M-US-00466

DRAGEN™ v4.3: Comprehensive coverage with leading edge innovations

Yi Lian

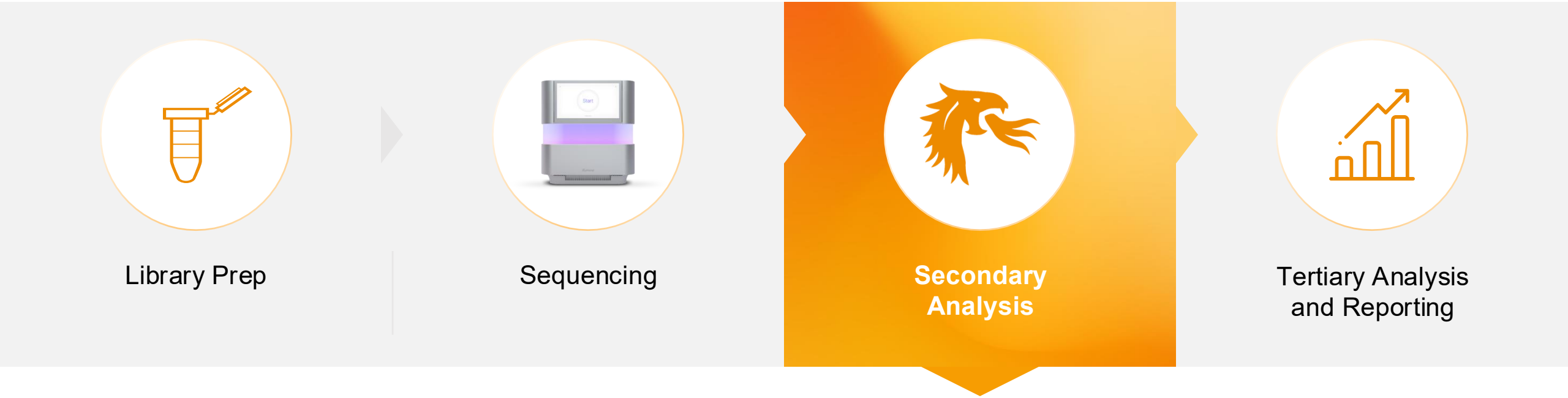
Staff Product Manager, Illumina

Shyamal Mehtalia

Principle Software Engineer, Illumina

illumina

DRAGEN is a collection of bioinformatics pipelines for secondary analysis



DRAGEN (Dynamic Read Analysis for GENomics)

Accurate, comprehensive, and efficient secondary analysis.

Award winning accuracy for germline and somatic variant calling

Germline

PrecisionFDA Truth Challenge V2

Announced August 2020

DRAGEN team won best performance in Difficult-to-Map Regions and All Benchmark Regions on Illumina Sequencing Data

Somatic

PrecisionFDA NCTR Indel Calling from Oncopanel Sequencing Data

Announced September 2022

	Best Recall (GOLD)	Best Precision (GOLD)	Best Overall (F1 Score) (GOLD)
Panel A	Novocraft	Yao-Chung Chen	Genetalks
Panel B	Genetalks	Yao-Chung Chen	Konstantinos Kyriakidis
Panel X	Chung-Tsai Su	DRAGEN team at Illumina	DRAGEN team at Illumina
			Best applicability

DRAGEN team won “Best Applicability”, Best Precision and Best Overall on Panel X

The right solution for your needs

Updated for DRAGEN v4.3

High

Medium

Low

DRAGEN on-instrument (NextSeq™ 1000/2000, NovaSeq™ X Series)

Push button analysis with instrument integration

DRAGEN on BaseSpace™ Sequence Hub (BSSH)

Push button solutions with easy-to-use graphical interface

DRAGEN on Illumina Connected Analytics™ (ICA)

Pre-packaged DRAGEN pipelines or tools to incorporate into custom pipelines

DRAGEN on-premise server

Appliance server with command-line interface

DRAGEN Multi-Cloud Bring Your Own License (AWS, Azure)

Command-line access with license provided by Illumina to run on cloud provider of choice

Configurability					
Frequency of updates					
Broadness of menu					
Push button analysis					
Integration with sequencer					

DRAGEN v4.3

A comprehensive genome with industry-leading innovations

More accurate than ever

Next generation multigenome (graph) with 128 samples

Capture more genetic diversity, reduce ancestry bias and improve SNV accuracy

Build your own multigenome (Illumina cloud)

Build custom multigenome reference from assemblies reducing bias in your population studies

Mosaic variant calling

Mosaic variant calling for low allele frequency variants

Comprehensive coverage

New specialized callers

Variant Calling in segmental duplication regions (PMS2, SMN1, STRC, NEB, TTN, IKBKG)

RNA

Improved gene fusion calling accuracy and Splice variant calling (beta)

AI powered annotations

Embedded Connected Annotations with PrimateAI-3D and SpliceAI

Enhanced efficiency & accessibility

ORA Compression

Expanded support for human methylation data and non-human data with high compression ratio

Unified targeted caller output

Streamline downstream workflow integration with unified targeted caller output, easier integration with Emedgene*

Population Genomics

Optimized performance for large cohort analysis

Emedgene supports unified targeted caller output for select callers

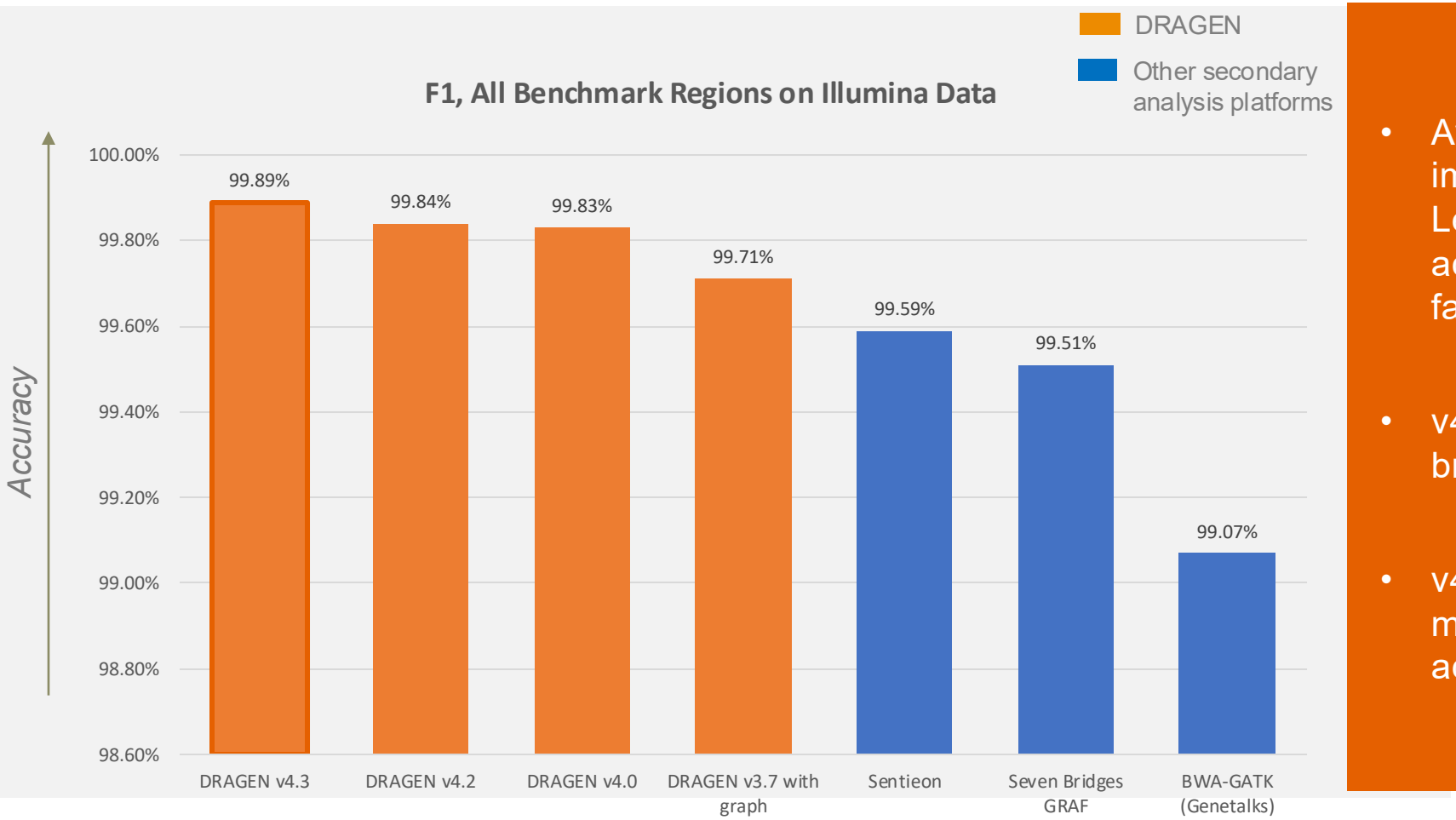
illumina®

DRAGEN Germline

illumina®

DRAGEN Secondary Analysis

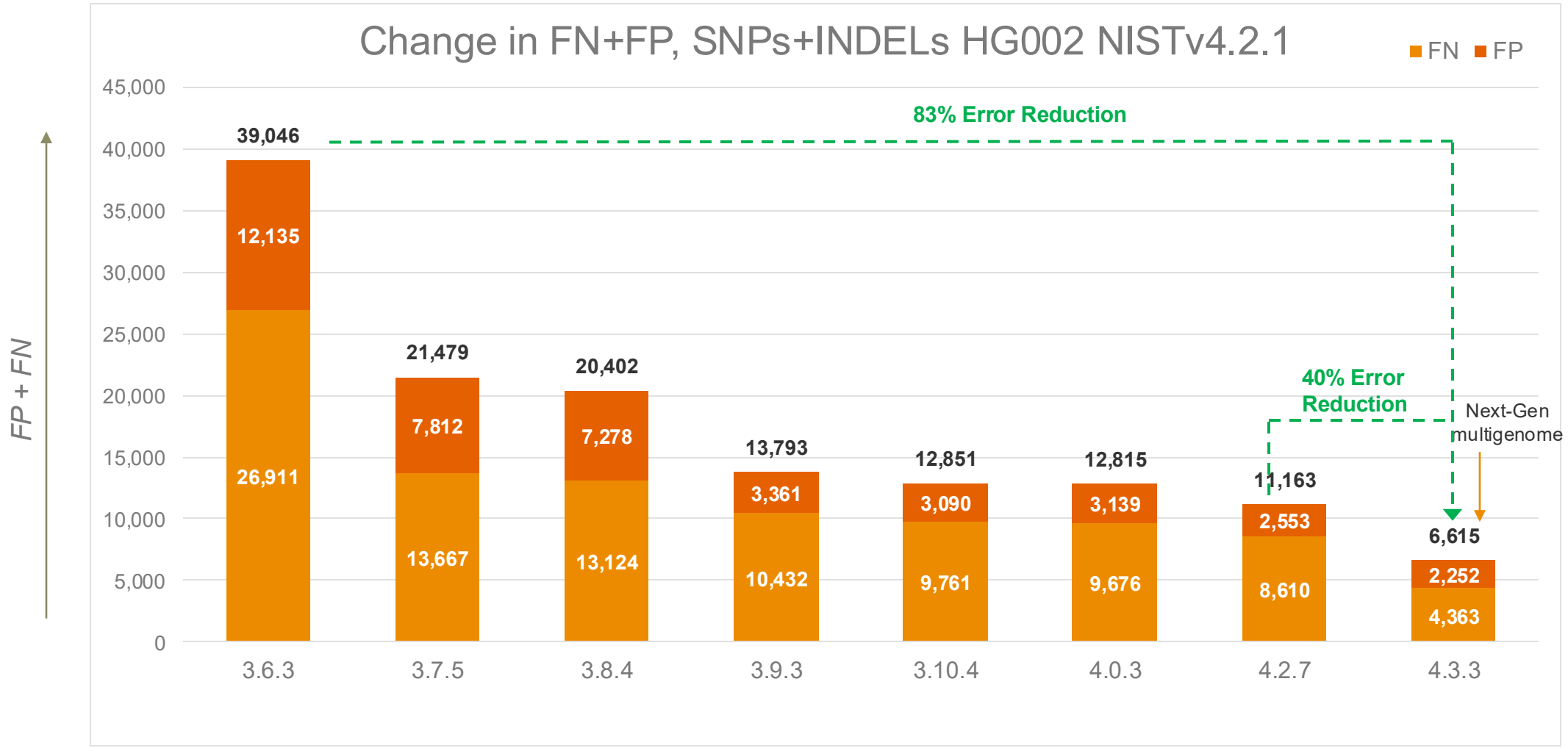
Most accurate secondary analysis in all-benchmark regions*



- Available since v4.0, continuous improvements of DRAGEN Machine Learning (ML) detects variants more accurately and effectively, reducing both false positive rate and false negative rate
- v4.3 - Germline integrated mosaic caller brings high sensitivity in mosaics
- v4.3 - New population aware multigenome mapper brings significant accuracy gains

*As compared against all participating solutions in F1 score using PrecisionFDA v2 Truth Challenge Benchmark Data (average between HG003 & HG004), data here, internal data on file for V4.

DRAGEN continues to set new standards for accuracy

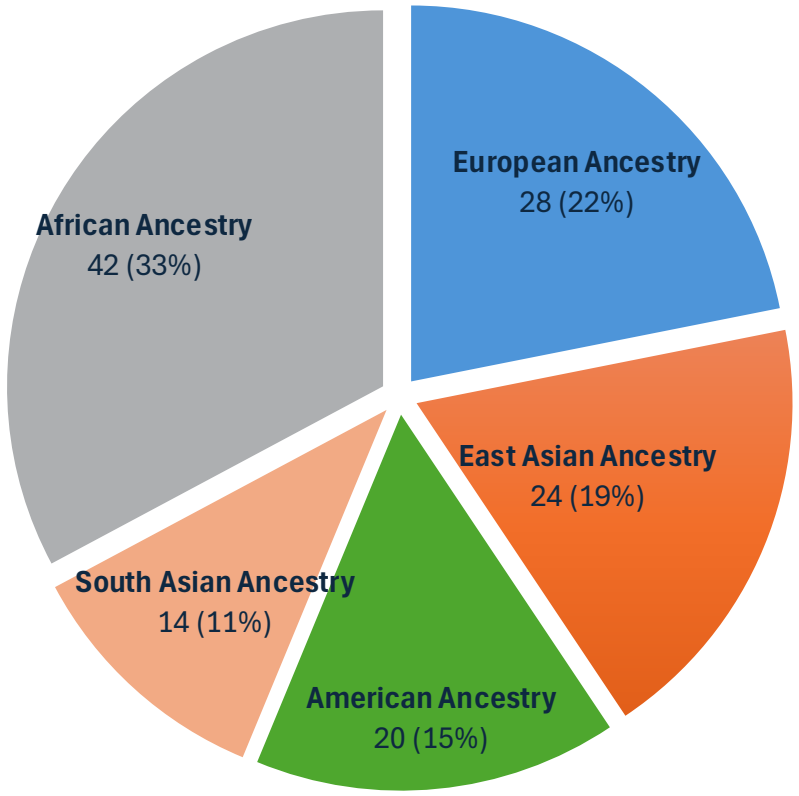


illumina®

For Research Use Only. Not for use in diagnostic procedures.

The most diverse set of data built-in in the next-gen multigenome reference

Enabling improved variant calling accuracy across populations



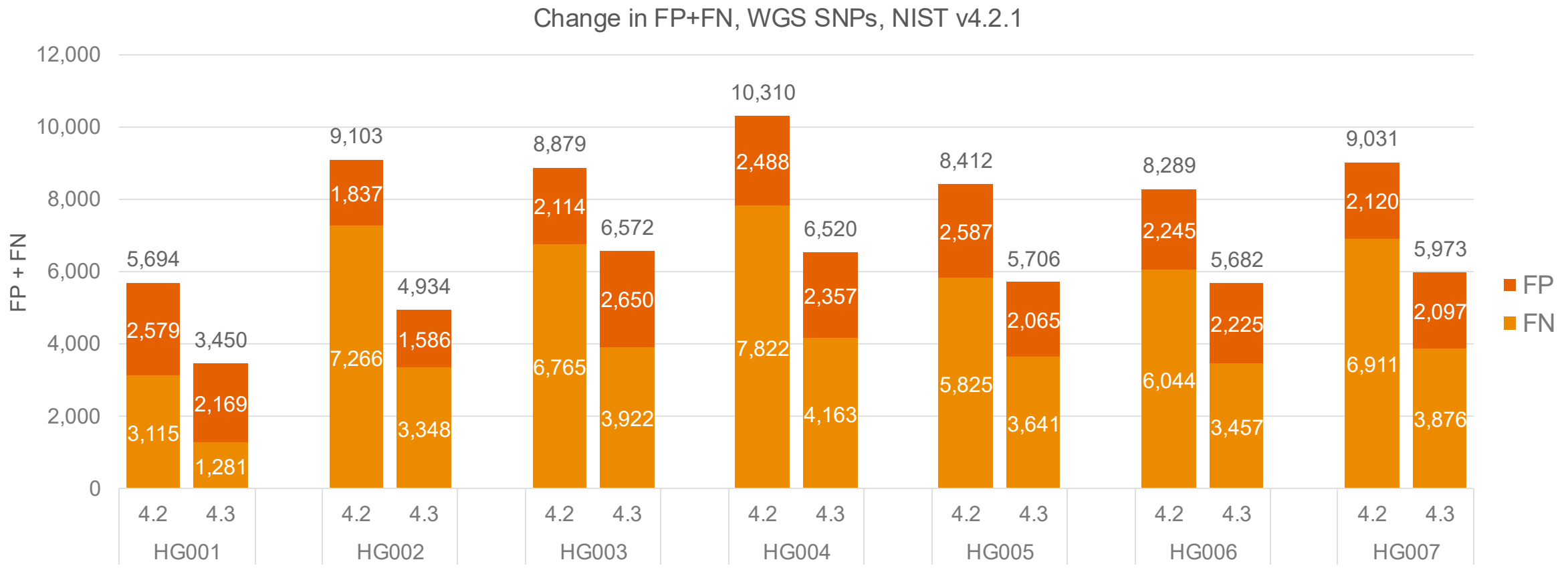
Built from 26 ancestries around the world

Setting standards for population representation across the globe

- Extends sample population to 128 samples of different ancestries around the globe
- Improved mutligenome hash table format, allows scaling to large population panels.
- Pre-built multigenome references for
  - Hg38
  - Hg19
  - CHM13v2.0
  - hs37d5

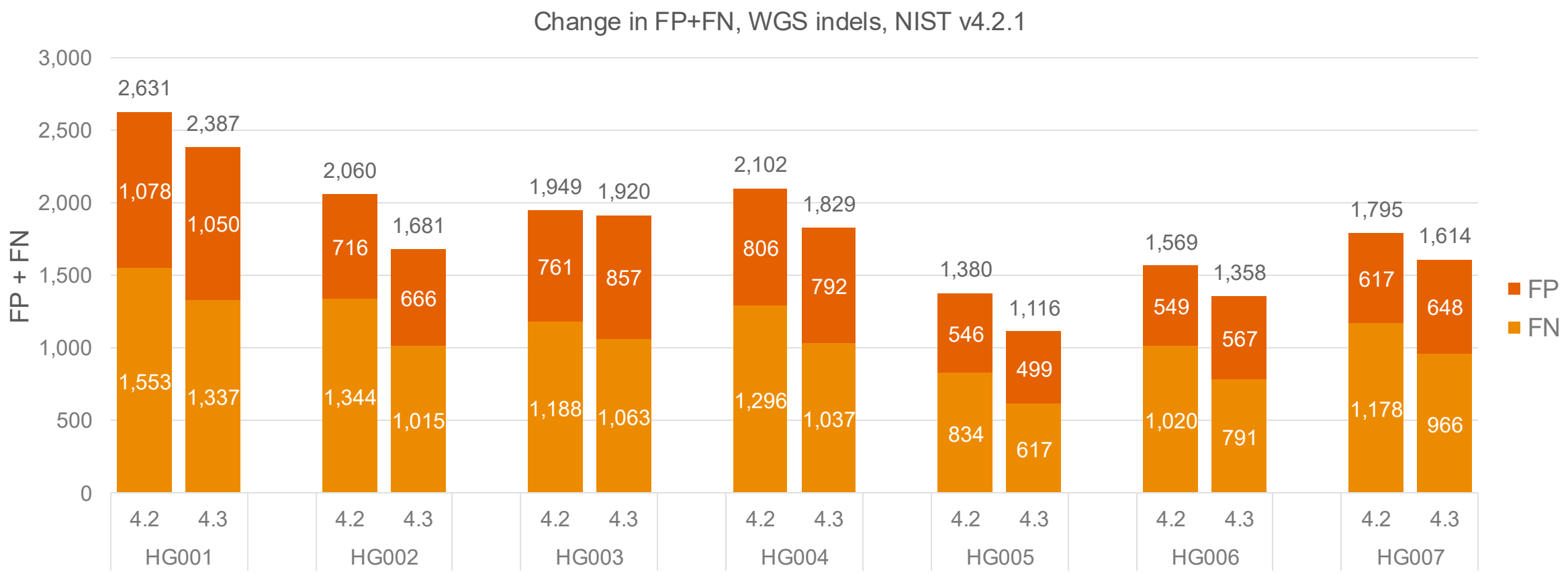
Germline small variant caller – SNP accuracy

