illumına®

DRAGEN v3.8.4
Software Release Notes

Introduction

These release notes detail the key changes to software components for the Illumina® DRAGEN™ Bio-IT Platform v3.8.4.

Changes are relative to DRAGEN™ v3.7.5. If you are upgrading from a version prior to DRAGEN™ v3.7.5, 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 3.8.4 software package includes:
• DRAGEN™ SW Intel Centos 7 - dragen-3.8.4-4.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)

Deprecated platforms:
• Support for IBM PPC has been deprecated since DRAGEN™ v3.7
• Support for Intel Centos 6 has been deprecated since DRAGEN™ v3.8

Contents

Overview

Below is a summary of the changes included in v3.8.4. DRAGEN™ v3.8 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 and Features

Biomarkers

- DRAGEN™ v3.8 includes Biomarker detection for TMB / HLA / MSI
- The callers are integrated into the DRAGEN™ end-to-end workflow, can be selectively enabled, and are output in addition to regular variant caller and DRAGEN™ outputs
- Tumor Mutational Burden (TMB)
    - o Integrated in a single execution end-to-end workflow with SNV calling, Nirvana annotation, and TMB calculation
    - o Flexible options: custom coding regions can be specified with a BED file, minimum required coverage can be set
    - o Supported with Tumor-Normal and Tumor-Only analysis
    - o TMB SNV and INDEL filtering VAF threshold can be configured
- Human Leukocyte Antigen (HLA) Typing
    - o DRAGEN™ v3.8 introduces a new fast and reliable HLA typing approach, relying on nucleotide-to-protein alignment of HLA alleles and uses ILP to find the sets of class I HLA alleles that can be aligned to most reads.
    - o Typing of six Class 1 HLA alleles (A1, A2, B1, B2, C1, C2)
    - o Supported with Germline and Tumor-Normal analysis
- Microsattelite Instability (MSI)
    - o Produces a MSI score (percentage of usable sites) on MSI sites that are at least 10bp long
    - o Supported with Tumor-Only and Tumor-Normal analysis

DNA Amplicon

- DRAGEN™ v3.8 now officially supports analysis of DNA Amplicon panels for small variants.
- Supports Germline and Somatic Tumor-Only DNA Amplicon pipelines
- Like other pipelines, DNA Amplicon is processed with one end-end workflow per sample starting with FASTQ and producing VCF
    - o Includes primer trimming
- Comparable SNV accuracy and superior Indel accuracy, compared to DNA Amplicon BaseSpace App v2.1.1

NA12878 AmpliSeq for Illumina Exome Panel on NextSeq2000
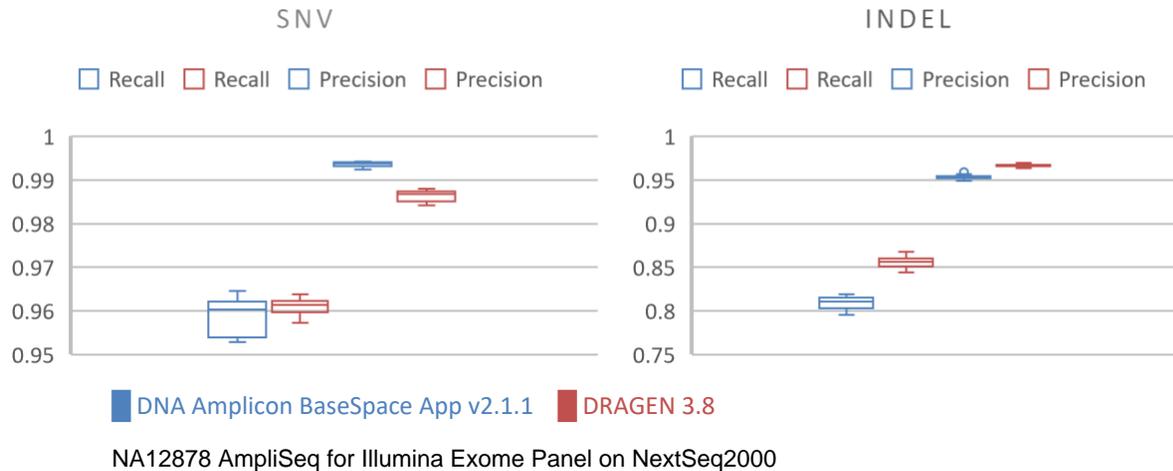
Hardware Trimming

