DRAGEN v3.6.3 Software Release Notes

June 19, 2020

Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform v3.6.3.

Changes are relative to DRAGEN™ v3.5.7. If you are upgrading from a version prior to DRAGEN™ v3.5.7, please review the release notes for a list of features and bug fixes introduced in subsequent versions.

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

The 3.6.3 software package includes:
- DRAGEN™ SW Intel Centos 6 - dragen-3.6.3.el6.x86_64
- DRAGEN™ SW Intel Centos 7 - dragen-3.6.3.el7.x86_64
- DRAGEN™ SW IBM PPC Centos 7 - dragen-3.6.3.el7.ppc64le.run

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

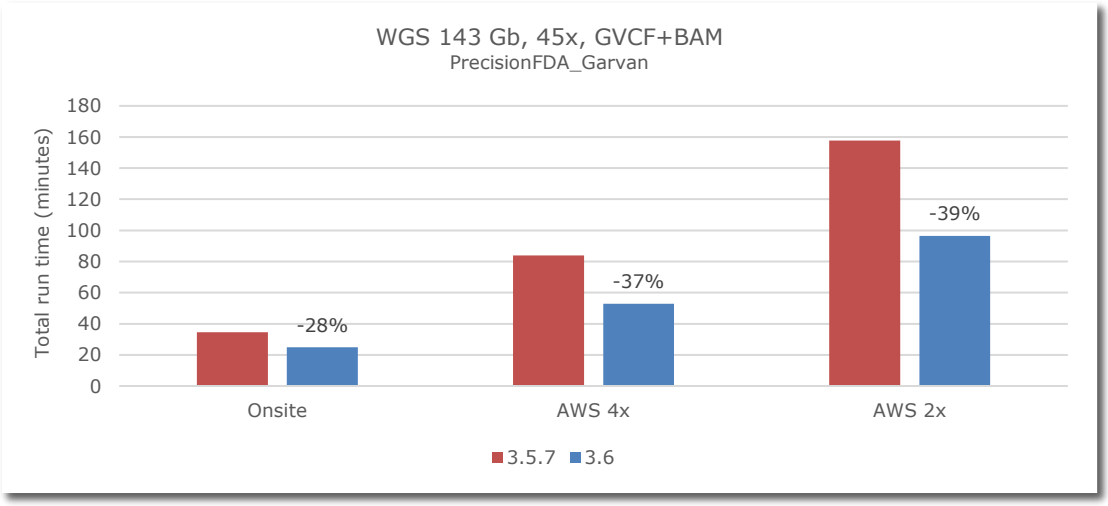
Contents

Highlights

Below is a summary of the changes included in v3.6.3. For full extensive details, please consult the Illumina DRAGEN™ Bio-IT Platform User Guide available on the support website at https://support.illumina.com/downloads/illumina-dragen-bio-it-platform-user-guide.html

HW GRAPH for Germline and Somatic Small Variant Calling

- New HW deBruijn graph module integrated in FPGA to accelerate graph construction used in the SNV caller
- Reduces SNV calling run time in the cloud, and on local servers with U200 FPGA cards, by up to 40%



