input-json=<JSON_path> --maf-transcript-
        source=<Refseq/Ensembl>
        ```

- **Systematic noise builder tool**

- DRAGEN has allowed for the generation of systematic noise BED files from normal samples that are library prep, sequencing system, and panel specific
- In v4.0 the command line interface for the systematic noise BED generation tool has been updated to make usage clearer and disambiguate the options for builder and the usage of noise files
- The following options have changed:

v3.10	v4.0	Description
`vc-systematic-noise-raw-input-list`	`build-sys-noise-vcfs-list`	List of normal VCF/GVCF files to be included in the systematic noise. One file per line
`vc-systematic-noise-germline-vaf-threshold`	`build-sys-noise-germline-vaf-threshold`	Minimum variant allele frequency threshold to define germline variants (Default 0.2)
`vc-systematic-noise-use-germline-tag`	`build-sys-noise-use-germline-tag`	If available, use germline tags to prevent germline calls from contributing to systematic noise. (Default: true)
`vc-systematic-noise-method`	`build-sys-noise-method`	Method to compute noise across samples ['mean'/'max']. For WGS 'max' is recommended, and for WES/panels 'mean' is recommended (default=mean)

    - The following optional settings are newly added
        - `--build-sys-noise-threads` Max number of threads used during noise generation. Each thread consumes approx. 50 GB of system memory. (Default 4)
        - `--build-sys-noise-merge-regions` Merge regions with the same systematic noise. This reduces the noise file size. (Default True)
        - `--build-sys-noise-decimal-precision` Number of decimal digits in noise file. Options are [3-6]. For typical WES/WGS with 50-500X coverage 3 decimal places should be sufficient. For deep UMI samples with low noise rates 5 decimal places are recommended. Lower precision may help reduce noise file size especially on WES/WGS. The default is set for accuracy (default=5)

Imputation

- New user-friendly tool to execute scalable whole gnome imputation in one single command line, completely end-to-end
- Able to impute a single sample or a list of VCF files on multiple chromosomes in one run (up to 100 samples simultaneously)
- Imputation improves sensitivity of 1x coverage data by >= 70% and reduces false positives by >= 80%, as measured on HG002-HG005 using GRCh38
- AVX accelerated software implementation of GLIMPSE [2020, Olivier Delaneau & Simone Rubinacci], a tool to infer bi-allelic SNP variants from low-coverage sequencing samples
    - DRAGEN is 1.7x faster than the original GLIMPSE code
- A Reference Panel package (IRPv1) for hg38 is provided for download
    - IRPv1 is an autosomal SNP reference panel containing the 2504 samples from the 1000 Genomes Project, which have been variant called from the ~50x NYGC data using DRAGEN v3.7.6 against hg38
    - The package contains both Genetic Maps (gmap.gz files) and Reference Panels (bcf and sites.vcf.gz files)

- Usage:
  - `dragen --enable-imputation true --imputation-ref-panel-dir <REF_PANEL_DIR> --imputation-chunk-input-region chr1 --imputation-ref-panel-prefix IRPv1 --imputation-phase-input-list <VCF_list.txt>  --imputation-genome-map-dir <MAP_DIR> --output-directory <OUT_DIR> --output-file-prefix <OUT_PREFIX>`
  - See the DRAGEN User Guide for a full description of the tool and its options

- Limitations:
  - Only human sequencing data is supported, on hg38
  - chrX, chrY, and chrM are not supported in the IRPv1 reference panel

SV Caller

- **Accuracy and runtime improvements**
  - Optimizations across the SV caller leading to improved accuracy and analysis times
  - Insertion recall improved by 2% for insertions. Recall now exceeds 72% and 64% for deletions and insertions respectively, while maintaining high precision, as measured on HG002 using NIST GIAB v0.6 tier1 SV truth
  - Significant reduction in run times across all workflows. Reduced by 20-60% across the board
- **Tumor-only scoring model**
  - Tumor-only pipeline previously reported all putative variants in the final VCF file, leading to low precision
  - Candidate events and now scored and assigned a QUAL score in the VCF for filtering and training, reducing the total number of reported SVs
  - The scoring model is consistent with the small variant caller
- **Systematic Noise filter**
  - Tumor only SV analysis can now accept a systematic noise file to significantly reduce the number of false positives reported
  - Without a matched normal, the tumor only workflow has challenges removing background noise from the analysis. The DRAGEN SV T/O systematic noise filter captures mapping and variant calling artifacts and recurrent structural variants, removing likely false positives. Any candidate events that match the entries in the systematic noise file will have the FILTER field set to "`SystematicNoise`".
  - A systematic noise file in BEDPE format in GRCh37 and hg38 is available for download
  - To enable, specify: `--sv-systematic-noise <PATH TO BEDPE FILE>`

