DRAGEN v3.7.5 Software Release Notes

October 21, 2020

Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform v3.7.5.

Changes are relative to DRAGEN™ v3.6.3. If you are upgrading from a version prior to DRAGEN™ v3.6.3, please review the release notes for a list of features and bug fixes introduced in subsequent versions.

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

The 3.7.5 software package includes:
- DRAGEN™ SW Intel Centos 6 - dragen-3.7.5-4.el6.x86_64.run
- DRAGEN™ SW Intel Centos 7 - dragen-3.7.5-4.el7.x86_64.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)

Support for IBM PPC has been deprecated and not available for DRAGEN™ v3.7 and later.

Contents

Overview

Below is a summary of the changes included in v3.7.5. DRAGEN™ v3.7 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 Callers

Single Cell RNA

- The DRAGEN™ single-cell RNA pipeline is a high-performance secondary analysis pipeline for single-cell gene expression data. The pipeline supports input from multiple common library prep systems, and the output is compatible with many downstream analysis tools.
- The pipeline supports the following operations:
    - Demultiplexing of separate sequencer base calls into individual samples
    - Use of cell barcodes to group reads to recover single-cell data
    - Sequence alignment of reads: Reads are mapped and aligned to a reference genome
    - Quantification of gene expression: UMIs for each cell and gene are counted and duplicates are removed after error correction
    - Dataset filtering and QC: Noncellular barcodes and low-quality cells are filtered, and QC metrics are calculated
    - Output of cell x gene expression matrix
- The output is compatible with open source single-cell analysis tools such as Scanpy / AnnDta, Seurat, etc. Downstream analysis such as UMAP projections, SAM / Leiden clustering, Marger gene discovery, can be run on the output. The output is highly consistent with established tools, yet 3x faster:
    - Correlation of per-cell gene expression results with STARsolo: r > 0.985
    - 1.4 billion reads, 8018 cells: 38 min, compared to 125 min for STARsolo
- DRAGEN™ Cloud App includes a graphical interactive HTML report with QC metrics, cell clustering and marker genes.
- Please refer to the DRAGEN™ User Guide for complete details and options

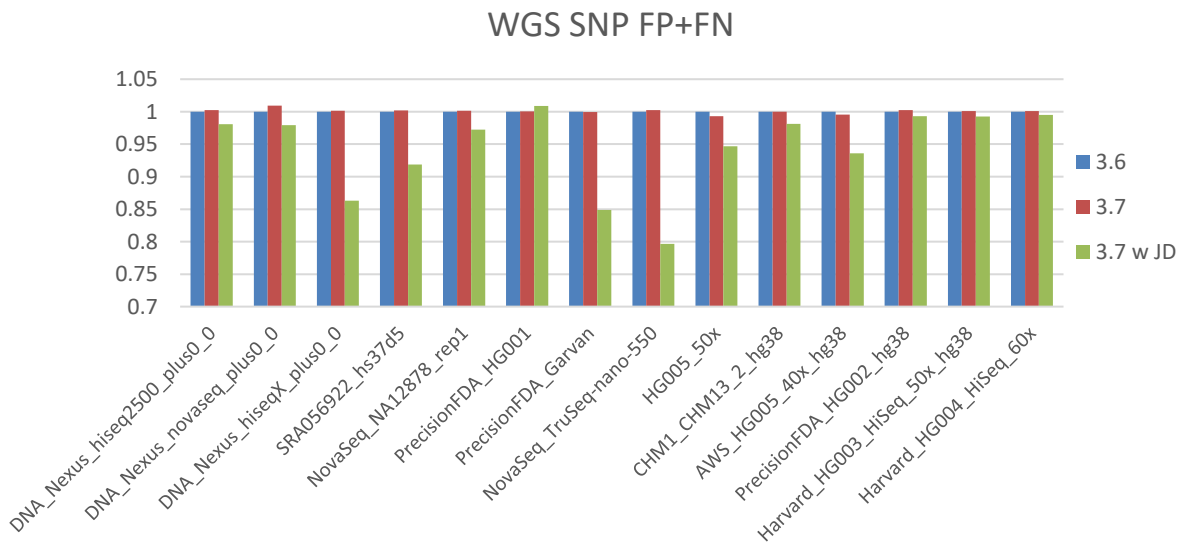
CYP2D6

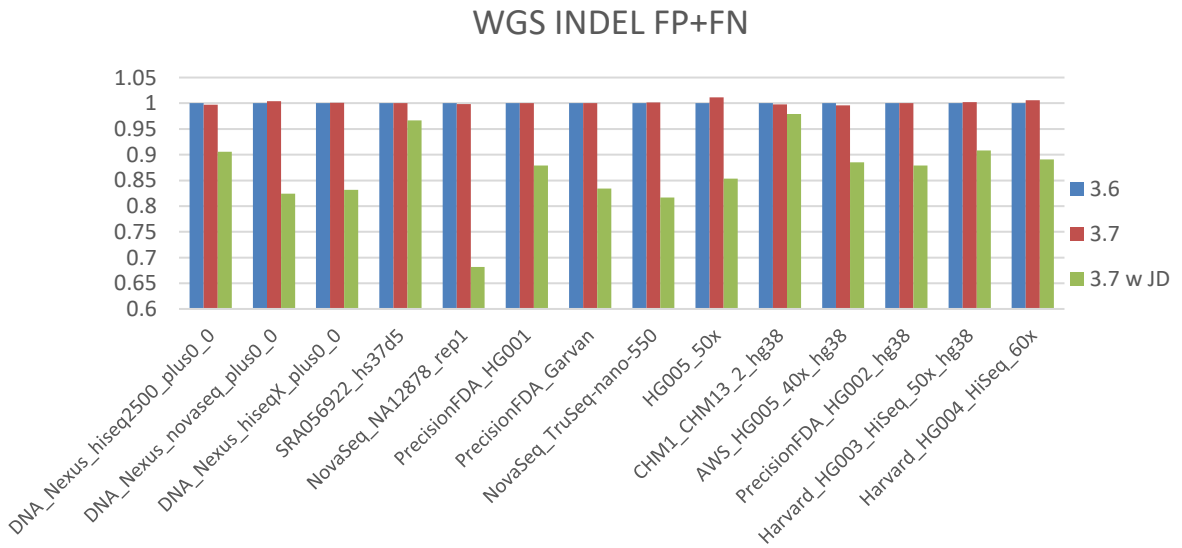
- CYP2D6 genotyping is integrated with DRAGEN™ for use with germline WGS data, and supports hg38, hg19 and GRCh37 references. The implementation is based on the open-source tool Cyrius and match the concordance.
- The CYP2D6 caller outputs star allele diplotype for each sample e.g. '`*1/*4+*68`'
    - Over 120 CYP2D6 star alleles are supported, allowing DRAGEN™ to call a definitive genotype in nearly all samples
- The caller can be enabled to run in parallel with other components in a germline WGS analysis. It can be enabled whether mapping from FASTQs, re-mapping from BAM/CRAM, or reading from pre-aligned BAM/CRAM input, by setting option `` `--enable-cyp2d6=true` ``.
- Please refer to the DRAGEN™ User Guide for complete details

