Introduction to Sentieon Genomics Software Suite

Sentieon Application Team

Sentieon

Introduction to Sentieon

- Mission: Enabling precision genomics data for precision medicine
  - Process genomic data at an affordable cost and time
  - Providing the most accurate results
- **Software-only** solutions on generic commodity hardware:
  - Easily scalable, deployable, and upgradable :
  - Any commodity hardware, local or cloud, no special hardware requirement
- Process short-reads data for reference-based alignment and variant calling
  - All sample sizes: WGS, WES, targeted panels of different sizes and depths
  - Sample types: germline, somatic, liquid biopsy with UMI
  - Variant types: SNV, Indel, SV, CNV

How we started?

- Sentieon's core strength:
  - Algorithm design
  - Software engineering
- Core-team's previous ventures:
  - Computational lithography: large-scale optimization problem in chip design and manufacturing. Acquired by ASML
  - Computation advertising: personal recommendation and real-time advertise serving. Acquired by Alibaba
  - Genomics: Solve data processing bottleneck
- Sentieon:
  - Founded in Mountain View, California in 2014
  - Funded by private and venture capitals

Where does Sentieon fit in NGS workflows?

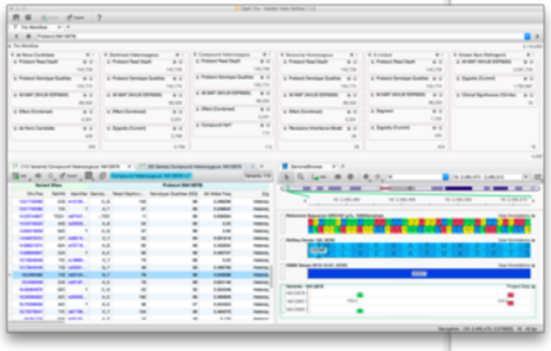
Sentieon

Data Generation

```
$("#word-list-out").e(" ");
var b = k();
h();
var c = l(), a = " ", d = parseInt($("#limit_val").a()), f = parseInt($("
function("LIMIT_total:" + d);
function("rand:" + f);
d < f && (f = d, function("check rand\u00f3\u00f3rand: " + f + "tops: " + d
var n = [], d = d - f, e;
if (0 < c.length) {
  for (var g = 0;g < c.length;g++) {
    e = m(b, c[g]), -1 < e && b.splice(e, 1);
  }
  for (g = 0;g < c.length;g++) {
    b.unshift({use_wystepuje:"parameter", word:c[g]});
```

Sentieon Secondary Analysis

BCL-FASTQ-BAM-VCF



Tertiary analysis

Software-only solution:

- Easily scalable, deployable, and upgradable
- Any commodity CPU hardware, local or cloud
- Licensed annually

