# DRAGEN v3.10.4 Software Release Notes

# Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform v3.10.4.

Changes are relative to DRAGEN™ v3.9.5. If you are upgrading from a version prior to DRAGEN™ v3.9.5, 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 3.10.4 software package includes installers for the on-site server:
- DRAGEN™ SW Intel Centos 7 - dragen-3.10.4-4.el7.x86_64.run
- DRAGEN™ SW Intel Oracle 8 - dragen-3.10.4-4.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 has 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 v3.10.4. DRAGEN™ v3.10 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

## Changes to Supported Platforms

### Red Hat Enterprise Linux (RHEL) v8

Support for the RHEL v8 operating system has been added since DRAGEN v3.9.5 for on-site servers. The .run file includes an RPM installer for el8. The installer now uses the Linux dkms service to manage the kernel driver. DRAGEN is tested and verified on the Oracle Linux Server release 8.4 distribution. AWS RPMs, AWS AMIs, and Microsoft Azure VMs are not available for el8.

### Dragen Server v1 FPGA

Support for the server v1 FPGA "eagle" card has been deprecated with v3.10. The installer still includes up to date FPGA bitstreams, but this software release has not been verified on server v1. Future releases will remove the FPGA bitstreams for the "eagle" card. A user can identify the FPGA version by running `dragen -V`. A software version starting with `01` indicates that the server has the deprecated "eagle" FPGA installed, where a software version starting with `07` indicates the supported "U200" FPGA is installed.

## New Product Files

New product files that accompanies the v3.10 software is available for download on the Illumina support website https://support.illumina.com/sequencing/sequencing_software/dragen-bio-it-platform/product_files.html

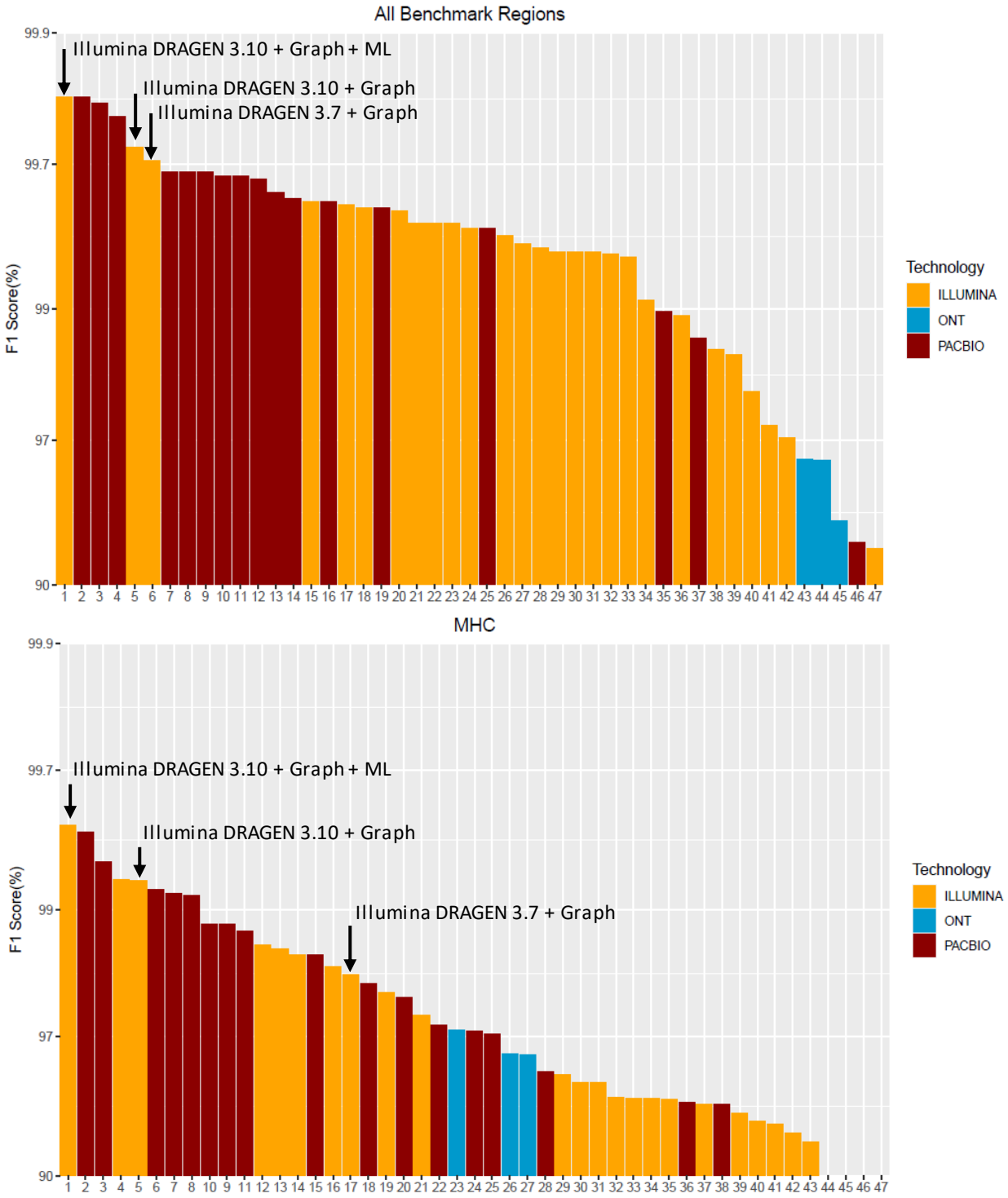
The following updated product files are available:

- Alt-masked v2 graph genomes for hg19 and hg38
- DRAGEM-ML model file version 3.1 for hg38
- Ora compression reference v2

## New Callers and Major Features

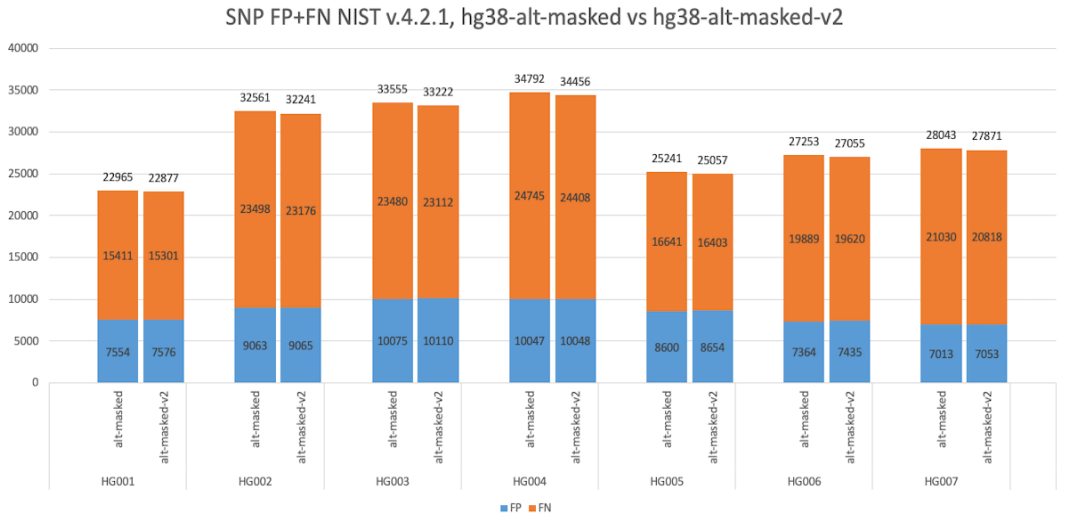
### Germline Small Variant Calling Accuracy Improvement