Improvements and Feature Additions

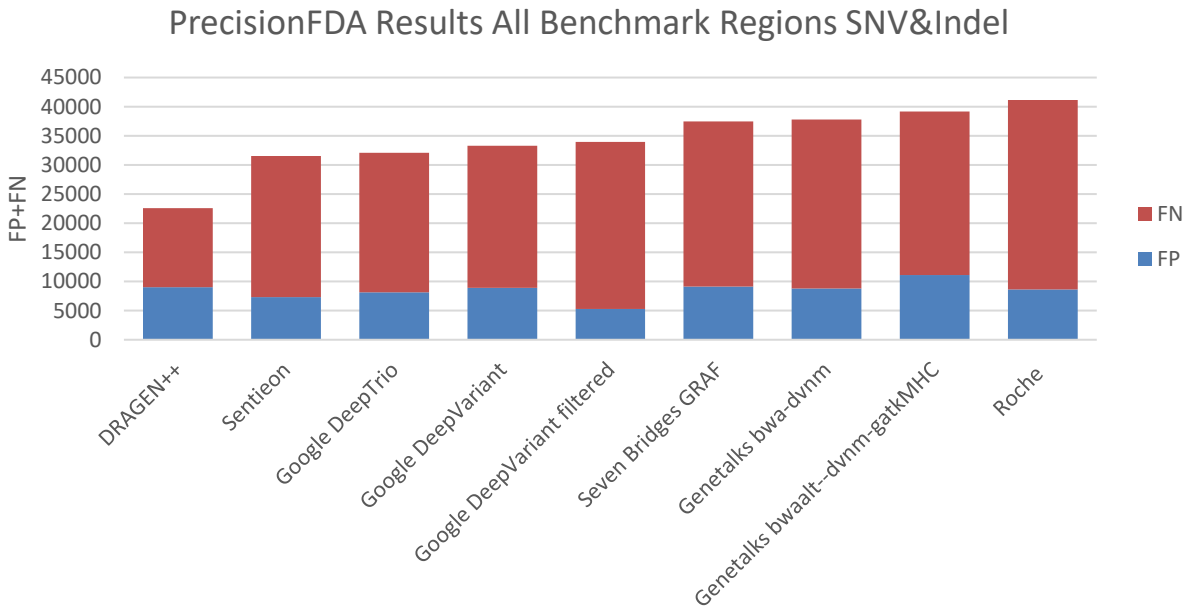
Small Variant Calling Improvements

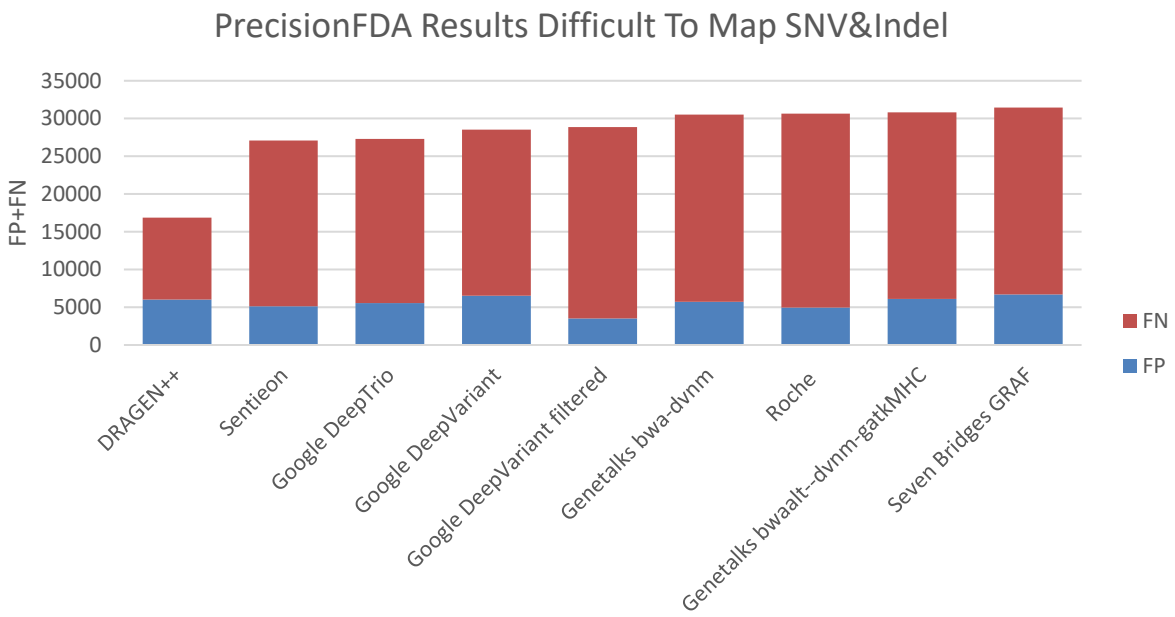
- **Germline**
  - **New Joint Detection of Overlapping Variants**
    - A significant accuracy boost is achieved when variants at multiple loci in a single active region are detected jointly. Joint Detection alters the variant caller algorithm in the area of localized haplotype assembly and genotyping under the appropriate conditions.
    - Accuracy of DRAGEN™ v3.7 with JD provides significant gains in INDEL, and some gains in SNP as well, especially for WGS
    - Enable Joint Detection by setting option `` `--vc-enable-joint-detection=true` ``