Our Approach to Secondary Analysis

1. mathematical methods

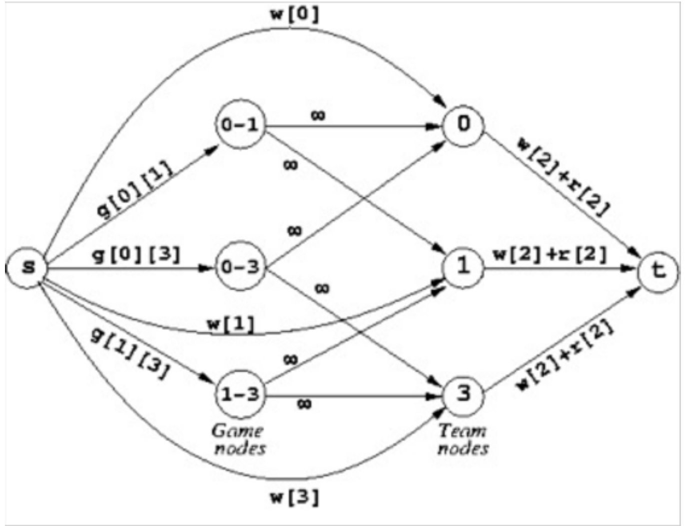
2. compute algorithms

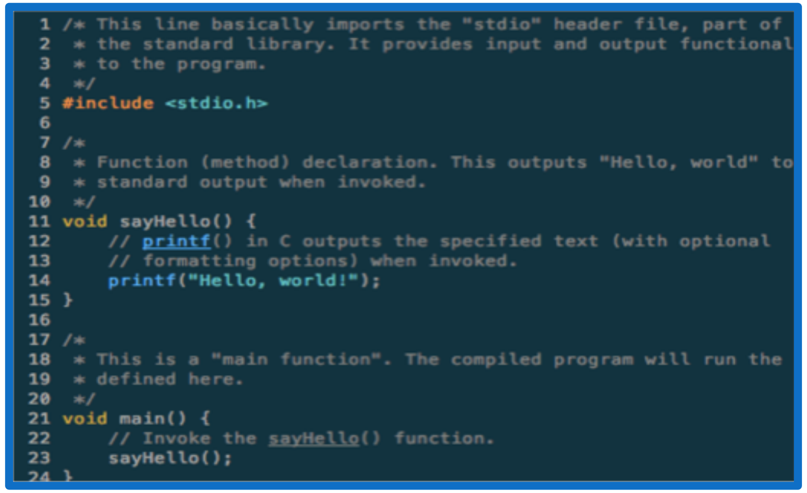
3. software implementation

$$\frac{D}{Dt}\overline{w'^i w'^j} + \overline{w'^i w'^a}\nabla_a \tilde u^j + \overline{w'^j w'^a}\nabla_a \tilde u^i - \alpha\left(g^{ia}\overline{w'^j \frac{T'}{\tilde T}} + g^{ja}\overline{w'^i \frac{T'}{\tilde T}}\right)\left(\nabla_a\Phi + \frac{D\tilde u_a}{Dt}\right)$$
$$+\frac{1}{\bar\rho}\nabla_a[\bar\rho\overline{u''^a w'^i w'^j} + \overline{(g^{ia}w'^j + g^{ja}w'^i)P'} - \overline{w'^i\sigma^{ja}(u')} - \overline{w'^j\sigma^{ia}(u')}]$$
$$+\frac{1}{\bar\rho}\overline{w'^i w'^j\nabla_a(\bar\rho u''^a)} - \overline{P'(g^{ia}\nabla_a w'^j + g^{ja}\nabla_a w'^i)} = -\frac{1}{\bar\rho}[\overline{\sigma^{ia}(u')\nabla_a w'^j} + \overline{\sigma^{ja}(u')\nabla_a w'^i}] = -\epsilon_2^{ij}, \quad (30)$$

$$(1+e_4)\frac{D}{Dt}\overline{\left(\frac{T'}{\tilde T}\right)^2} - 2f(t)\overline{\left(\frac{T'}{\tilde T}\right)^2} - 2\overline{w'^a\frac{T'}{\tilde T}}D_a + \frac{1}{(1+e_4)\bar\rho C_p^2}\nabla_a\left[(1+e_4)^2C_p^2\bar\rho\overline{w'^a\left(\frac{T'}{\tilde T}\right)^2}\right] + \frac{1+e_4}{\bar\rho}\overline{\left(\frac{T'}{\tilde T}\right)^2\nabla_a(\rho u''^a)}$$
$$+\frac{2}{\bar\rho\tilde T C_p}\overline{\frac{T'}{\tilde T}\left[P'\nabla_a w''^a - \nabla_a(P'_g w''^a) - \frac{DP'_g}{Dt}\right]} = \frac{2}{\bar\rho\tilde T C_p}\overline{\frac{T'}{\tilde T}[\sigma^{ab}(u')\nabla_a u'_b - \nabla_a F_r'^a]} = -\epsilon_3, \quad (31)$$