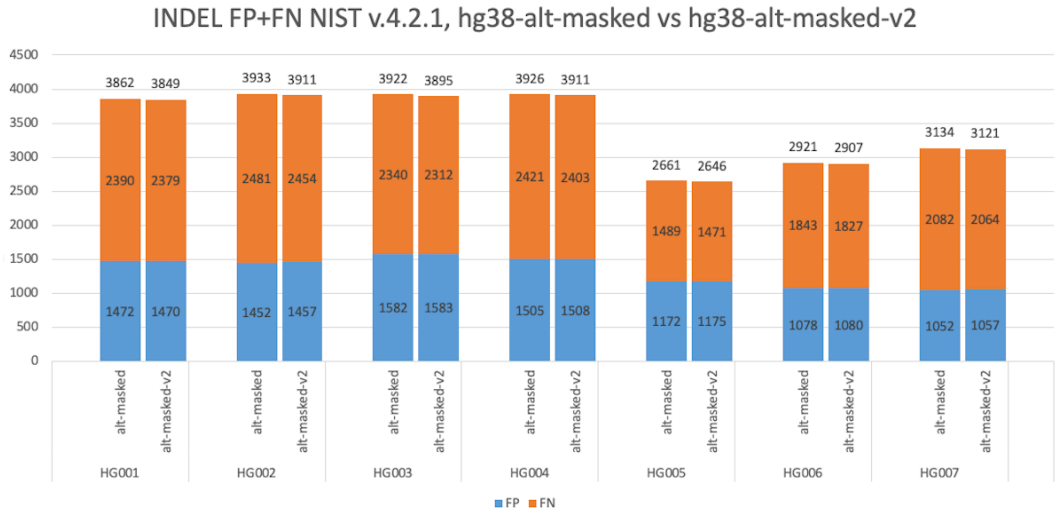
- DRAGEN v3.10 sets new standards for data accuracy on Precision FDA benchmark data. When employing the Graph and ML improvements, it results in lead accuracy across all read technologies in all benchmarks and MHC regions
- The graphs below show the accuracy of DRAGEN v3.10 Graph and ML compared to the Precision FDA v2 submissions in the 'All benchmark regions', as well as the accuracy of DRAGEN v3.10 Graph and ML compared to the Precision FDA v2 submissions in the MHC region


All Benchmark Regions
Illumina DRAGEN 3.10 + Graph + ML
Illumina DRAGEN 3.10 + Graph
Illumina DRAGEN 3.7 + Graph
F1 Score(%)
99.9
99.7
99
97
90
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
Technology
ILLUMINA
ONT
PACBIO
MHC
Illumina DRAGEN 3.10 + Graph + ML
Illumina DRAGEN 3.10 + Graph
Illumina DRAGEN 3.7 + Graph
F1 Score(%)
99.9
99.7
99
97
90
1 2 3 4 5 6 7 8 10 9 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
Technology
ILLUMINA
ONT
PACBIO

## Reference Updates

- Regions have been identified where there are false positive duplications in hg38. New reference was released by GIAB which masked the false duplications by N's. The updated reference rescues genes, including CRYAA, CBS and KCNE1.
- DRAGEN v3.10 automatically masks these regions when the hash tables are built, therefore no changes to fasta are needed. When alt-masked references are re-built with DRAGEN v3.10, these updates are available. The result is an improvement in mapping and variant calling accuracy for hg38 genomes
- These updates are referred to as Version 2 of alt-masking (denoted v2) in this release note and communication.
- A comprehensive technical guide to DRAGEN reference genomes is available here: https://www.illumina.com/science/genomics-research/articles/dragen-demystifying-reference-genomes.html