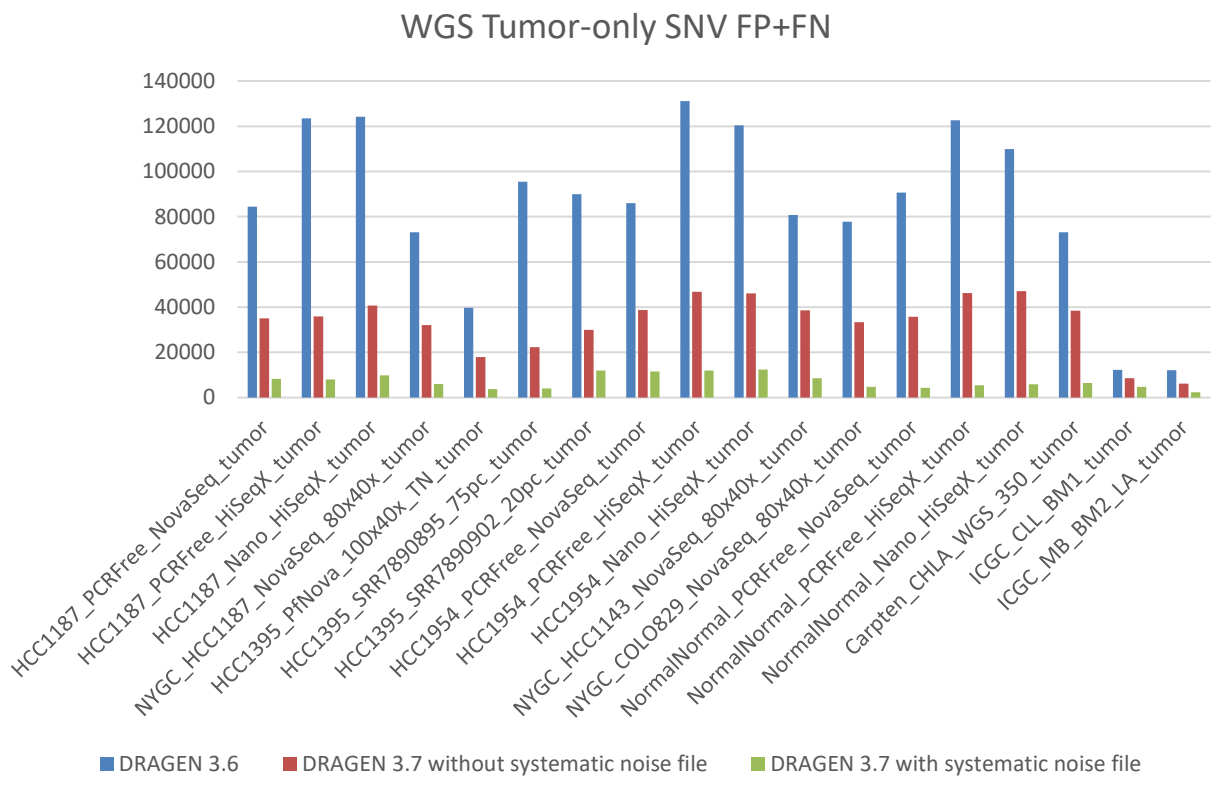


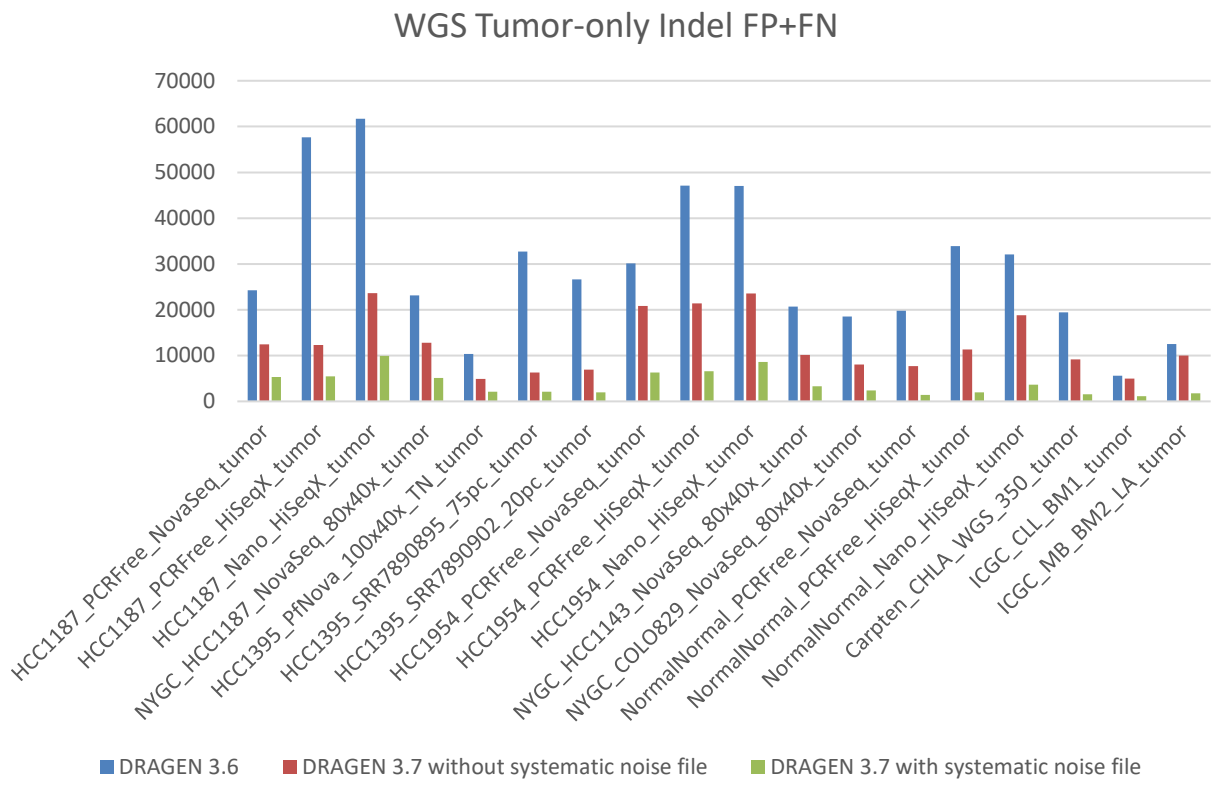
- **New Graph-Capable Mapping**
  - DRAGEN™ v3.7 includes an Alpha release of a graph-capable mapper
  - The graph-capable mapper in DRAGEN™ is a feature that is a key enabler in improving variant calling accuracy in segmental duplications and other regions previously difficult to map with Illumina reads. DRAGEN's graph-based method uses alt-aware mapping for population haplotypes stitched into the reference with known alignments, effectively establishing alternate graph paths that reads could seed-map and align to. This reduces mapping ambiguity because reads containing population variants are attracted to the specific regions where those variants are observed.
    - Resolves regions of the genome that are inaccessible to short reads due to repeat sequences
    - Increase the coverage of clinically important genes
    - Fully compatible with existing hg38 reference and existing BA<.VCF format
    - Enables SNV/SV/CNV variant calling in difficult-to-map regions
  - While building the DRAGEN™ Hash Table, the FASTA reference is augmented with several hundred thousand short alternate contigs derived from population haplotypes of phased variants.
  - Currently, this Alpha release supports graph-capable mapping on the hg38 reference. Population haplotypes of phased variants, and population SNPs for hg38 are provided and available under /opt/edico/liftover/ after installation. Support for hg19/hs37d5 references will be added to future releases.