- A new hardware-based trimmer is added to the DRAGEN™ mapper, and v3.8 releases a Beta version of DRAGEN-Trimmer.
- The DRAGEN-Trimmer pre-processes FASTQ-formatted sequences to remove low-quality or otherwise undesirable bases and reads from a dataset before it passes through the DRAGEN-FastQC and DRAGEN™ Mapper/Aligner modules.
- Similar functionality to Atropos and CutAdapt tools, but multiple times faster and integrated in the DRAGEN™ flow.
- Available for Phase2 and Phase 3 DRAGEN™ servers (u200) and AWS
- Supported modes
  - Hard and Soft trimming. Soft-trimming allows mapping with trimmed reads but mapper output untrimmed reads.
  - Trimming without mapping
  - Trimming and filtering before mapping
- Trimming options include, fixed length, Poly-G, Low Qual, Adapter Sequence, Bisulfite-T, Ambiguous Base

Ora Compression

- ORA is a compression format used for efficient compression of FASTQ files. It supports all FASTQ files generated by Illumina instruments. It performs completely lossless compression, the md5 checksum of the FASTQ content is preserved after a compression / decompression cycle.
- On human data generated by Novaseq / Nextseq 2000 sequencers, the compression ratio is expected to be between 4x and 6x compared to the .fastq.gz file. The compressed file will have extension *.fastq.ora*
- DRAGEN™ provides new functionality for
  - The compression of files in *.fastq* or *.fastq.gz* format into *.fastq.ora*, and decompression of *.fastq.ora* files to *.fastq.gz* format.
  - Native support for input of Ora compressed FASTQ files to DRAGEN™ for map/align
- Supports AWS S3 streaming and pre-signed URL inputs
- A Compression license is required for compression of FASTQs

Variant Deduplication

- A new pipeline that takes two VCFs and removes duplicate reads
- Supports the removal of small variants from SV VCF
  - Runs as optional step in end-end DRAGEN™ analysis when both small VC and SV callers are enabled
  - Can run as standalone tool with small VC and SV VCF inputs
  - A filtered copy of the SV VCF is produced
- Supports the filtering of small VC with SMN variants form Expansion Hunter
  - Run as standalone tool with small VC and EH VCF inputs
  - A filtered copy of the small VC VCF is produced, containing only SMN variant

Improvements and Feature Additions

Small Variant Calling

- **Germline**
  - **Joint Detection of Overlapping Variants**
    - Available since DRAGEN™ v3.7 as option, now enabled by default
    - Results in significant accuracy improvements in SNP and INDEL
    - Run time optimizations in DRAGEN™ v3.8 to allow default enabling:
      - U200 on-site runs are now ~4% faster than v3.7
      - AWS f1.4x runs are now ~10% slower vs v3.7
  - **Graph-Capable Mapping**
    - DRAGEN™ v3.7 introduced a beta version of the graph-capable mapper with support for hg38
    - 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.
    - DRAGEN™ v3.8 adds:
      - Graph-capable mapper officially supported and fully validated with all downstream callers
      - More references supported. Added GRCh37 and hg19
      - Similar gains in accuracy achieved on GRCh37 and hg19. ~30% reduction in SNP errors and ~7% reduction in INDEL errors
      - Pre-built Graph Hash Tables available for user download on the Illumina DRAGEN™ Bio-IT Platform support website

HG002 FP + FN (hg38)