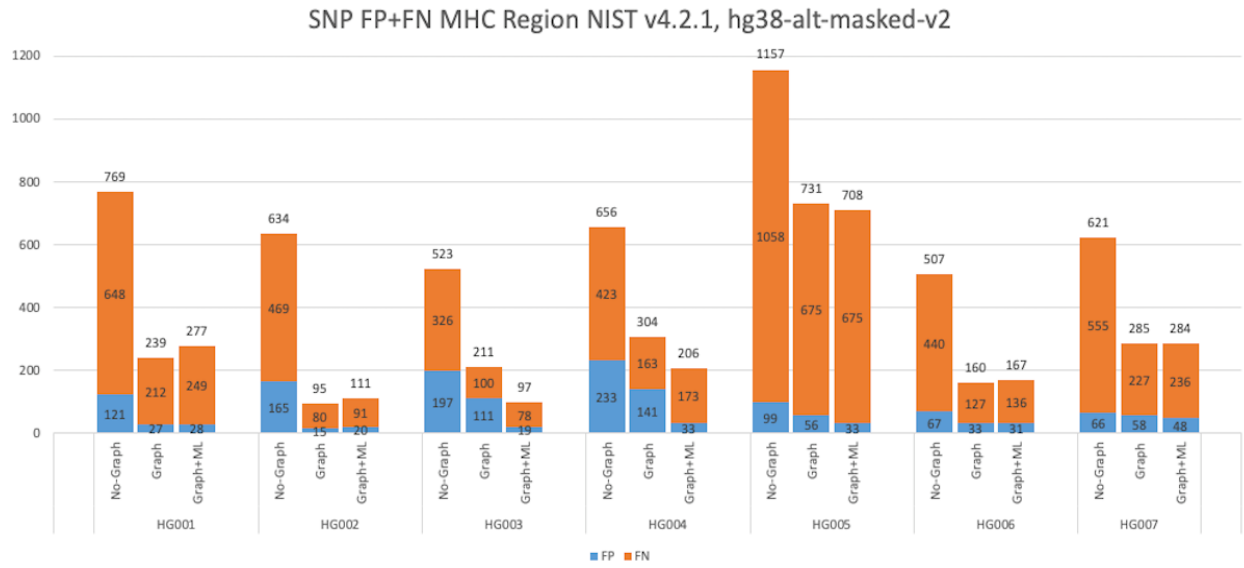



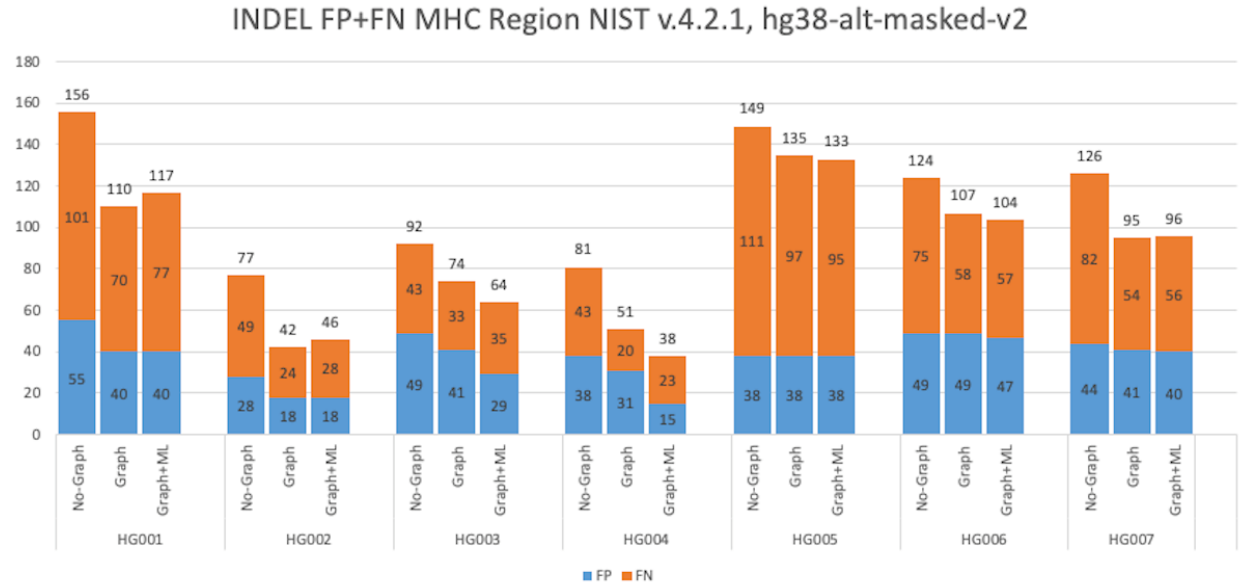

| Gene | Total variants | FP alt-masked | FP alt-masked-v2 | FN alt-masked | FN alt-masked-v2 |
|---|---|---|---|---|---|
| CRYAA | 9 | 1 | 0 | 8 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| CBS | 53 | 12 | 0 | 46 | 0 |
| KCNE1 | 106 | 1 | 0 | 80 | 0 |

## Graph Genome Improvements

- Updated Graph Genome now covers a wider portion of the MHC regions, with a greater diversity of population ALT haplotypes
- The updated graph genomes include both MHC region and reference updates, and are named *hg19 alt-masked-graph-v2* and *hg38 alt-masked-graph-v2*



## Machine Learning (ML) Updates

DRAGEN v3.10 offers a Beta release of the ML based recalibration of variant parameters, which improves the functionality and accuracy with the following updates:

- Added support for multi-allelic variant recalibration. Multi-allelic positions were skipped with DRAGEN v3.9 ML Alpha
- The model has been extended to support recalibration in haploid sex chromosomes. The performance matches diploid positions
- Added support for hom-ref/force GT. The model now supports QUAL/GQ recalibration in positions where VC output a 0/0 genotype, recovering more FNs
- Performance Improvements. All variants are recalibrated in v3.10. Accuracy improved 5% on SNP FP+FN, with smaller gains on Indels. Indel ROC curve is improved

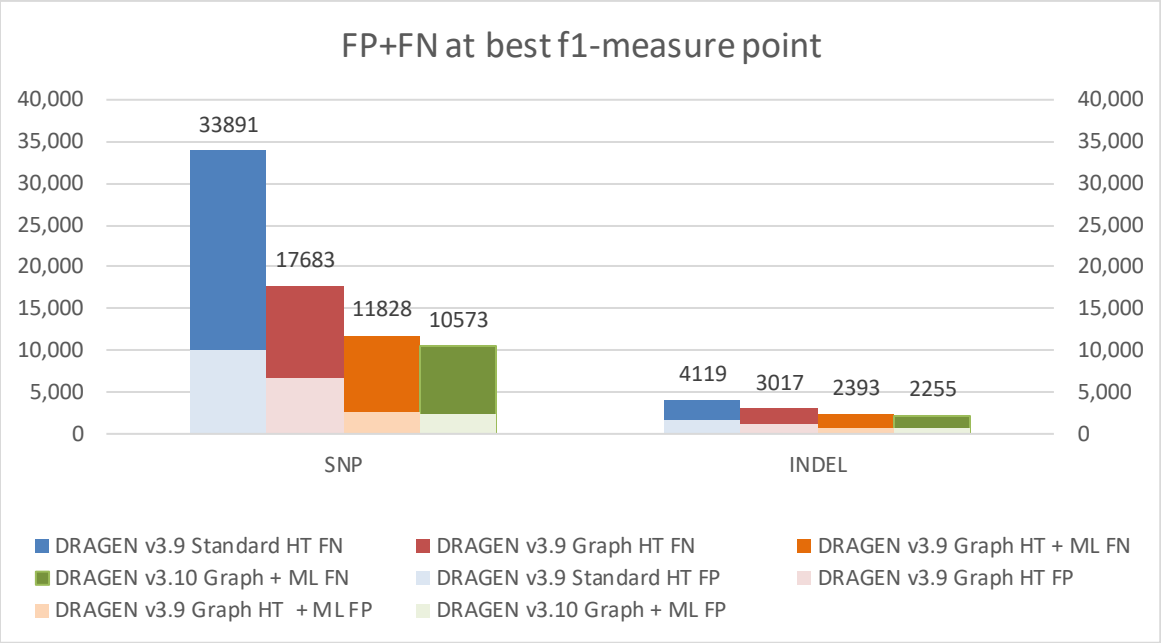
Notes for ML usage

- Can be optionally enabled as part of germline small variant calling by setting
  - `--vc-ml-enable-recalibration=true`, and
  - `--vc-ml-dir=</path/to/package/directory>`, the path to the ML model file
- ML runs automatically with the small variant caller
- The ML model file has been updated to version 3.1 for DRAGEN v3.10 and must be downloaded from the Illumina support site. The download is a self-extracting relocatable installer
- hg38 is the only supported reference in this release
- Minimal extra processing time of 5 to 7 minutes for a whole human genome
- Can be used for WGS or WES samples
- The workflow does not support re-calibration of existing VCF files

Important updates to fields and filtering

- QUAL, GT, GQ fields are updated in the VCF. NOTE: The values will be different between ML enabled or disabled
- Original VC scores and genotypes are preserved and output in the VCF under DQUAL, DGT, DGQ fields
- Hard-filtering is applied at QUAL 3 when ML is enabled for both SNP and INDEL

The graph below shows the accuracy improvement of ML in v3.10 compared to v3.9




## Targeted Caller for GBA

- GBA variants cause the autosomal recessive disorder Gaucher disease, and carriers are at increased risk of Parkinson's disease and Lewy body dementia. GBA is one of the 113 genes recommended for carrier screening by ACMG and its analysis is complicated by sequence homology with its pseudogene paralog GBAP1
- The DRAGEN GBA caller can detect both recombinant and non-recombinant variants in *GBA* by differentiating between reads mapping to the *GBA* gene and its paralog *GBAP1*
- *GBA* variants are reported along with the carrier status
- The GBA Caller supports GRCh37, hg19 and hg38 references
- Usage:
  - Optionally enabled as part of germline pipeline by using `--enable-gba=true`
  - Outputs a file *.gba.tsv file containing GBA variants and carrier status (example)

    ```
    #Sample    is_carrier    recombinant_variants
    HG00115    True          c.1263del+RecTL
    ```

  - Requires WGS input data with at least 30x coverage
- Performance
  - Validation is done against 24 samples, both clinical and from 1000 Genomes Project.
  - The caller output is concordant with orthogonal methods. Those methods include digital PCR and targeted long-read sequencing