Reduce SNV FP+FN by 35% on average cross population



Germline small variant caller - INDEL accuracy

Reduce indel FP+FN by 12.1% on average cross population



Personalized Reference (beta) for small variant calling

DRAGEN builds **a 2-haplotypes personalized reference to impute variants**, use as priors in the Variant Caller, create new personalized ML model

- Reduces FP+FN by 20-30 % for SNPs, 7-8 % for INDELs
- Easy to use end-to end workflow with single flag
- Supports both WGS and WES pipelines

To enable personalized reference:
--enable-personalization=true (default to false)

Causes increased runtime for WGS and WES pipelines. Run time on 30X WGS approx. 45 minutes (small VC only)

Availability


BSSH

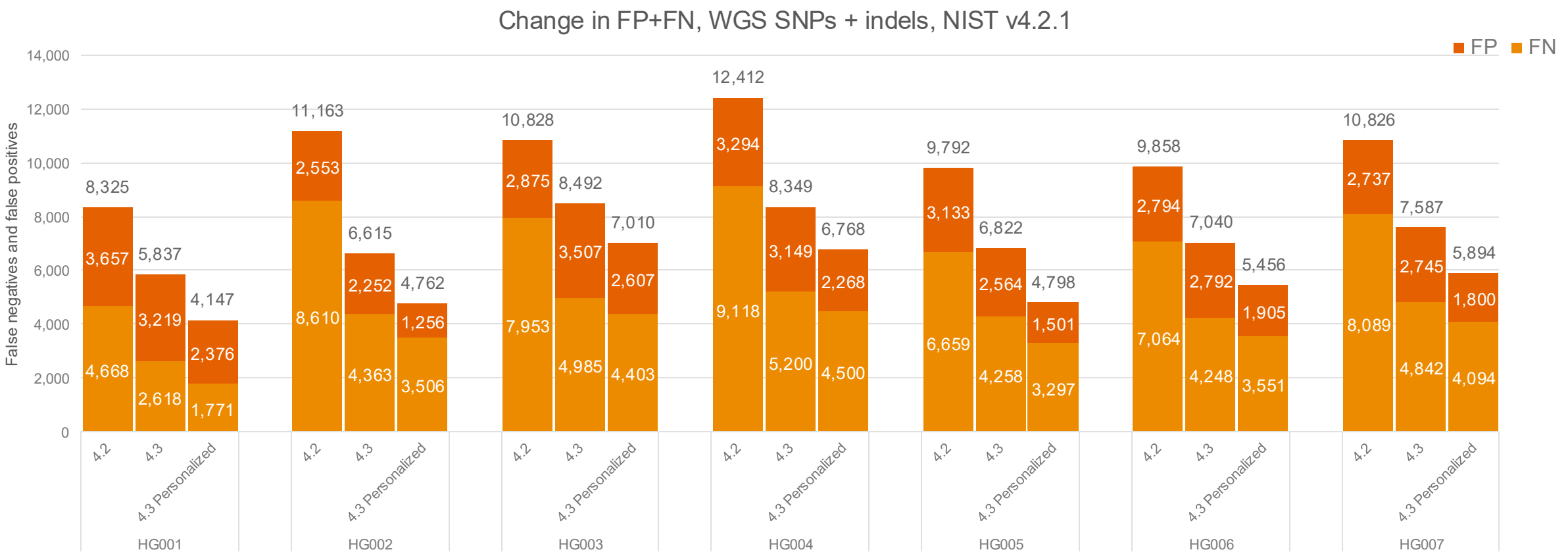

ICA


Multi-Cloud


On-Premise Server

Personalized Germline Small Variant Caller – Accuracy

Personalization reduces FP+FN by 28.7% for SNPs, and 7.9% for indels compared to 4.3 *



* Average FP+FN reductions for 7 standard samples (HG001-HG007), compared DRAGEN v4.3 standard

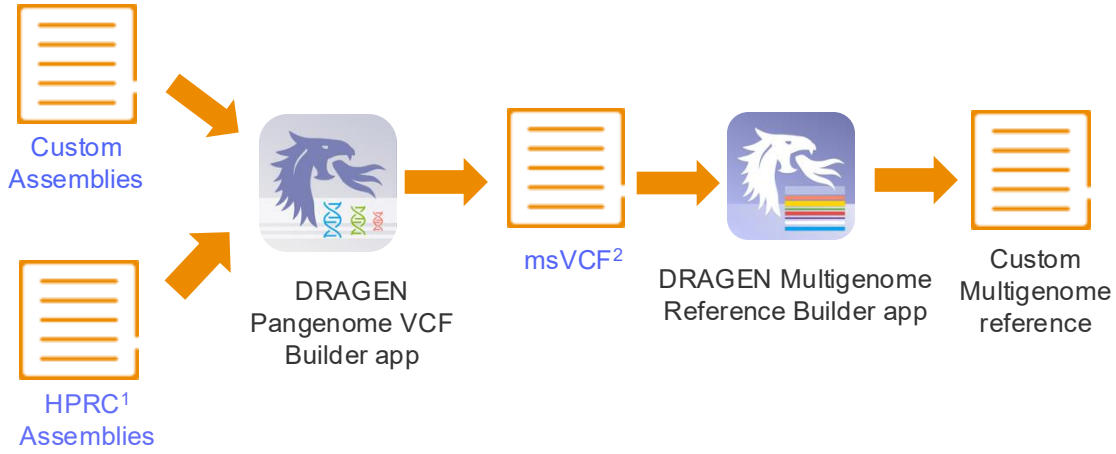
illumina®

DRAGEN provides tools to build custom multigenome references

Build reference genome tailored to your population

- Customers can build high-quality custom multigenome references from pangenome data
- Reduces ethnicity bias and allows customization of multigenome references for population-specific studies
- Two app workflow that allows for graph customization:

1. DRAGEN Pangenome VCF Builder App - generate phased msVCFs from custom assemblies (FASTA)
2. DRAGEN Multigenome Reference Builder App –
   1. Select prebuilt datasets from HPRC
   2. Combine prebuilt HPRC samples and custom samples to build DRAGEN multigenome reference



1. Liao W.-W. et al. A draft pangenome reference. Nature 617, 312–324 (2023).
2. Prebuilt HPRC msVCF is available for selection in DRAGEN Multigenome Reference Builder app

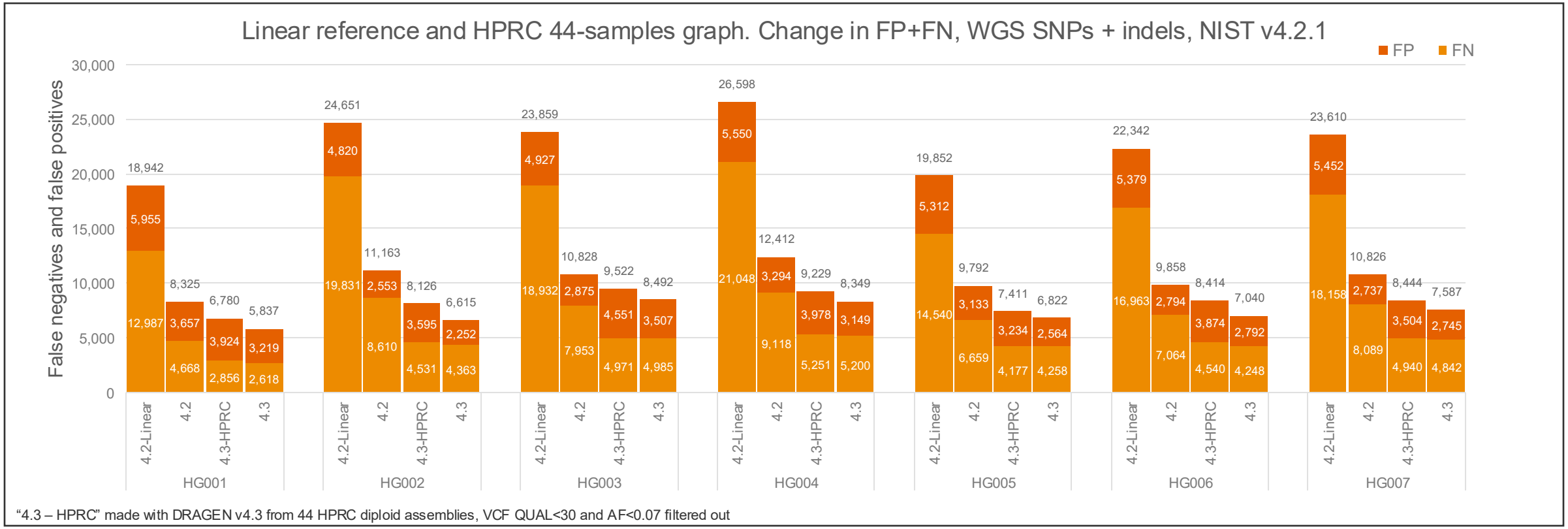
Blue color text indicates customer inputs, custom haplotype-resolved assemblies FASTA, UI input for HPRC

Availability (Illumina cloud)

BSSH ICA

Improve small variant calling accuracy with custom references

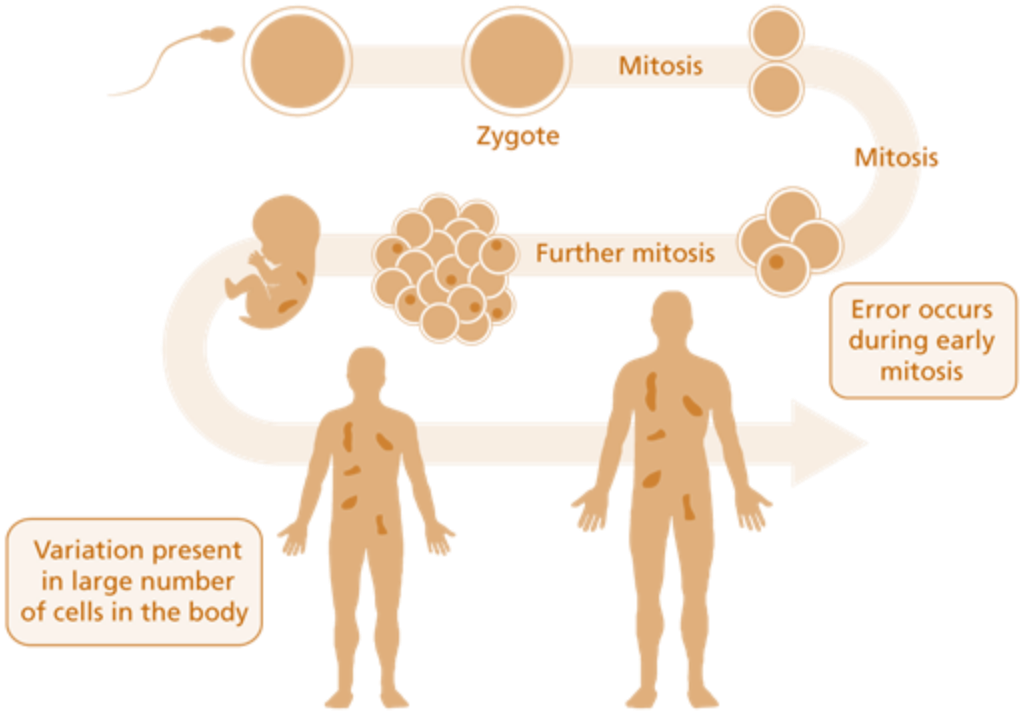
DRAGEN HPRC-based multigenome reference yields better accuracy results than DRAGEN v4.2



"4.3 – HPRC" made with DRAGEN v4.3 from 44 HPRC diploid assemblies, VCF QUAL<30 and AF<0.07 filtered out

Introducing mosaic variant caller

- **Mosaicism** is a postzygotic mutation that leads to some cells in the **body having different DNA than others**
- Mosaic **variants occur at low allele frequencies**, making them hard to detect
- To study the effects of mosaicism on biology and disease, mosaic variants need to be mapped at **high sensitivity and high depth, is compute intensive**



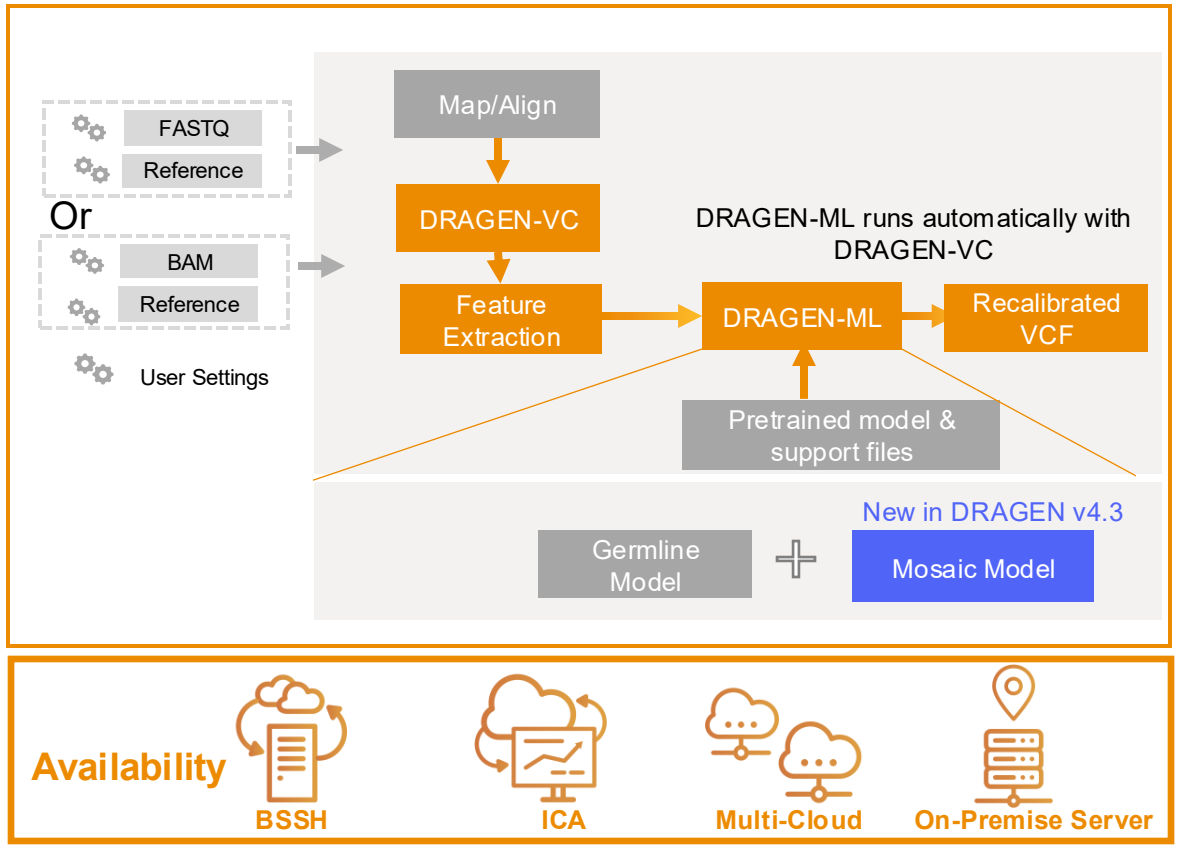
Dr. Siobhan Simpson, Errors during embryonic development, NHS England

Detect Mosaic variants at low allele frequency

Enable **Mosaic variant** detection at low allele frequency

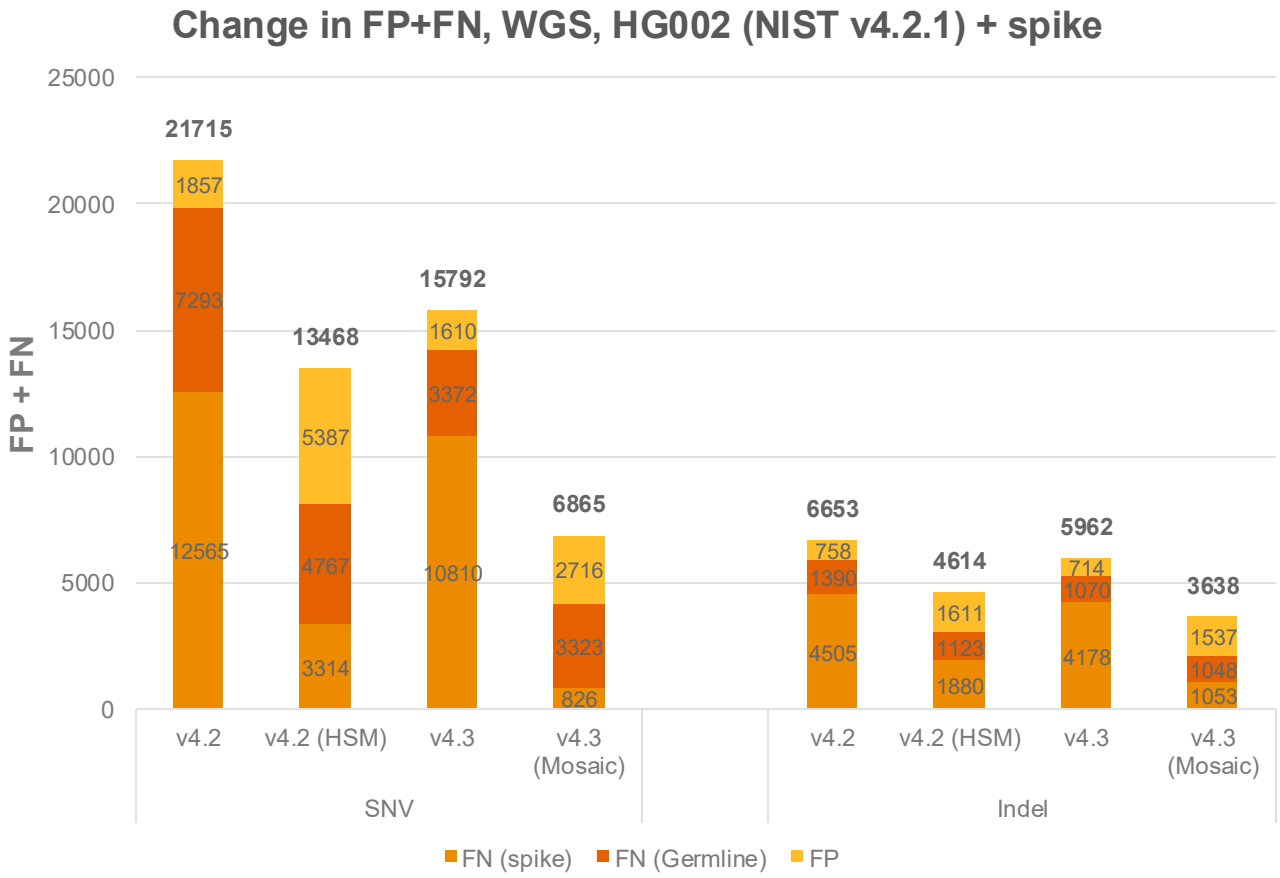
New in v4.3 - integrated mosaic caller

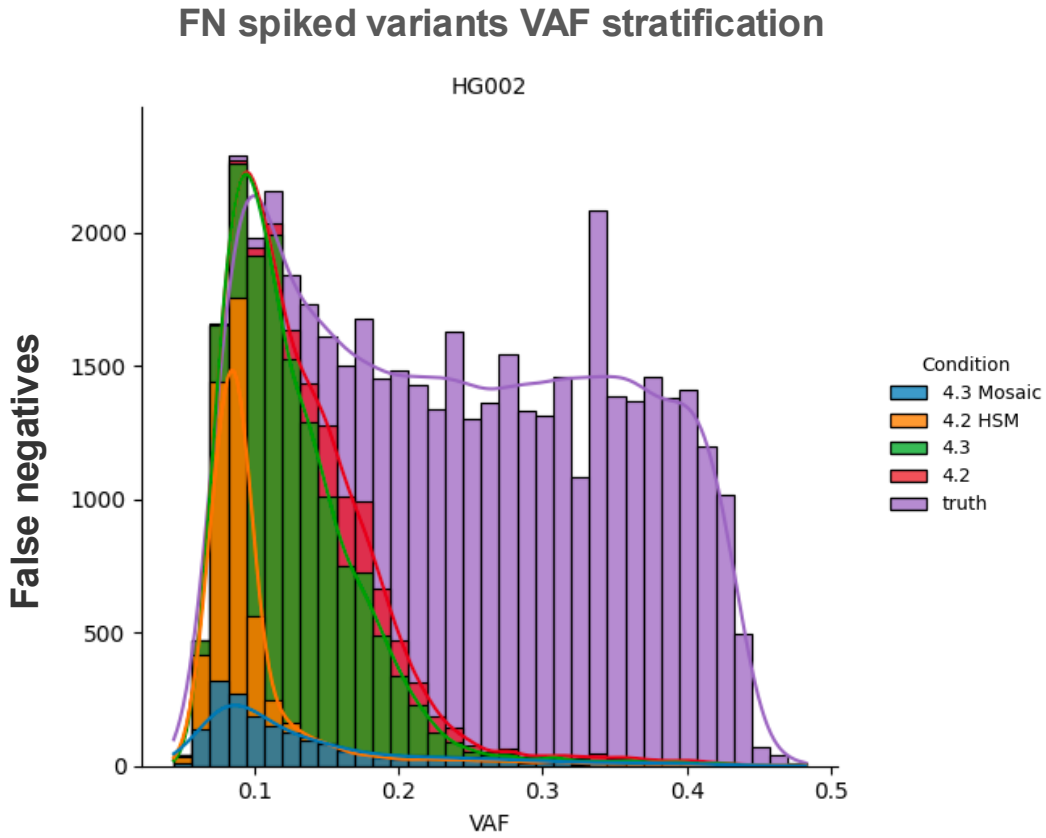
- Integrated single workflow with germline variant calling. Facilitated by Machine Learning.
- By default, small VC reports mosaic variants with >20% allele frequency.
- Low-AF Mosaic Detection Mode (optional): Enables calling of mosaic variant with low allele frequency (custom defined threshold).
- Enabled for both WGS and Enrichment pipelines