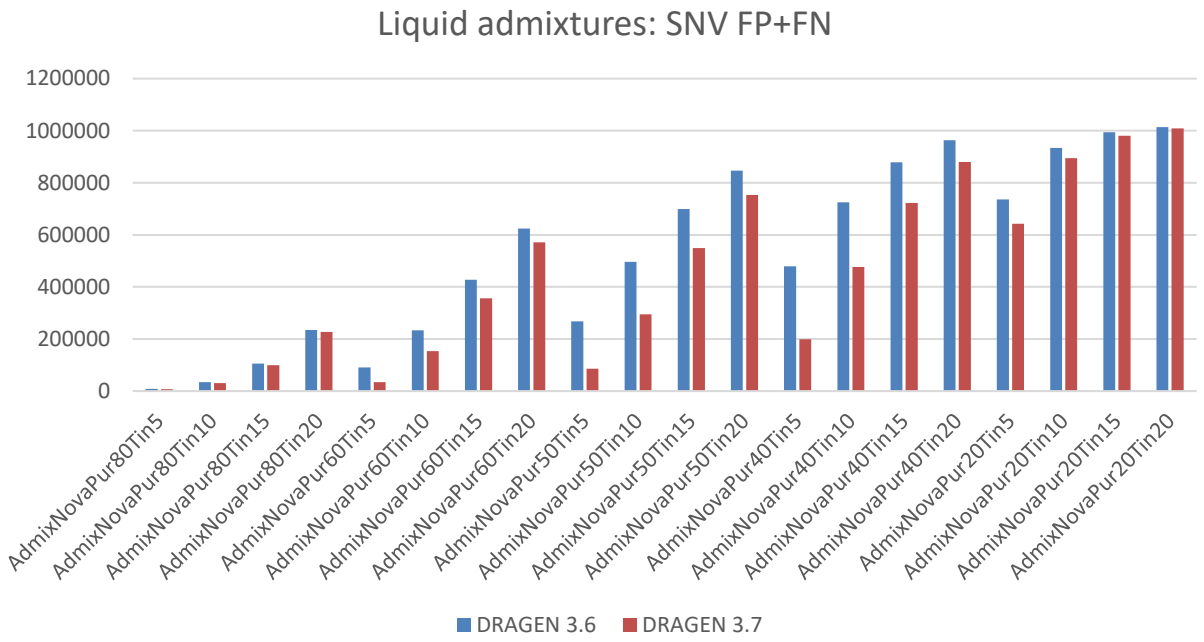

PrecisionFDA Results Difficult To Map SNV&Indel
FP+FN
35000
30000
25000
20000
15000
10000
5000
0
DRAGEN++
Sentieon
Google DeepTrio
Google DeepVariant
Google DeepVariant filtered
Genetalks bwa-dvnm
Roche
Genetalks bwaalt--dvnm-gatkMHC
Seven Bridges GRAF
FN
FP

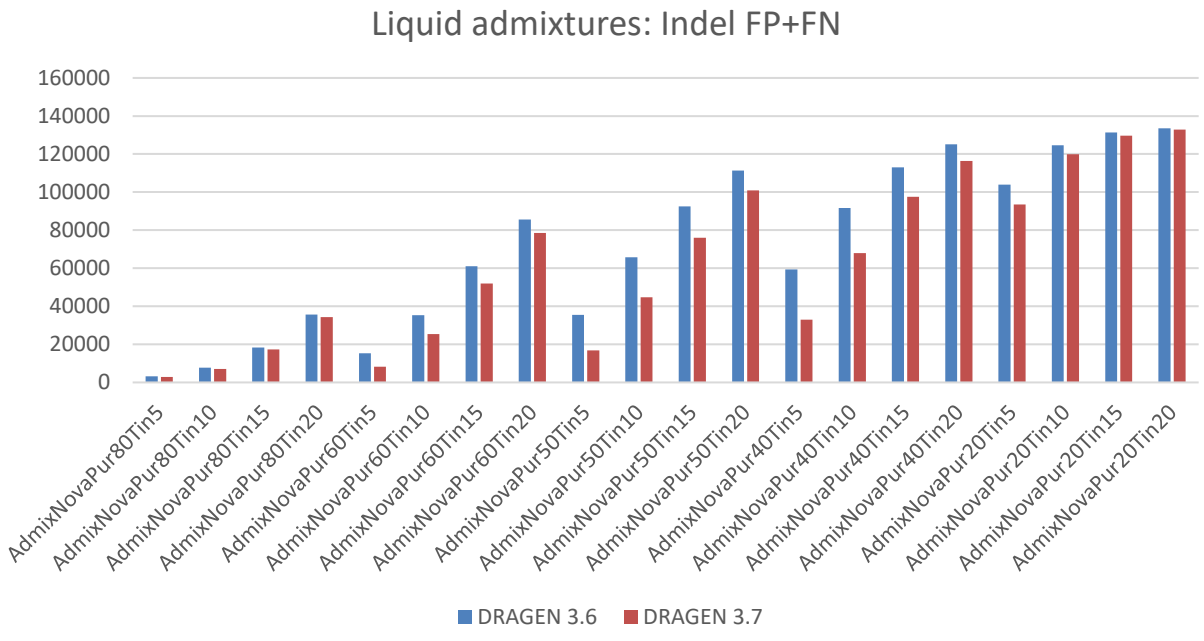
- **Somatic**
  - **New Support for Tumor-Only pipelines**
    - Significant accuracy improvement for Tumor-Only analysis, with 5x reduction in FPs compared to DRAGEN™ v3.6, and sensitivity > 95%
    - The genotyping approach for Tumor-Only has been updated to be the same as DRAGEN™ T/N calling, leading to improved precision. The pipeline uses Systematic Noise Filtering and % Non-Informative Reads Filtering to improve false positives
    - Support for WGS, WES and Panels
  - **New Systematic Noise Filtering feature**
    - Systematic Noise Filtering BED files are provided to the DRAGEN™ variant caller for the purpose of filtering out sequencing / systematic noise
    - Applicable to all pipelines as an optional input to improve precision. Particularly important for Tumor-Only variant calling (either with or without UMIs) where a matched normal sample is not available for detection of systematic noise artifacts, but is also helpful for Tumor-Normal calling.
    - Pre-built systematic noise files are available for download from the Illumina DRAGEN™ support website and a public AWS s3 bucket
    - Please refer to the DRAGEN™ User Guide for complete details on usage and generating your own Systematic Noise files