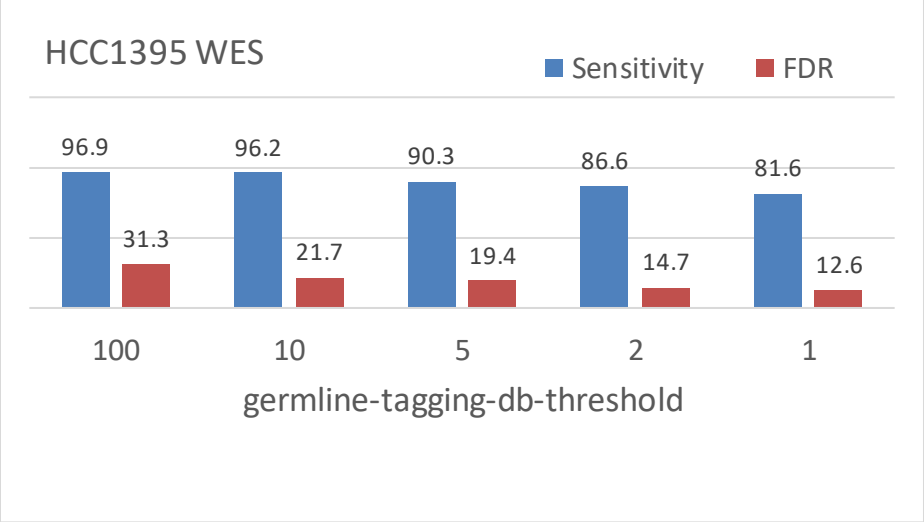
| Copy number change | Recombinant variant | Total samples | Concordant |
|---|---|---|---|
| **Gain** | None | 10 | 10 (**100%**) |
| **Loss** | None | 2 | 2 (**100%**) |
| | RecNciI | 1 | 1 (**100%**) |
| **Neutral** | A495P | 1 | 1 (**100%**) |
| | L483P | 7 | 7 (**100%**) |
| | c.1263del | 1 | 1 (**100%**) |
| | c.1263del+RecTL | 2 | 2 (**100%**) |
| **Total** | | 24 | 24 (**100%**) |

## Somatic Small VC Updates

- New Germline filtering feature
  - In somatic tumor-only variant calling, potential germlines variants can be annotated in the VCF file by enabling the Nirvana variant annotation. The Nirvana annotation runs as integrated part of the workflow. It annotates germline variants by using its population-database. The workflow automatically processes the annotations and updates the somatic VCF output file with tags called `GermlineStatus`
  - Usage:
    - *`--vc-enable-germline-tagging true --enable-variant-annotation true --variant-annotation-data <Nirvana data file> --variant-annotation-assembly <GRCh37/GRch38>`*
    - Pre-requisites: User must download the Nirvana annotation data file. See user guide.
    - Additional parameters:
      - `--germline-tagging-db-threshold` The minimum alternative allele count in database for a variant to be defined as germline. The default is 50.
      - `--germline-tagging-pop-af-threshold` The minimum population allele frequency for a variant to be defined as germline. Once specified, this will ignore the input from `--germline-tagging-db-threshold`.
    - Example output:

```
chr1 11301714 . A G . PASS
DP=3626;MQ=249.61;FractionInformativeReads=0.974;AQ=100.00;GermlineStatus=Germline_DB GT:SQ:AD:AF:F1R2:F2R1:DP:SB:MB
0/1:64.73:1772,1758:0.498:872,901:900,857:3530:846,926,843,915:894,878,874,884
chr6 29910675 . G A . PASS
DP=1066;MQ=171.30;FractionInformativeReads=0.946;AQ=100.00;GermlineStatus=Somatic GT:SQ:AD:AF:F1R2:F2R1:DP:SB:MB
0/1:11.79:934,74:0.073:492,43:442,31:1008:444,490,31,43:462,472,32,42
```

  - The following graph shows an example of the impact to Sensitivity and FDR on a WES sample for different values of the alt allele count used to classify a variant as germline

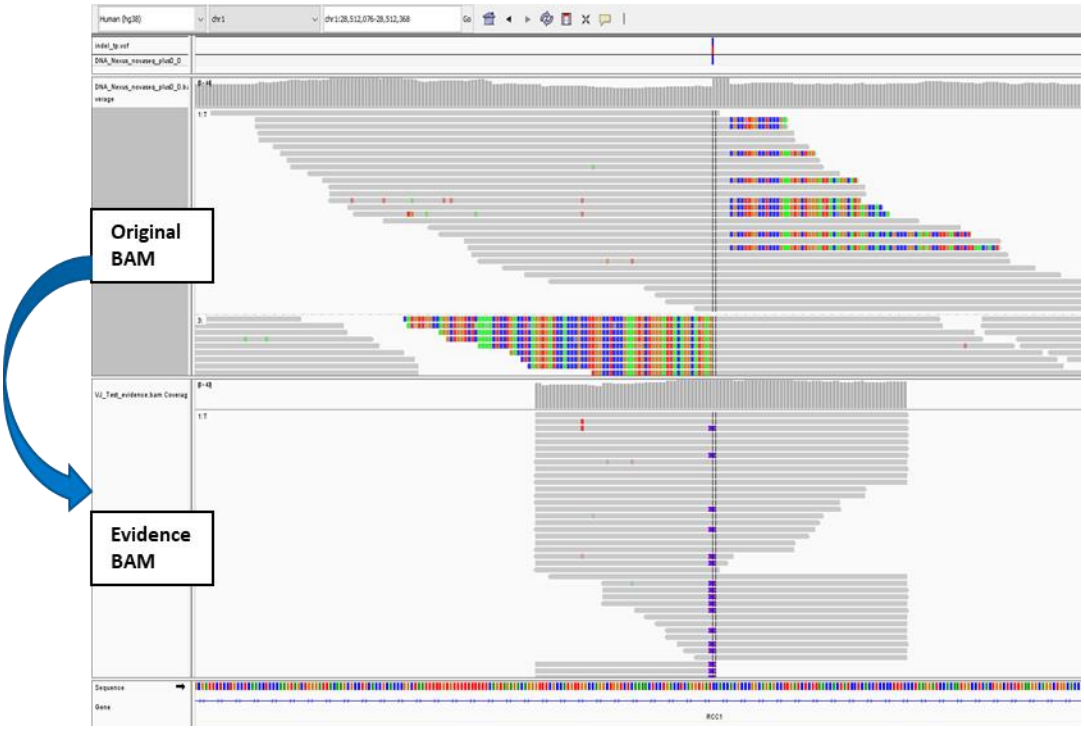



- New Nucleotide error bias correction (NTD filter)

  - Somatic VC can now also compensate for biased nucleotide error rates due to oxidation/deamination. In addition to correcting orientation bias (F1R2 vs F2R1 difference), it can also correct bias in any SNV type substitution. As a result, the NTD filter can rescue some FP and FN that is not corrected by existing orientation bias filtering.
  - Usage:
    - `--vc-enable-unequal-ntd-errors=true --vc-enable-trimer-context=true`

### VC Evidence BAM

- The DRAGEN small variant caller is a haplotype-based caller and performs local assembly of all reads in an active region using a deBruijn graph. The assembly process uses all the read bases including the soft-clip bases. The soft-clip bases provide evidence for the presence of variants, specifically longer insertions and deletions which are not present in the read cigar and hence cannot be directly viewed in IGV. The assembly and HMM (combines all alignment paths between read and haplotype) modules in the variant caller effectively realign the read locally, reducing alignment errors, and improving the variant caller accuracy.
- A new Evidence BAM can be generated by DRAGEN v3.10 to provide user insight into how the small variant caller realigns and sees the read evidence.
- Usage:
  - Set `--vc-output-evidence-bam=true` along with other standard variant calling options to enable evidence BAM output
  - Outputs:
    - Realigned reads are represented in a bam file in the output. The file suffix is `_evidence.bam`
    - By default, the read realignment is performed only on a subset of the genome, regions which have evidence for indel variants, and some percentages of reads are soft-clipped. A bed file is generated containing all the regions that are realigned and output in the evidence BAM. The file suffix is `realigned-regions.bed`
  - Set `--vc-evidence-bam-force-output=true` to force read realignment on all active regions
- The figure below shows an example of the differences in the pileup of the post-sort BAM and the post-vc evidence BAM, helping provide insight into why a call is made