HG002 FP+FN (hs37d5)
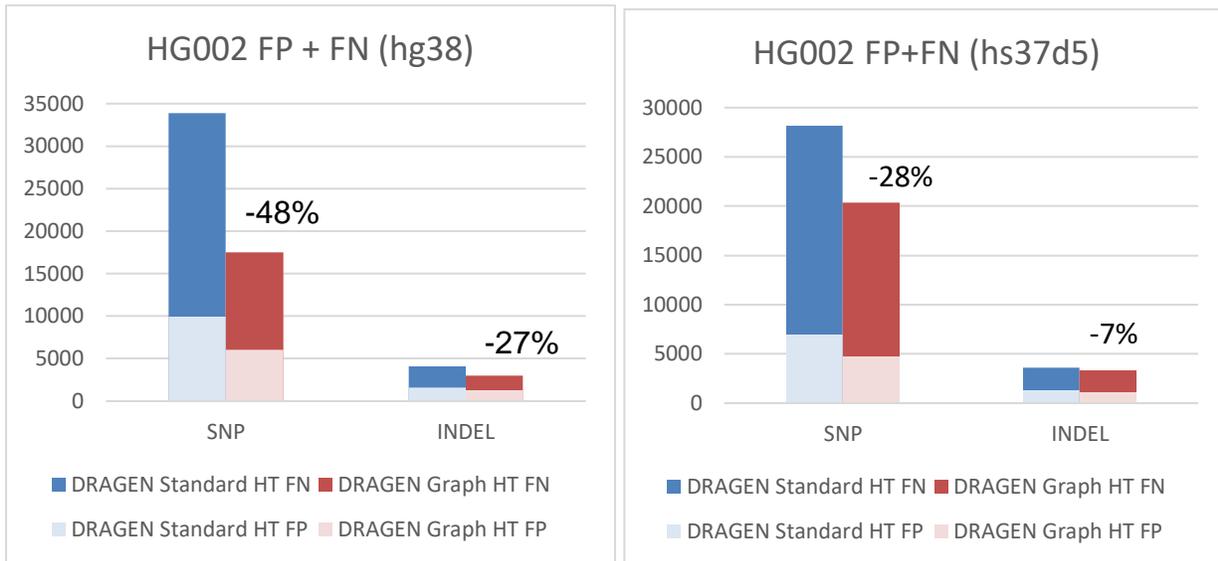
- **Somatic**
  - **Auto detected hotspot regions**
    - DRAGEN™ auto detects hotspot regions for hg19/hs37d5/hg38
    - Identifies hotspot regions where the risk for genetic mutations are elevated
    - Improves SNV caller's sensitivity with limited impact on specificity
    - Enabled by default. Can be disabled or overridden
 - **New Force Genotyping (FGT) for small variants**
    - Key positions are genotyped and reported in vcf even if there is low or no evidence of variants
      - Allows for reporting on known cancer mutations (cosmic positions)
      - Allows tracking of mutations across multiple cancer samples from same patient
    - Supported for Tumor-Normal and Tumor-Only
    - Enabled with command line options
  - **FLT3-ITD Calling**
    - DRAGEN™ can call FLT-ITDs with the standard somatic pipelines
    - Depending on event size, calls are made by either small variant caller or structural variant caller
    - Specialized caller like Pindel is not needed to call ITDs, and DRAGEN™ achieves better recall and lower FP than Pindel
  - **Combining of phased variants**
    - Phased variants from the same phase set can be combined
    - Phase variance combination distance can be specified

Performance on 25 TCGA Samples

	Recall	Precision
DRAGEN	0.76	0.90
Pindel	0.40	0.08
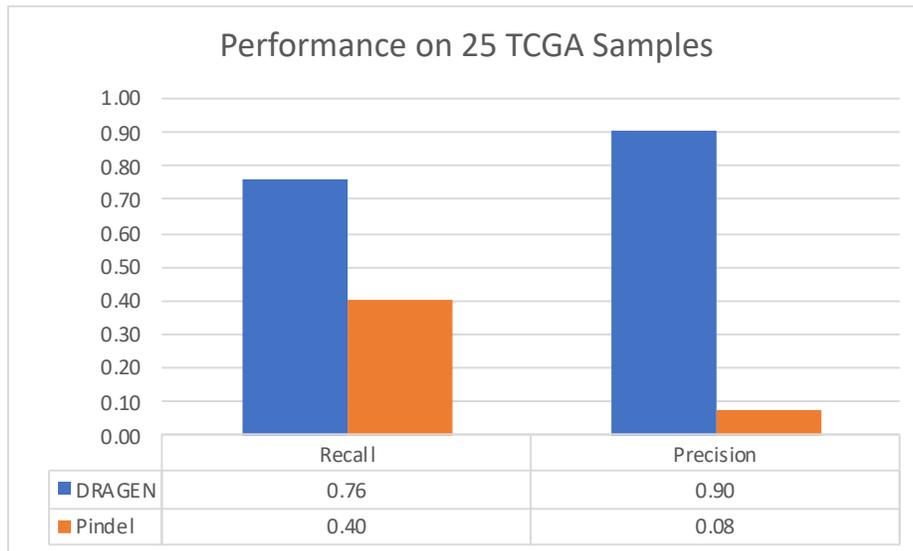
CNV Caller Updates