Methylation

- **End-to-end deduplication**
  - To output a sort/deduped methylation CX report used to require two runs: alignment (fastq→bam) followed by methyl-calling (bam→ CX report)
  - Now an end to end run where fastq->bam (deduped) ->CX report (deduped) is possible when:
    - `--enable-duplicate-marking true`
    - `--methylation-generate-cytosine-report true`
    - `--methylation-mapping-implementation single-pass` (default)

  - This leads to a 50% faster run time and lower overhead costs
- **UMI support**
  - UMI enables more accurate methylation calling by removing PCR and sequencing errors
  - Methyl-UMI has the same principle and requirement as DNA UMI:
    - UMI needs pre-methylation to prevent bisulfite conversion
    - In the sample sheet, specify UMI length and position:
      - eg. `OverrideCycles, U7N1Y143;I10;I10;U7N1Y143`
    - Input fastq will have UMI barcode in QNAME (7th field)
      - `@NS500561:434:H5LC2BGXJ:1:11101:10798:1359:CACATGA+ACATTC 1:N:0:TGGTACCTAA+AGTACTCATG`
  - Read collapsing
    - As a result, reads with the same UMI will be collapsed if mapped to the same genomic location on the same strand, either from top (OT/CTOT) or bottom (OB/CTOB) strand
  - To enable methyl-UMI read collapsing:
    - Add `--umi-enable true` for random UMI (common), OR
    - Add `--tso500-solid-umi true` for non-random UMI from TSO500
    - Only supported in `--methylation-mapping-implementation single-pass` (default)
- **Deprecation notice**
  - In a next DRAGEN release, the option `--methylation-mapping-implementation single-pass | multi-pass (default: single-pass)` will be removed, leaving single-pass as the only implementation

HLA Biomarker

- New HLA genotyping component replaces the legacy version
  - nucleotide-based HLA alignment and genotyping replaces legacy amino-acid based method
  - Usage:
    - HLA is enabled as in previous versions, with `--enable-hla=true`
    - TSO500-solid input: HLA is enabled with `--tso500-solid-hla=true`
    - TSO500-liquid input: HLA is enabled with `--tso500-liquid-hla=true`
- HLA accuracy improvements
  - Overall accuracy in WES/WGS test cohorts increased to >99%
  - Overall accuracy in TSO500-solid/TSO500-liquid increased to >97%
- Interface changes
  - Removed command-line options `--hla-bed-file, --hla-bed-padding, --hla-reference-file, --hla-allele-frequency-file, --hla-tiebreaker-threshold, --hla-zygosity-threshold and --hla-min-reads`
  - Removed `<prefix>.freq_tiebreaking_candidates.tsv` output file
  - Added new output file `<prefix>.hla_2field_EM.tsv`, which reports a list of candidate alleles at two-field resolution and an Expectation-Maximization posterior probability for each of them
  - Output file `<prefix>.hla_metrics.csv` changed format: it now reports the number of reads supporting each allele result, and the total number of HLA reads analyzed
- Known limitations for HLA component
  - HLA must run with map-align enabled
  - HLA with Tumor-normal paired input from BAM is not currently supported
  - Does not support single-ended inputs