## Indel Realignment

- Genomic mappers consider reads independently and may favor mismatches over indels. Indel Realignment attempts to correct alignments close to indels using a consensus model that considers all reads in that region.
- DRAGEN v3.10 offers an optional integrated indel realignment step. The indel realignment step works on reads upstream of sort/dedup. Downstream callers would see re-aligned reads as if they were originally aligned that way.
- No major impact on accuracy in DRAGEN. The feature can be used to produce indel-realigned BAM/CRAM files for third-party analysis of mapped read data from DRAGEN.
- Usage:
  - Set `--enable-indel-realigner` true
  - Can run integrated in the DRAGEN end-end pipeline
  - Can run as standalone tool, from pre-aligned input (BAM, CRAM)
  - Realigned reads are represented in the BAM generated by DRAGEN. Original alignments are stored in the `OA` BAM tag
  - Options to control the functionality of the tool
    - `--ir-write-intervals-file` When Indel Realigner is enabled, output a file with the reference intervals that contain evidence for realignment (default=false)
    - `--ir-max-num-candidates` Max number of realignment candidates in an interval for realignment (default 256)
    - `--ir-max-num-consensus` Max number of consensus reads in an interval for realignment (default 256)
    - `--ir-max-num-reads` Max number of reads in an interval for realignment (default 20000)
    - `--ir-max-distance-between-mates` Max distance between the alignment positions of mates to be considered for realignment (default 100000) Less restrictive than GATK

- `--ir-realignment-threshold` Minimal improvement of sum of mismatching base qualities to merit realignment (default 50)

## HRD BRCA LR Biomarker

- Within gene large-rearrangement (LR) events can be detected via tumor-only CNV calling. Performance of BRCA LR detection has been verified with Illumina's TruSight Oncology TSO500 panel. The same cmd line options can be tested on other Tumor-Only pipelines and panels
- Usage:
  - `--enable-cnv true --cnv-normals-list <list of panel of normals> --cnv-target-bed <targeted bed file> --cnv-within-gene-lr-bed <BRCA_region.bed> --tumor-bam-input <BAM> [additional CNV settings]`
  - Additional CNV settings should fine-tuned per pipeline. Contact your technical representative for further details.
  - Output:
    - A JSON file is output. The suffix is `cnv.LR.json`
    - Example JSON output

```
"Breakpoints": {

    "BRCA1": {

        "nSegs": "3",

        "segments": [

         { "id": "BRCA1.1", "chromosome": "chr17", "start":
         "41197309", "stop": "41228630", "nBin": "40", "segmentMean":
         "0.729", "segmentMeanLog2": "-0.455" },
```

## RNA Amplicon

- DRAGEN v3.10 now supports accurate Gene Fusion calling of Amplicon sequencing outputs
- RNA Amplicon sequencing performs deep sequencing in targeted regions using custom probes and PCR products (amplicons) to efficiently validate and screen gene fusions. In DRAGEN 3.10, we add special processing of RNA-seq reads within the gene fusion caller to identify RNA amplicon targeted fusions.
- DRAGEN's TP fusion calls are concordant with Illumina's legacy BaseSpace RNA Amplicon application
- Gene expression can be quantified using target region read counts reported in the QC metrics file
- Usage:
  - Required options: `--enable-rna-amplicon=true --amplicon-target-bed=<amplicon.bed> --enable-rna-gene-fusion=true`
  - Optional: `--rna-gf-enriched-genes=<fusion_genes.txt>` File with list of genes included in enrichment panel, 1 per line
  - Notes:
    - RNA variant calling is not available for RNA amplicon
    - RNA quantification is not recommended for RNA amplicon. Read-level coverage for each region in the amplicon target BED is output as part of the mapper QC metrics (in file `target_bed_read_cov_report.bed`). Gene expression can be quantified from this coverage BED
    - The amplicon target BED is required. Example structure of the target BED

```
#chr1  11810241  11810242  T00001.AGTRAP.ENST00000314340 AGTRAP Fusion
#chr1  49332862  49332863  T00002.AGBL4.ENST00000371839  AGBL4  Fusion
#chr1  114165615 114165616 T00003.MAGI3.ENST00000369617  MAGI3  Fusion
……
#chrX  13754596  13754597  T00261.OFD1.ENST00000340096   OFD1   Fusion
#chrX  64956786  64956787  T00347.MSN.ENST00000360270    MSN    Fusion
```

        - A *targetType* set to "`Fusion`" is used to identify fusion targets. The gene IDs for fusion targets are collected and written to an output file
- Method notes
    - De-duplication of fusion supporting reads is disabled to fully utilize the information in the targeted amplicons. NOTE: This can lead to long run times if the sample coverage is too high.
    - A transcript breakpoint that overlaps both genes of the fusion is allowed for this pipeline, to enable specific amplicon targeted fusions such as FIPL1--PDGFRA to be called
    - The scoring model used for filtering fusion calls is customized based on amplicon re-training
    - Both genes in a fusion need to be present in the enriched gene list (which can be derived from amplicon bed file or directly specified) in order to pass through filtering