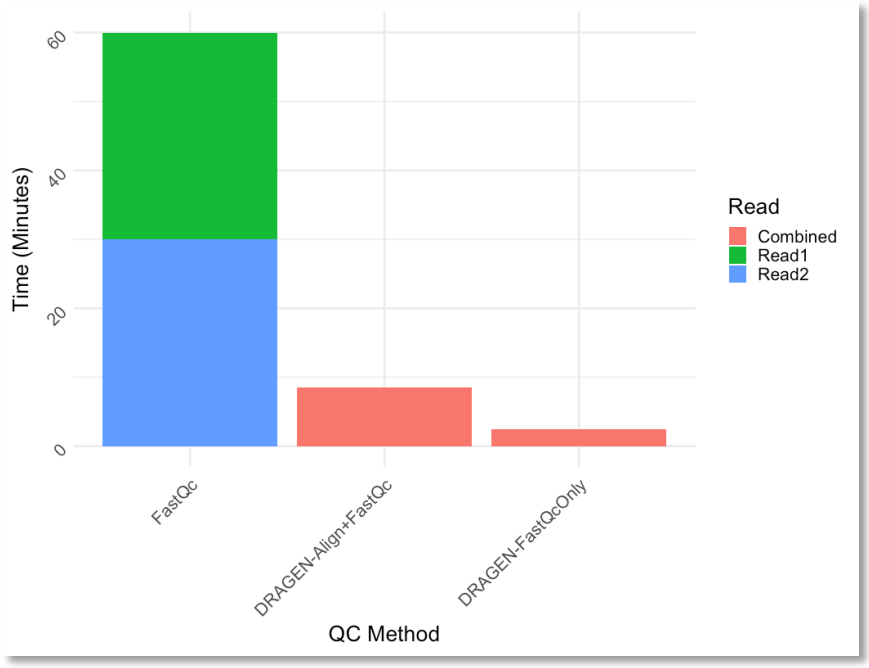
HW Trimmer for Poly-G

- DRAGEN™ includes hardware-accelerated Poly-G read trimming available as part of the mapper, which adds no additional run time or cost when enabled. Available on the cloud and on local servers with U200 FPGA cards
- Includes a novel lossless soft-trimming mode, which uses Poly-G trimmed reads for mapping, but does not strip the Gs from the output reads. This improves mapping accuracy in the presence of Poly-G artifacts without any loss of original information. Enabled by default
- Optional hard-trimming mode to permanently trim Poly-G from incoming reads
- Metrics:
    - The trimmer generates a metrics file <prefix>.trimmer_metrics.csv, which contain
        - Statistics on the number of reads trimmed
        - Statistics on R1/R2 3' ends containing likely poly-G artifacts after trimming
    - Metrics allows a user to spot the potential impact of trimming, and also see how many reads may be affected by Poly-G.
- The following options can be used to configure read trimming in DRAGEN™:
    - `--soft-read-trimmers` Set to "`polyg`" to enable soft Poly-G trimming, or "`none`" to disable. During mapping, reads will be aligned as though they had been trimmed, but no bases will be removed from the reads. Enabled by default.

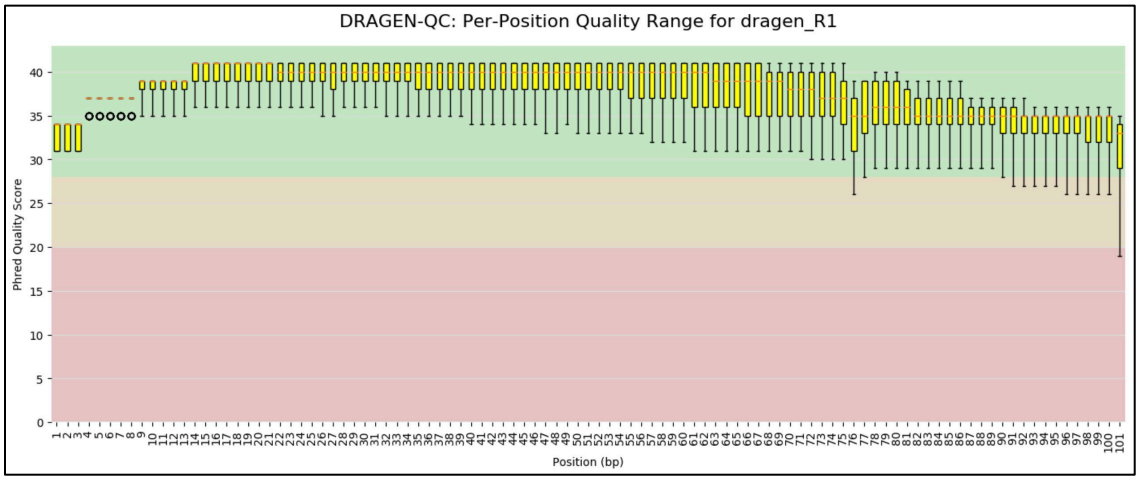
- `--read-trimmers` Set to "`polyg`" to enable Poly-G trimming, or "`none`" to disable. During mapping, artifacts will be removed from all reads, and the reads will be mapped accordingly.