- **Somatic WGS**
  - Added support for detection of subclonal CNVs
    - Subclonal regions are now identified by a special HET(erogeneous) state
  - Mismatched normal detection with new metric: OutlierBafFraction
    - Measures the fraction of B-allele fractions that are incompatible with the segment they belong to
    - Allows identification of mismatched normal, or cross-sample contamination, or other source of mosaic genome
- **WES / Small Panels**
  - Added support for mixed gender panel of normals
  - Added support for DeNovo CNV from WES / Small Panels
 - Ability to define genes or regions of interest to segment with a BED file

SV Caller

- Expanded forced genotyping capability
  - Now supports FGT of tandem duplications and breakends, in addition to the support for insertions and deletions already introduced in v3.7
- Germline accuracy improvements
  - Recall exceeds 72% for deletions and 61% for insertions, while maintaining high precision. v3.8 now greater than double the insertion recall of Manta

Methylation

- Improved methylation mapper
  - The multiple map/align pass implementation is replaced with a single map/align stage, resulting in significantly faster run time while maintaining the same accuracy and features
  - Single pass in v3.8 is 3x faster than v3.7
  - A single pass methylation hash table need to be built
  - Multi pass mode is still supported when old hash table is used

- Option to let the CX_report keep all the cytosines from reference even if absent from input

GM12878
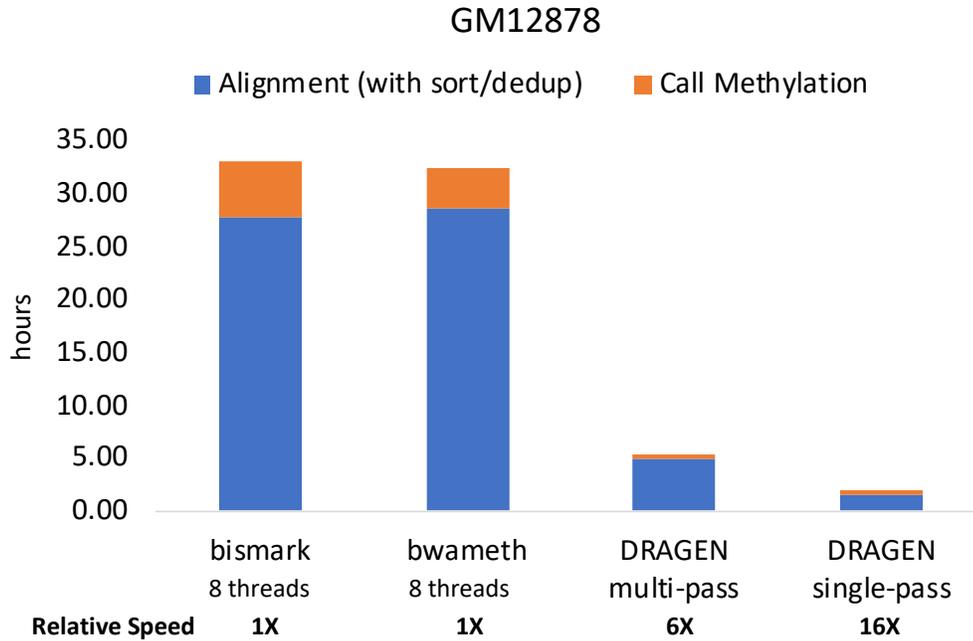
Expansion Hunter (EH)