PGx

- **Star Allele Caller**
  - The Star allele caller genotypes 16 Tier 1 genes that have the highest (Level A) designation from CPIC i.e., they have moderate to high levels of evidence in favor of changing a drug prescription
  - The genes are as follows: CACNA1S, CFTR, CYP2C19, CYP2C9, CYP3A5, CYP4F2, IFNL3, RYR1, NUDT15, SLCO1B1, TPMT, UGT1A1, VKORC1, DPYD, G6PD and MT-RNR1
  - The Star allele caller can be run in stand-alone mode from a DRAGEN gVCF file which contains variants and genotypes for all locations on the genome
  - Alternatively, the caller can be run in end-to-end mode from a FASTQ/BAM file. In this case, upstream DRAGEN processes such as read mapping and variant calling are executed first to generate the gVCF file
  - If the genotype is missing for one or more relevant positions, the caller provides a low confidence call for that gene, under the assumption that these positions are homref. The missing positions are noted in the output
  - In addition to the genotype call, the caller also provides the corresponding metabolism status based on resource files from an external PGx caller, PharmCAT
  - The caller supports different versions of the human reference hg38
  - Usage:
    - Optionally enabled in end-to-end mode on a BAM input file with `--enable-star-allele true`, `--enable-variant-caller true` and `--vc-emit-ref-confidence gvcf`
    - Optionally enabled in stand-alone mode on a DRAGEN gVCF file with `--enable-star-allele true` and `--star-allele-gvcf <path_to_gvcf_file>`
    - Outputs a `*.star_allele.json` file containing sample name, star allele genotype and metabolism status (among other relevant fields) for all 16 genes
- **Targeted Caller for CYP2B6**
  - CYP2B6 is a member of the cytochrome P450 family of important pharmacogenes. Detecting variants in CYP2B6 is complicated by sequence homology with its pseudogene paralog CYP2B7
  - The DRAGEN CYP2B6 caller is capable of detecting copy number variants and common small variants in CYP2B6 and reports the star allele diplotype identified from those variants
  - The CYP2B6 targeted caller supports GRCh37, hg19 and hg38 references
  - Usage:
    - Optionally enabled on the germline pipeline with `--enable-cyp2b6 true`
    - Outputs a `*.cyp2b6.tsv` file containing the sample name, star allele diplotype and filter status
      `HG00554 *17/*6 PASS`
    - Requires WGS input data with at least 30x coverage
  - Issues resolved
    - Improved reproducibility of variant detection by filtering low quality base calls
    - Report multiple possible star allele diplotypes when there are ambiguous sets of variant calls
    - Detect rare version of *13 where breakpoint is detected in intron 6
    - Do not consider the *5 haplotype when the corresponding duplication star allele haplotype is unknown in the population
    - Do not fail with an assertion on low coverage (< 30x) data
- **General**

- A batch option is available for enabling all PGx callers (e.g., Star Allele, CYP2D6, CYP2B6), VC will also be enabled: `--enable-pgx true`

Expansion Hunter

- Expansion Hunter v4 in DRAGEN v3.10 supported genotyping of 30 STR loci by default
- DRAGEN v4.0 integrates Expansion Hunter v5 and expands the default catalog supported to 60 pathogenic STR loci (including 30 from gnomAD)
- Optionally, users can also call repeats on an expanded catalog with ~50K polymorphic STR loci in DRAGEN v4.0 with minor runtime impact. The 50K STRs were selected based on their proximity to coding regions (exons)
- Usage:
  - `--repeat-genotype-enable true` Enable repeat genotyping using Expansion Hunter
  - New option:
    - `--repeat-genotype-use-catalog` Variant catalog type to use (`default|default_plus_smn|expanded`). (Default=`default`).
    - The repeat catalogs are packaged with DRAGEN and autodetected based on reference and `repeat-genotype-use-catalog` setting
    - Alternately, `--repeat-genotype-specs` can be used to specify the catalog to use
- Runtime impact
  - With other variant callers, EH adds negligible overhead with default catalog and ~8-10 min overhead with expanded catalog on AWS f1.4xlarge

CNV

- Improved error handling during CNV file parsing if a column heading is missing, instead of vague message "missing expected column header..." in the middle of execution
- `--cnv-enable-plots` is no longer supported nor recommended. CNV visualization should make use of bigwig files and IGV session XMLs which has been supported since DRAGEN v3.7
- Passing DeNovo calls now should also have the `FILTER` equal to `PASS` in order to be reported as PASS in the output metrics - previously based solely on the presence proband `DN=DeNovo` flag
- HRDScore step is now skipped if upstream CNV step does not complete, for example due to insufficient coverage
- `--cnv-bypass-contig-check` can be used to bypass contig checks for self-normalization
- Issues resolved:
  - Somatic WGS
    - Fixed a bug on a minimum purity boundary condition found running DRAGEN on a 300x WGS dataset.
    - Fixed output issue on cnv_metrics, where tumor purity was not indicated as NA when DRAGEN was not able to find a good confidence model.
  - Sex genotyper confidence is now printed as NA if confidence is NaN
  - Added guard against divide by 0 during logging in KmerChecker
  - Fixed hang in DRAGEN CNV due to invalid input without newlines
  - Fixed segfault in CNV using a boost::thread member
  - Added check for empty file given for --cnv-normal-cnv-vcf