$$(1+e_4)\left[\frac{D}{Dt}\left(\overline{w'^i\frac{T'}{\tilde T}}\right) + \overline{w'^a\frac{T'}{\tilde T}}\nabla_a\tilde u^i - \alpha\overline{\left(\frac{T'}{\tilde T}\right)^2}g^{ia}\left(\nabla_a\Phi + \frac{D\tilde u_a}{Dt}\right)\right] - f(t)\overline{w'^i\frac{T'}{\tilde T}} - \overline{w'^i w'^a}D_a$$
$$+\frac{1}{\bar\rho C_p}\nabla_a\left[(1+e_4)C_p\bar\rho\overline{w'^i w'^a\frac{T'}{\tilde T}}\right] + \frac{1+e_4}{\bar\rho}\overline{w'^i\frac{T'}{\tilde T}\nabla_a(\rho u''^a)} + \frac{1}{\bar\rho\tilde T C_p}\overline{w'^i\left[P'\nabla_a w''^a - \nabla_a(P'_g w''^a) - \frac{DP'_g}{Dt}\right]}$$
$$-\frac{1+e_4}{\bar\rho}\overline{\frac{T'}{\tilde T}\nabla_a\sigma^{ia}(u')} + \frac{1}{\bar\rho\tilde T C_p}\overline{w'^i[\sigma^{ab}(u')\nabla_a u'_b - \nabla_a F_r'^a]} = -\epsilon_3^i, \quad (32)$$





```
/* This line basically imports the "stdio" header file, part of
 * the standard library. It provides input and output functional
 * to the program.
 */
#include <stdio.h>

/*
 * Function (method) declaration. This outputs "Hello, world" to
 * standard output when invoked.
 */
void sayHello() {
    // printf() in C outputs the specified text (with optional
    // formatting options) when invoked.
    printf("Hello, world!");
}

/*
 * This is a "main function". The compiled program will run the
 * defined here.
 */
void main() {
    // Invoke the sayHello() function.
    sayHello();
}
```

Identical mathematical models as the Broad Institute

More efficient compute algorithms

Enterprise strength software implementation

Sentieon

Where are we now?

- Deployed worldwide: hundreds of deployment
  - Commercial companies: major pharmaceutical and biotech companies, sequencing service providers, and molecular diagnostic companies
  - Hospitals and cancer centers: Orien Project, a joint collaboration of pharmaceutical companies and about 15 cancer centers in the United States
  - Government agencies: like US NIH and FDA
  - Academia: Universities and research institutes in US and China for genomic studies in human, plants and animals
- Large usage:
  - About 80 million core-hours accumulated usage, and ~50 million core-hours in 2018 alone
  - ~37 petabytes of data processed
- References in the literature:
  - Universities: Stanford, Cornell, Columbia, UCSF, Fudan, Shanghai Jiaotong, OSU, U of Utah,
  - Government organization: NIST, NCI, CAS
  - Commercial: BGI, Mayo, Intermountain, DNANexus, BMS, AbbVie, H3

Sentieon Genomics Software products

- Two product families for germline and somatic variant calling:
  - **seq** products: DNAseq & TNseq
    - Drop in Improvement/Replacement for Broad Institute tools
    - 10x more efficient, deterministic results
    - No down-sampling to correctly process any sequencing depth
  - **scope** products: DNAscope & TNscope
    - Sentieon developed math, with improved accuracy beyond GATK/MuTect/MuTect2(v3.8)/Mutect2(v4)
    - Also includes structural variations

	Germline	Somatic
Seq product series (Matching Broad Institute)	DNAseq: • match Haplotyper	TNseq: • TNsnv: Match MuTect • TNhaplotyper: MuTect2 (GATK3.8) • TNhaplotyper2: Mutect2 (GATK4)
Scope product series (Own Improved variant caller & Structural variant calling)	DNAscope	TNscope