Greatly Improved Recall for Low Allele Frequency Variants

Higher sensitivity for variant calls at low AF, with low added FPs



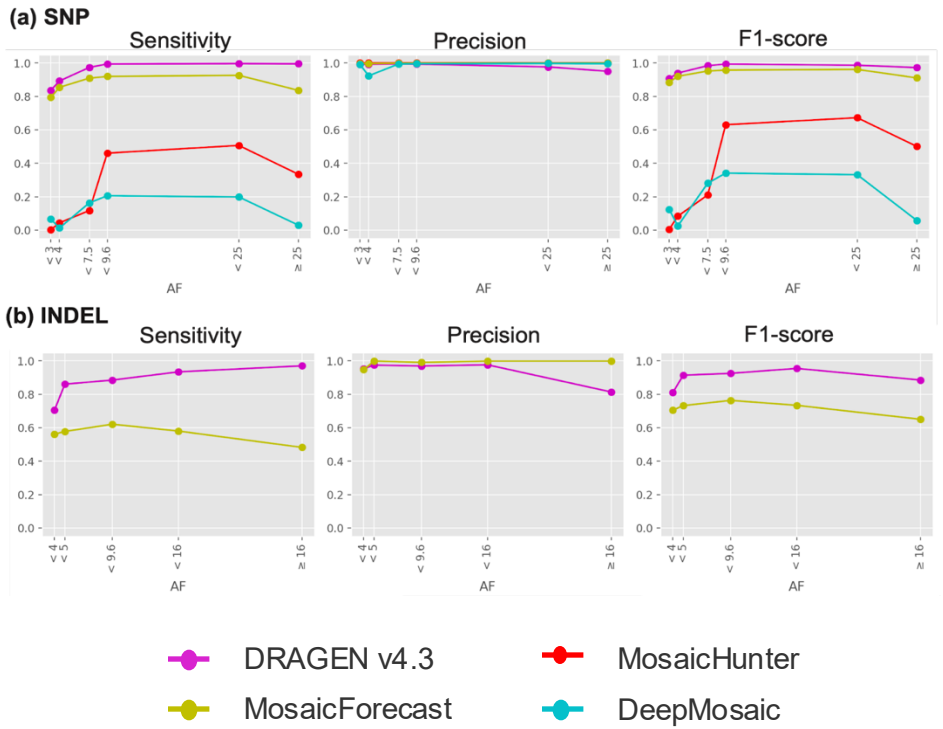


* Tested with DRAGEN v4.2 on NIST HG002 dataset, with 3,936,174 variants of which 45,581 uniformly distributed spiked variants between 5% and 45% VAF at 35x coverage

DRAGEN is more accurate than other mosaic callers

End-to-end analysis with fast turnaround time

Mosaic variant calling accuracy comparison

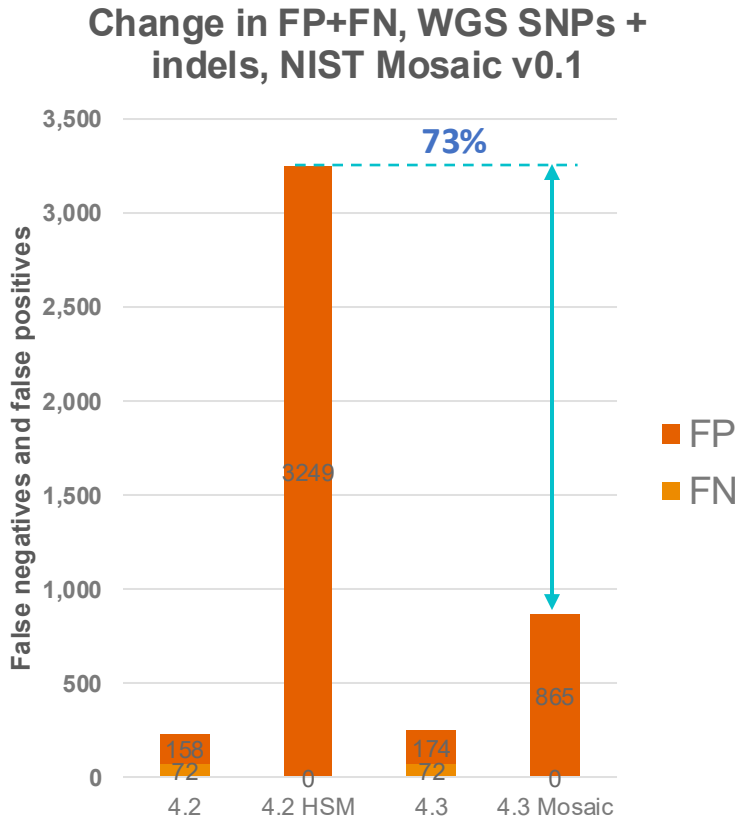


Mosaic variant calling runtime comparison

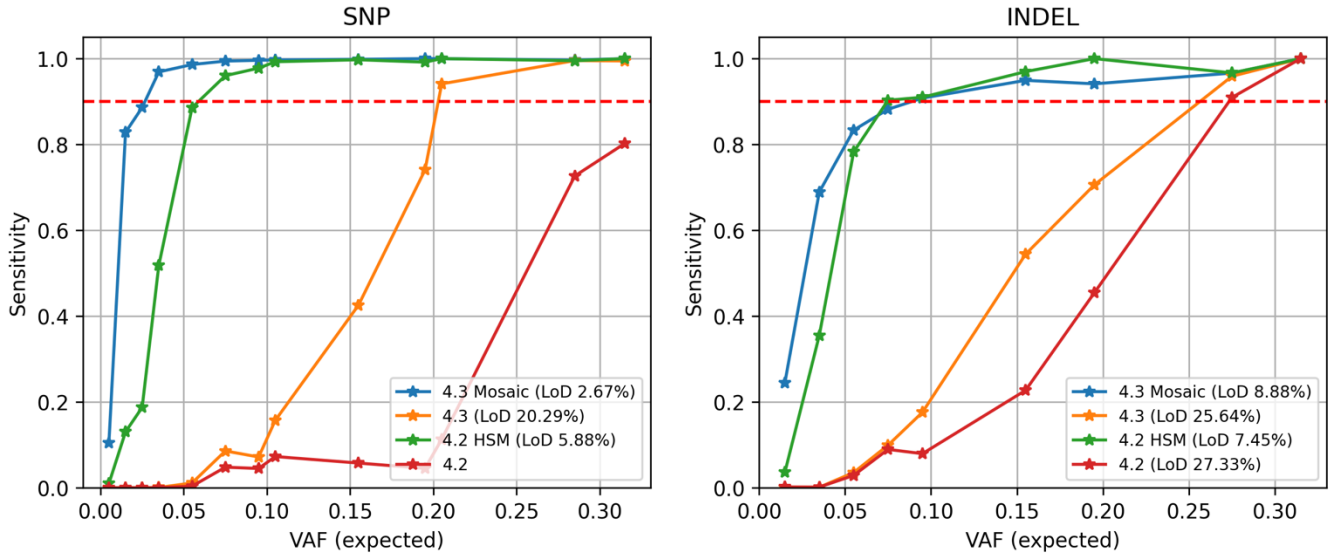
	Runtime (hr)	Hardware
DRAGEN v4.3	0.3	DRAGEN Server v4
DeepMosaic	5.5	Multi-core CPU + NVIDIA A100 GPU
MosaicForecast	12.8	Multi-core CPU

* DRAGEN runtime includes map/align. Other runtimes were for mosaic variant calling only.

Greatly improved precision for high-depth samples with fast turnaround time



Tested on HG002 WGS dataset at 300x with NIST/MDIC *draft* mosaic truth set



Limit of Detection at 2.67% AF for SNPs and 8.8% AF for Indels on 1100x WES admixtures*.

	Runtime
300x WGS	6 hours
1100x WES	20 minutes

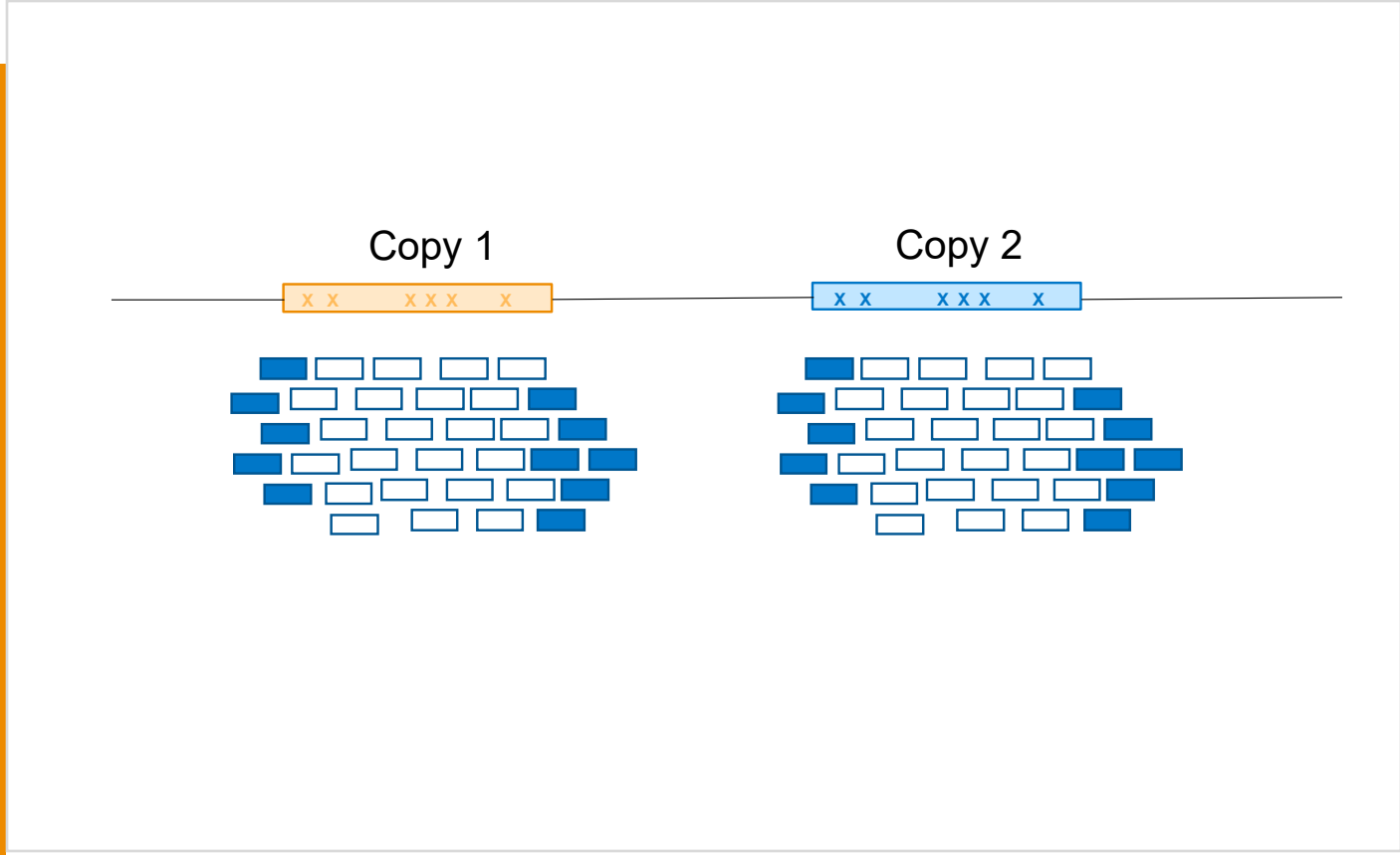
*Ha, YJ., Kang, S., Kim, J. *et al.* Comprehensive benchmarking and guidelines of mosaic variant calling strategies. ***Nat Methods* 20**, 2058–2067 (2023).
*Ha, YJ., Oh, M.J., Kim, J. et al. Establishment of reference standards for multifaceted mosaic variant analysis. **Sci Data 9**, 35 (2022).*

DRAGEN provides accurate genotyping of clinically relevant genes in segmental duplications

- **Segmental duplications** (SegDup) are 1 kb+ regions with ≥ 90% similarity
- Contain sequences that are relevant to disease
- Represent 5% of the genome, and have poor mappability

Genotyping in segmental duplications with DRAGEN v4.3

- ✓ SNV, indel
- ✓ CNV



Multi-region joint detection (MRJD) enables de novo germline small variant calling in paralogous regions

DRAGEN MRJD Caller (*NEW!*)

- ✓ Haplotype-based variant calling from collected reads potentially mapped to SegDup regions
- ✓ Genotyping for 7 medically relevant genes in SegDup regions
- ✓ Runs as standalone pipeline on DRAGEN server (integration with Germline workflow in the future)

Availability


BSSH


ICA

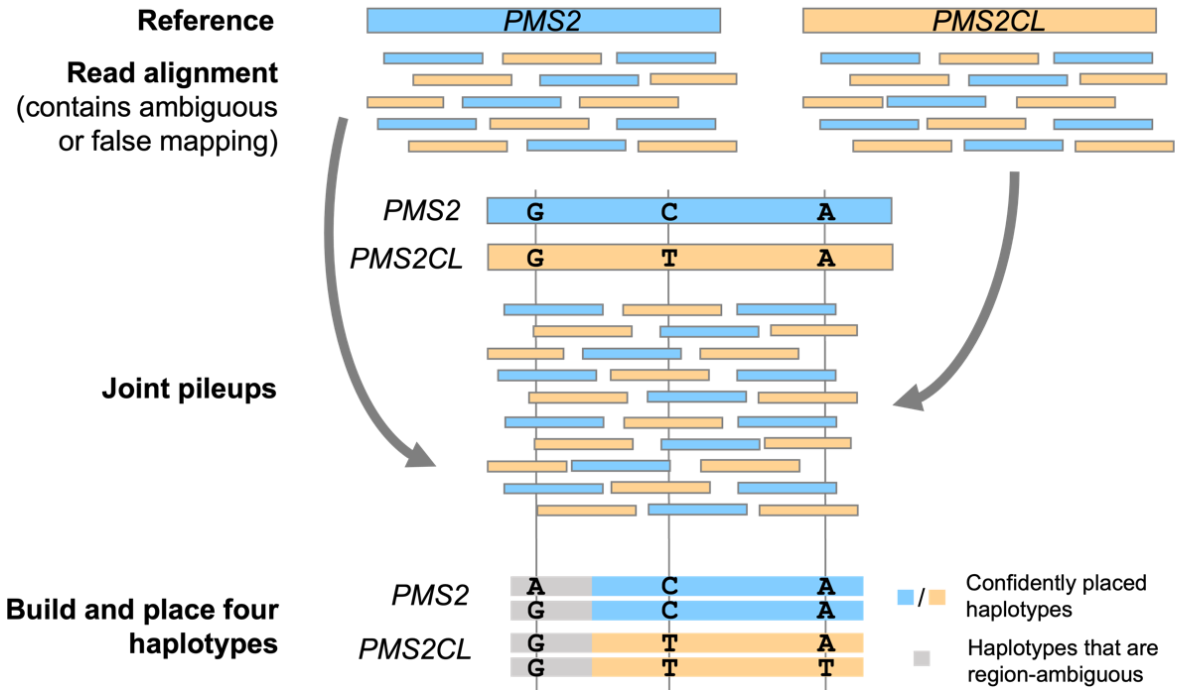

Multi-Cloud


On-Premise Server

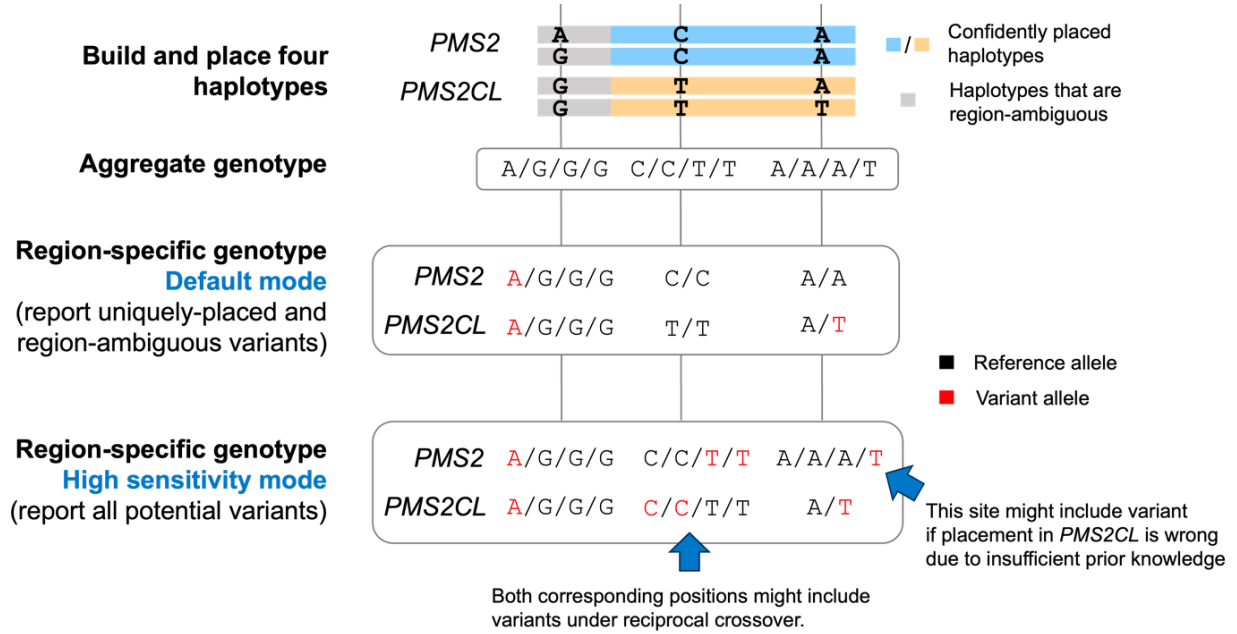
Gene	Application | Conditions
PMS2	Hereditary Cancer Screening | e.g. Lynch Syndrome for Colorectal/Endometrial Cancer
SMN1/SMN2 (small variants)	Carrier Screening | Spinal Muscular Atrophy
STRC	Carrier Screening | Nonsyndromic hearing loss
NEB	Carrier Screening | Nemaline myopathy
TTN	Newborn Screening & Rare Diseases | Cardiomyopathy
IKBKG	Newborn Screening | Incontinentia pigmenti, hypohidrotic ectodermal dysplasia

Multi-region joint detection (MRJD) enables de novo germline small variant calling in paralogous regions

1 Build and place all haplotypes using all reads

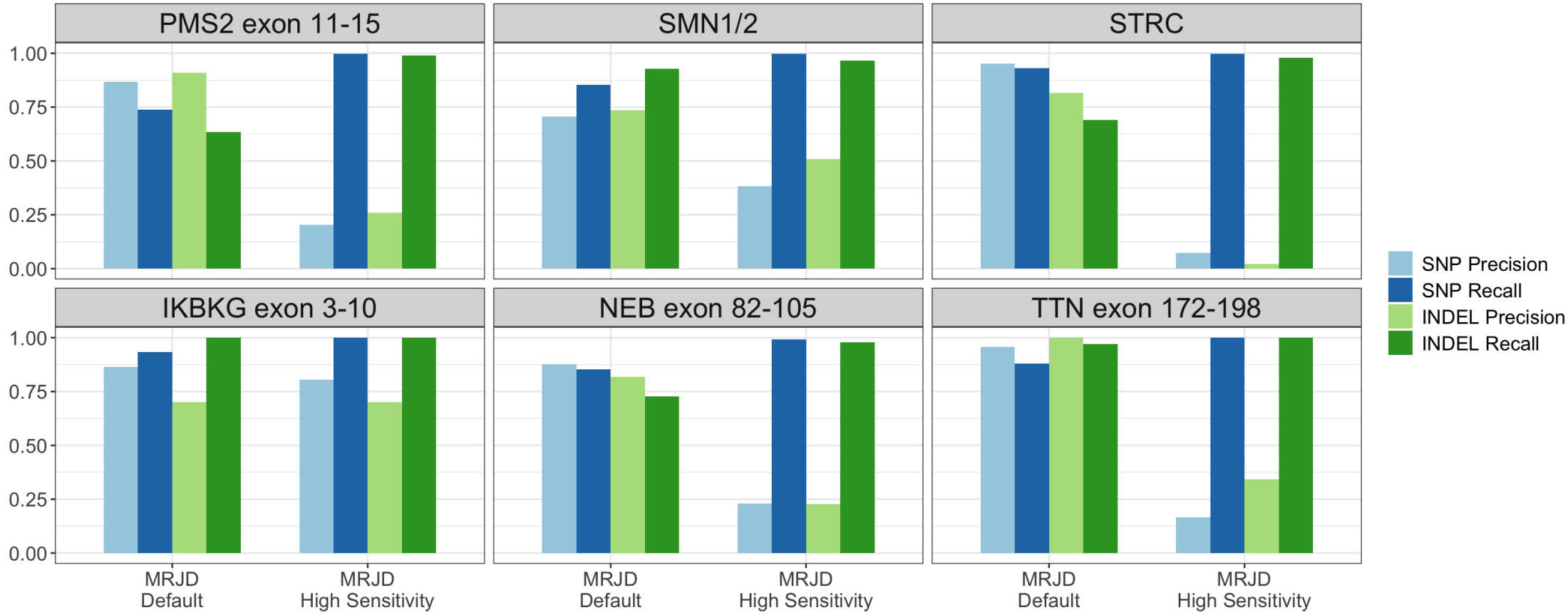


2 Report uniquely placed and region-ambiguous variants



MRJD empowers researches on clinically relevant and challenging genes

- Benchmarked on more than 100 cell lines per gene (using RTG tools squash ploidy mode)
- MRJD High Sensitivity mode offers high recall at the expense of precision, with low spurious call rate



Detect copy number variations in segmental duplications

DRAGEN v4.3 rescues ~1Mbp of CNV bins previously excluded from analysis.