Single-Cell

- **scRNA**
  - Variable size blocks in single-cell barcodes are now correctly processed
  - Memory usage during single cell RNA processing reduced by upwards of 20%
  - Single-cell barcodes can now be marked as reverse complemented
  - Cell barcodes of up to 40bp are now supported
  - A filtered version of the cell-by-gene matrix containing only valid barcodes is now also generated
- **scATAC**
  - The new single-cell ATAC pipeline accepts as input sets of FASTQ files (2 files with reads and 1 file with cell barcodes) and produces cell-by-peak chromatin accessibility matrix in `mtx` format. See the user guide for complete details
  - Usage:
    - Enabled by setting the option `--enable-single-cell-atac` to true.
  - Outputs
    - Cell-by-peak matrices in `mtx` format
      - Full matrix with all cell barcodes
      - Filtered matrix with cell barcodes corresponding to called cells
    - Metrics
      - Per-sample metrics
      - Per-barcode metrics
    - Peak annotations - `TSV` file with information about genes associated with each called peak. Peaks are classified as promoter, distal or intergenic
    - Transcription Factor Motif Enrichment analysis matrix - motif-by-cell mtx matrix
- **Single Cell Multiomics**
  - The new single cell multiomics pipeline can process data sets from single-cell RNA-Seq and ATAC-Seq reads in a single DRAGEN invocation, generating a cell-by-feature counts matrix
  - The pipeline is compatible with library designs that have:
    - For single-cell RNA: one read in a fragment match to a transcript and the other containing a cell-barcode and UMI
    - For single-cell ATAC: two reads matching to the genome and the third one containing cell-barcode
  - Usage:
    - Enabled by setting `--enable-rna=true --enable-single-cell-rna=true --enable-single-cell-atac=true`
  - Outputs:
    - The single-cell multiomics pipeline outputs a combined cell-by-feature count matrix in `mtx` format (feature can be a gene or a peak)
    - Per-sample and per-barcode summaries from RNA and ATAC are combined
- **Interface changes**
  - Command line options for scRNA has changed in DRAGEN v4.0 and may not be backward compatible. This has been done to accurately distinguish common vs specific settings for RNA, ATAC, Multiomics
  - The following options have changed:

v3.10	v4.0	Description
`single-cell-barcode-position`	`scrna-barcode-position`	Cell barcode locations in barcode read (blocks '+'-separated)

`single-cell-umi-position`	`scrna-umi-position`	UMI location in barcode read
`single-cell-barcode-sequence-whitelist`	`scrna-barcode-sequence-list`	File with valid cell barcode sequences
`single-cell-count-introns`	`scrna-count-intron`	Count reads matching gene introns
`scrna-cell-barcode-tag`	`single-cell-barcode-tag`	BAM tag for cell barcode in single-cell (default tag is assumed to be XB)
`scrna-umi-tag`	`single-cell-umi-tag`	BAM tag for UMI in single-cell (default tag is assumed to be RX)

RNA

- Poly-A tail trimming is supported using option `--read-trimmers polya` for hard trimming or `--soft-read-trimmers polyg,polya` for soft trimming
- Gene fusion scoring model is updated based on an improved truth set and aligner
- Fusions involving chrM genes are filtered by default. ChrM can be reported using option `--rna-gf-filter-chrm false`
- Fusion candidates with multiple overlapping genes are reported as separate entries by default. To merge into a single-entry set `--rna-gf-merge-calls true`
- Added `PSPH` to the list of repetitive genes prioritized over paralogs for gene fusion calling
- Structural variants matching fusions may be reported using input VCF of SV calls using the option `--rna-gf-sv-vcf <VCF FILE>`
- Fixed a bug on GTF parser to handle whitespaces in attributes - they are not treated as attribute delimiters when surrounded by quotations

Fragmentomics

- New in v4.0, Fragmentomics metrics calculation for ctDNA assays
- Output three Fragmentomics metrics (fragmentation profile, end motif frequency and window protection score) at user-defined bin sizes, motif length, or target regions
- Support WGS/WES/TSO500 cfDNA assays in tumor only or germline mode
- Usage:
    - Enabled by setting `--enable-fragmentomics=true` (default true)
    - Enable end motif frequency calculation by setting `--fragmentomics-end-motif-len` > 0 (default 0)
    - Enable window protection score calculation by setting `--fragmentomics-wps-target-file` (default NULL)
- Fragmentation profile normalization by GC contents with options `--fragmentomics-num-gc-bins` and `--fragmentomics-gc-enable-smoothing`
- Support read filtering at exclude regions with option `--fragmentomics-exclude-bed`
- Over 30x acceleration over existing tools while accuracy remains the same
- Known issues: window protection score estimation skewed for short read length (e.g., 36bp)