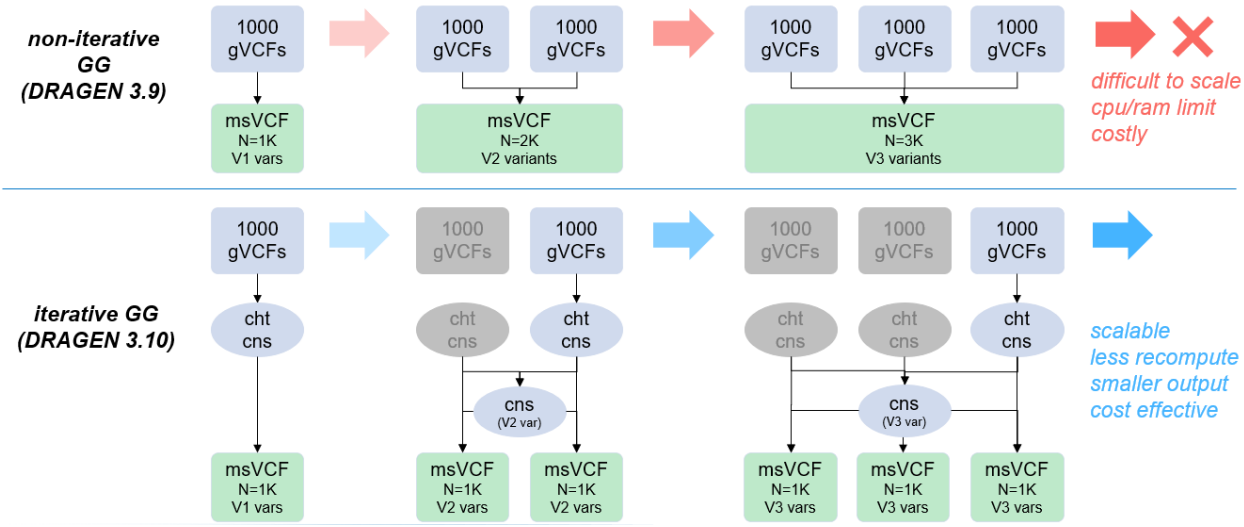
### Integrated Down-sampling

- Trimming, QC, down-sampling, alignment, and analysis are now all possible in a single workflow, single command line.
- Down-sampling is seeded, pseudo-random
- Compatible with all Germline and RNA workflows
- Applications:
    - Ease-of-use, to enable simplified means of doing comparisons
    - Example:
        - Between samples in high-multiplex RNA differential expression analysis, perform RNA analysis and quantification on exactly 15 million reads, but only on reads >=100bp after trimming
        - Use: `--read-trimmers=quality --trim-min-quality=20 --trim-min-length=100 --enable-rna=true --enable-rna-quantification=true --enable-down-sampler=true --down-sampler-reads=15000000`
    - Produce a down-sampled FASTQ output
        - Set `--enable-down-sampler-output=true`
        - NOTE: this feature is not compatible with RNA map/align.

### Popgen Updates

- New Iterative Gvcf Genotyper design, allows for scalable aggregation of gVCFs beyond 100K scale
- N+1 support is added. Aggregation of batches without re-processing existing batches, update msVCF for all batches
- Global statistics are calculated. Multi-allelic SNP and Indel AC, AF, sample coverage info across all samples.
- Highly parallelizable. Parallelization by genome shards and sample batches

- The diagram below depicts the advantage of the new Iterative Gvcf Genotyper workflow in DRAGEN v3.10 over the legacy design of v3.9




- Usage is detailed in the table below

| | Description | Commands |
|---|---|---|
| | Use these options for all executions | `--enable-gvcf-genotyper-iterative true --sw-mode --ht-reference <ref> [--shard \| --gg-regions]` |
| Multiple batches | | |
| Step1 | aggregate gVCF files from one batch of samples into a Cohort file and a Census file | `--gvcfs-to-cohort-census true --variant-list` |
| Step2 | aggregate a list of per batch Census files into a global Census file | `--aggregate-censuses true --input-census-list` |
| Step3 | generate a multi-sample VCF for one batch of samples from per batch Cohort file, per batch Census file and global Census file | `--generate-msvcf true --input-cohort-file --input-census-file --input-global-census-file` |
| Single batch | | |
| Step 1 | aggregate gVCF files from one batch of samples into a Cohort file and a Census file and generate multi-sample VCF | `--gvcfs-to-msvcf true --variant-list` |

- Very large cohort analysis is typically not suitable for single server execution due to the large computational load. It requires a multi-node scatter-gather type of analysis, the ability to split processing jobs into reasonable chunks and complete the analysis in reasonable time frames
- The following table lists the analysis types and the recommended platform to use for large cohort analysis

| Compute platform | Type of analysis | Max cohort size | Notes |
|---|---|---|---|

| | | | |
|---|---|---|---|
| DRAGEN server | Non-iterative | 10K hard limit,<br>1k samples/24h | Can distribute compute across several DRAGEN servers using `gg-regions` parameter |
| | Iterative | 1k / batch hard limit, can run several batches in sequence | Can distribute by genomic region using `--shard` command line option |
| Illumina Connected Analytics (ICA) | Non-iterative | 10K | Using CLI |
| | iterative | *>100K batch-wise* | *Not yet available* |
| HPC | Non-iterative | 50-70K | HPC binary and template workflow available on request |
| | Iterative | > 100K, batch-wise | HPC binary and template workflow available on request |

## Ora Compression Updates

- CPU performance has been optimized, and compression ratio is slightly improved
  - A new ORA reference (index v2) has been made available. With the new reference, a 30-35% reduction in CPU usage and 5% improvement in output size is achieved when compressing FASTQ.gz into FASTQ.ora
- DRAGEN BCL can generate FASTQ.ora output
  - A beta version of direct BCL to FASTQ.ora is available to trial.
  - Usage:
    - Add below setting to the sample sheet under the `[BCLConvert_Settings]` section to select the "dragen" Fastq compression format
      - `FastqCompressionFormat,dragen`
    - Command line options for DRAGEN BCL unchanged. Example
      - `dragen --bcl-conversion-only true --bcl-input-directory=/data/Files --sample-sheet /data/Files/SampleSheet_v2.csv --output-directory /staging/result --ora-reference=/staging/lenadata`
- Interleaved FASTQ option.
  - A beta version of interleaved compression can be used. The output is a single FASTQ.ORA file per paired reads (R1 and R2), per sample across lanes. This reduces the total file size by up to 10% with no impact on run time.
  - The feature can be optionally enabled through a sample sheet setting in the `[BCLConvert_Settings]` section:
    - `FastqCompressionFormat,dragen-interleaved`
- Output integrity check
  - DRAGEN ORA has had an internal integrity check to ensure losslessness since it has been available. DRAGEN 3.10 now allows the user to also perform the integrity check again, with printed output.
  - Method:
    - compares checksum of decompressed FASTQ.GZ and decompressed FASTQ.ORA. If values are equal outputs: "ORA integrity check successful", if not outputs: "integrity check failed for file x"
  - Usage:
    - Add `--ora-check-file-integrity=true` when running a compression job

## Other

- **Alt-masked Hash Tables Update**
  - DRAGEN v3.9 introduced alt-masked hash tables. Masking was applied to the internal reference, leading to potential issues where downstream integrated components like CRAM

compression uses the reference that is different from the fasta (it is masked). In v3.10 this is fixed. We have changed the way alt-masking is done during hash table generation, so that the internal reference matches the fasta used in the hash table build. Therefore, any masking operation on the fasta does not impact output files and downstream tools.
    - Customers are advised to rebuild their alt-masked hash tables using DRAGEN v3.10, or download the alt-masked-v2 graph hash tables from the Illumina support site
- **Somatic SNV Calling with UMI**
    - DRAGEN installer now packages SNP error calibration BED files for exome targeted regions, and automatically uses them during somatic variant calling when UMI collapsing is enabled. SNP error calibration files are added for hg19, hg38, hs37d5
- **CNV Caller**
    - A new command line option to allow users to bypass the human reference genome check for self-normalization `--cnv-bypass-contig-check`
    - Improve CNV Hast Table Generation to scale better with number of contigs
    - Update T/N CNV/SNV to use intermediate somatic SNV VCF `--cnv-use-somatic-vc-baf`
    - Made target count variance estimation more robust to wide targets
    - Support T+N+PON CNV WES, when running with CNV with SNV or SV
    - Change command line option for specifying the minimum fraction overlap between target intervals and exclude-bed to exclude from the list `--cnv-exclude-bed-min-overlap`
- **SV Caller**
    - New early exclusion filter. Allow a BED file containing the set of exclusion regions for SV, calling (optionally gzip or bzip compressed) by using `--sv-exclusion-bed`. Cannot be used if any inclusion BED file is passed using `--sv-call-regions-bed` and vice versa
- **RNA**
    - RNA variant caller is forced to use the somatic VC mode, regardless of whether the normal or tumor inputs options are used
    - Gene Fusion caller improves error checking on file formats for all input types
- **BCL**
    - Improve accuracy metrics, by increasing the number of decimal places of the "Percent Totals" aggregate
    - Allow TopUnknownBarcodes to be set to any number, including 'all', through new command line option `--num-unknown-barcodes-reported`
    - New option to validate RunInfo.xml & SampleSheet files without running the conversion `--bcl-validate-sample-sheet-only`
    - Improved robustness to missing and corrupt input files
    - Bug fixes
- **UMI**
    - A new batch mode option `--enable-positional-collapsing`, which initializes a set of parameters needed when doing positional correction, to improve ease-of-use
- **Gvcf Genotyper**
    - Allow writing of only forced sites (from `--gg-sites-list`) but no other variants, by enabling option `--gg-only-forced-sites`
- **Single Cell**
    - Option to select filtering on either UMIs or total reads counts distribution ("umi" or "read") by using `--single-cell-threshold-filterby` (default umi)
- **UMI-Aware Methylation**
    - Enable UMI collapsing + single-pass methylation to run together
- **QC Contamination Metrics**
    - The minimum required coverage for a pileup to be used in contamination is made configurable via the command line, by using `--qc-contam-min-cov` (default 10)
        - The default value for somatic contamination remains unchanged at 10X. The default value for germline contamination increases from 5x to 10X. Some small samples may now report "`NA`" rather than a best estimate.
    - New options for pileup filtering, and a new pileup output file