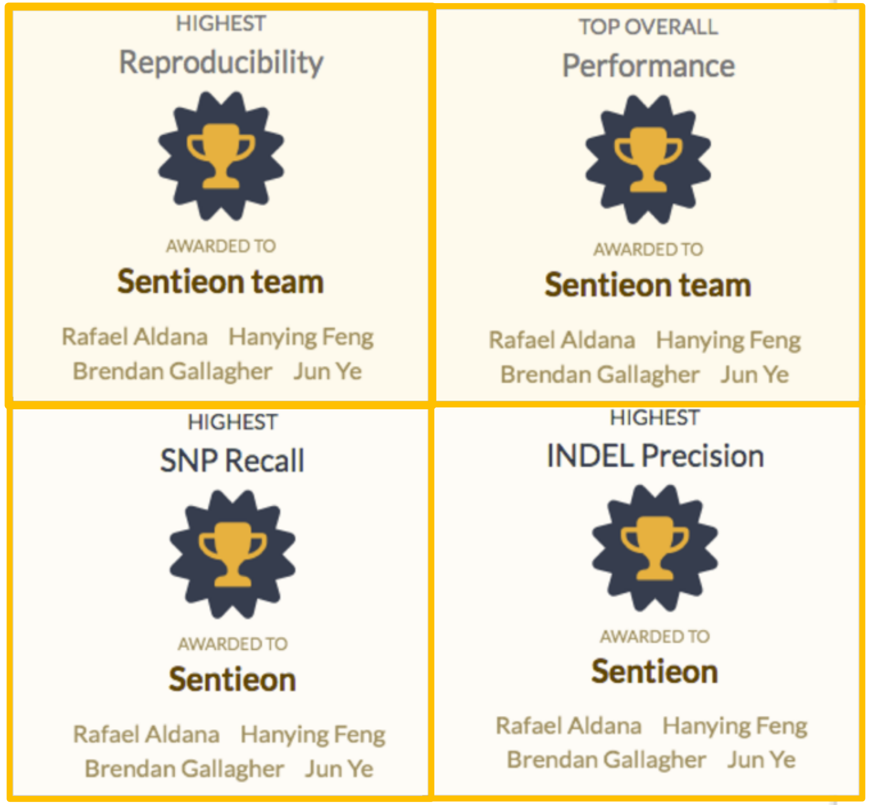
Sentieon DNAseq: precisionFDA challenge winner

Consistency Challenge
(2/25-4/25/2016)
-Top Overall Performance
-Highest Reproducibility

Truth Challenge
(4/26-5/26/2016)
-Highest INDEL Precision
-Highest SNP Recall

Speed: 30X WGS fastq-to-vcf run on baseline-32 single server: ~4 hours

precisionFDA



HIGHEST
Reproducibility
AWARDED TO
Sentieon team
Rafael Aldana Hanying Feng
Brendan Gallagher Jun Ye