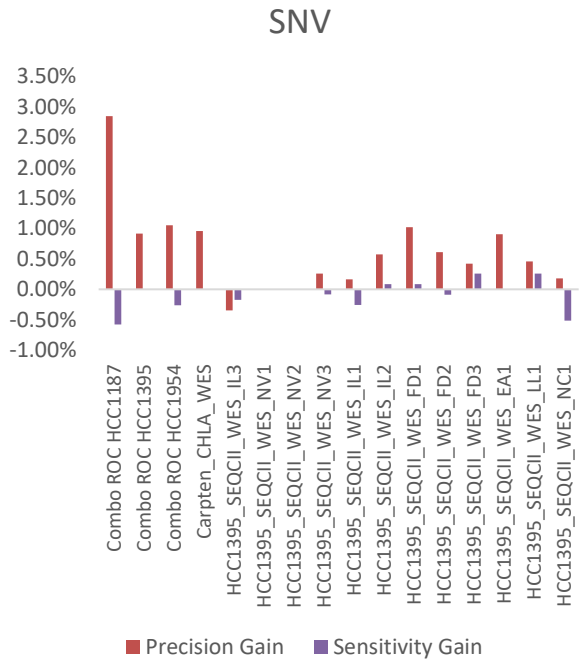

WGS Tumor-only Indel FP+FN
70000
60000
50000
40000
30000
20000
10000
0
HCC1187_PCRFree_NovaSeq_tumor
HCC1187_PCRFree_HiSeqX_tumor
HCC1187_Nano_HiSeqX_tumor
HCC1187_NovaSeq_80x40x_tumor
NYGC_HCC1187_NovaSeq_80x40x_TN_tumor
HCC1395_PfNova_100x40x_75pc_tumor
HCC1395_SRR7890895_20pc_tumor
HCC1395_SRR7890902_NovaSeq_tumor
HCC1954_PCRFree_HiSeqX_tumor
HCC1954_PCRFree_NovaSeq_tumor
HCC1954_Nano_HiSeqX_tumor
NYGC_HCC1143_NovaSeq_80x40x_tumor
NYGC_COLO829_NovaSeq_80x40x_tumor
NormalNormal_PCRFree_NovaSeq_tumor
NormalNormal_PCRFree_HiSeqX_tumor
NormalNormal_Nano_HiSeqX_tumor
Carpten_CHLA_WGS_350_tumor
ICGC_CLL_BM1_tumor
ICGC_MB_BM2_LA_tumor
DRAGEN 3.6
DRAGEN 3.7 without systematic noise file
DRAGEN 3.7 with systematic noise file

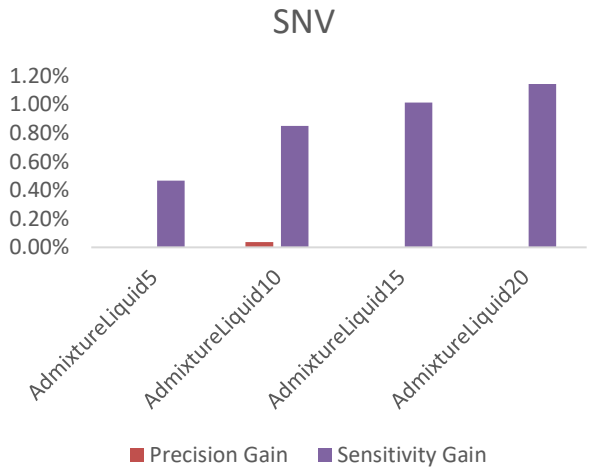
- **WGS T/N Liquid Tumor accuracy improvements**
  - Improved FP+FN for a range of tumor purities and tumor-in-normal (TIN) contamination levels

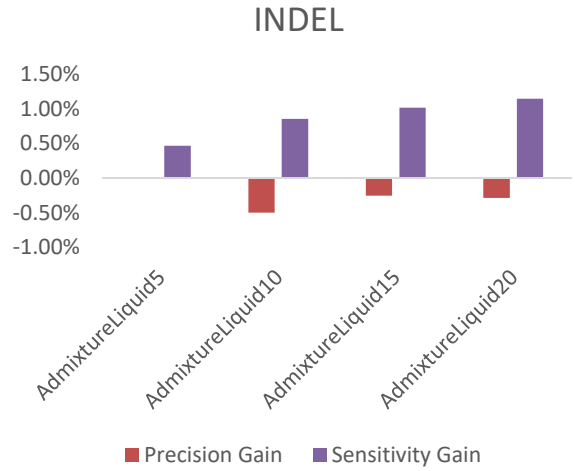




- **WES T/N accuracy improvements**
  - v3.7 has significant precision improvements on **solid tumor** compared to v3.6
  - v3.7 has significant sensitivity improvements on **liquid tumor** compared to v3.6







- **New Support for Liquid Biopsy Pipelines**
  - Small VC for UMI libraries
- Internal benchmarking to show that DRAGEN™ somatic caller meets clinical reproduction requirements