HW FastQC Statistics

- New HW FASTQ statistics module integrated in FPGA that collects FASTQ read metrics in real time while mapper/aligner is running, at no additional run time or cost.
- Available on the cloud and on local servers with U200 FPGA cards. The output is enabled by default
- FastQC is used to perform quick quality control checks of raw sequence data prior to further downstream analysis, and without the need to run additional 3rd party tools to process the FASTQs.
- DRAGEN™ performs calculations and generates metrics like ones produced by Babraham Bioinformatics' FastQC
- A FastQC-only mode can be used to produce metrics more quickly without having to run the mapper/aligner and without consuming bases from the license quota. Use the option `--fastqc-only=true`



- Output
  - The *.fastqc_metrics.csv generated by DRAGEN™ FastQC can be used to generate plots with a customized standalone tool based on the bioinformatics results aggregation and summary tool, MultiQC
  - The tool and usage instructions can be found at https://github.com/bnbowman/MultiQC
  - For customers using Illumina BSSH, the charts are now available in all the applicable DRAGEN™ applications.

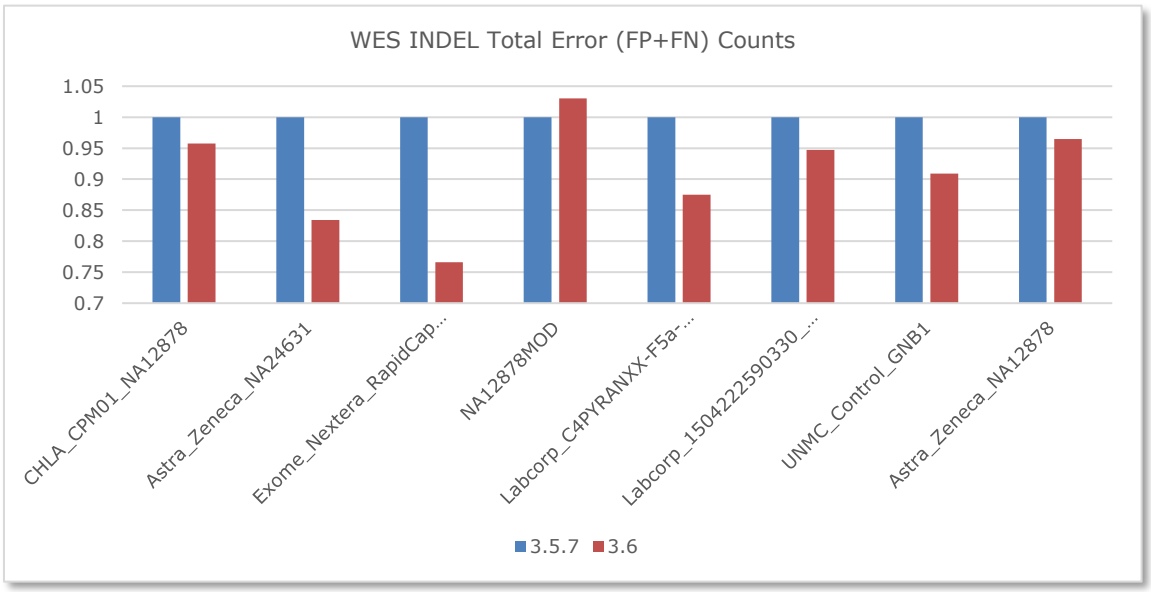


Small Variant Caller Accuracy Improvements

- Germline:
  - New default hard-filtering QUAL threshold for INDELs
- Somatic:
  - Improved FRD Algorithm (making use of MAPQ score > 60)
  - Filter optimizations
  - Support for Hotspot VCF
  - Accuracy improvement for orientation bias filter (FFPE V3)
- Germline & Somatic
  - Improvements in the PCR Error Model, benefiting targeted libraries (e.g. WES)
  - Better handling of overlapping mates
- These improvements result in gains in precision, with no loss in sensitivity