- Improves CNV detection across 43 clinically relevant genes
- Enabled by default for DRAGEN germline WGS analysis with Hg38 reference

Example Command Line:

```
dragen \
-r <HASHTABLE> \
--output-directory <OUTPUT> \
--output-file-prefix <SAMPLE> \
--ref-dir <HASHTABLE> \
--bam-input <BAM> \
--enable-map-align false \
--enable-cnv true \
--cnv-enable-self-normalization true
```

Availability


BSSH

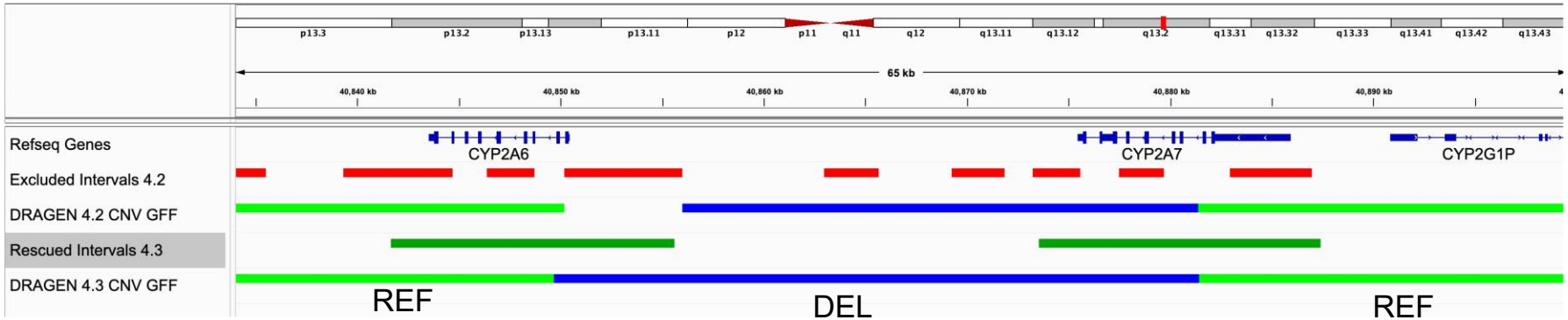

ICA


Multi-Cloud


On-Premise Server

CNV bin rescuing allows for more accurate CNV calls in segmental duplication regions

Experimentally confirmed CYP2A6 – CYP2A7 gene fusion (HG01755) identified by rescuing previously excluded bins



Comparison against experimentally validated CNV calls in segdup genes

	No. samples	Correct by DRAGEN	Method & Citation
CYP2A6	20	20 (100%)	getRM Pratt et al. 2016
RHD/RHCE	40	38 + 1* (97.5%)	Molecular Inversion Probes (MIP) Nuttle et al. 2013
FCGR3A/B	40	39 (97.5%)	TaqMan qPCR Qi et al. 2016

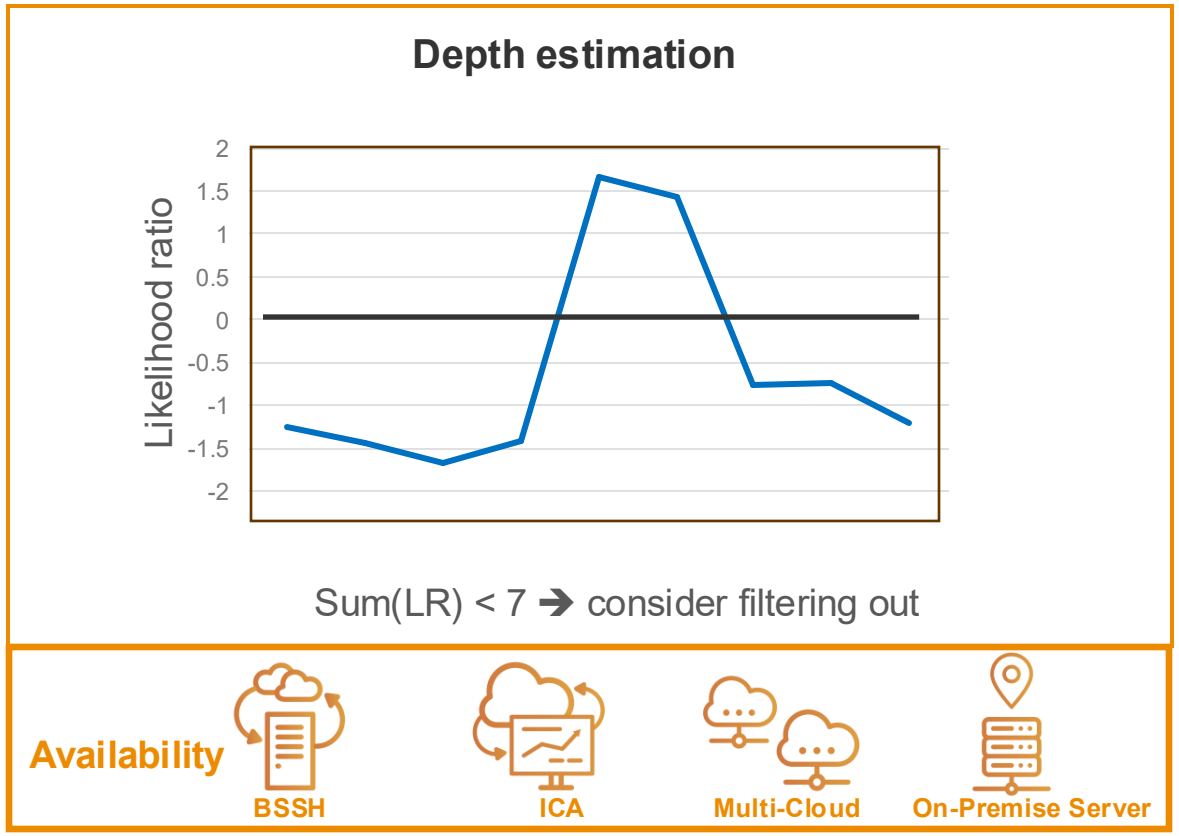
Enhanced CNV calling using WES data

New filters reduce false positive rates

- **Likelihood Ratio** (FORMAT: LR) – Log10 likelihood ratio of ALT to REF
- **Dinucleotide Biases** (FORMAT: GC/CT/AC) – Measure of dinucleotide biases that are outside of typical ranges will be filtered out.

Panel of Normals (PoN) updates

- **Improved PoN validation** – DRAGEN cross checks critical options against the case sample under analysis to ensure matched parameters.
- **PoN Metrics**: Additional panel of normals statistics calculated per target interval, originally introduced in 4.2.



Improved structural variant calling

SV accuracy improvements

- Robust assembly methods in repeat regions **improves insertion recall**[1]
- DRAGEN SV **retains higher recall in Mobile Element Insertions (MEI)** when compared to dedicated methods[2]

1 - Compared against NISTv1.0 HG002 truthset.
2 - Insertions stratified by annotation from Delage et al on NISTv0.6

Run time improvements

- Additional assembler time adds minimal runtime compared to v4.2.
- Significant **reduction in runtime** with CRAM library updates.

Method	MEI Recall
DRAGEN 4.3	**0.885**
DRAGEN 4.2	0.867
Mobster 0.2.4.1	0.756

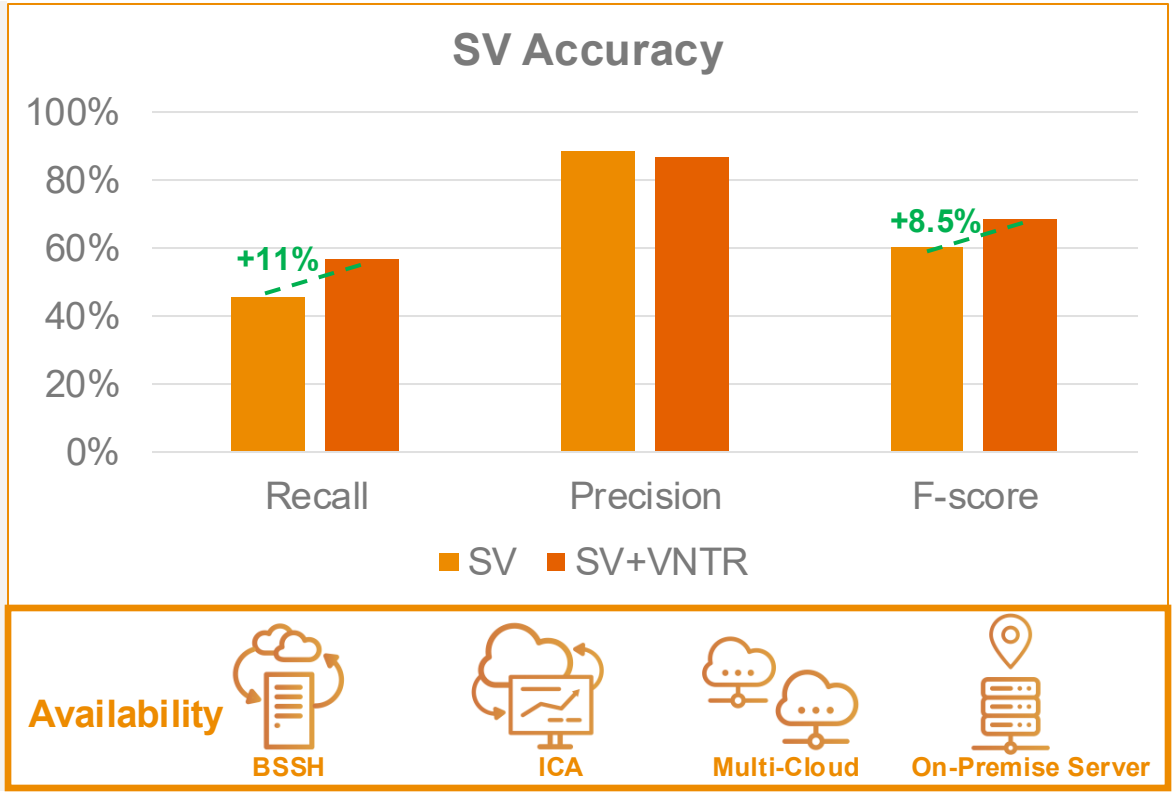
Availability

BSSH | ICA | Multi-Cloud | On-Premise Server

Improve structural variant recall with new VNTR caller

- **Variable Number Tandem Repeats (VNTRs)** – tandem repeat (TR) with pattern size ≥10 bp, larger than STRs.
- New **VNTR caller** for large variant calls in TRs
  - Supports hg38, hg19, GRCh37 references with curated VNTR catalog included
  - Automatically integrates calls into SV VCF
- Improves SV recall by **>10%** *

* Benchmarked using HG002 dataset and hg38 reference



VNTR Caller - Detects expansions and contractions in tandem repeat regions

- Enabled with option *"--enable-vntr true"*
- Reports calls for all TR regions in input catalog bed
- Specialized calling of copy number per region (haplotype-resolved when possible)
- Utilizes full read-fragment information
- Realigns reads to TR sequence as necessary
- Genotyping of top-scoring haplotypes from Bayesian model using classifications of read fragments

DRAGEN-VNTR reports calls following the VCFv4.4 spec:

```
chr1    2141862 vntr_00000216   A
<CNV:TR>,<CNV:TR>       0.946           PASS
SVTYPE=CNV;EVENTTYPE=VNTR,VNTR;SVLEN=634,634;END=2142
496;REFRUC=15.1;RUS=CAGCCGAAGCGCCTCCTTTCACTCCAGGGTCCA
CACATCCAG;RUL=42;RB=508,549;CN=0.801,0.866;CNVTRLEN=-
126,-85;RUC=12.1,13.1;RUCCHANGE=-3.0,-2.0
GT:SVFT:VQ:CN:TOTALRB:TOTALCNVTRLEN:TOTALRUC
1/2:PASS:0.946:1.667:1057:-211:25.2
```

- REFRUC: reference copy number
- RUC: total copy number found per haplotype
- RB: total length of each haplotype

Discover new STR expansions with the STR profiler module

ExpansionHunter de novo is now in DRAGEN with faster compute times and improved ease of use

- Compute sample profiles while mapping:
  - --enable-str-profiler=true
- Compare cohorts of profiled samples:
  - --enable-str-profiler=true
  - --str-profiler-analysis = < casecontrol | outlier >

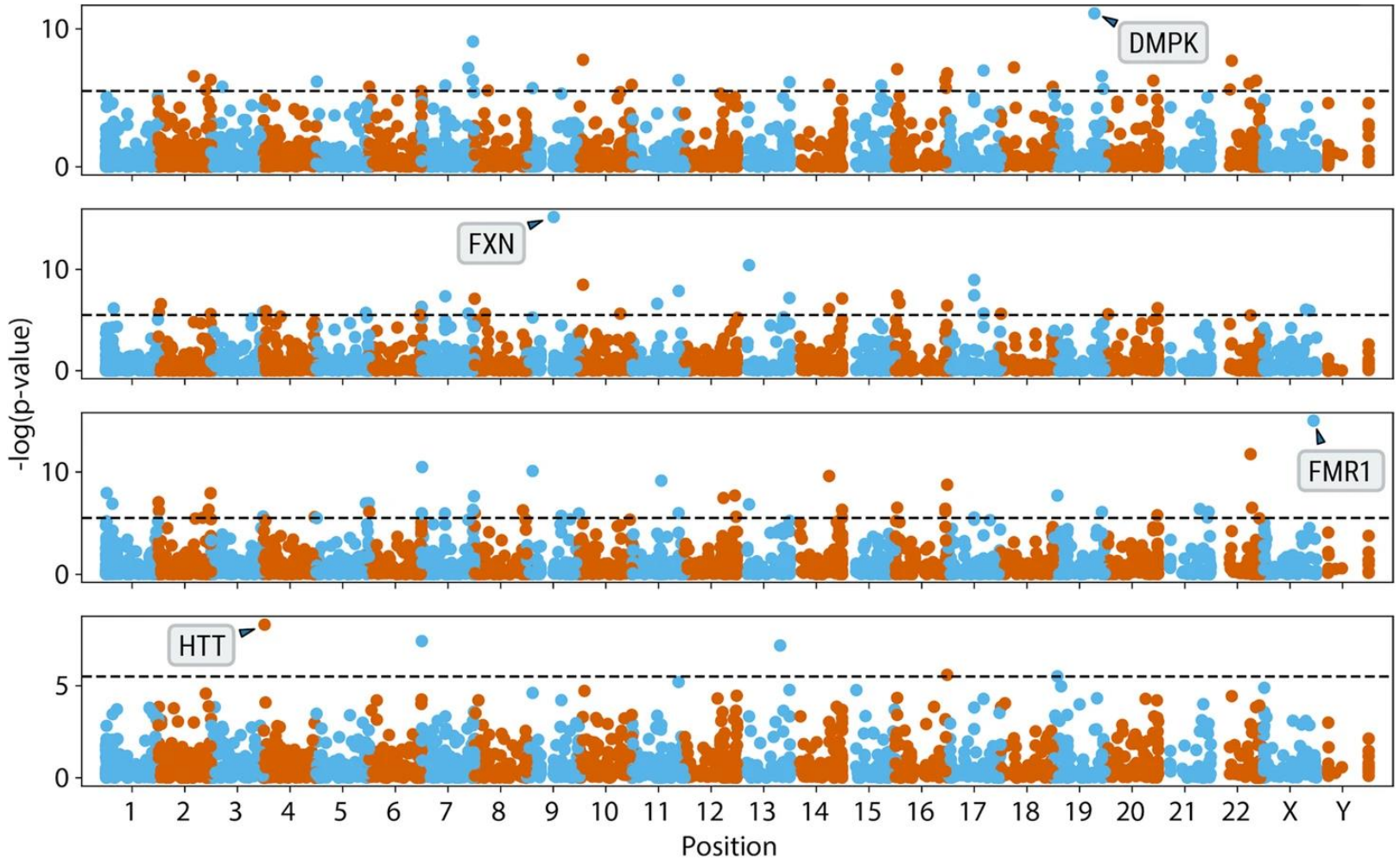
Cases vs controls

contig	start	end	motif	pvalue	bonf pvalue	counts
chr1	1246	1347	AAAG	0.3356	1	[...]

Outlier

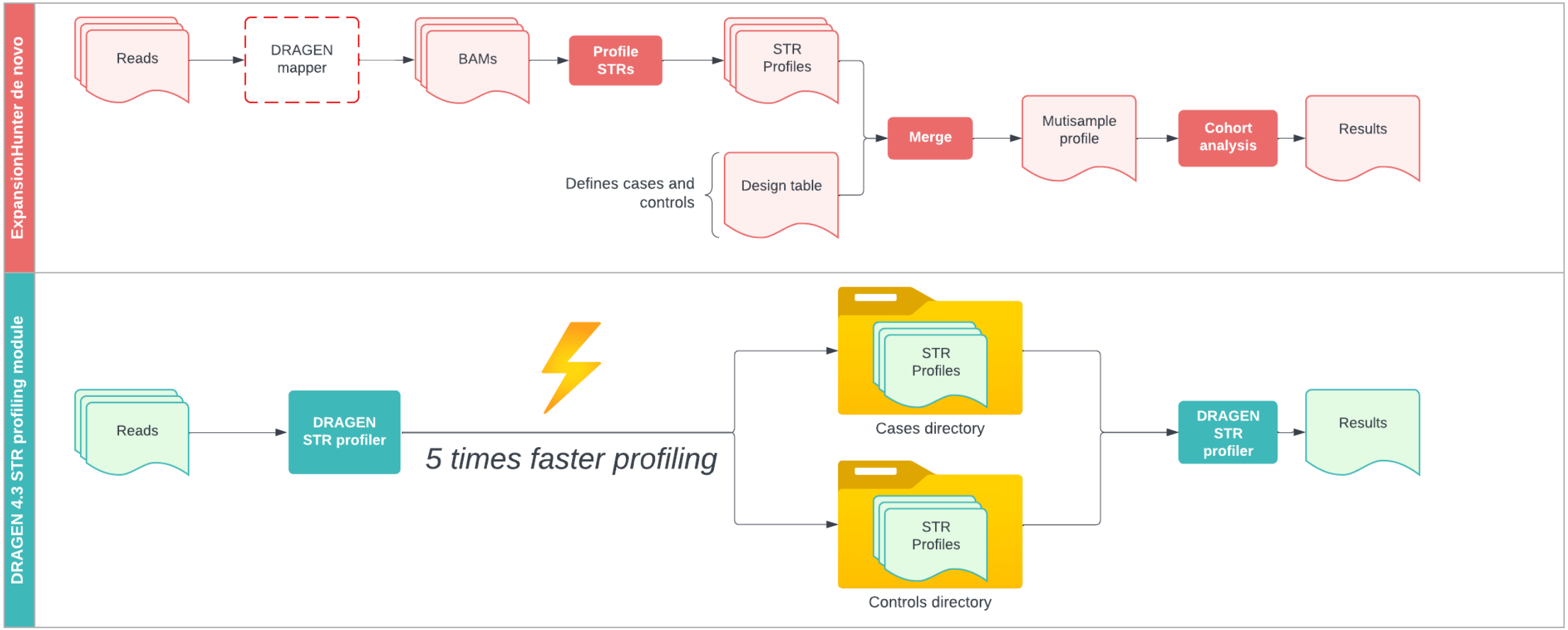
contig	start	end	motif	Top Zscore	Top counts	counts
chr1	1114	1249	ACAG	2.46	5.12	[...]