BCL Conversion

- Per tile and per cycle primary statistics output to report files
  - `Adapter_Cycle_Metrics`: provides, for each cycle and sample, the number of reads mapping to the sample where the adapter was detected beginning at that cycle.
  - `Demultiplex_Tile_Stats.csv`: includes all columns that exist in the Demultiplex_Stats.csv output file but at the per sample and per tile level
  - `Quality_Tile_Stats.csv`: includes all columns that exist in the Quality_Stats.csv output file but at the per sample and per tile level
- Support `Sample_Name` column in the Sample Sheet Data section
  - Only allowed when `--sample-name-column-enabled` is enabled
  - Must be specified for every sample when enabled
  - Fastq filename will be named according to `Sample_Name` in sample sheet.
  - Fastq files will be output to subdirectory name by `Sample_ID` when `--sample-name-column-enabled` is enabled and `--bcl-sampleproject-subdirectories` is enabled
  - Reports will include `Sample_Name` and `Sample_Project` values when these features are enabled on the command line.
- Allow legacy `FindAdaptersWithIndels` sample sheet setting (default off)
  - Provides identical output to bcl2fastq2 2.20 for default adapter trim settings.
- Allow for no sample sheet to be provided using the `--no-sample-sheet` command
  - Default is disabled (sample sheet required)
  - When enabled, all sequences will go to the 'Undetermined' fastq files
  - Cannot be enabled with any of the following options
    - `bcl-sampleproject-subdirectories`
    - `sample-name-column-enabled`
    - `bcl-only-matched-reads`
    - `num-unknown-barcodes-reported`
    - `bcl-validate-sample-sheet-only`
    - `sample-sheet`
  - Can specify fastq gzip compression level
    - Command line option is `--fastq-gzip-compression-level` and default is 1
    - 0 through 9 are allowed to be specified
  - Corruption detection for BCI input files
    - In strict mode, will abort conversion with an error
    - In robust mode, will skip processing lane
- Resolved Issues
  - `--tiles` command line option did not work correctly with bgzf inputs, resulting in fastq files with incorrect tile header. This will now be correct.
  - Barcode Collision Error and Solution: Previous versions of BCL Convert allowed a conversion to continue if either index (i7 or i5) had sufficient hamming distance from all other samples in the lane. DRAGEN v3.10 and v4.0 use stricter barcode collision logic to support increased high-throughput and complex sample pooling. Each index in a dual setup must individually meet the hamming distance requirements set by the BarcodeMismatchesIndex# value. If either i7 or i5 does not meet the hamming distance requirements the program will error. For default BarcodeMismatchesIndex1 and BarcodeMismatchesIndex2:
    - Barcodes must differ by at least three bases.
    - If any two samples in i7 differ by fewer than 3 bases, an error is produced and the run will not proceed, regardless of their i5 values.
    - If any two samples in i5 differ by fewer than 3 bases, an error is produced regardless of their i7 values.

If you receive errors with current versions of DRAGEN or BCL Convert, lower the mismatch tolerance for the index producing the error by using the BarcodeMismatchesIndex1 or BarcodeMismatchesIndex2 sample sheet settings.

DRAGEN ORA compression

- **ORA output from DRAGEN BCL**
  - No longer a Beta feature
  - Can be enabled from the command line, no need to edit sample sheet
  - Processing options are configurable
  - Usage example:

    ```
    dragen --bcl-conversion-only true --bcl-input-directory <DIR> --ora-reference <path to ora reference files> --fastq-compression-format
    dragen --bcl-num-ora-compression-threads-per-file 16 --output-directory <DIR>

    --sample-sheet #only needed if SampleSheet.csv not in --bcl-input-directory
    ```

  - New command line options in DRAGEN v4.0:
    - `--fastq-compression-format` specify the type of compression:
      - `dragen` (regular ora compression) or
      - `dragen-interleaved` (paired read interleaved compression)
    - `--bcl-num-ora-compression-threads-per-file` Optionally set the number of threads used per file files. Default is 10
    - `--bcl-num-ora-compression-parallel-files` Optionally set the number of files processed in parallel. Default is 6
  - Known issue: The BCL to FASTQ.ORA interleaved mode compresses the paired read files into one single interleaved `.ora` file and output the files in the output directory, but the `fastq-list.csv` file created does not update to proper path for these single interleaved files

- **ORA interleaved compression**
  - The interleaved mode can be enabled with a list of paired read files as input. Use `--fastq-list` in conjunction with `--ora-interleaved-compression` set to true
  - Ora files generated with the interleaved mode are labelled "interleaved" in filename when the ORA compression is enabled from BCL or from FASTQ.GZ
  - Usage example:
    - `dragen --enable-map-align false --enable-ora true --fastq-list <file.csv> --ora-interleaved-compression true --ora-reference <...> --output-directory <...>`
- New features (only when compression is enabled from FASTQs)
  - Option to compute md5 checksum of fastq.ora files during compression with `--ora-enable-md5` (false by default)
  - Option to automatically delete the input FASTQ file from the disk upon completion of compression with `--ora-delete-input-files` (false by default)
  - Tool to check ora file integrity. Run a separate job on ora file with option `--ora-check-file-integrity=true`, to validate that the decompressed ora file checksum matches that of the expected decompressed fastq.gz