Somatic WGS Concurrent Multi-caller Support

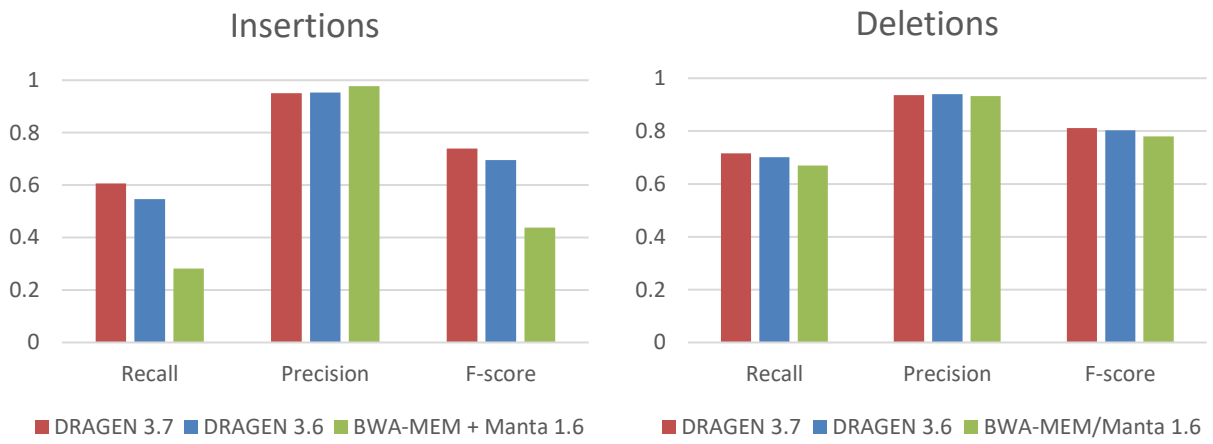
- Additional somatic concurrency support has been added to simplify somatic workflows and reduce overall analysis time. Concurrent multi-caller analysis now allows all combinations of SNV, SV, CNV callers to be enabled at the same time with one execution. All metrics and outputs will be available from one run. Efficient multi-threaded parallel execution of the callers reduces I/O and eliminates duplication in processing, leading to up to 20% lower execution times and cost for multi-caller somatic workflows
- Multiple input formats are also supported: FASTQ with mapping/aligning, or BAM/CRAM input without aligning. Multi-caller BAM/CRAM with re-mapping supported for Tumor only, but not yet for T/N.
- New supported concurrency for DRAGEN™ v3.7 in one execution
    - SNV+CNV+SV from BAM/CRAM or FASTQ
    - SNV+CNV from BAM/CRAM or FASTQ
    - CNV+SV from FASTQ

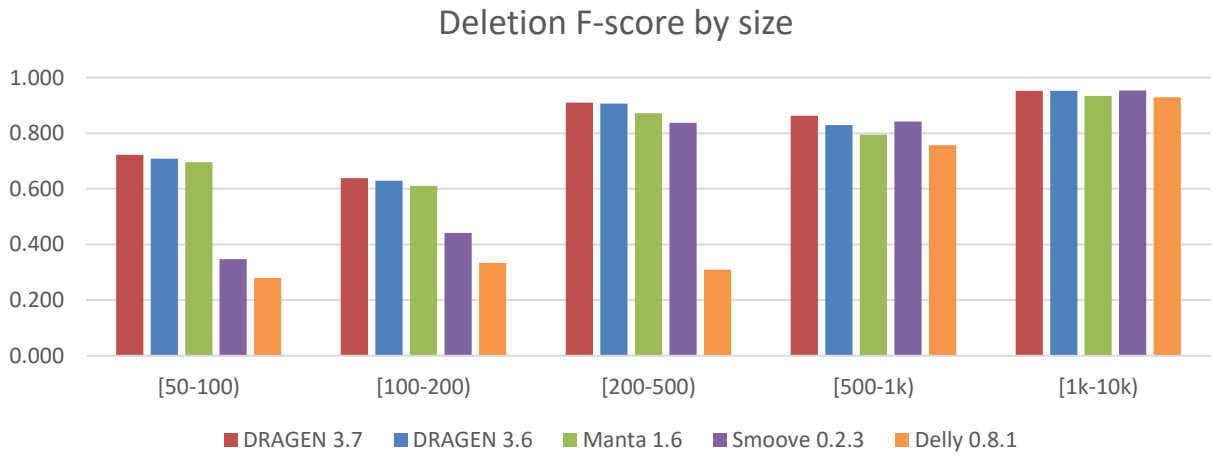
CNV Updates

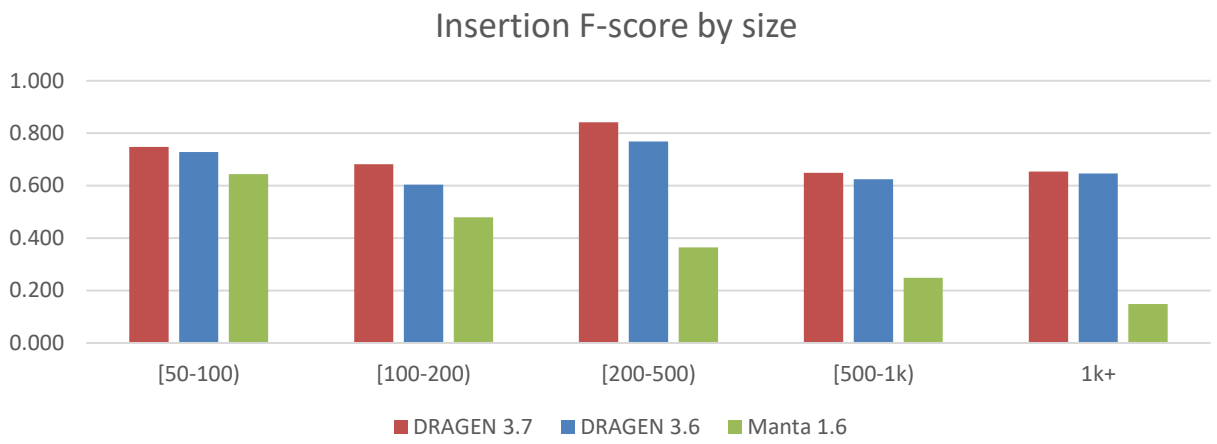
- Somatic WGS T/N Purity Estimation
    - Somatic purity/ploidy model now leverages somatic VAFs when CNV+SNV are run together. This improves the resolution of liquid tumors, and high ploidy (WGD) vs heterogeneity
- WES CNV Calling
    - Improved segmentation for WES CNV, leading to better accuracy for both large and small events. Threshold fine tuning to achieve fewer reported total events
- Germline WGS Coverage Uniformity Metric
    - Identifies low quality samples with possible sequencing issues. Helps to save time during analysis.
- CNV Blacklist BED
    - Blacklist BED support now filters out intervals from analysis, at the target intervals stage. Useful for regions of the genome which are known to be problematic due to library prep, sequencing or mapping issues.
- IGV Session XML
    - Auto generated igv_session.xml file for easier visualization and analysis of CNV tracks.
    - Intermediate BED-like files are now gzipped compressed to minimize output file sizes.