Example: Germline WES

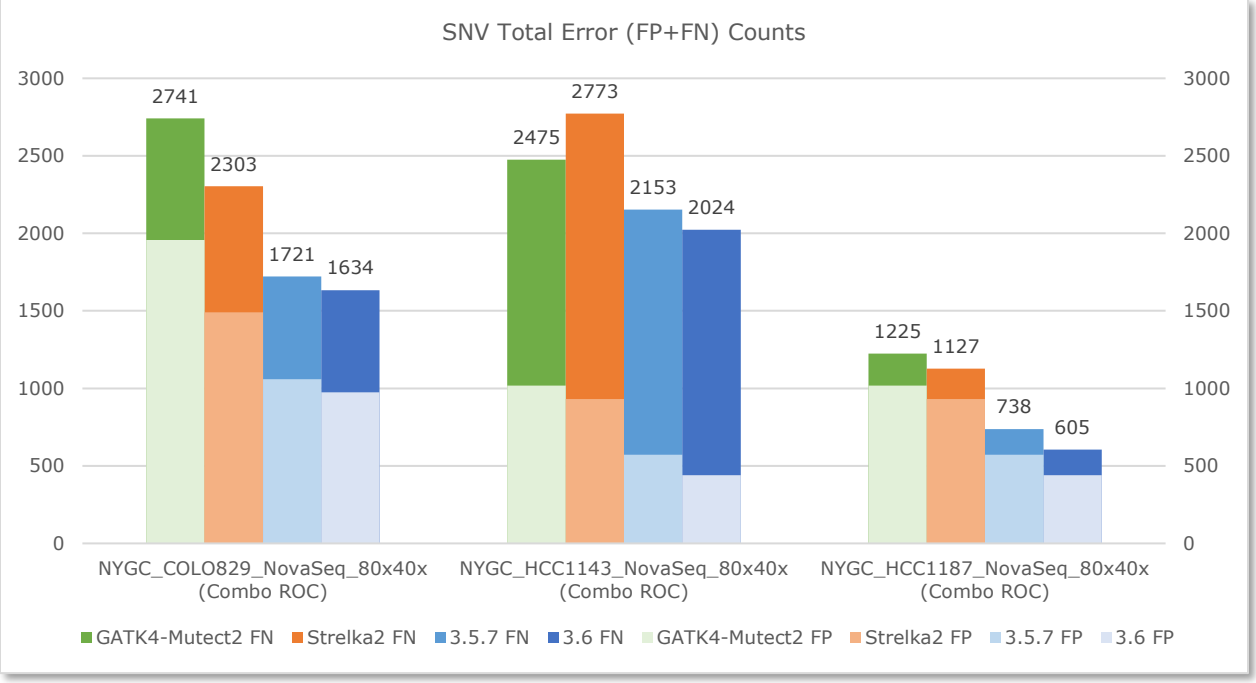
Gain from 3.5.7	SNP		INDEL	
	Sensitivity	FP+FN	Sensitivity	FP+FN
WGS	+0.0%	-0.87%	+0.0%	-1.68%
WES	+0.0%	-2.53%	-0.3%	-8.94%