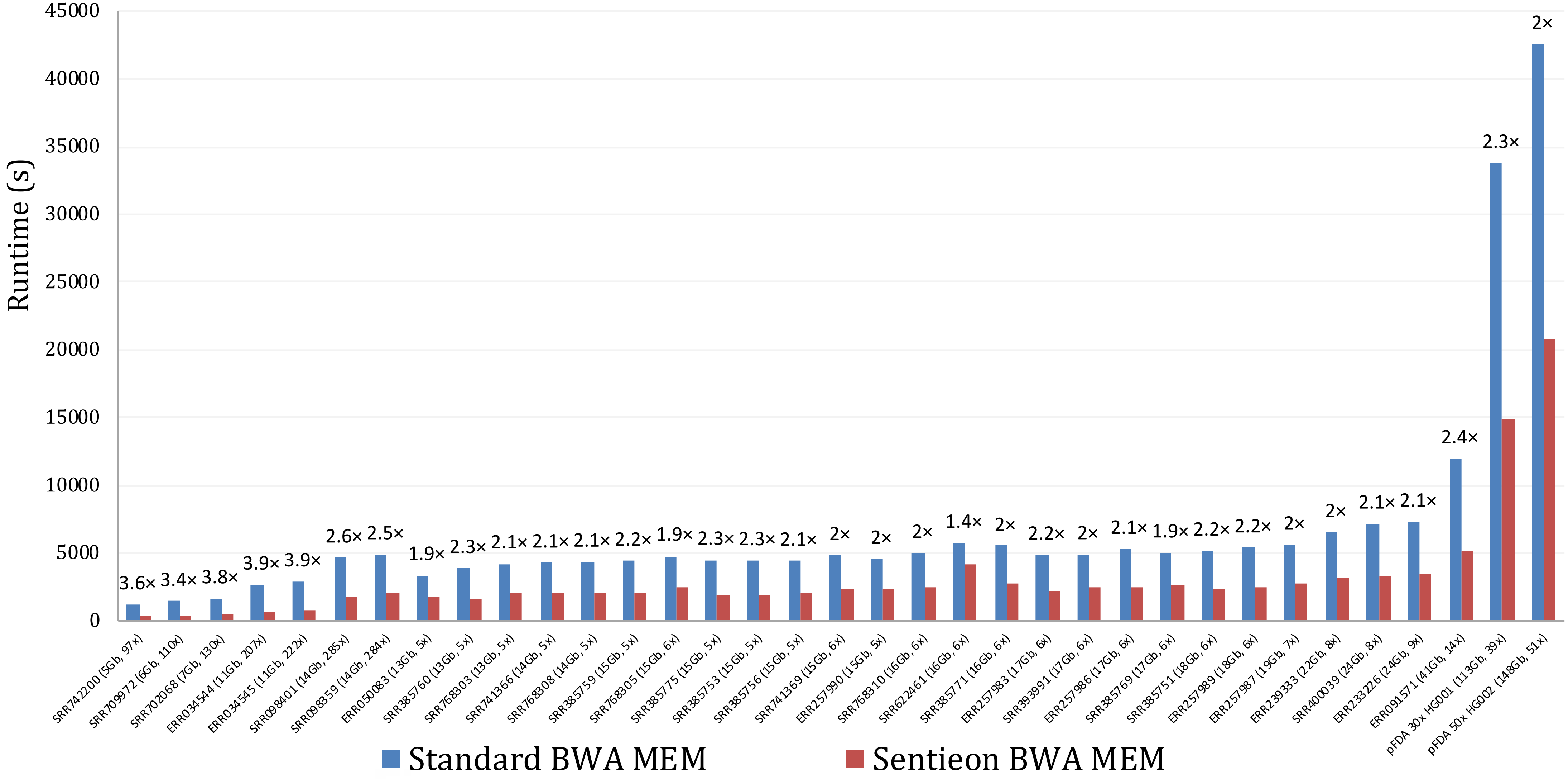
TOP OVERALL
Performance
AWARDED TO
Sentieon team
Rafael Aldana Hanying Feng
Brendan Gallagher Jun Ye

HIGHEST
SNP Recall
AWARDED TO
Sentieon
Rafael Aldana Hanying Feng
Brendan Gallagher Jun Ye