- Option to generate ora compressed files in parallel of mapping step. When running map/align with fastq input, set `--enable-ora=true` to compress the files as part of the DRAGEN analysis. Bases are deducted from both the Germline and Compression licenses

gVCF Genotyper

- New features in v4.0
  - 2x speed improvement compared to DRAGEN v3.10
  - `mimalloc` wrapper script, results in an additional ~50% faster runtime when used (~3x faster than DRAGEN v3.10)
  - New command to combine per-batch msVCF. 12x faster than `bcftools merge`
  - Concise msVCF output: up to 40% space saving
  - HPC wrapper scripts and demo workflow available on request
- Command line parameters added
  - `--gvcfs-to-msvcf` end-to-end mode for a single batch (default=true if no other step option is set)
  - `--merge-batches` merge msVCF files from different sample batches into a single msVCF (default=false)
  - `--input-census-list` file specifying list of census files for input (applicable when aggregate-censuses is true)
  - `--input-batch-list` file specifying list of sample batch msVCF files for merging (applicable when merge-batches is true)
  - `--gg-enable-indexing` set to true to generate a tabix index for the merged msVCF (default false)

Hash Table builder

- **New anchored hash table for HLA**
  - To run the HLA caller, an HLA-specific anchored reference hash table must be built. Set `--ht-build-hla-hashtable=true`. The command will create an `anchored_hla` subdirectory inside the `--output-directory`. The HLA-specific reference subdirectory can be built at the same time as the primary reference construction
  - An HLA resource file is packaged with DRAGEN and located at the following path after installation: `/opt/edico/resources/hla/HLA_resource.v1.fasta.gz`. This file is used by default when building the HLA-specific anchored hash table. A custom file can be specified with `--ht-hla-reference`. See the HLA section in the DRAGEN User Guide for more information
  - Custom HLA reference files might require customized memory allocation, which can be specified with an argument to the command-line option `--ht-hla-ext-table-alloc`

- **Build a custom multigenome human reference**
  - A new tool is made available for building a multigenome reference hash table from a set of population variants (VCF). Improves the accuracy results for a given population
  - The method introduces alternate graph paths to the reference hash table to represent more broadly the allelic diversity of the population in specific regions
  - Resource file packages containing a mask bed, graph bed and reference genome FASTA, are provided for download for hg38, hg19 and hs37d5 respectively
  - Usage:

- `dragen --build-hash-table=true --output-directory=<output_dir> --ht-reference=<path>--ht-graph-vcf-list=<path> --ht-graph-bed=<path> --ht-mask-bed=<path>`
- Key commands to be used within DRAGEN
  - `--build-hash-table=true`
  - `--ht-graph-vcf-list` path to text file containing VCF files to build custom multigenome hash table
  - `--ht-reference` path to reference genome FASTA file
  - `--ht-graph-bed` path to bed file, which filters `--ht-graph-vcf-list` regions. This acts as an allowed list
  - `--ht-mask-bed` path to bed file, which filters `--ht-graph-vcf-list` regions and masks specific regions in FASTA seq provided with `--ht-reference`

- Limitations/known issues
  - Support only for human genome builds hg38, hg19 and hs37d5
    - only diploid calls are supported even for sex chromosomes
    - for optimal results VCFs input provided shall be phased, no warning is outputted if not
    - reference genome build (FASTA) provided shall be the same build as the one used to generate the input VCF, no warning/error is outputted if not

Other DRAGEN Updates and Features

- **Backgrounding and Signal Handling**
  - Users who want to background a `dragen` process should make use of a terminal multiplexer like 'screen' or 'tmux'.
  - Traditional methods of backgrounding, like 'nohup' or the ^Z hotkey, are not supported. DRAGEN is known to crash or hang when backgrounded in these ways.
  - To stop a running `dragen` process, users must send either the SIGINT (kill -2) or SIGTERM (kill -15) signal.
    - Using the SIGKILL (kill -9) signal is harmful and must never be used. A `dragen` process interupted by SIGKILL may damage the system's FPGA.
    - Users of job queues like slurm and SGE should ensure that the job queue never sends SIGKILL to a `dragen` process.
- **AWS**
  - New AWS FPGA shell. When re-using an AWS FPGA f1 instance, and when running different versions of `dragen` on the re-used instance, a SW version older than v4.0 will not recognize the shell left by v4.0 and ASSERT thinking something is wrong
  - DRAGEN v4.0 `dragen_reset` has been updated to clear the shell, so that instances can be re-used by older versions
- **CRAM Decompression**