STR expansions are “rediscovered” in affected individuals



Dolzhenko et al., 2020

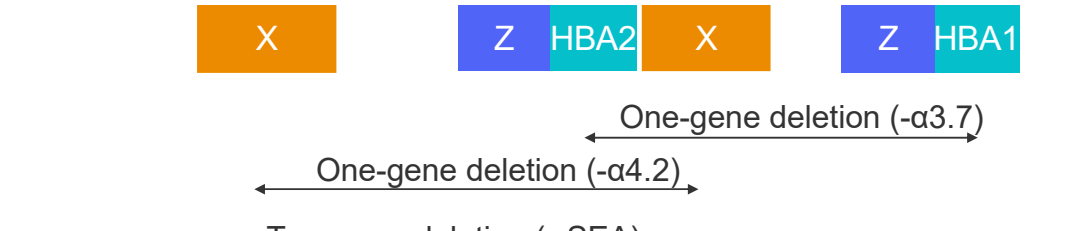
STR profiling module – streamlining the workflow



Unified VCF output for three targeted callers

Improves efficiency for downstream interpretations

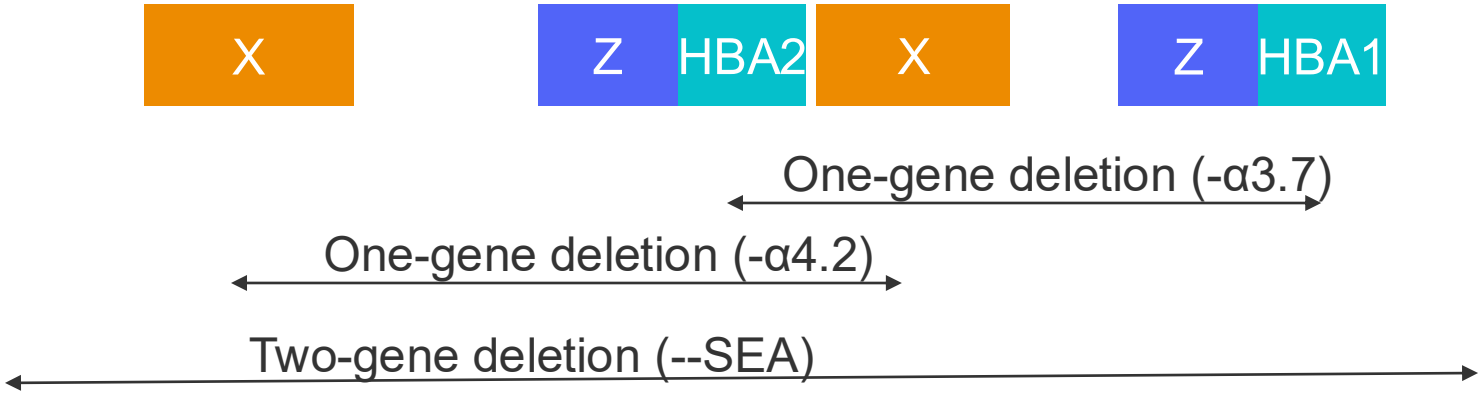
- Enabled for HBA, GBA and CYP21A2
- Records with ambiguously placed variants can be PASS even with low QUAL and low GQ values
- Variants require further assay if placement in target gene is required
- Small Variant VCF Output for Recombinant Haplotypes for GBA and CYP21A2




```
chr6  32008271 . C T 7.65PASS .                                  GT:JGT 0/1:0/0/1/1
chr6  32041006 . C T 7.65PASS ALLELE_ID=.,NM_000500.9:c.1360C>T
GT:JGT 0/1:0/0/1/1
```

HBA Structural Variant VCF Output

- Forced genotyping of all 5 supported structural variants:
  - -α3.7 DEL and ααα3.7 DUP
  - -α4.2 DEL and ααα4.2 DUP
  - --SEA DEL
- ALLELE_ID INFO field used to label variant alleles
- SVModelQual FILTER applied when data does not fit any of the supported genotypes



Targeted calling for these 3 genes is enabled by default when small VC is enabled (only for WGS).


```
{
 "sampleId": "NA19138",
 "hba": {
  "totalCopyNumber": 2,
  "genotype": "-a3.7/-a4.2",
  "genotypeQuality": 64,
  "genotypeFilter": "PASS",
  "variants": []
 }
}
```


```
chr16  165397 . A <DEL>      0.00  LowQUAL END=184700;ALLELE_ID=.,--         GT 0/0
chr16  170262 . G <DEL>,<DUP> 86.17 PASS    END=174517;ALLELE_ID=.,-a4.2,aaa4.2 GT 0/1
chr16  173301 . A <DEL>,<DUP> 86.17 PASS    END=177104;ALLELE_ID=.,-a3.7,aaa3.7 GT 0/1
```

VCF Output for Region-Ambiguous Small Variants

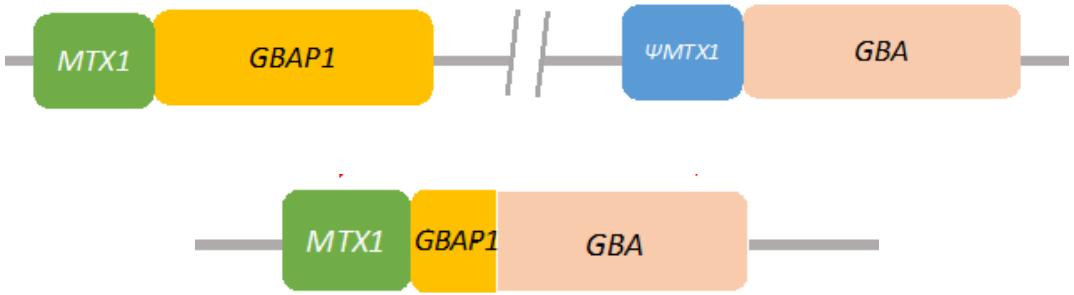
- Enabled for HBA, GBA and CYP21A2
- Records can be PASS even with low QUAL and low GQ values
- These variants require further assay if placement in target gene is required
- VCF format consistent with MRJD region-ambiguous variants
- HGVS identifiers reported in ALLELE_ID INFO field
- Reported genotypes are consistent with any overlapping SV DELs detected by the caller
- Polyploid genotypes and associated quality scores for "joint" analysis of homologous sites are reported in separate VCF fields (e.g. JGT, JGQ, JPL)


```
{
 "sampleId": "WGS1000-P4-A5",
 "cyp21a2": {
  "totalCopyNumber": 4,
  "variants": [ {
   "alleleId": "NM_000500.9:c.1360C>T",
   "alleleCopyNumber": 2,
   "genotypeQuality": 18,
   "filter": "PASS"
  } ]
}
```


```
chr6  32008271 . C T 7.65     PASS .                          GT:JGT 0/1:0/0/1/1
chr6  32041006 . C T 7.65     PASS ALLELE_ID=.,NM_000500.9:c.1360C>T  GT:JGT 0/1:0/0/1/1
```

Small Variant VCF Output for Recombinant Haplotypes

- Enabled for GBA and CYP21A2
- Gene conversions reported as small variants in target gene rather than SV breakends
- Deletion-like recombinant haplotypes reported as small variants in target gene with overlapping deletion alleles in pseudogene rather than SV DEL




```
{
 "sampleId": "HG00422",
 "gba": {
  "totalCopyNumber": 3,
  "deletionBreakpointInGene": true,
  "recombinantHaplotypes": [
   "RecNcil",
   ""
  ],
 }
}
```


GBAP1 deletions (note: variant allele G at first site due to wrong reference base in hg38):

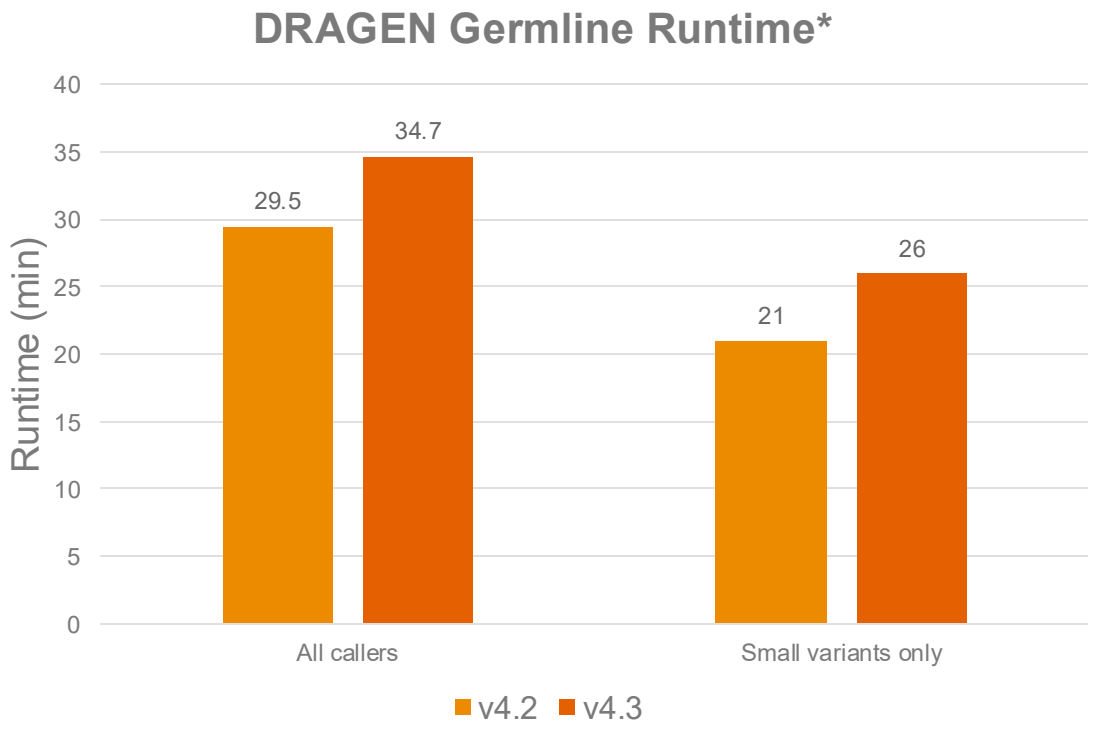
```
chr1 155214590 . C G,* 150.00 PASS    .                    GT 1/2
chr1 155214625 . G A,* 150.00 LowVQL .                     GT 0/2
```

GBA1 variants:

```
chr1 155235217 . C G,* 150.00 PASS ALLELE_ID=.,NM_000157.4:c.1483G>C,. GT 0/1
Chr1 155235252 . A G,* 150.00 PASS ALLELE_ID=.,NM_000157.4:c.1448T>C,. GT 0/1
```

DRAGEN runtime updates

More accurate, more comprehensive, with small runtime increase



* Average runtime for HG001-HG007 (NIST v4.2.1)

Approx. 15-20% runtime increase compared to v4.2
Includes mapping with multigenome reference and ML calibration

- **26 minutes*-** Map, align, and small variant call a 40x human genome
- **34 minutes -** Call the full suite of structural variants, repeat expansion, and targeted callers

Cohort analysis using gVCF Genotyper

illumina®

Improved multi-sample VCF output

New options to compress and customize multi-sample VCF (msVCF) output

- Sparse compression of msVCF output (Michael Lin, 2020), yields 20-30x size reduction for large cohorts
- Easy customization of msVCF output, easier ingestion into third-party tools
- Many more options available, please refer to user guide

To enable sparse msVCF output :

```
--gg-output-type=spVCF
--gg-squeeze-msvcf=true
```

How to customize msVCF output:

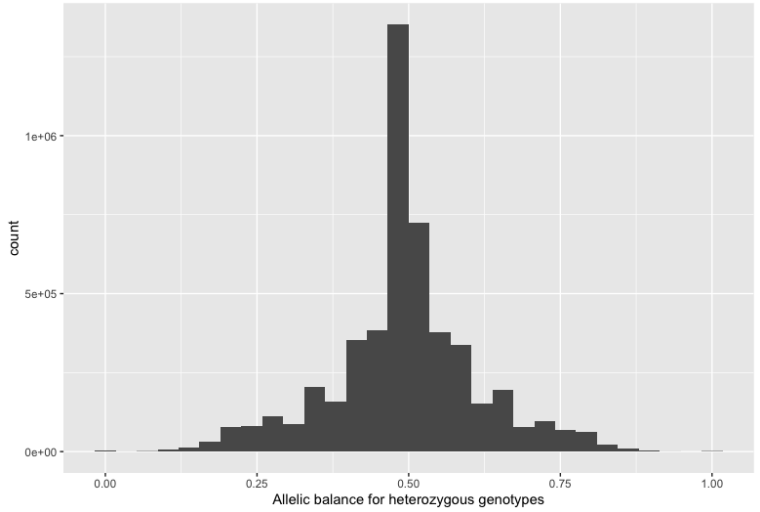
```
--gg-msvcf-info-fields=AC;AN;NS;NS_GT;NS_NOGT
--gg-msvcf-format-fields=GT:LAD:LPL:LAA:QL
```

Availability

New msVCF metrics in gVCF Genotyper

New metrics enable filtering for highly accurate call sets

Allelic balance can identify incorrect Genotype calls



Genotypes with unusual allelic balance are candidates for filtering

To enable output of allelic balance in the msVCF:

--gg-msvcf-info-fields=ABHom;ABHet;ABHetP

To filter msVCF based on maximum P-value of allelic balance:

--gg-hard-filter="ABfilter:all:ABHetP < 0.05"

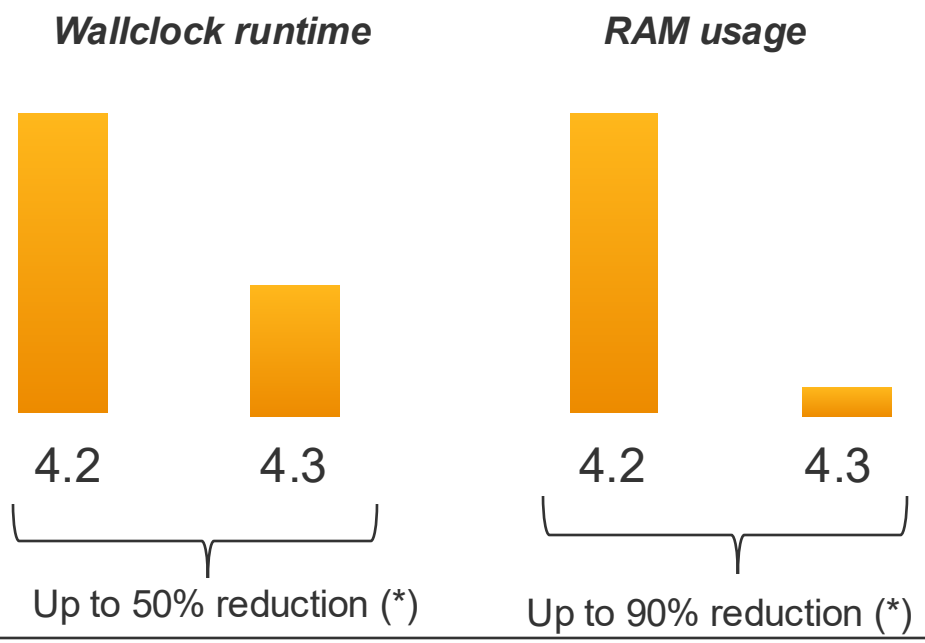
For further details, please refer to the DRAGEN user guide, section "Joint Analysis of Multiple Samples"

illumina®

Performance improvements enable Biobank analysis

New gVCF Genotyper callset released for the UK Biobank

Improved msVCF generation in DRAGEN 4.3



- 500k UKBB WGS samples were re-processed with DRAGEN ML recalibration and gVCF Genotyper on ICA
- Aggregated 1.4 billion variants. 44% were singleton SNPs with 57% novel to dbSNP and 81% with AF< 1/100K
- Workflow ran in ICA with 874K jobs and took less than 3 months

See also:
https://www.illumina.com/company/news-center/feature-articles/uk-biobank-500000-milestone.html

(*) best case for large cohorts

illumina®

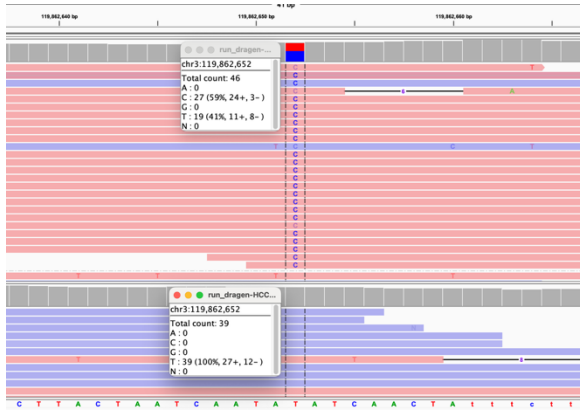
DRAGEN Somatic

illumina®

General somatic VC improvements

- **Columnwise detection** is now enabled by default, improving small variant recalls in repetitive regions
- Somatic hotspots are now allele-specific, reducing FP impact of the hotspots feature

Sample variant rescued by columnwise detection:



- Repetitive local sequence leads to cycles in graph assembly of reads and may result in missed variants
- Columnwise detection identifies variants directly from the BAM pileups to supplement the graph

Availability

BSSH | ICA | Multi-Cloud | On-Premise Server

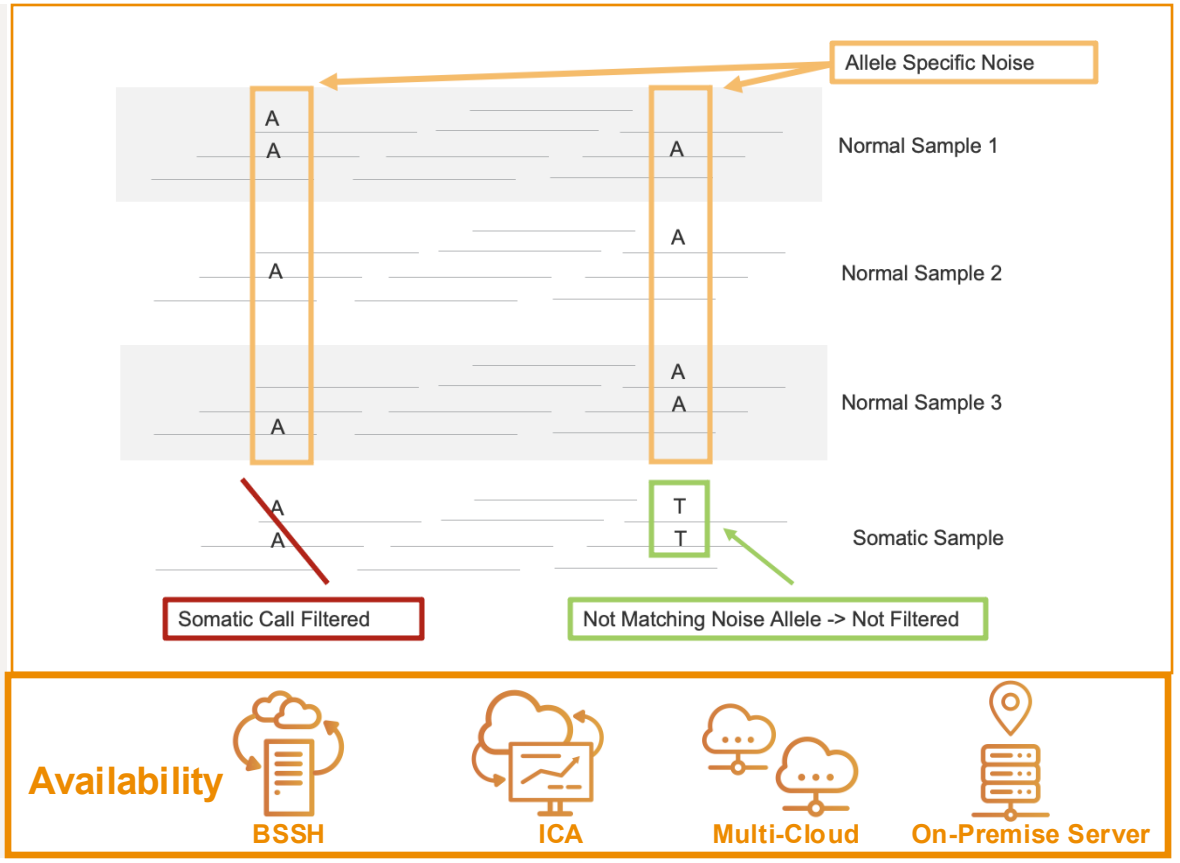
Allele specific systematic noise files for somatic variant calling