SV caller

- New forced genotyping capability for SV insertions and deletions
- Faster execution from CRAM: SV calling from CRAM runs in 66% of the time required for DRAGEN™ v3.6
- Improved accuracy compared with DRAGEN™ v3.6
    - Improved recall for both insertions and deletions compared to DRAGEN™ v3.6
    - DRAGEN™ v3.7 now achieve greater than double the insertion recall of Manta, and superior recall compared to other tools while retaining high precision
    - Deletion and Insertion accuracy is improved across all size groups





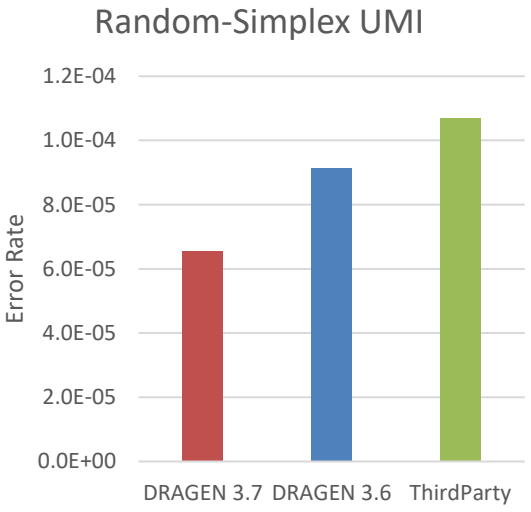


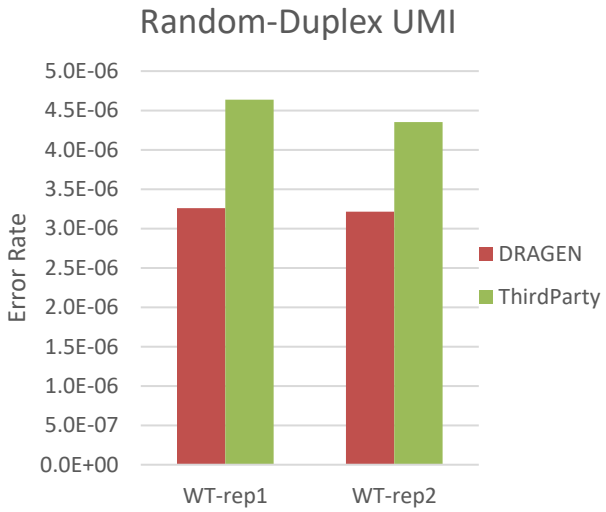
Comparison to NIST GIAB v0.6 tier1 SV truth set: https://jimb.stanford.edu/giab-resources. PrecisionFDA Truth Challenge HG002 sequencing data. Assessment using witty.er: https://github.com/Illumina/witty.er

- High Mobile Element Insertion (MEI) Accuracy
  - Higher precision than dedicated MEI methods, and best-in-class recall for MEIs

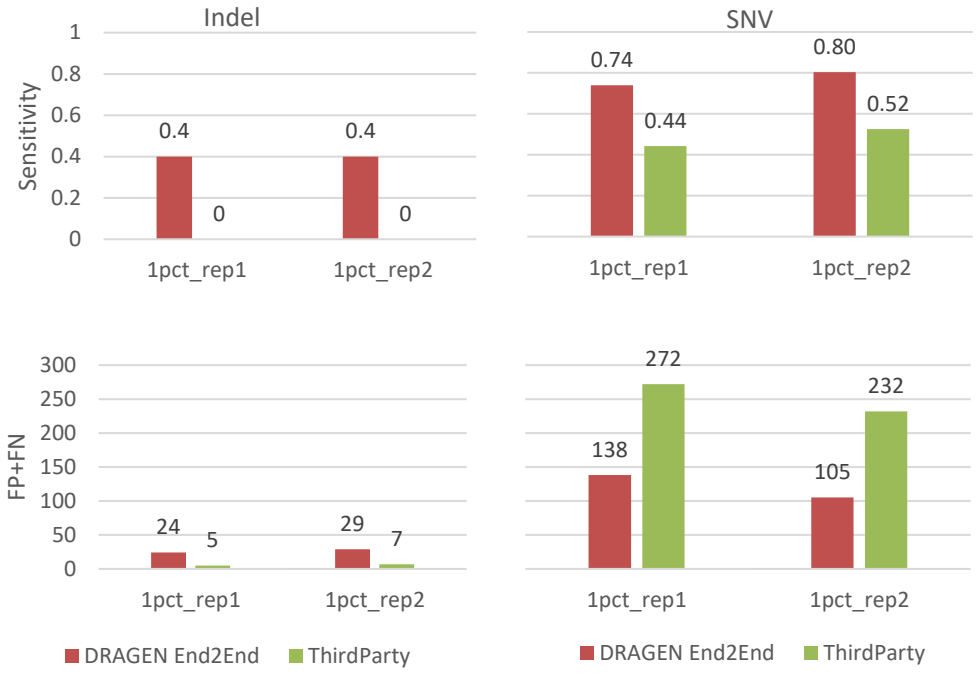
UMI

- Support for all UMI types: random-simplex (Agilent HS1), random-duplex (IDT xGen Duplex Seq), non-random duplex (TruSight UMI)
- New probabilistic UMI collapsing model with improved accuracy and best in class error rates
  - Ability to solve UMI jumping artifact
  - Enable duplex collapsing with simplex UMI
  - Improved error reduction on random UMI
  - Enable customized nonrandom UMI with valid sequence as input
  - Improved robustness on high coverage datasets
  - Enable positional collapsing on non-UMI data
  - Equivalent bam-level support for SNV and Indel
- DRAGEN™ Variant Caller is UMI aware for targeted panels
- Performance
  - Run time: 15-25X faster than third-party tools
  - Random-Duplex UMI: ~30% lower error rate than third-party tools
  - Random-Dual UMI: >30% higher sensitivity than third-party tools
  - Random-Single UMI: higher sensitivity and precision than third-party tools