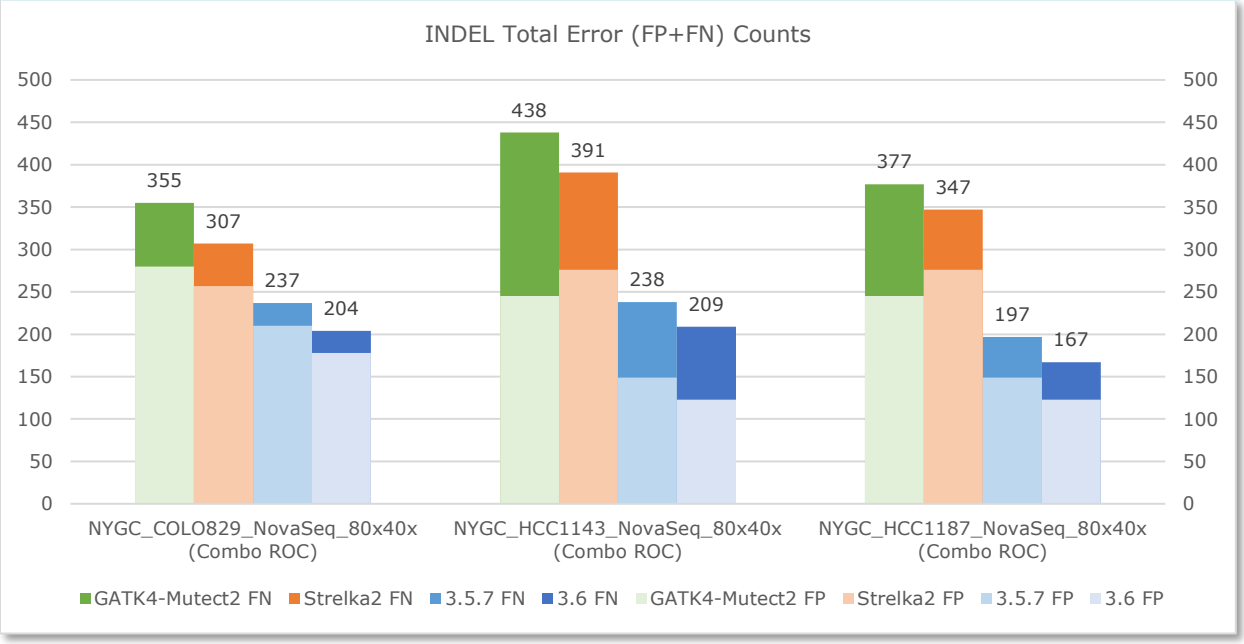
Example: Somatic WGS T/N

Gain from 3.5.7	SNV			INDEL		
	Sensitivity	Precision	FP+FN	Sensitivity	Precision	FP+FN
WGS	-0.04%	+0.47%	-3.59%	-0.19%	+2.87%	-6.43%

DRAGEN yields better SNV sensitivity, precision and total error count than 3rd party tools



DRAGEN yields better INDEL sensitivity, precision and total error count than 3rd party tools



Germline and Somatic Small VC Hotspots VCF

- DRAGEN™ somatic hotspots allows you to provide a VCF that acts as a whitelist with positions where the risk for genetic mutations is significantly elevated. This prior information improves the SNV caller's sensitivity with limited impact on specificity
- `--vc-somatic-hotspots <path to VCF type file with hotspot locations> --vc-hotspot-log10-prior-boost <prior adjustment for somatic sites present in hotspot vcf (Default=4.0)>`

CNV Caller

- General Updates
  - All segmentation algorithms now aligned to Shifting Levels Model (SLM) algorithms
    - Germline WGS: SLM
    - Somatic WGS: Adaptive SLM
    - WES: Heterogeneous SLM
- Somatic CNV:
  - Improvements to model scoring, with ability to distinguish between samples without any events and samples with very few events.
  - Improvements to segmentation to better handle noisy depth segments (FFPE).
  - Added ability to reduce fragmentation of reported events by smoothing over negligible failing events.
  - Germline-aware mode of operation: Matched normal germline CNVs are considered during tumor modeling.
  - Speed optimizations for processing the B-allele input VCFs.
- Exome CNV:
  - More robust error handling of Panel of Normals samples and generation.
  - Updated default segmentation to use HSLM, to align with WGS applications and SLM.