- New v2 noise file format encodes mean and max noise levels
- Allele-specific filtering
- Generate custom systematic noise files
  - ***Run normals using tumor-only workflow in sensitive mode***

```
--vc-detect-systematic-noise
# optional
--vc-detect-systematic-noise-mode=UMI
```

  - ***Aggregate Noise***

```
--build-sys-noise-vcf-list
# panels & WES: mean, WGS: max
--build-sys-noise-method=max
```



ASCN calling in somatic WES analysis

Extends ASCN detection in tumor-only mode

Allele-specific copy number (ASCN) analysis extends the utility of copy number alteration by enabling the detection of loss-of-heterozygosity (LOH)

- Available in both **tumor-only (new!)** and **tumor-normal** analysis, enabling **purity estimation**, reporting of **LOH** regions, and **HRD** Scoring
- Requires panel-of-normals for exome analysis

Example DRAGEN command:

```
--tumor-bam-input=<TUMOR_BAM>
--cnv-population-b-allele-vcf=<CNV_POP_VCF>
--enable-cnv=true
--cnv-target-bed=<BED>
--cnv-normals-list=<PON>
```

HCC1395	Metric	DRAGEN 4.2 WES T/O	DRAGEN 4.3 WES T/O	DRAGEN 4.3 WES T/N
Deletions	Recall	0.397	**0.982**	**0.987**
	Precision	0.017	**0.912**	**0.958**
	F-score	0.033	**0.946**	**0.972**
Duplications	Recall	0.400	**0.972**	**0.968**
	Precision	0.996	**0.991**	**0.990**
	Fscore	0.571	**0.981**	**0.979**

Availability

BSSH

ICA

Multi-Cloud

On-Premise Server

New DUX4 caller in somatic WGS analysis

DUX4 gene related rearrangements detection

DUX4 gene (DUX4-r) rearrangements are involved in a subtype of acute lymphoblastic leukemia (ALL)

- DUX4 rearrangement detection enabled by machine learning
- Demonstrated high sensitivity and specificity
  - **100% Recall** (52 cases of IGH::DUX4 fusions and 1 case of IGH::QSOX1::DUX4 fusion*)
  - **100% Precision** (FP = 0)

* Validated with collaborators

To enable DUX4 caller :

--enable-dux4-caller=true

Example output from dux4.vcf.gz:

```
chr3    75667931        ExtendedDUX4:IGH:Bnd_W  T       T[chr14:105586938[      .       .
    SVTYPE=BND;CIPOS=.,.;EVENTTYPE=TRA;MATEID=ExtendedDUX4:IGH:Bnd_X;TotalReadsNum=9061
GT:PR:SR:SRPB   1/1:1:0:110363.09
chr3    75667932        ExtendedDUX4:IGH:Bnd_V  G       ]chr14:105586937]G      .       .
    SVTYPE=BND;CIPOS=.,.;EVENTTYPE=TRA;MATEID=ExtendedDUX4:IGH:Bnd_U;TotalReadsNum=9061
GT:PR:SR:SRPB   1/1:1:0:110363.09
```

```
* Supported references: hg38
* Supports tumor-only analysis
```

Availability

BSSH

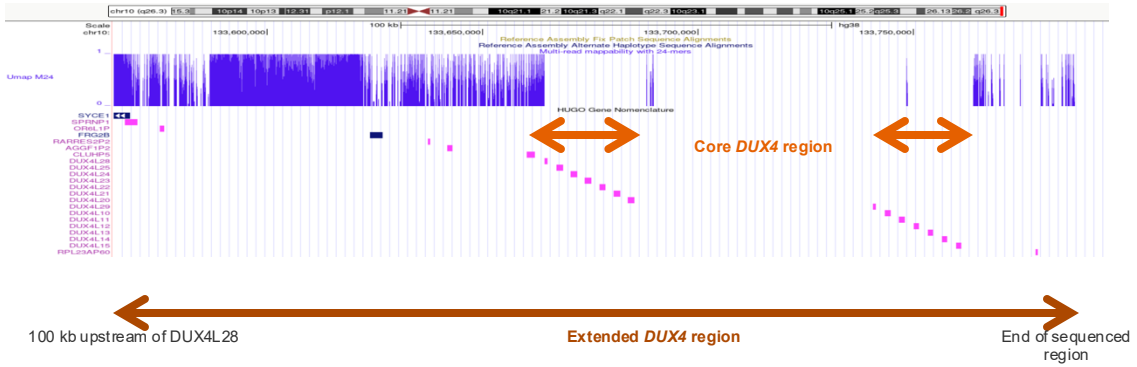
ICA

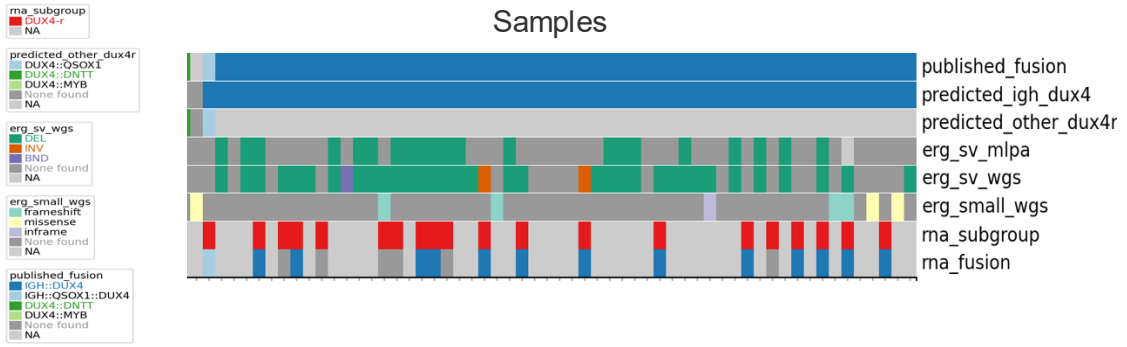
Multi-Cloud

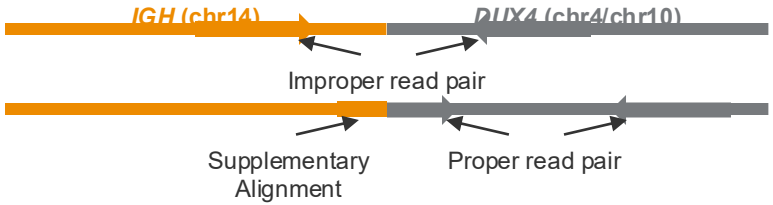
On-Premise Server

DUX4 caller

DUX4 gene related rearrangements detection from WGS data







Paired spanning read

Split spanning read

- Define customized DUX4 calling regions
- Generate hypothesis for somatic-only WGS by detecting evidential reads spanning affected genes
- Call DUX4 events with ML models that are validated by orthogonal experiments

illumina®

DRAGEN RNA

illumina®

Improved RNA fusion detection by updated machine learning

Extensively trained machine learning model improves RNA fusion detection accuracy

- Updated non-linear model with more features
- Significantly enhanced truth set, using simulations and real data
- Additional datasets cross library preps (WTS, mRNA, Panels), and read lengths, tissue types, sample types, etc.
- Pure score-based filtering enables better recall/precision tradeoffs

Attribute	Old Model (pre-4.3)	New Model (4.3)
Model	Linear	Non-linear
Features Used	8	39
Truth Description	Gene-pair only	Genes + breakpts
Total Truth Set Size	223	>12,000
PASS/FAIL Filters	Score + 5 post-filters	Score only

Availability

BSSH

ICA

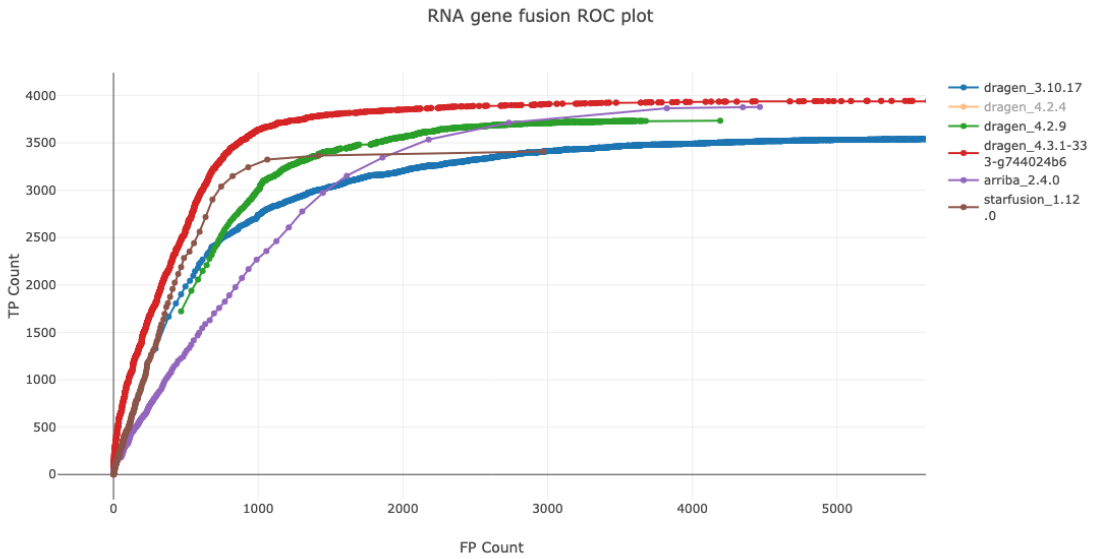
Multi-Cloud

On-Premise Server

Improved RNA fusion detection by updated machine learning

Extensively trained machine learning model improves RNA fusion detection accuracy

- Updated non-linear model with more features
- Significantly enhanced truth set, using simulations + real data
- Additional datasets cross library preps (WTS, mRNA, Panels), as well as read lengths, tissue types, sample types, etc.
- Pure score-based filtering enables better recall/precision tradeoffs



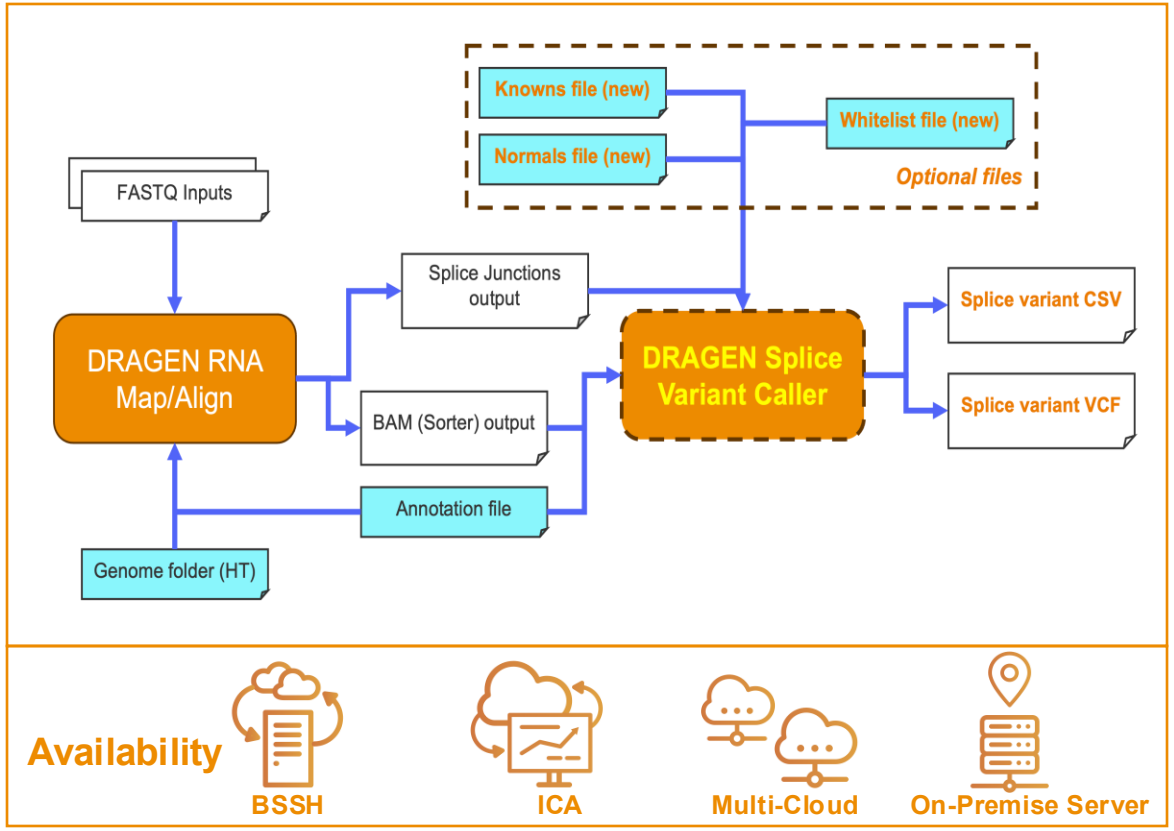
Availability

BSSH | ICA | Multi-Cloud | On-Premise Server

New DRAGEN RNA Splice Variant Caller (beta)

Integrates the caller used in DRAGEN TSO500

- Enables detection of alternative splicing or intra-genic fusions
- Targets well-known splice variant biomarker events*:
  - EGFR VIII
  - MET exon14
  - AR-V
- General (non-targeted) option available with increased false positives



* Requires resource files. See user guide for details

RNA Splice Variant Caller Resource Files

Required Resource Files

- Anchored-RNA reference HT (for mapper)
- Gene Annotation file (i.e. GTF or GFF)

Optional (recommended) New Resource Files

- *rna-splice-variant-knowns*: Knowns SJ file (“whitelist”)
- *rna-splice-variant-normals*: Normals SJ file (”blacklist”)
- *rna-splice-variant-regions:* BED file for making calls

Outputs

Column	Column Name	Description
1	gene_start	Semicolon-delimited list of genes containing junction start
2	gene_end	Semicolon-delimited list of genes containing junction end
3	chromosome	Chromosome for the junction
4	start	Junction start position
5	end	Junction end position
6	strand	Junction strand
7	motif	Junction's intron motif
8	annotated	True only if the splice junction is annotated
9	total_reads_ref	Unique non-supporting reads
10	total_reads_alt	Unique variant (ALT) supporting reads
11	max_spliced_alignment_overhang	Maximum spliced alignment overhang
12	score	Junctions score based on read evidence

<prefix>.splice_variants.csv

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE
chrX	66905968	.	G	<DEL>	0	LowQ;LowUniqueAlignments	SVTYPE=DEL;END=66913996;ALTDEDUP=0;ALTDUP=0;I	AD:DP	0:00
chrX	66905968	.	G	<DEL>	1	PASS	SVTYPE=DEL;END=66914514;ALTDEDUP=411;ALTDUP=	AD:DP	411:9240
chrX	66905968	.	G	<DEL>	0	LowQ:LowUniqueAlignments	SVTYPE=DEL:END=66915499:ALTDEDUP=3:ALTDUP=3:I	AD:DP	0.17291667

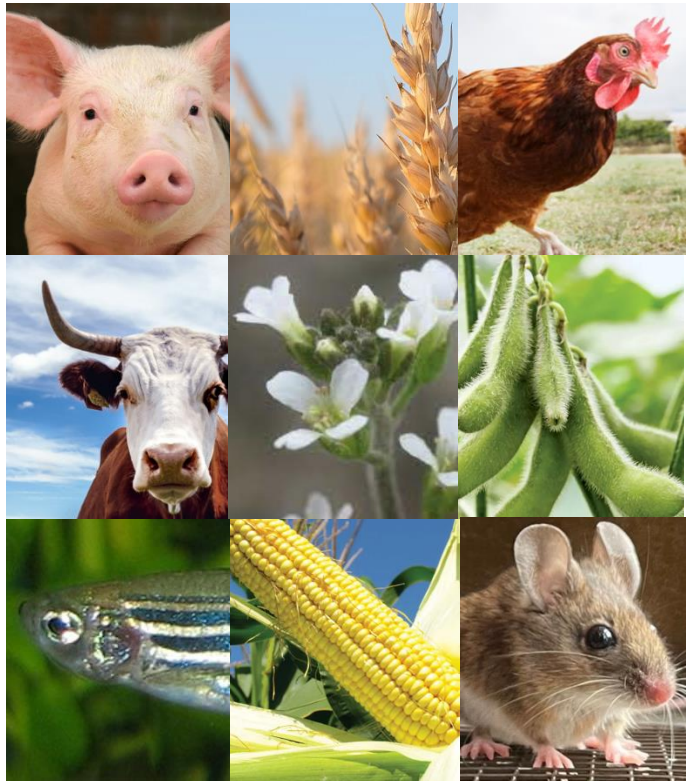
<prefix>.splice_variants.vcf.gz

Availability

BSSH | ICA | Multi-Cloud | On-Premise Server

ORA Compression

illumina



Compress non-human data at high compression ratio

List of supported species :

- ✓ Homo sapiens – bisulfite (methylated data)
- ✓ Sus scrofa (pig)
- ✓ Gallus gallus (chicken)
- ✓ Oryza sativa (Japanese rice)
- ✓ Arabidopsis thaliana
- ✓ Triticum aestivum (bread wheat)
- ✓ Bos taurus (cattle)
- ✓ Glycine max (soybean)
- ✓ Rattus norvegicus (Norway rat)
- ✓ Zea mays (maize)
- ✓ Danio rerio (zebrafish)
- ✓ Mus Musculus (house mouse)
- ✓ Caenorhabditis Elegans (roundworm)

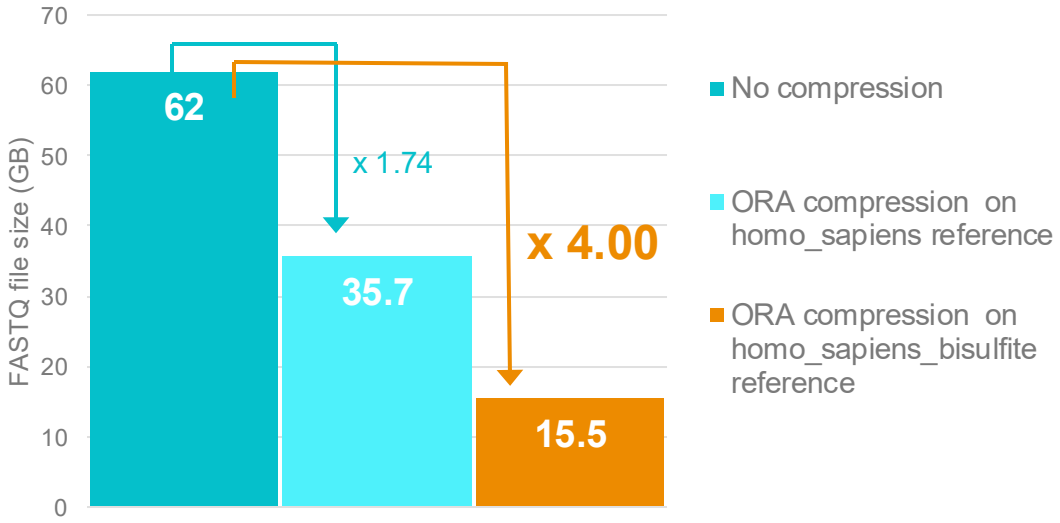
Reference files are available on DRAGEN product file page