- Updated EH to v4.0.2
- Improved alignment accuracy for difficult-to-align reads
  - In-repeat reads located in regions containing multiple long STRs with the same motif.
  - Enables better AR repeat genotyping
- Repeat catalog updates: Added NIPA1, GLS, RFC1, PABPN1 (all catalogs), NOTCH2NL (hg38 only) repeats
- When outputting the joint EH multi-sample VCF, the samples are now written in sample-name-sorted order, sorting alphanumerically
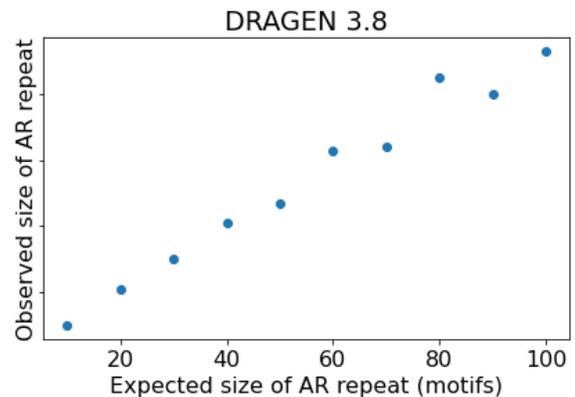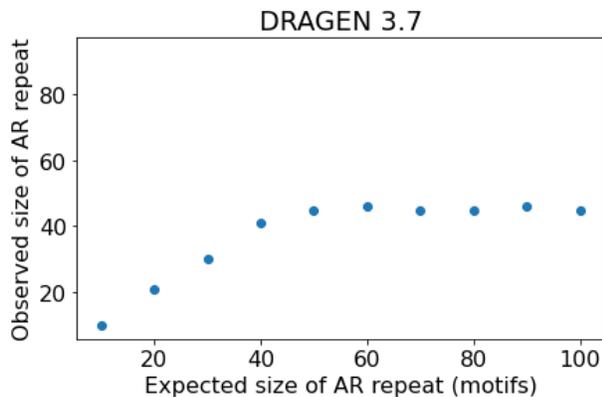
CYP2D6 Caller

- Substantial increase in genotype accuracy
  - Concordance against 212 samples with validated truth increased from 94.8% in v3.7 to 99.5% in v3.8
- Caller improvements
  - New structural variant model to handle rare subtypes
  - Updated star allele haplotype prior probabilities
  - Added a genotype likelihood model to select the most likely star allele diplotype given:
    - structural variant calls
    - small variant calls
    - population haplotype frequencies

RNA-seq

- DRAGEN™ 3.8 implements optional ribosomal RNA (rRNA) filtering during RNA-seq alignment
  - Reports a %rRNA metric that can be used for sample and library QC
- Ribosomal RNAs (rRNA) can make up a large fraction of reads in RNA-seq experiments
- DRAGEN™ rRNA filters any read pair with a significant alignment to rRNA sequence
  - Uses a copy of the rRNA sequence on an extra/decoy contig in the reference genome
  - Reads are unaligned and marked *rRNA* in the output BAM
  - Faster DRAGEN™ runtimes, and smaller output file sizes
  - Avoids very deep read pileup at genomic rRNA repeat loci, improving performance and stability of downstream tools on the BAM
- Run time speed up by 10-30% with rRNA filtering

Single-cell RNA

- DRAGEN™ 3.8 implements genotype-based demultiplexing
  - Supports demultiplexing of mixed cells from multiple individuals in one single-cell RNA library run
  - Infers sample identities of cells based on genotypes provided as input VCF file
  - Detects Doublets, droplets containing tow or more cells from different individuals
  - Comparable with demuxlet tool (Kang et al., 2018) but much faster (<10 minutes vs hours)

PopGen Workflow

- Gvcf Genotyper
  - Supports new sparse VCF output format (spVCF)
    - Reduces size by 34% for N=2504 samples
  - Supports force-genotyping of sites
    - Specify a list of sites at which genotype information will be produced
  - Adds carry-over of phasing information from the input gVCF files