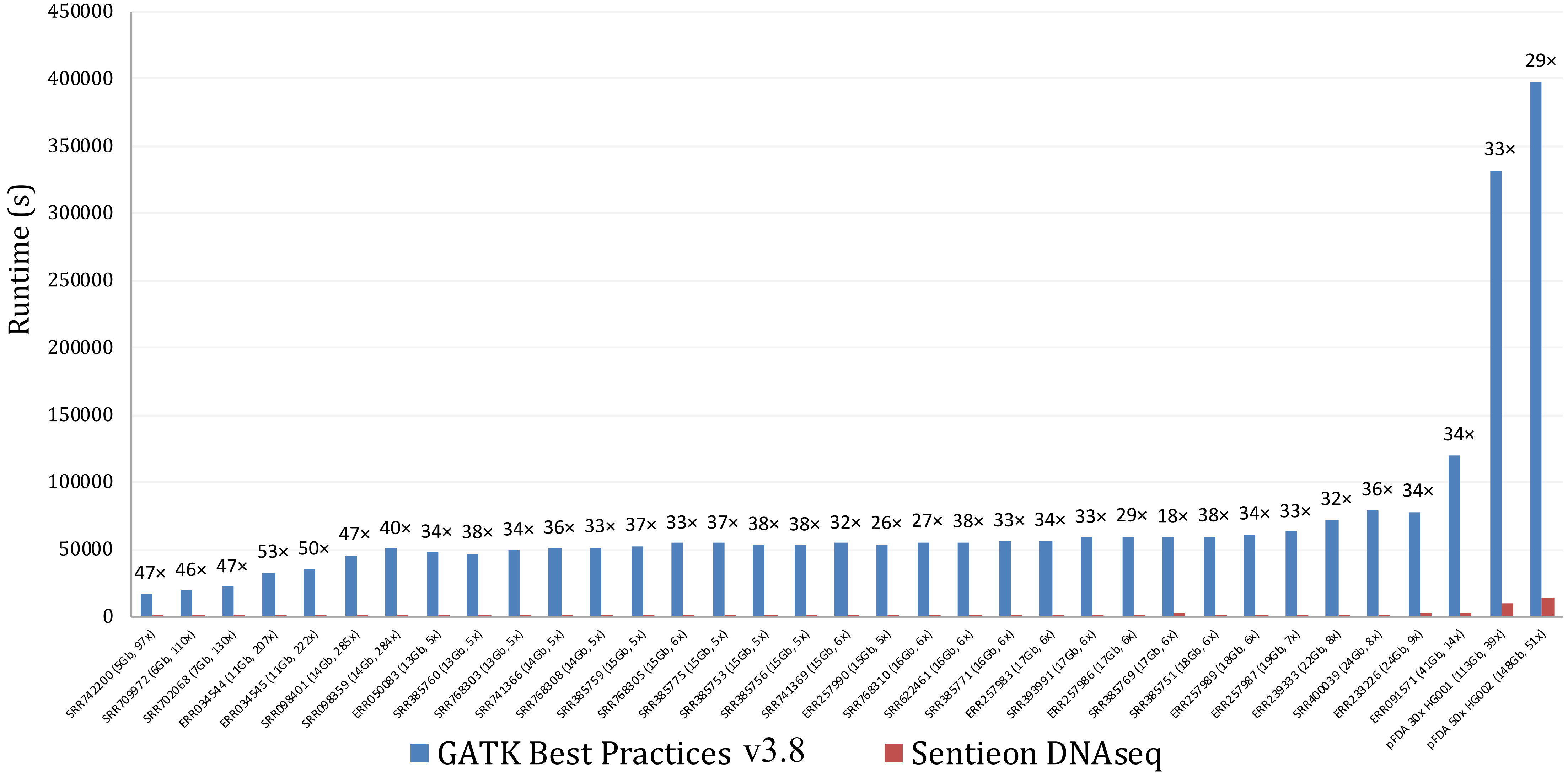
HIGHEST
INDEL Precision
AWARDED TO
Sentieon
Rafael Aldana Hanying Feng
Brendan Gallagher Jun Ye