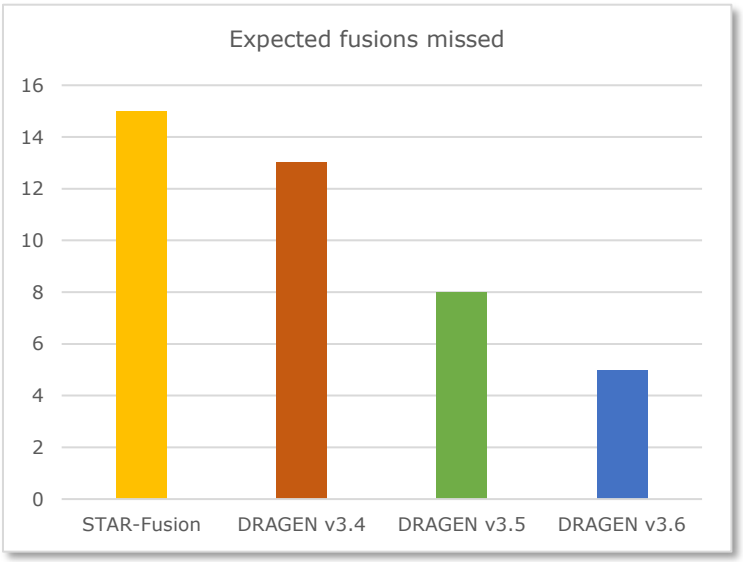
Note that the experimental 'fpop' segmentation mode will no longer be supported.

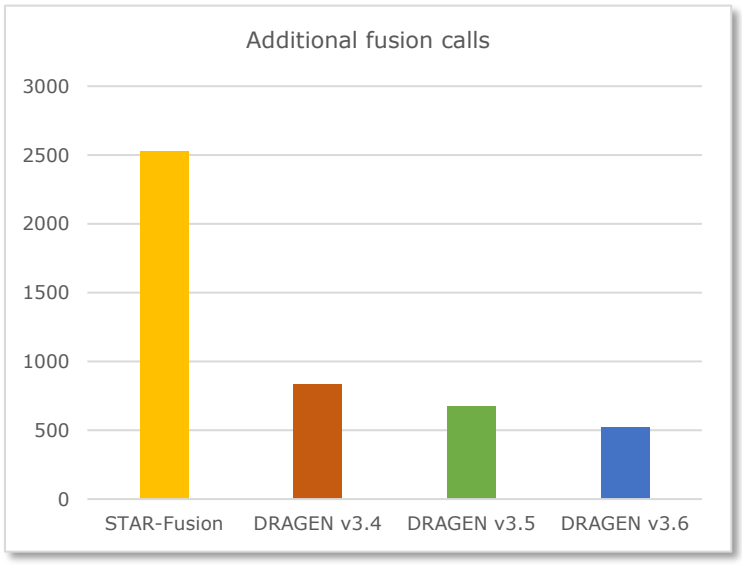
Ploidy Estimator

- A ploidy.vcf is now emitted for all WGS runs, indicating whole chromosome mosaic losses or duplications. This is an extension of the ploidy_metrics.csv but with proper call thresholding, QUAL scores, and depth annotations in a VCF format.

RNA Gene Fusion Calling

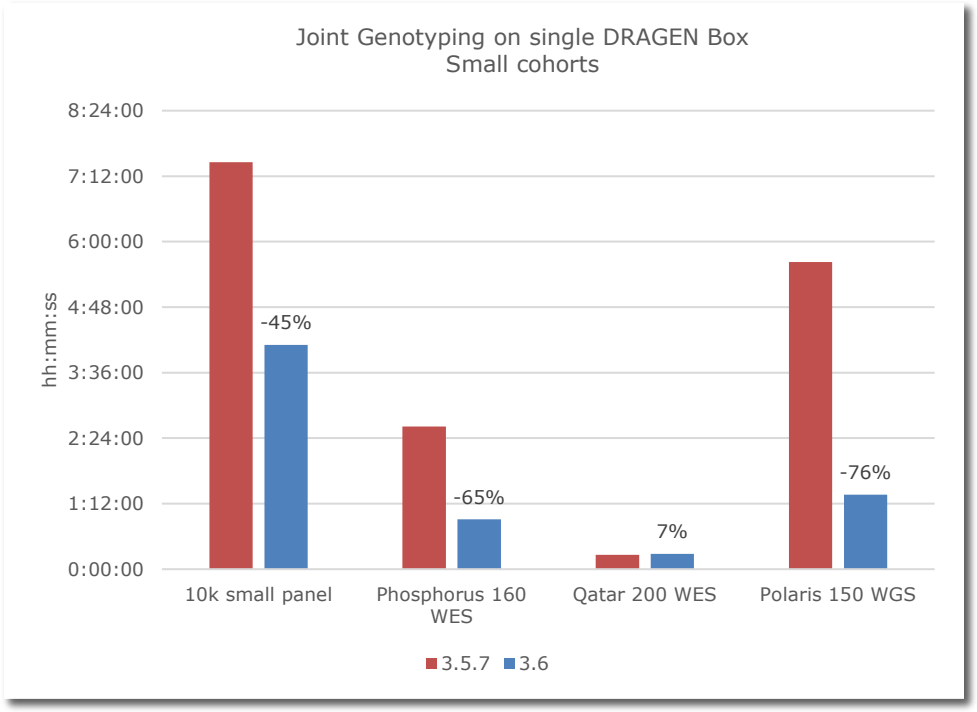
- Added a new novel probabilistic scoring of fusion candidates
- Added support for gene enrichment assays
- Fusion calls with share breakpoints merged into one
- Improved fusion detection
    - Measured on 153 sample set:
    - 38% fewer missed fusions than v3.5
    - 23% fewer additional calls than v3.5