BCL Conversion

- Robust support for very high sample counts (100K+)
- New features
  - Support for '`no-lane-splitting`'

Nirvana Annotation

- Support for Nirvana annotation now integrated in end-end DRAGEN™ workflow
  - When enabled, output VCFs are annotated automatically by Nirvana and JSON outputs are produced
- Nirvana updated to v3.14

Issues Resolved

Issues found on DRAGEN™ v3.7.5 that are fixed in v3.8.4

Defect ID	Component	Description
DRAGEN-10394	CNV	Fix for watchdog timeout during VAF modeler, encountered when processing large tumor samples
DRAGEN-10415	CNV	Improve the run time of CNV caller for highly unstable tumor samples
DRAGEN-10435	BCL	Fix rare segfault seen during BCL conversion
DRAGEN-10551/2	BCL	Improved error messaging when invalid command line option used for `bcl-only-matched-reads` and `bcl-write-no-fastqs`
DRAGEN-10943	CNV	Fix for segfault when accessing kmerMap with invalid contig. When self-normalization is run with CNV target bed, and the contig mismatch. This fix improves error logging for the issue that is encountered in a non-recommended use case.
DRAGEN-10947	QC Metrics	Fix for issue where invalid path to `qc-coverage-region-n` BED file is not detected during argument parsing, and results in later VC crash. The fix adds input checks and argument error
DRAGEN-10968	Somatic VC	Fix for Tumor-Only pipeline filtering some calls as weak_evidence, which are likely germline but possibly somatic. Seen on variants with AF close to 50% and when SQ threshold is raised above default.
DRAGEN-11070	Ploidy Estimator	Fix for issue where ploidy estimates are inaccurate on certain sample kits due to different emphasis on chromosomes. Options added to allow user to provide coverage factors to match the sample kit characteristics
DRAGEN-11496	Sos-report	Fix for missing files in sosreport. /var/log/dragen and /var/log/dragen.log were missing from sosreports in v3.7.5
DRAGEN-11565	Methyl reports	Fix for issue where methyl-reports-only mode outputs wrong results on v3.7.5. The issue was introduced in v3.7.5 with methyl sort-dedup.

Known Issues

Defect ID	Component	Issue Type	Description	Remedy / Workaround
DRAGEN-11376	System	Support	Support for CentoOS6 has been deprecated and not available for DRAGEN™ v3.8 and later	Upgrade system to CentOS7
DRAGEN-11718	Small VC	Run time	Germline WGS runtime is ~10% slower than v3.7 on AWS instances, due to the enabling of Joint Detection of Overlapping Variants	None
DRAGEN-11729	BCL	Usability	BCL conversion for very high sample counts (>50k) may require >128GB system RAM	256GB RAM is recommended for very high sample count BCL conversion up to 150k samples
DRAGEN-11800	Graph Mapper	Run time	Graph based mapper run time up to ~5% slower than legacy	None. Improved accuracy may lead to more work for the small VC.
DRAGEN-11801	DNA Amplicon	Bug	Amplicon caller does not exit with non-zero code when `vc-target-bed` argument is missing. Run completes with calls outside target region.	User must ensure to specify both `vc-target-bed` and `amplicon-target-bed`
DRAGEN-11636	License	Bug	License manager crash due to license .gbin file corruption has been observed for on-site systems. Rare occurrence	Delete gbin files and re-install licenses
DRAGEN-11514	Installer	Stability	Installation failure has been observed, due to inability to allocate hugepage memory buffers. Rare occurrence	Stop running programs. Sync memory. Re-run installer
DRAGEN-11511	HWAL	Stability	Timeout writing to hardware crash has been observed on Phase 1 on-site servers only. Rare occurrence that coincide only with concurrent dragen_info usage.	System power cycle is required

SW Installation Procedure

- Download the desired installer from the Illumina support website and unzip the package
- The archive integrity can be checked using: `./<DRAGEN 3.8.4 .run file> --check`
- Install the appropriate release based on your Linux OS with the command: `sudo sh <DRAGEN 3.8.4 .run file>`

Template No: 15048849 Rev A

- 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