ORA Compression supports more datatypes and species

Reduces file size and storage cost by up to 75%

Compression improvement on human bisulfite data

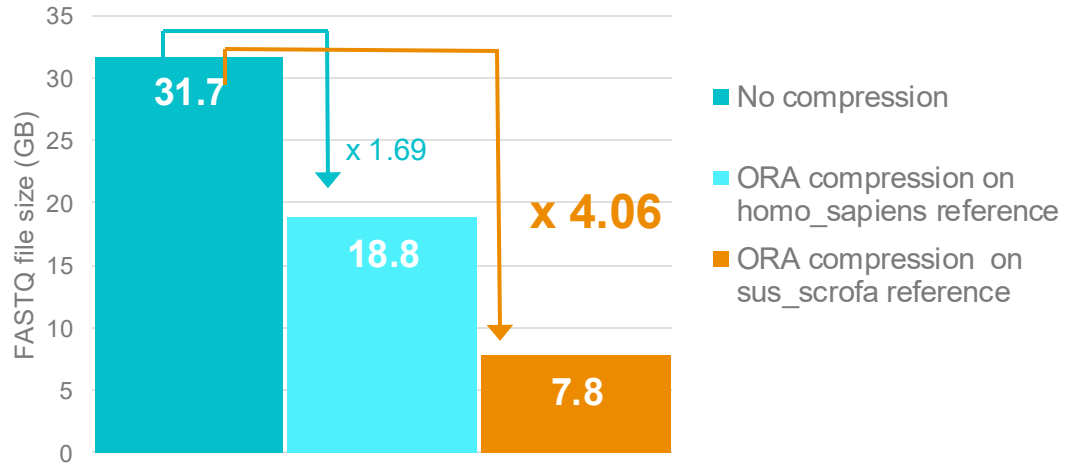
dataset: B-Hela_S2 WGS 45x sequenced on NovaSeq6K, compression on an interleaved mode



Runtime ORA compression on homo_sapiens_bisulfite reference, DRAGEN 4.3, 16 threads: **10 min 25s**

Compression improvement on pig data

dataset: SRR14901873 WGS 25x sequenced on NovaSeq6K, compression on an interleaved mode



Runtime ORA compression on sus_scrofa reference, DRAGEN 4.3, 16 threads: **4 min 35 s**

DRAGEN Imputation & Population Haplotyping

illumina

Boost small variant calling accuracy from low-pass data

- New reference panel (IRPv2.1, human hg38) improves indel imputation accuracy
- Supports up to 100 input fastq samples or 1000 input vcf samples[1]
- New option to batch imputation samples for a quicker turnaround time
- Build custom reference panels[2]

1. FASTQ inputs are supported by pipelines on BaseSpace Sequence Hub and Illumina Connected Analytics
2. Available on ICA only (Population Haplotyping pipeline)

	IRPv2.1	IRPv2.0
Data source[1]	3,202 samples	3,202 samples
Multi-allelic SNP positions	Yes	Yes
INDELs	**Yes – all INDELs**	Yes - AF>3%
ChrX	Yes	Yes
Total Number of Variants	**125,715,255**	111,279,429

1. Data from 1000 Genome Project, processed with DRAGEN gVCF Genotyper v4.0

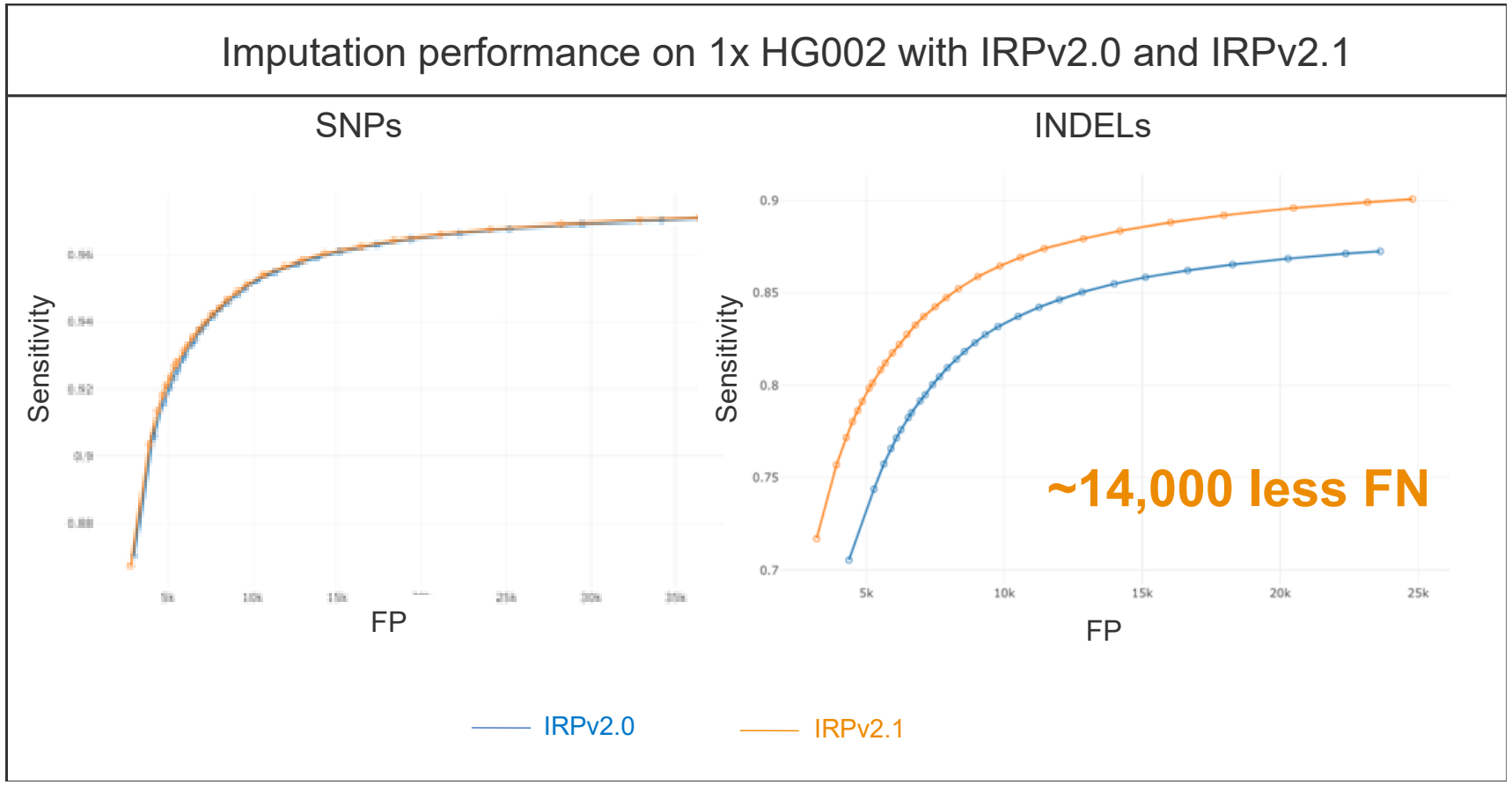
Availability

BSSH

ICA

On-Premise Server

Greatly improved indel imputation accuracy



DRAGEN Imputation and Population Haplotyping

Handle batch of samples through an end-to-end workflow on BSSH and ICA

Population Haplotyping pipeline (ICA)

builds haplotypes from a set of population samples

(Integration of ShapeIT5[1] into DRAGEN)

Use case: build a custom reference panel for Imputation

- User-friendly: multi-step pipeline integrated into an end-to-end workflow on BSSH and ICA, tuned for optimization of runtime and accuracy
- Leverage the multi-node infrastructure of ICA

1. Hofmeister, R.J., Ribeiro, D.M., Rubinacci, S. *et al.* Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet* **55**, 1243–1249 (2023). https://doi.org/10.1038/s41588-023-01415-w

Availability

ICA

On-Premise Server[2]

Imputation pipeline (BSSH/ICA)

infers variants of low-pass sequencing data

- Supports 100 input fastq samples or 1000 input vcf samples.
- Pre-built human reference panels (hg38) or custom reference panel
- Infer multi-allelic SNPs and INDELs on autosomes and chromosome X
- New option to batch imputation samples for a quicker turnaround time
- New IRPv2.1 available on 4.3 to with better accuracy on INDELs vs IRPv2.0
- Estimated cost for imputation pipeline from fastq (incl. FGT): 5 icredits/sample

Availability

BSSH

ICA

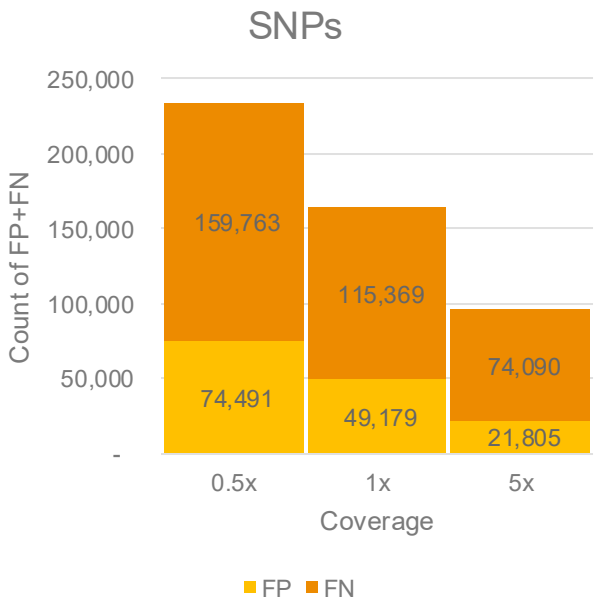
On-Premise Server

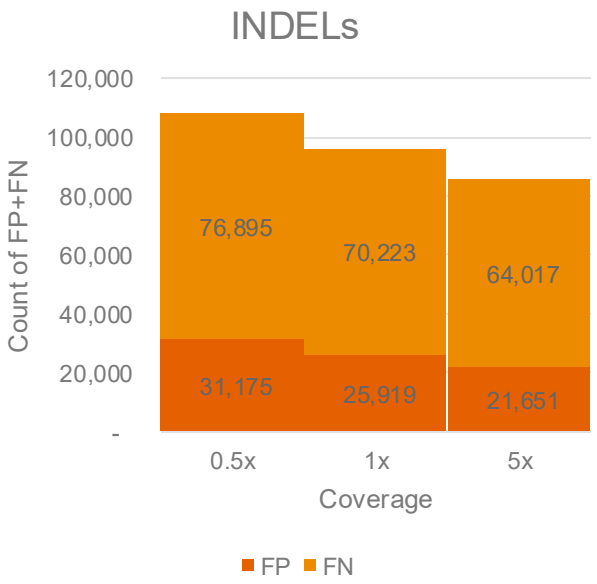
2. not available in an end-to-end mode On-Premise

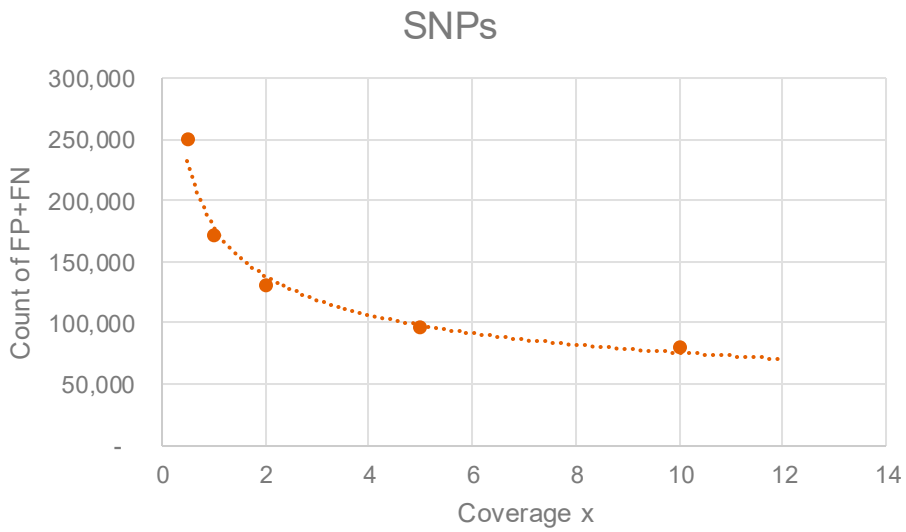
illumina®

DRAGEN Imputation

Higher gain of accuracy for coverage between 0.5-4x

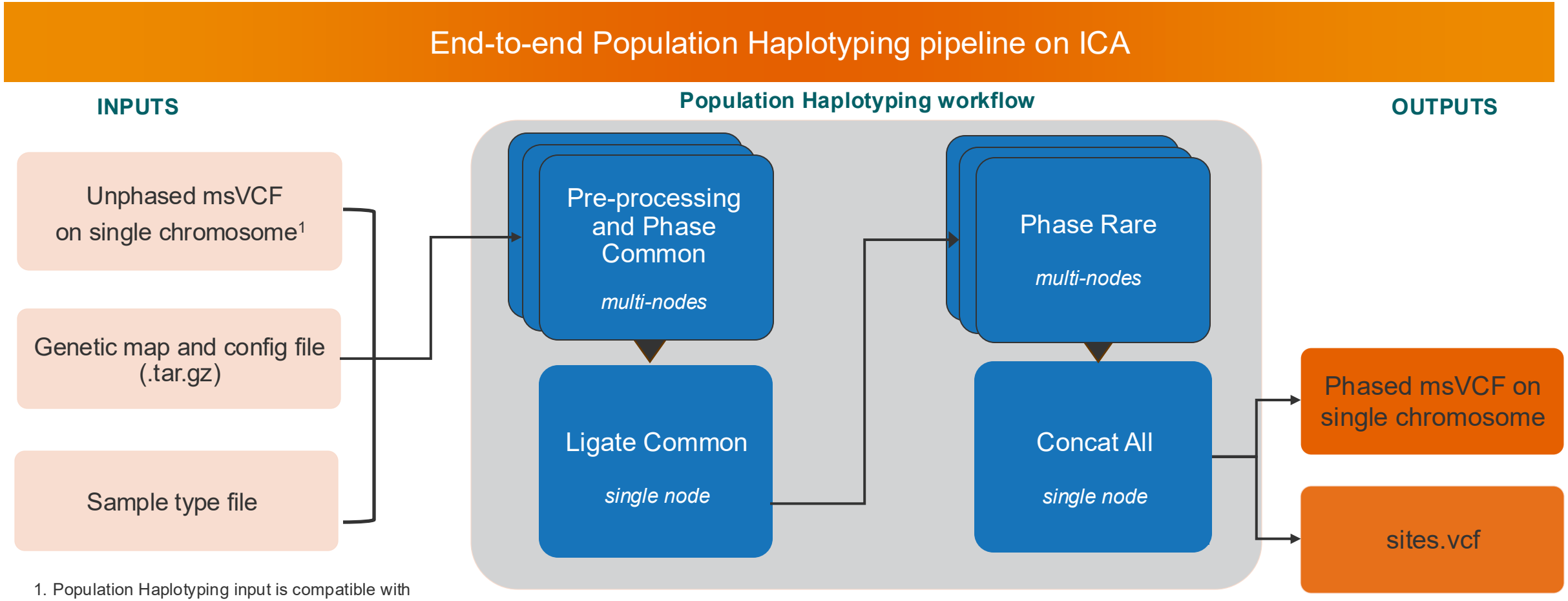






Imputation pipeline ran with IRPv2.0 on HG002 at different coverage, accuracy test with NIST 4.2.1

DRAGEN Population Haplotyping

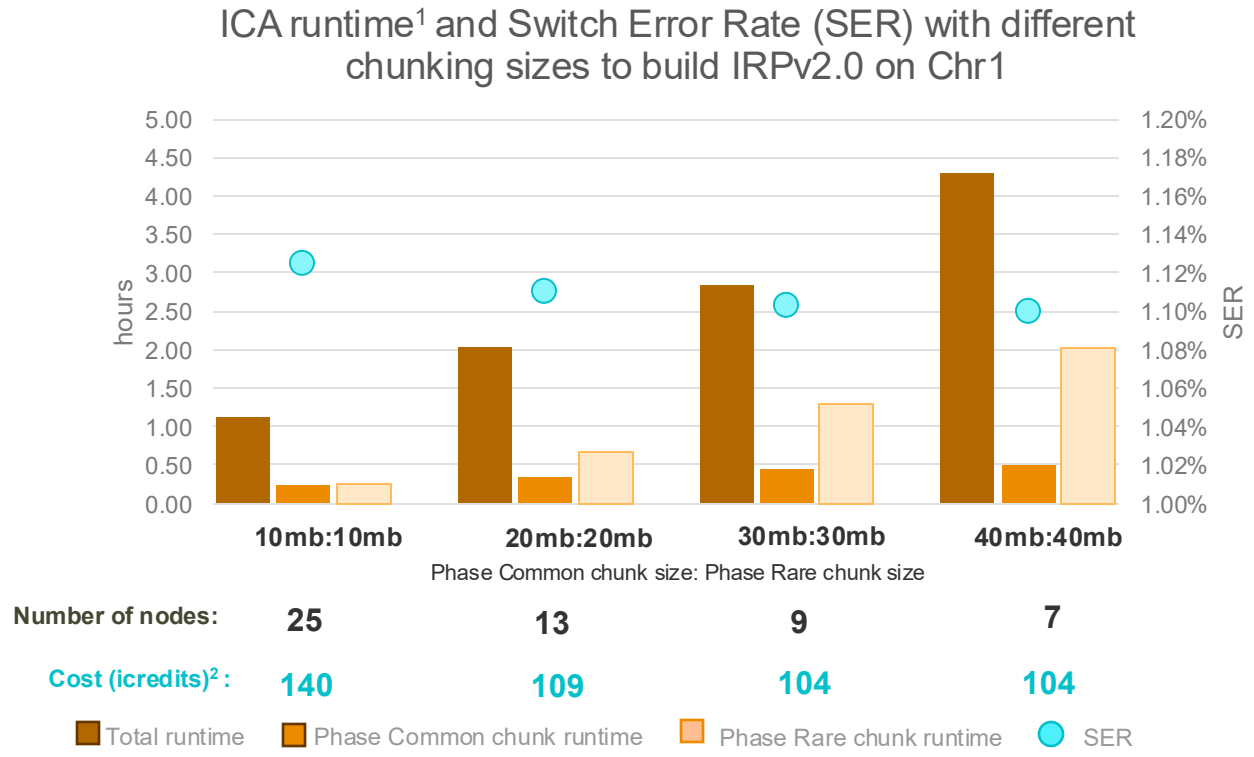


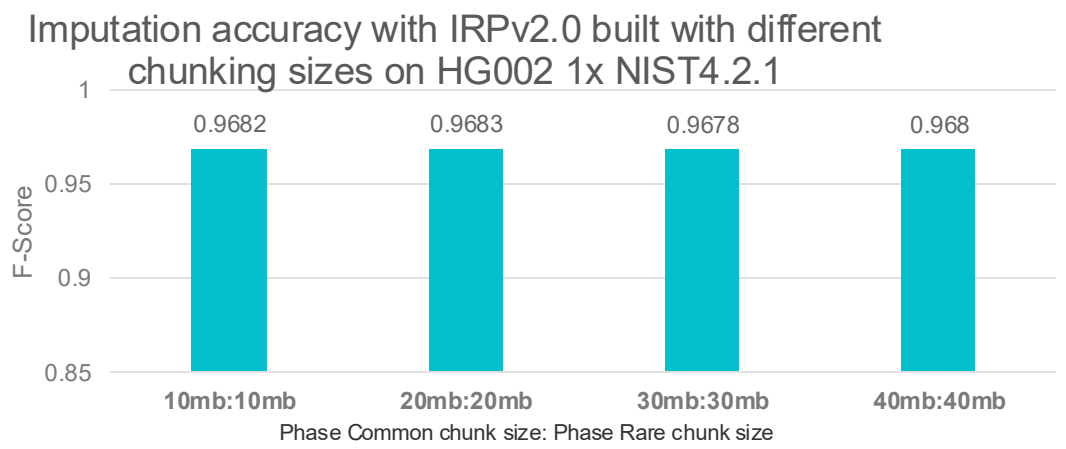
1. Population Haplotyping input is compatible with DRAGEN Iterative gVCF Genotyper output.

illumina®

DRAGEN Population haplotyping

Chunking size impacts accuracy, runtime, and cost – ICA workflow includes optimized settings





Optimized default chunking sizes:
20 mb Phase Common: 20 mb Phase Rare

Can be changed to 10 mb Phase Common: 10 mb Phase rare for a faster analysis (higher cost)

1. Input of Population Haplotyping: 3,202 samples from 1k Genome Project, pipeline parallelized on ICA nodes, SER computed with WhatsHap with truth: 1kg phased panel on HG00096
2. Cost average on 5 runs (compute and storage), 3,202 samples on Chr1

illumina®

Fractional Downsampler

illumina

New DRAGEN Fractional Downsampler Method

Downsample high coverage samples to simulate different amounts of sequencing

All input formats supported by DRAGEN can be used

Downsampling rate is provided by the user and is applied to raw reads coming out of primary analysis without modification

New Fractional Downsampler

Simulates different amount of sequencing for high coverage samples

- Fully integrated workflow with downstream analysis
- Subsampling based on user-defined percentage of reads
- Applied to raw reads with no modification (no trimming, no filtering, pre-dedupped)
- Reduce runtime and cost of analysis using high-depth samples

To enable the fractional downsampler:

--enable-fractional-down-sampler=true

To set percentages of number of reads to keep:

--down-sampler-normal-subsample=<float>
--down-sampler-tumor-subsample=<float>

* <float> is the approximate percentage of reads to keep (e.g. 0.05 = 5%)