- DRAGEN v4.0 now supports the re-alignment with a CRAM input that was created with a different reference, in one step
- Use the `--cram-reference` option to make the CRAM decompressor use the specified reference
- `--cram-reference` can be either a fasta file, or a DRAGEN hash table folder. If pointing to a fasta file, the fasta .fai index file must be present next to the fasta file
- It only applies to decompression, and when map/align is enabled. CRAM output will always be compressed using the reference specified with `--ref-dir`
- Usage:
    - Examples: CRAM was created with hg19, re-analysis with hg38

```
dragen -r <ref_dir HG38> --cram-input <cram> --output-directory
<out_dir> --output-file-prefix <out_prefix> --cram-reference
<ref_dir HG19>
```

```
dragen -r <ref_dir HG38> --cram-input <cram> --output-directory
<out_dir> --output-file-prefix <out_prefix> --cram-reference
<hg19.fa>
```

- **Other**
    - `bam2cram` utility was not functional and has been removed
    - `dragen` will report a non-zero exit code on the std output and runlog and syslogs, with "DRAGEN finished with exit code " x message in such a case
    - DRAGEN now handles `vcf.bgz` (bgzipped) inputs correctly. Previous versions required the file extension to match `vcf.gz`
    - The ploidy estimation and ploidy caller modules now take GC bias into account for normalization, improving the accuracy of aneuploidy calls

Issues Resolved

Issues found on DRAGEN™ v3.10 or older that are fixed in v4.0

Component/s	Defect ID	Issue Description
Amplicon, Gene Fusion	DRAGEN-16254	Fix for RNA Amplicon runtime ~15 Hours on large sample
Azure	DRAGEN-16335	Fix for popen exception on azure cloud suites
BCL	DRAGEN-17031	Fix for DRAGEN BCL not displaying copyright info when -h or --help command line options are used
BCL	DRAGEN-15757	Fix a segfault in BCL conversion seen on v3.9.3
BCL	DRAGEN-14936	Fix for wrong output when per-sample-settings are combined with per-sample-cbcl workflow
BCL	DRAGEN-14888	Fix for per-sample settings output being incorrect when either i7 or i5 index is completely masked out
BCL	DRAGEN-14842	Fix for multi-adapter-per-read settings producing an incorrect error after per-sample-settings support
CNV VC	DRAGEN-17017	Fix CNV failure: requested priors for a custom germline state with tumor specifics that has not been constructed
Compression, Inputs	DRAGEN-17533	Fix Assertion when UMI FASTQ is an Ora file
Downsampler	DRAGEN-16681	Fix for issue where enabling downsampling disabled the ploidy estimator
Dupmarking	DRAGEN-14765	Fix dragen dedup hang in hash-based dedup mode, on corrupted input data. Detect and crash
Force GT, SNV VC, Somatic	DRAGEN-16279	Fix for variants not being force genotyped when ForceGT and CNV are run together
GVCF Genotyper	DRAGEN-18147	Fix for force-gt failure with --gg-only-forced-sites
GVCF Genotyper	DRAGEN-17905	Fix excessive flushing when writing bgzip files inflating output file size
GVCF Genotyper	DRAGEN-17403	Throw exception when inputs has duplicate sample names
GVCF Genotyper	DRAGEN-17139	Optimize BgzTbiAsciiReader::readHeader to not read the entire file
GVCF Genotyper	DRAGEN-17060	Fix for cohort & census headers not being preloaded
GVCF Genotyper	DRAGEN-16771	Fix segfault on chr21 of 1000 genomes data
Hash Table Builder	DRAGEN-9474	Fix error building hash table on small custom references