*Screenshots from *https://precision.fda.gov/*

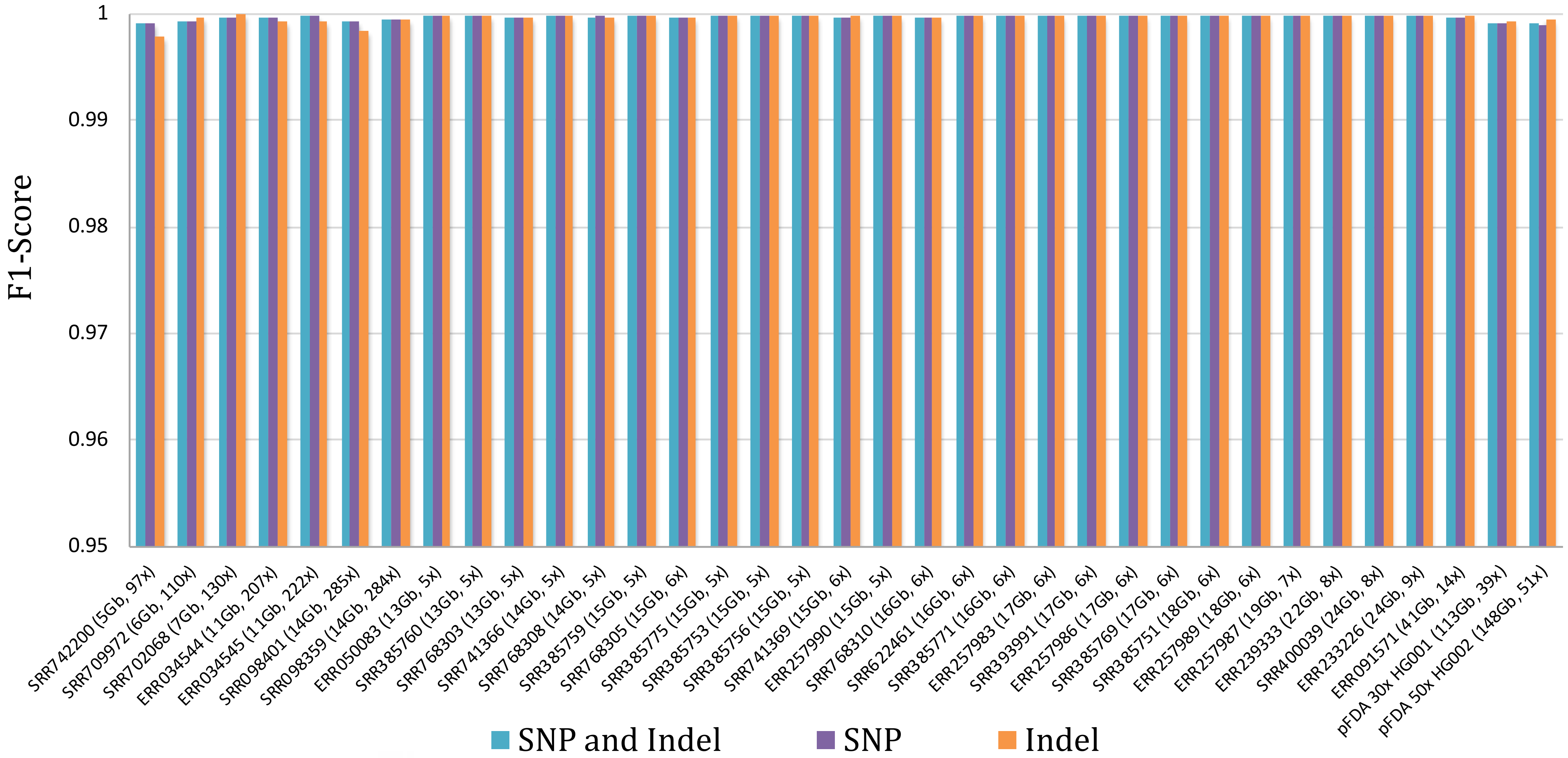
Sentieon DNAseq: BWA Performance (FASTQ -> BAM)



Sentieon DNAseq: Overall Pipeline Performance (BAM -> VCF)



Sentieon DNAseq: High concordance with BWA/GATK



Sentieon DNAseq: Accuracy and Consistency

- BWA/GATK has been considered the de facto standard of the germline variant discovery pipeline
- High concordance with BWA/GATK:
  - F1 score between Sentieon DNAseq and BWA/GATK is about 0.999
  - Difference comes from:
    - Rounding error
    - Run-to-run difference of GATK due to thread dependency, random sampling in high-depth region
    - GATK down-samples reads in high-depth (>500) region, while Sentieon DNAseq does NOT
- Perfect run-to-run consistency:
  - Guaranteed by rigorous algorithm design and software engineering
  - Identical results regardless of multi-threading or distributed processing
- Proven by awards in the PrecisionFDA Truth and Consistency challenges
- Comparison with truth set on HG002 (50x):

	Precision	Sensitivity	F1
SNP	0.9992	0.9991	0.9992
InDel	0.9977	0.9979	0.9978

Comments on GATK 4.0

- What is in GATK4.0?
  - Improved ease of deployment in various environment, in particular Google cloud
  - Changed Multi-threading to single-threading + distributed processing
  - Re-implementation of MuTect2
- What has not changed:
  - Core-implementation of GATK and thus same speed as in core-hours, except for MuTect2
  - GATK4 to match GATK3.8
  - Still down-sampling at high-depth region, but now with fixed random seed
- Sentieon:
  - Sentieon tools match GATK3.8 and thus GATK4 on the algo/tool level, but does not down-sample or cut corner for better speed
  - TNhaplotyper2 matches MuTect2 of GATK4