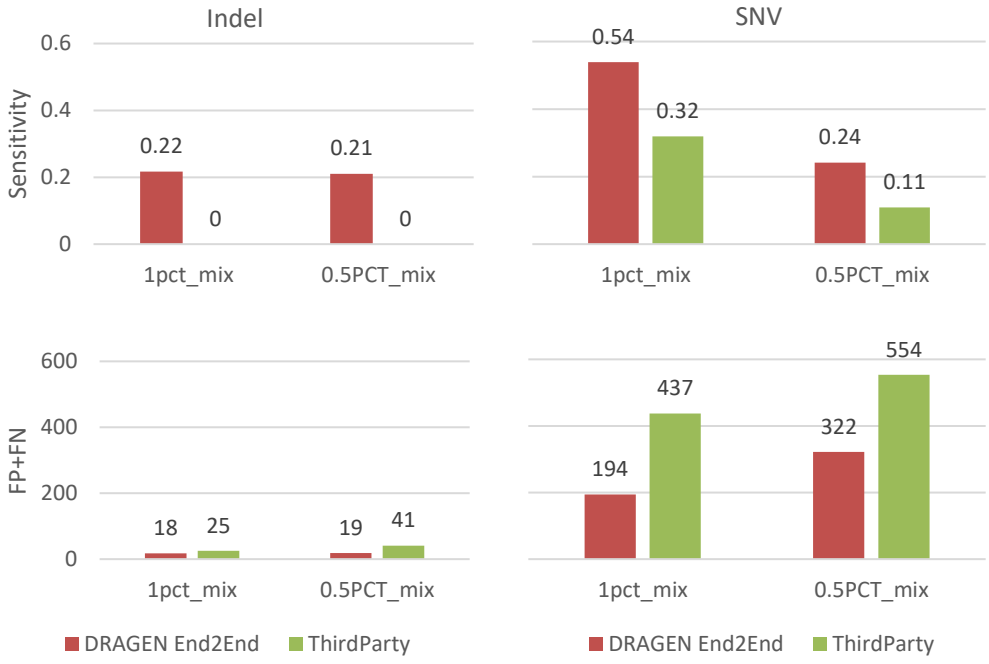




Random Dual UMI Accuracy



Random Single UMI Accuracy



PopGen Workflow

- Gvcf Genotyper
  - Accuracy improvements at multi-allelic sites leading to gains in INDEL recall
  - Support for streaming from S3 bucket
  - Bugfixes to multi-sample gGVCF input parsing that lead to errors for sites with mixed ploidy
  - New option to remove <NON_REF> symbolic allele from GvcfGenotyper output
  - New option to filter out calls with low depth. No loss in true variants from the input
- Joint Genotyping
  - Now uses phasing information for chromosome M calls
- Cloud based workflow
  - For large cohort analysis, an end-to-end popgen workflow is available in the cloud on the Illumina Analytics Platform (IAP)
    - Simple CLI based scheduling for the execution of large cohort analysis. No need to develop a workflow
    - Supports input from IAP GDS, S3, HTTPS
    - Built in robustness features: retry, output validation, abort and resume, job monitoring
  - 5K WGS cohort benchmark using 100 parallel nodes
    - Gvcf Genotyper: 3 hr
    - Joint Genotyper: 14 hr

BCL Conversion

- DRAGEN™ v3.7 includes an Alpha release of support for very high sample counts (100K+)
- New features
  - Support for '`no-lane-splitting`': fastq files merged across lanes. Requires lack of "`Lane`" column in Sample Sheet `[Data]` section
  - Support for turning off trimming of UMI sequences via Sample Sheet '`TrimUMI,0`'
  - Support for output of index reads into fastq format
- Additional thread and I/O controls via command line

Expansion Hunter

- New STR genotyping algorithm achieves higher accuracy through better handling of ambiguously aligning reads
- The new algorithm eliminates PHOC2B false positive calls

Methylation

- New sorting and duplicate marking/removal for Methylation caller
- Added support for alignment output in CRAM and SAM
- Improved metrics reporting
- Improved robustness

HT builder

- Improved the speed of hash table generation for references with large number of contigs

Misc.

- Improved stability for Somatic runs
- Improve robustness of BED file handling

Issues Resolved

- Fix for rare invalid CIGAR output by RNA mapper, leading to crash in RNA processing
- Fix for incorrect genotyping in GVCF on the Mitochondrial chromosome, caused by a logic difference in the calculation of likelihood scores between VCF and GVCF paths for chrM.
- Fix for intermittent license server communications timeout on AWS platforms caused by temporary network outages. The timeouts are extended and number of retries are increased.
- bam_list.csv output is disabled for v3.7. The RGID<->BAM mapping had been incorrect in the bam_list.csv output in prior releases.
- Fix for Tumor/Normal CNV caller crash in v3.6, when `cnv-normal-b-allele-vcf` is supplied with a GVCF from SNV caller that had been run with ForceGT enabled. The dots in GT field from ForceGT calls were processed as integers leading to bad lexical cast.
- Fix for Methylation pipeline ignoring the output format CRAM option and writing BAM instead.
- Fix for insert size metrics for Tumor/Normal mode being reported as aggregate of all reads, instead of being split for Tumor and Normal
- Fix for a rare run-to-run variation in coverage metrics output due to a multi-thread race condition, when using `ignore-overlaps=true` for depth of coverage metrics
- Fix for a rare crash "`Fatal error: boost: mutex lock failed`", caused by unsafe order of objects being deleted in threads.
- Fix for intermittent hang in Hash Table Builder with CNV enabled, caused by long CNV kmer building on non-human genomes. Note that the recommendation is still to set `--enable-cnv=false` when building hash table for non-human genomes because only human samples are supported for CNV WGS.
- Fix for issue when orientation bias filtering was run on an existing VCF, without running VC, the OBPa, OBParc, and OBPsnp format fields were not present and did not have descriptions in the VCF header.

Known Issues and/or Impacts

- Increased memory usage for BCL conversion on certain flowcells. A minimum of 64GB RAM is recommended
- Default system ulimits have changed for this release. A reboot or session closure will be required when installing
- Support for IBM PPC has been deprecated and not available for DRAGEN™ v3.7 and later
- CNV Somatic VAF model fitting may crash due to watchdog timeout on AWS when the number of somatic SNVs are extremely large (> 1M). Due to lower number of available threads, and extremely large input for the model fitting, the run time exceeds a watchdog limit timeout and the module did not indicate that calculation is in progress. Workaround is to disable watchdog for in such cases.
- Nirvana doesn't support regular gzip VCF files. DRAGEN™ outputs are bgzipped and indexed
- Nirvana Downloader unintentionally removes genomic reference file. When using the Downloader to download GRCh37, it will erase the GRCh38 reference. Use Downloader mode "Both" which will download both GRCh37 and GRCh38

SW Installation Procedure

- Download the desired installer from the Illumina support website and unzip the package
- The archive integrity can be checked using: `./<DRAGEN 3.7.5 .run file> --check`
- Install the appropriate release based on your Linux OS with the command: `sudo sh <DRAGEN 3.7.5 .run file>`
- Please follow the installer instructions. Cold boot may be required after installation, depending on the currently installed version. A cold boot is a hard reset or power cycle. 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