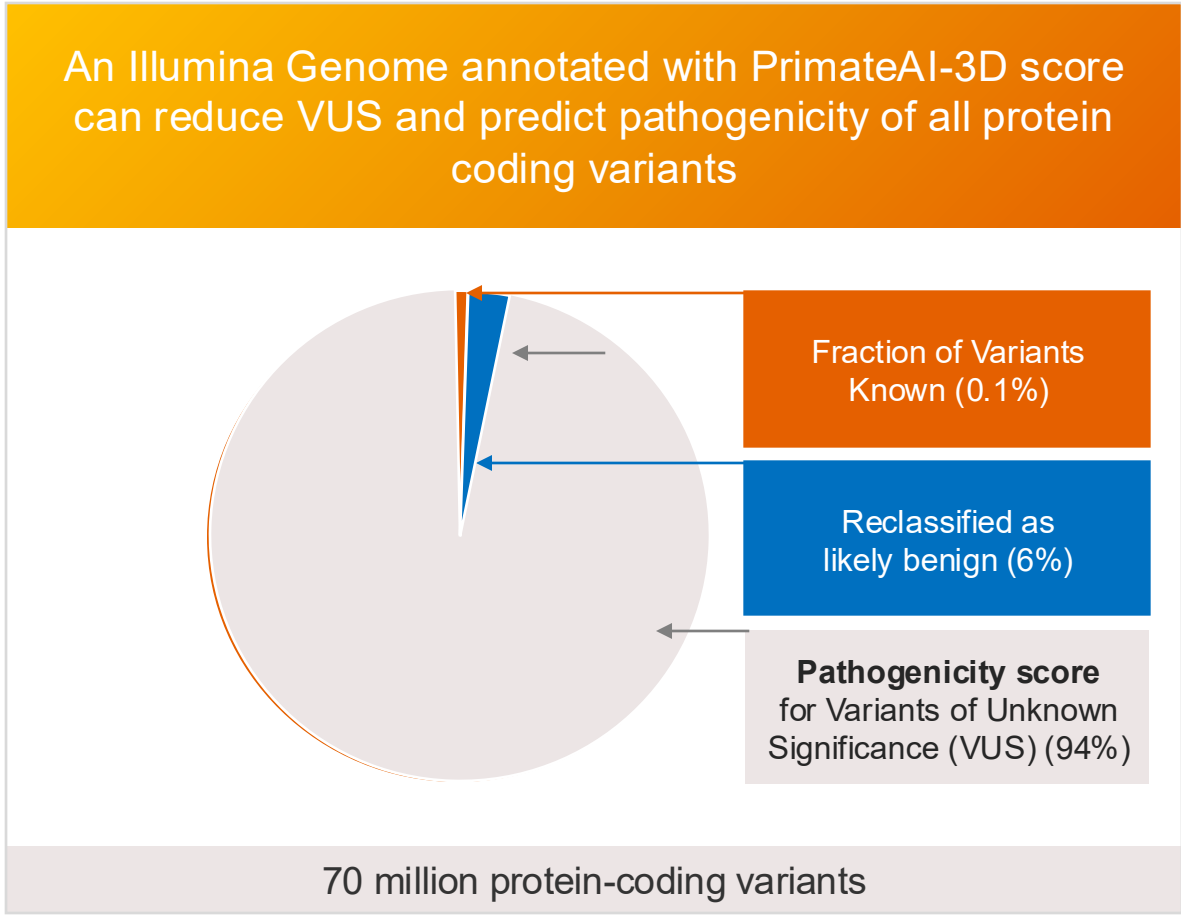
Availability

Illumina Connected Annotations (Nirvana)

illumina

Connected Annotations

- Build on the foundation of a previous tool called Nirvana **Illumina Connected annotations** is the latest version of the annotation engine, **incorporating PrimateAI-3D as a key dataset**
- **Leveraging Primate AI-3D reduces variants of unknown significance and predict pathogenicity of all protein coding variants**
- Provides predicted impact on coding sequence and protein sequence following Human Genome Variation Society (HGVS) standards. Provides consequences relevant to each variant using Sequence Ontology standard nomenclature.
- Connected Annotations **is extremely accurate at an extraordinary speed:** 17.29 minutes run time for 6.5 million DNA variants with 99.9983% average accuracy

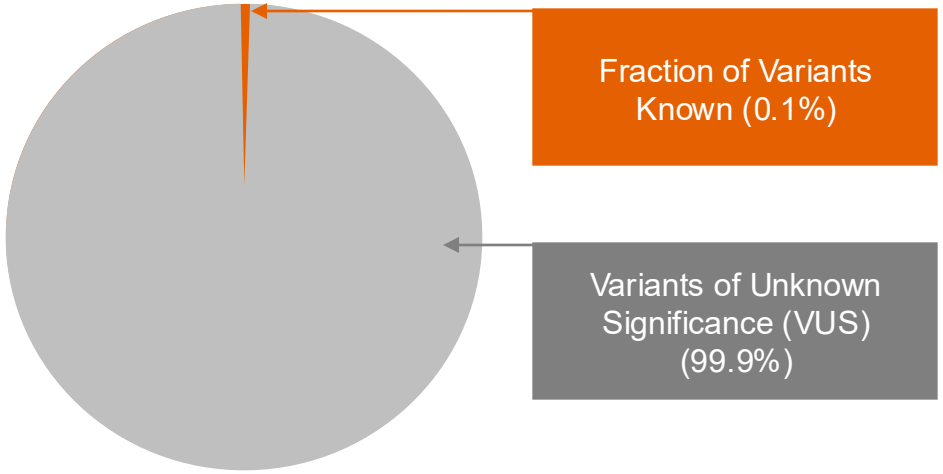


https://www.illumina.com/science/genomics-research/articles/Connected-Annotations-blog.html

https://www.illumina.com/science/genomics-research/articles/primateai-3d.html

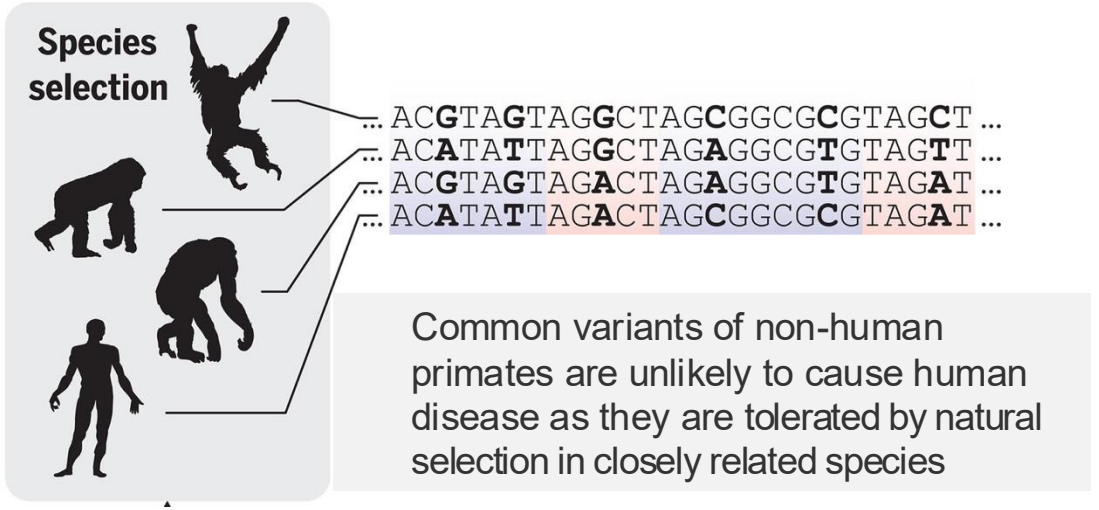
PrimateAI-3D utilizes primate variants and protein structures to solve VUS in human genome coding regions

Our current knowledge of the clinical effects of genetic variants is nascent



70 million protein-coding variants

PrimateAI-3D, A deep learning model trained on millions of benign primate variants

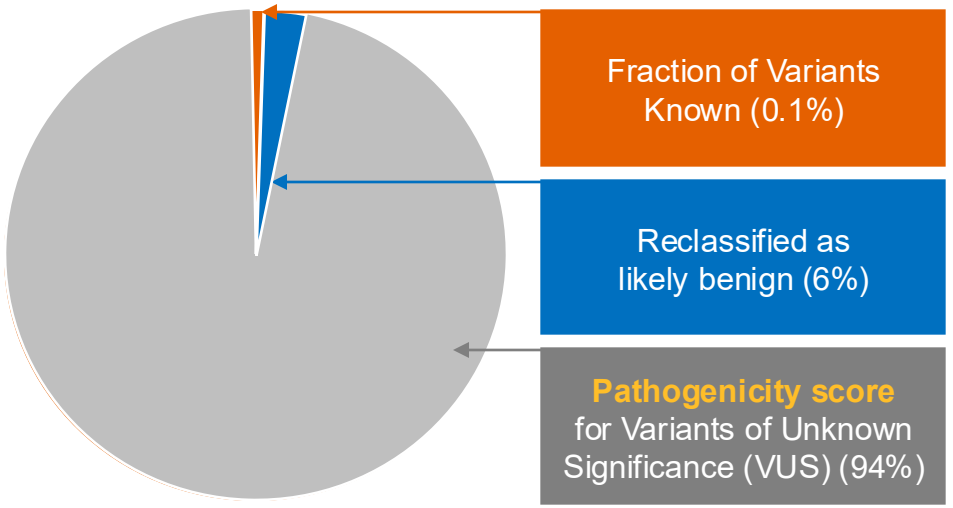


Common variants of non-human primates are unlikely to cause human disease as they are tolerated by natural selection in closely related species

PrimateAI-3D analyzes the 3D structure of proteins and identifies the location of these benign variants. When presented with a new human genetic variant, the model imputes the pathogenicity of the new variant.

An Illumina Genome annotated with PrimateAI-3D score can reduce VUS and predict pathogenicity of all protein coding variants

Benefits customers in genetic disease and oncology researches



70 million protein-coding variants

PrimateAI-3D reclassified >4 million human missense variants of previously unknown consequence as likely benign, resulting in a **> 50-fold increase** in the number of annotated missense variants compared to existing clinical databases.

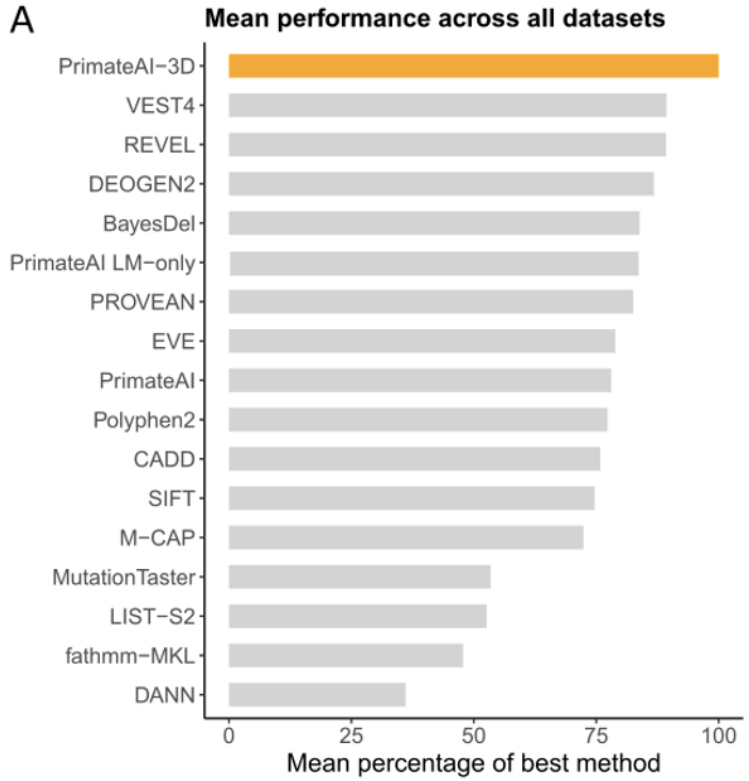
The pathogenicity of the remaining 94% of variants were computed with deep learning, achieving state-of-the-art accuracy for diagnosing pathogenic variants in patients with genetic disease.

Validated to have the top classifier performance in six different benchmarks based on real-world rare and common disease patient cohorts.

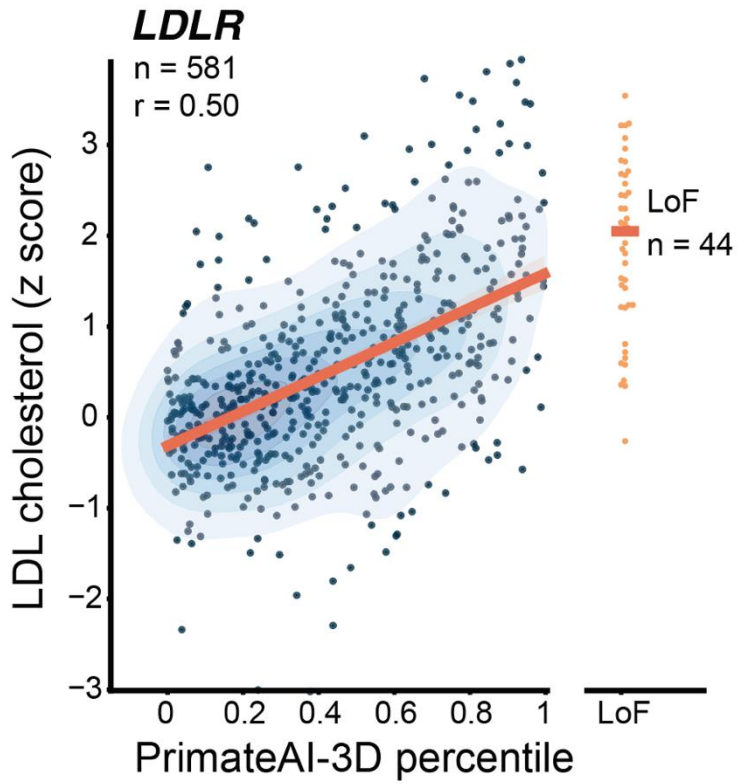
PrimateAI-3D improves genetic risk prediction using rare variants

PrimateAI-3D PRS is highly accurate in both European and non-European populations

PrimateAI-3D achieves top performance across six benchmarks for variant interpretation



PrimateAI-3D accurately predicts phenotypes in 450,000 individuals from the UK Biobank



illumina®

For Research Use Only. Not for use in diagnostic procedures.

SCIENCE 2 Jun 2023, Vol 380, Issue 6648

Thank you!

illumina®

Appendix

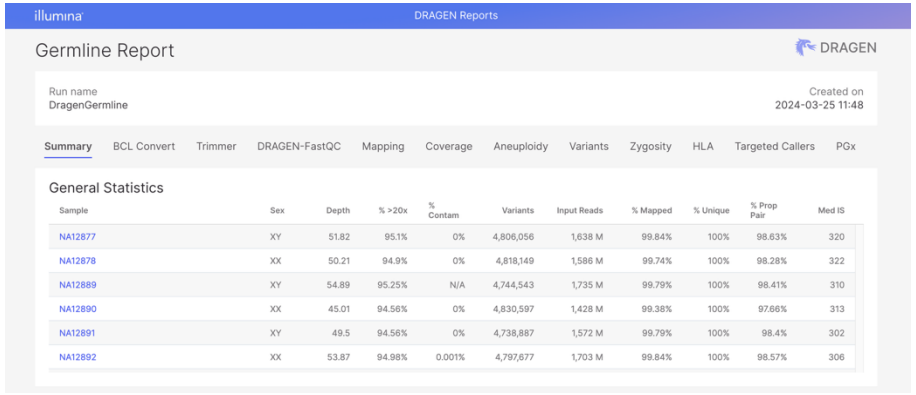
illumina®

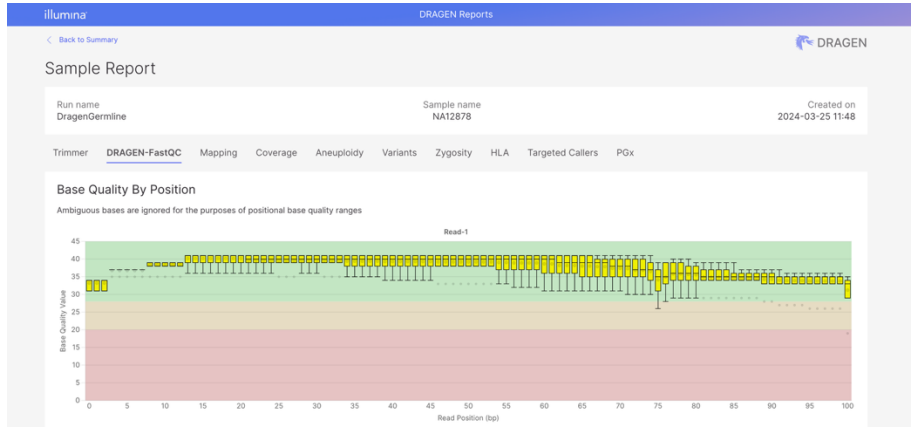
New DRAGEN Report tool

Generate comprehensive, interactive HTML reports from DRAGEN's output files

Flexible deployment options with standalone docker image

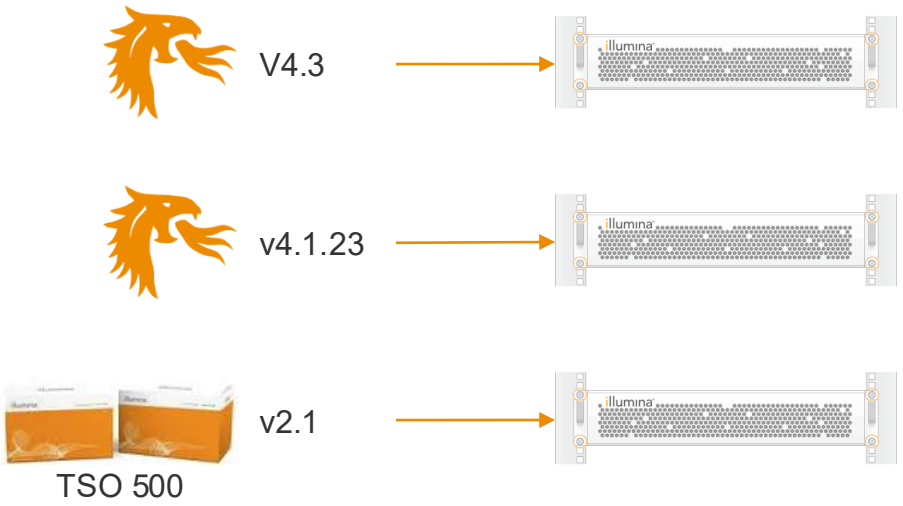
- Generate reports without re-running a whole workflow
- Mix and match samples from across workflows into one report
- Supports current features and new callers in DRAGEN v4.3
- Drop-in Replacement for FastQC / MultiQC



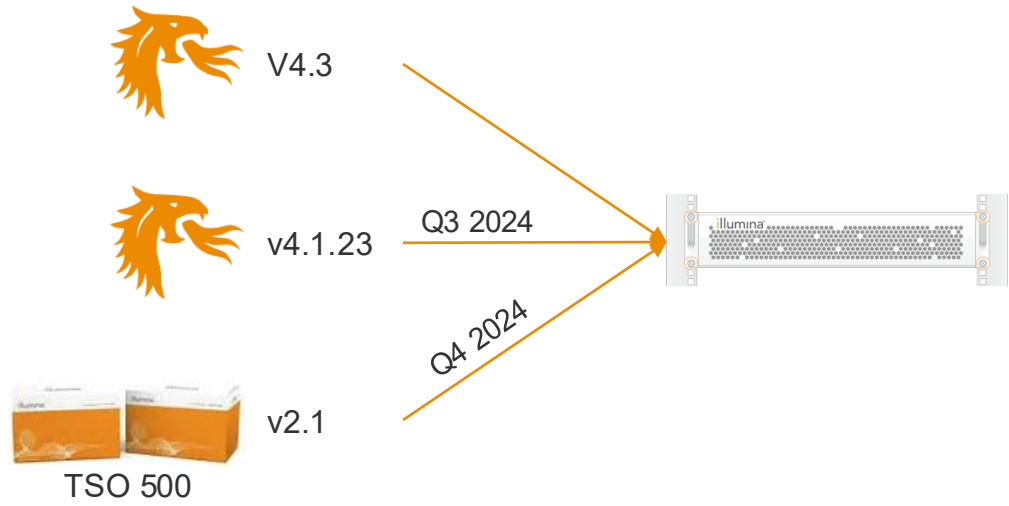


Coming soon – multi-version management on DRAGEN Server

Current – one DRAGEN version on one DRAGEN server at a time



Future – multiple versions on each DRAGEN Server



- New DRAGEN installer allows simultaneous hosting of DRAGEN software v4.3 and later
- Backward compatibility with previous versions – coming soon*

* Requires re-installation of current DRAGEN software using new installers