Popgen – Gvcf Genotyper and Joint Genotyper

- Improved run time of Joint Genotyper for popgen by factor 3x for large scale population calling
- Added support for GATK v4.1 input gVCFs to Gvcf Genotyper. Use `--vc-enable-gatk-acceleration` option
- Added support for multi-sample pedigree gVCFs to Gvcf Genotyper. The output of a family joint called run can be input as part of the popgen cohort



Popgen in Cloud

- A Popgen cloud solution based on DRAGEN™ v3.6.3 + Illumina Analytics Platform (IAP) is available to IAP customers. Supports processing of up to 10K samples.

BCL

- New features
  - Support a new sample sheet setting: `MinimumAdapterOverlap`
    - Possible values: 1, 2, 3. Default1
    - If the detected adapter sequence impinges upon a genomic read by less than the stated # of bases, do not trim any bases.
    - This can be used to prevent skew caused by the 25% chance to trim one base by chance by setting value to '2'
  - Support for fastq.gz placement into per-project subdirectories during BCL conversion
    - Support for `Sample_Project` column in sample sheet
    - Added cmd-line option `--bcl-sampleproject-subdirectories` (defaults to 'false')
    - Requires use of both sample sheet column and command line option to activate
    - Supported in *bcl-convert* software-only executable as well
    -
- Bug fixes
  - See section below
- Enhancements
  - Illumina v2 sample sheet processing used on the NextSeq 1000/2000 sequencer
    - Updated support for processing v2 sample sheets
    - Add strict check that Version 2 Sample Sheets must have a `[BCLConvert_Settings]` section
    - New sample sheet validation mode
    - New sample sheet validation of `[Read]` section of v2 sample sheets on startup
  - Sample Sheet Validation:
    - Use `--bcl-validate-sample-sheet-only` option to do no conversion, only validate RunInfo.xml and sample sheet files.
    - Fixes and refinements to error messages relating to sample sheet inputs
    - Detect barcode collisions (based upon mismatch tolerance)
  - Eliminated an outdated warning message regarding 0-cluster tiles
  - Added detailed I/O error reporting when writing metrics files at end of BCL conversion run

SV Caller

- SV caller supports a list of contig patterns to skip during SV calling. Per default the list contains all unplaced/unlocalized and alt contigs from hg38 references [*_random, chrUn_*, *_alt, HLA-*]. Contigs can be filtered from the SV-caller with `--sv-filtered-contigs` option. These filtration patterns are applied in addition to those in `--vc-decoy-contigs`.
- This improves the run time of the SV caller on GRCh38 alt-aware reference by up to 20%.