Component/s	Defect ID	Issue Description
HWAL/Infra	DRAGEN-13503	[XLNX] ERROR: Xilinx I2C timeout waiting for bus busy. Fix with more robust retry mechanism
Infrastructure, Inputs	DRAGEN-14371	Update dragen return code and print message, when input vcf is not sorted
Joint Genotyping	DRAGEN-17588	Fix a regression in denovo calling accuracy when ML is enabled
Paralog Caller	DRAGEN-17108	Fix CYP2D6 reproducibility for Coriell NA24217 replicates
Paralog Caller	DRAGEN-15473	Fix for CYP2D6 calls discordant with GeTRM
QC Metrics, scRNA	DRAGEN-15950	Fix run-to-run variation in scRNA on bdbio
RNA	DRAGEN-16360	Fix for DRAGEN on NextSeq 2000: Single Cell RNA hang-up
RNA Alignment, Infrastructure	DRAGEN-16472	Fix cloud UL streaming test fail: Dragen assumed the gene annotation file should stay in the local dir for downstream step
scRNA	DRAGEN-15329	Fix the number of non-null entries in the matrix.mtx file (feature counting)
SNV VC	DRAGEN-17676	Use reads likelihoods when ploidy = 1
SNV VC	DRAGEN-15895	Fix for Germline MNV calls printing 0.00 for QUAL
SNV VC	DRAGEN-15845	Update handling of A,C,G,T bases with baseQ=0 in DRAGEN
SNV VC	DRAGEN-15694	Add a non primary allelic filter to remove new FP in TSO500 Solid
SNV VC	DRAGEN-14890	Fix segmentation fault in MNV
SNV VC	DRAGEN-16395	Fix Double free crash in PcrModelSampler
Somatic, GVCF, SNV VC	DRAGEN-16085	Fix unexpectedly high homref score at variant site in somatic caller
Somatic, SNV VC	DRAGEN-15908	Realign reads with soft-clips for UMI-based VC to improve read position filter
Somatic, SNV VC	DRAGEN-15798	Fix GVCF diff between GVCF-only mode and GVCF+VCF mode with MNV enabled
Somatic, SNV VC	DRAGEN-15180	Fix for MNV dropping homref score in GVCF mode
Somatic, SNV VC, GVCF	DRAGEN-15482	Fix DRAGEN Somatic crash when running in gVCF Mode with gVCF+VCF output and a Depth Annotation Threshold
SV	DRAGEN-16760	Fix WatchDog hang on 200x T/N sample
SV	DRAGEN-14956	Fix for TSO500 filter failing on a LoB dataset
TMB	DRAGEN-16869	Update TMB parsing to be more robust

Component/s	Defect ID	Issue Description
UMI	DRAGEN-17082	Fix for BAM not having SA tag when downsampling
UMI	DRAGEN-16412	Fix for UMI output being different from run-to-run

Known Issues

Known issues of the DRAGEN™ v4.0.3 release

Component/s	Defect ID	Issue Description	Remedy/Workaround
BCL	DRAGEN-19115	bcl-convert hang but does not crash	None. A watchdog mechanism does not exist for sw-only tools
BCL, Ora Compression	DRAGEN-19251	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-19185	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	DRAGEN-19157	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.
Compression	DRAGEN-19223	When CRAM is decompressed with a different HashTable ref than the map/align, AND CRAM is 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
DNA Alignment	DRAGEN-16308	When running different read trimmers back to back, sigabort during RecomputeTags computeTags	Workaround: Issue a dragen_reset in between the runs.
HLA	DRAGEN-17786	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	DRAGEN-13717	RNA VC hits ERROR: Invalid node flags	Rare hardware error. Re-run sample as it is expected to pass
Inputs, Parallog Caller	DRAGEN-11094	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
Joint Genotyping	DRAGEN-16531	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

Component/s	Defect ID	Issue Description	Remedy/Workaround
SNV VC	DRAGEN-16467	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	DRAGEN-15927	Gene panel accuracy regression, one missing 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.
SNV VC, ML	DRAGEN-18342	With ML enabled, some of the mapq=0 reads have 'N' bases which when comparing to the ALT allele gives a different result between BAM and FASTQ input.	None. A fix for this issue will be released
Somatic	DRAGEN-19052	Variant caller run time slowdown of up to 33% on certain types of somatic tumor only samples on v4.0 compared to v3.10	Run time impact of NTD error estimation
Somatic	DRAGEN-16319	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
Somatic	DRAGEN-19218	Increased Indel FP on Carpten_CHLA_WGS_350_tumor T-Only test dataset (INDEL FP+FN +10.7% relative to v3.10)	None
SV	DRAGEN-18913	Extra SV call in FLT3-ITD hotspot region for FLT3_C317.TCGA-AB-2830 T/N sample	None
VC ML	DRAGEN-18807	Accuracy regression of up to 16% with ML vs non-ML when the ref is hs37d5	None. The use of hg38 reference is advised. A fix for this regression will be released

SW Installation Procedure

- Download the desired installer from the Illumina support website and unzip the package
- The archive integrity can be checked using: `./<DRAGEN 4.0.3 .run file> --check`
- Install the appropriate release based on your Linux OS with the command: `sudo sh <DRAGEN 4.0.3 .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