    - `--qc-contam-min-basequal` Minimum base quality for a read to be used in contamination detection (default=10)
    - `--qc-contam-min-mapq` Minimum mapping quality for a read to be used in contamination detection (default=10)
    - `--qc-contam-min-valid-read-ratio` Min ratio of reads in pileup that must be clean and match the population allele to be a valid pileup. (default=0.95)
- **QC Coverage Metrics**
    - Update the coverage metrics output. The coverage calculations reported in MAP/ALIGN metrics and COVERAGE metrics were based on different formulae, leading to confusion. This change deprecates the "`Average sequenced coverage`" from the MAP/ALIGN metrics. Instead, COVERAGE metrics are now output to the stdout also.
    - The discrepancy showed up particularly when running with Graph Genome, and the calculation of `Average sequenced coverage` was technically incorrect
    - Users are recommended to now use the "`Average alignment coverage`", which is output in the COVERAGE metrics.
        - For background, these are the differences in the calculations
        - `Average sequenced coverage = [total number of ACTG bases in input reads] / [number of bases in the reference genome (excludes liftover contigs and Ns)]`
        - `Average alignment coverage = [[number of bases input - duplicate bases - clipped bases] AND [MAPQ > 0] AND [BQ >= 0]] / [sum of bases in the reference contigs where alignments were made]`
- **AWS AMI Cloud**
    - New option to specify the geographical region of AWS S3 buckets: `--aws-s3-region` to allow streaming from buckets in different location as the instance. Argument is the string, e.g. "`us-east-1`"
    - AMIs are available for new regions: Tokyo, Seoul, Singapore and Canada
- **Microsoft Azure Cloud**
    - Authentication has been improved. Credential management is now controlled via environment variables and used in the azure SDK, allowing support for Azure SAS streaming, and BLOB identity-based credential management. Managed identity authentication is supported.
        - The need for and use of `~/.azure-credentials` file has been deprecated. Two new environment variables are required: `AZ_ACCOUNT_NAME` and `AZ_ACCESS_KEY`. Environment variables remove the need for file parsing logic. When using an access key, it improves security since the key will live in memory only
        - Example usage: `$ AZ_ACCOUNT_NAME=dragen ./bin/dragen --output-directory=https://dragen.blob.core.windows.net/container/ --intermediate-results-dir=/staging/dir [..]`
    - Azure upload / output streaming direct to BLOB storage is now supported
- **Cloud Logs**
    - DRAGEN logs are now automatically saved to the output-directory when running in the cloud, instead of on the host node. This allows customer to have logs available when contacting support, to ensure faster time to answer.

## Interface Changes

- When Machine Learning (ML) is enabled in the germline small variant caller, QUAL, GT, GQ fields are updated in the VCF. NOTE: The values will be different between ML enabled or disabled. Hard-filtering is also applied at QUAL 3 when ML is enabled
- Users are advised to rebuild alt-masked Hash Tables with v3.10 and/or download the updated alt-masked Graph Hash Tables

## Issues Resolved

Issues found on DRAGEN™ v3.9 that are fixed in v3.10.4

| Component/s | Defect ID | Issue Description |
|---|---|---|
| **SNV VC** | DRAGEN-15706 | Integer overflow when using vc-dragstr-pcr-params on very large input files |
| **SNV VC, UMI** | DRAGEN-13477 | When MAP/A is run in UMI mode, VCF is also run in UMI mode |
| **HW GRAPH** | DRAGEN-14087 | Abort during custom reference genome build |
| **Somatic** | DRAGEN-14676 | Somatic forced calls with SOM tag applied are being filtered |
| **Somatic** | DRAGEN-15023 | BSSH Enrichment App germline variant caller with MNV failing in post-dragen processing |
| **Somatic** | DRAGEN-13752 | Forced GT somatic run has high SQ FGT calls |
| **Somatic** | DRAGEN-13366 | Incorrect germline variants in prefilter VCF with combined VCF/GVCF output |
| **Somatic** | DRAGEN-11904 | Forced GT somatic run output a baf file that contains positions not found in final vcf |
| **Somatic Force GT** | DRAGEN-14114 | DRAGEN Tumor-Normal pipeline: Force Genotyping issues |
| **Somatic GVCF** | DRAGEN-15225 | Calls missing from GVCF that are in VCF |
| **CNV VC** | DRAGEN-16211 | Different count for WES bam-input vs tumor-bam-input |
| **CNV VC** | DRAGEN-15284 | CNV germline aware mode not compatible with BAF from VC |
| **CNV VC** | DRAGEN-14743 | Sample excluded from sex genotyper metrics file |
| **CNV VC** | DRAGEN-14575 | Option cnv-exclude-min-overlap has no effect |
| **SV** | DRAGEN-13990 | Running the DRAGEN-SV forced genotyping failed for a SV at the end of chrM |
| **Gene Fusion** | DRAGEN-13204 | DRAGEN RNA job killed in the cloud (out of memory) |
| **Gene Fusion** | DRAGEN-14755 | RNA gene fusion output md5sum difference between U200 P2 and P4 servers due to precision |
| **Gene Fusion** | DRAGEN-12779 | PML--RARA not called by DRAGEN but called by RNA Amplicon |
| **BCL** | DRAGEN-14606 | BCL stats that are ratios against total samples should have more accuracy |