Metrics

- Coverage metrics
  - Added an option to avoid double-counting bases overlapping read 1 and read 2. Use `qc-coverage-ignore-overlaps=true`
  - Added an option to count soft clipped bases. Use `qc-coverage-count-soft-clipped-bases=true`

Nirvana Annotation

- Nirvana binaries are packaged with the DRAGEN™ installer. They are in the `/opt/edico/share/nirvana` directory after installation
- The binaries work on CentOS 6 and CentOS 7 using x64 processors only
- Annotation data files can be downloaded using the `Downloader` tool

Methylation

- Added support for TAPS methylation sequencing data. Use `--methylation-TAPS=true` if data is generated from a TAPS assay.
- Improved the cytosine report generation. Records are only emitted in the Cx_report for bases with read coverage >= 0. Reports are smaller and generated faster. (Used with `--methylation-generate-cytosine-report=true`)
- Since DRAGEN™ v3.5, several default parameters were optimized to reduce the rate of spurious alignments in noisy datasets. This release fixes a bug where the new default could not be overridden with command line options.

Issues Resolved

- DRAGEN™ v3.5 Limitation: if `--methylation-protocol` is set to something other than '`none`', then the `Aligner.match-score` and `Aligner.min-score-coeff` were hardcoded and command line options would not take effect. This limitation has been removed.
- Fix for BCL conversion crash if (tilesize % 256K (buffersize)) == 0 issue present in DRAGEN™ 3.5.7. This bug fix is also available in DRAGEN™ v3.5.7b patch.
- Fix for crashing on corrupt input CBCL gzip regions. Now more robust to corrupt input.
- Fix a corner case where BCL conversion could zombify DRAGEN™ process if access to output directory is interrupted during conversion
- Fix in Demultiplex_Stats.csv: Undetermined reads were not all counted as 'Perfect match'
- Fix to DRAGEN™ inputting fastq to secondary analysis via fastq_list option. An empty fastq file caused subsequent fastqs to be skipped
- Fix for invalid VCF fields when enabling Orientation Bias Filtering on a sample run. The format fields did not match the values, and the field descriptions were missing from the header.
- Fix where RNA analysis aborts with bad::alloc error on BSSH
- Fix for issue that prevents license manager to recognize dark sites, and attempts to upload records, leading to freeze of dragen_lic
- Fix to improve robustness of BED file handling, by reading the file header instead of the file extension to determine if a file is zipped.
- Fix for previously known discordances between VCF and GVCF outputs. VCF and GVCF outputs are now concordant.

Known Issues and/or Impacts

- Somatic T/N Prefilter VC Metrics SNP Transitions/Transversions Mismatch RTG. The Difference is < 0.03%
- INDEL accuracy degradation on high-purity solid admixture datasets. The degradation is expected, and this tradeoff is made for major improvements on most datasets as detailed above. In general, we are aware that FRD increases FN and that this can lead to increased FP+FN on admixtures
- 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

- Hash Table Builder with CNV enabled intermittently hang when building the CNV kmer on non-human genomes. Recommendation is the set `--enable-cnv=false` when building hash table for non-human genomes. Note that only human samples are supported for CNV WGS.
- When orientation bias filtering is enabled, the OBPa, OBParc, and OBPsnp format fields are not present and do not have descriptions in the header.
- When orientation bias (OB) filtering is enabled, the B-Allele Frequency (BAF) output contains entries for PASS variants prior to OB filtering. The BAF is not updated with the OB filtered variants.
- Remove previously deprecated command line options `--vc-enable-depth-of-coverage`, `--vc-enable-histogram-of-coverage`, `--vc-depth-intervals-bed`.

SW Installation Procedure

- Download the desired installer from the Illumina support website and unzip the package
- The archive integrity can be checked using: `./<DRAGEN 3.6.3 .run file> --check`
- Install the appropriate release based on your Linux OS with the command: `sudo sh <DRAGEN 3.6.3 .run file>`
- Cold boot (hard reset or power cycle) may be required after installation, depending on the currently installed version. Please follow the installer instructions. An updated FPGA shell image needs to load from flash, this is only achieved with cold boot.
- Installing prior releases after v3.4.5 was installed:
    - Installing a prior release, v3.3.7 or older, will require the following two steps. The prior .mcs file needs to be flashed manually:
        - Install the prior release: `sudo sh <DRAGEN 3.3.7 .run file>`
        - `program_flash /opt/edico/bitstream/07*/*.mcs`
        - Power cycle