| Component/s | Defect ID | Issue Description |
|---|---|---|
| **BCL** | DRAGEN-14502 | Top Unknown barcodes summary metrics has incorrect values in % of unknown barcodes |
| **Paralog Caller** | DRAGEN-15617 | Assertion failed in RawBuffer2DbamBlockTransformer.cpp: expected new fragment after suppressed record |
| **Compression** | DRAGEN-14966 | CRAM generated with alt-masking HT may not be decodable by 3rd party |
| **Compression** | DRAGEN-14824 | ORA compression from unzipped fastq not working |
| **Compression** | DRAGEN-14390 | DRAGEN hits segmentation fault on CRAM input mismatched with reference |
| **Combine GVCF** | DRAGEN-12776 | DRAGEN combine GT: -- false -- ERROR: There are multiple non-ForceGT lines in contig MT at position 302 |
| **scRNA** | DRAGEN-15412 | Single cell RNA doesn't support variable-length cell barcode blocks |
| **Dedup/UMI** | DRAGEN-15353 | Very large log files created during UMI processing |
| **Hash Table** | DRAGEN-14843 | Reference masking does not work as expected |
| **GVCF Genotyper** | DRAGEN-15038 | gVCF Genotyper: --gg-drop-genotypes=true fails on chrM |
| **GVCF Genotyper** | DRAGEN-14884 | Regression on runtime for 10k small panel dataset |
| **GVCF Genotyper** | DRAGEN-14781 | GvcfGenotyper missing LowDepth Filter meta data line in header |
| **GVCF Genotyper** | DRAGEN-13303 | gVCF Genotyper Account for missing values in FORMAT/LOD (3.8 gVCF) |
| **Oracle8, Licensing** | DRAGEN-16026 | License Installation Failure on el8 from Empty Curl Response |
| **Infrastructure** | DRAGEN-15430 | Auto-detect for primary reference genome |
| **HWAL/Infra** | DRAGEN-13242 | CTRL-C is not blocked during PR |
| **Nirvana** | DRAGEN-15282 | Nirvana Aborts with Error: ERROR: Unable to find the block GZip compression library on DRAGEN Ph3 Server |

## Known Issues

Known issues of the DRAGEN™ v3.10.4 release

| Component/s | Defect ID | Issue Description | Remedy / Workaround |
|---|---|---|---|
| **BCL** | DRAGEN-16555 | Minimum Adapter Overlap setting not working | Whether it's set to a valid or invalid value, an error message is displayed. No workaround |
| **Infrastructure** | DRAGEN-16498 | AWS f1.4x l LICENSE_MSG\| Challenge get token error: Get instance ID failed (Unable to retrieve AWS identity signature) | Timeout while retrieving AWS instance ID has been observed. The rate of occurrence has been too low to measure. This leads to failure in the licensing and dragen job exits, run fails. Re-run would pass |
| **DNA Alignment** | DRAGEN-16468 | Bam generated from file conversion CRAM -> BAM with hg19_alt_masked_v2 reference has invalid header | Reproducible issue that has been shown to have existed since v3.7 or earlier. When converting from CRAM to BAM using dragen "file-conversion" method, the BAM has an invalid header due to a bug in the CRAM reader. No workaround. Re-header the file |
| **DNA Alignment** | DRAGEN-16467 | Germline workflow is slower with graph hash table | Dragen 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. No workaround |
| **Dedup/UMI** | DRAGEN-16412 | Probabilistic UMI output is different from run to run | There is a run-run variation in the UMI probabilistic model. Non-prob model (non-random UMI) does not have run-to-run variation. The variation leads to ~2 reads being missing from output. This impact shall be a very small fraction. |
| **BCL** | DRAGEN-16406 | DRAGEN BCL does not abort when AdapterRead2 is specified for a sequencing run with single genomic read | Failure of a strict negative test to validate that the settings from sample sheet and input data matches. BCL does not detect this user error. Intended usage works as expected. The additional setting `AdapterRead2` is just ignored |
| **BCL** | DRAGEN-16405 | DRAGEN BCL does not abort when global Adapter settings are used without AdapterRead1/2 | Validation checks for Adapter Behavior strict negative tests (user error) are skipped. BCL does not detect this user error. Intended usage works as expected |

| Component/s | Defect ID | Issue Description | Remedy / Workaround |
|---|---|---|---|
| **BCL** | DRAGEN-16404 | DRAGEN BCL requires AdapterBehavior for all samples, even when AdapterRead1/2 missing | An adapter behavior check is failing for a strict negative test case in the new "per-sample settings" feature. A variant of the negative test may also lead to segfault. This a corner case. Intended usage works as expected. |
| **DupMarking** | DRAGEN-16399 | Assertion `pos < m_num_bits' failed, in Dupmark:: DupmarkTable:: getDuplicates() | Crash in duplicate marking when there are more than 4G read pairs, which can happen when reads of multiple replicates are combined into one read group through manual BAM file editing. The system has a physical limit. Not a regression from prior releases |
| **Cloud / Azure** | DRAGEN-16335 | popen exception on azure cloud suites | A very long running workflow such as TSO500 ctDNA crashes on Azure due to eventual failure in popen() calls. The issue is reproducible. No workaround |
| **Somatic** | DRAGEN-16319 | Elevated SNP and INDEL FP on ICGC datasets | The impact of the issue is an elevated number of FPs for ICGC datasets in 3.10 compared to 3.9: a 5-6% increase in the SNP FPs and a 25%-30% increase in the INDEL FPs. |
| **DNA Alignment** | DRAGEN-16308 | read trimmer adapter trimming sigabort during RecomputeTags:: computeTags | Reproducible when running different read trimmers back-back. Workaround to run dragen_reset |
| **Installer** | DRAGEN-16276 | Unsafe to call dragen installer during dragen run | Calling the dragen installer while a run is in progress breaks the run. No workaround |
| **Amplicon Gene Fusion** | DRAGEN-16254 | Excessive RNA Amplicon runtime on large samples | RNA Amplicon run time is very long when the coverage is significantly higher than expected for typical Amplicon samples |
| **BCL** | DRAGEN-16245 | DRAGEN BCL assumes Minimum Trimmed Read Length = 35 when using MaskShortReads in SampleSheet | A setting of the value "Minimum Trimmed Read Length" does not take effect. It is a regression from 3.9. No workaround |
| **SNV VC Somatic** | DRAGEN-16149 | Germline MNV - phased calls with same PS and GT and within distance | Some phased calls are not getting combined into MNVs to MNV output. |

| Component/s | Defect ID | Issue Description | Remedy / Workaround |
|---|---|---|---|
| | | threshold are not getting combined into MNVs | Full support for germline MNV is planned for a subsequent release |
| **Metrics scRNA** | DRAGEN-15950 | A run-run variation in scRNA output | Some datasets have run-to-run variations in the mapping metrics Q30 metrics field. The issue affects only the metrics output and not the caller output |
| **BCL** | DRAGEN-15944 | DRAGEN BCL logs insufficient warning when corrupt files supplied | In the rare event of a corrupt aggregated bcl. bgzf input file, the customer will correctly receive an error message of the lane and the cycle that is corrupted, but not the specific file name. |
| **Methyl-Seq** | DRAGEN-15796 | md5sum discordance b/w cloud and local runs | Impacts multi-pass mode and specific dataset. Single pass mode has been the recommended mode and does not have the issue. Multi-pass will be deprecated in future. |
| **DNA Alignment EH** | DRAGEN-15151 | Large run to run variation of mapper run time for EH | Up to 20% run time variation seen for mapper phase |
| **HW GRAPH RNA VC** | DRAGEN-13717 | RNA VC hits ERROR: Invalid node flags | Issue is a HW graph error and rare (happens once every 6-9 months in routine VC testing). The assertion check / trap will remain in place so that invalid results will not be produced for end user. If seen in field, recommendation is to re-run sample as it is expected to pass. |
| **Compression** | DRAGEN-10783 | BAM input to DNA mapper: Deflate engine error: 0x9080 on stream 1 | Extremely low repeatability. A re-run will pass |

## SW Installation Procedure

- Download the desired installer from the Illumina support website and unzip the package
- The archive integrity can be checked using: `./<DRAGEN 3.10.4 .run file> --check`
- Install the appropriate release based on your Linux OS with the command: `sudo sh <DRAGEN 3.10.4 .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