Sentieon DNAseq: Large-cohort Joint Calling

- Joint-calling provides:
  - Population-wide statistical background
  - Much improved sensitivity to identify variants, and reduce false positive
- GATK solution:
  - Very slow and difficult to scale to beyond hundreds of WGS
  - Needs to many intermediate file merging steps
- Sentieon DNAseq joint-calling:
  - Highly efficient in computation and resource demand
  - One-step solution without intermediate steps, on a single server or distributed
  - Scalable: joint calling over 100,000 WGS
  - Expected runtime/memory:
    - About 2 core*hour per input GVCF WGS sample
    - About 3G memory usage per thousand input GVCF WGS samples

DNAscope: Sentieon propriety variant caller

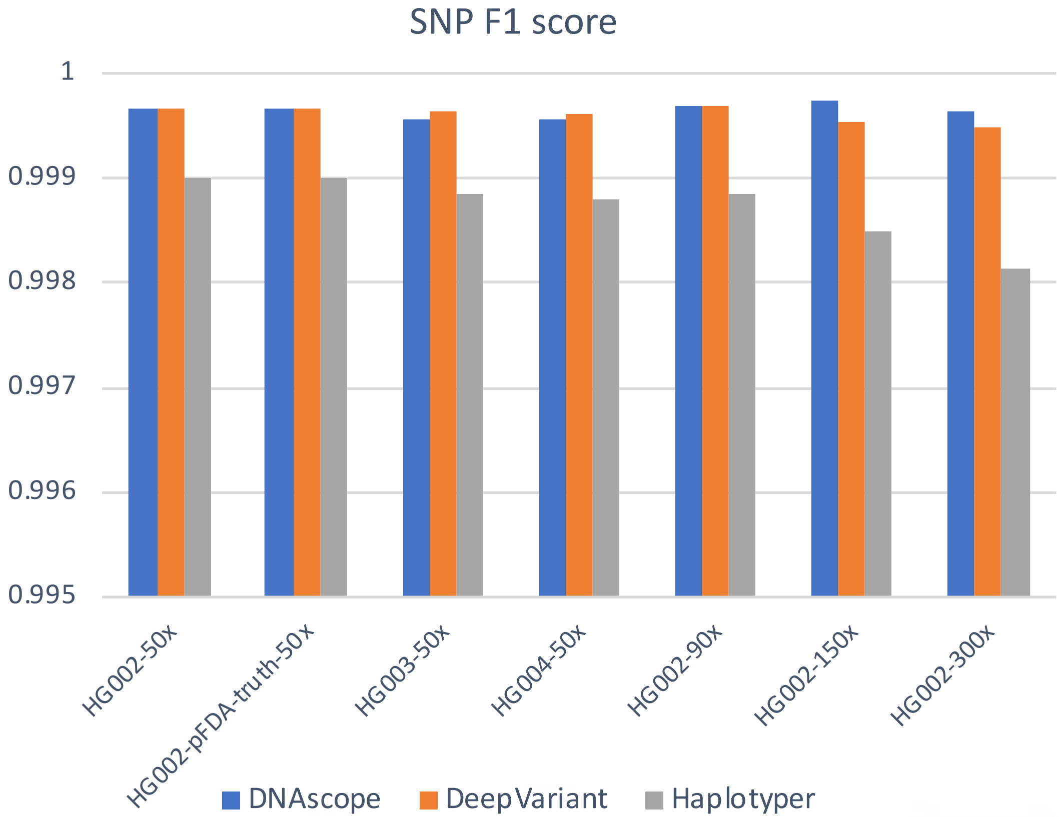
- Improved local assembler in DNAscope solves limitations of GATK assembler:
  - GATK fails assembly due to complexity in pileup
  - GATK fails assembly in specific reference regions: sample independent blind spots
  - GATK looses haplotype connectivity information, generating imperfect haplotypes
- DNAscope improvements:
  - More stable, better behaved in high complexity regions
  - More sensitive, able to detect weaker signal
  - Supports structural variant calling
  - Better treatment of soft-clips, which are more abundant in non-European samples
- Structural variant calling

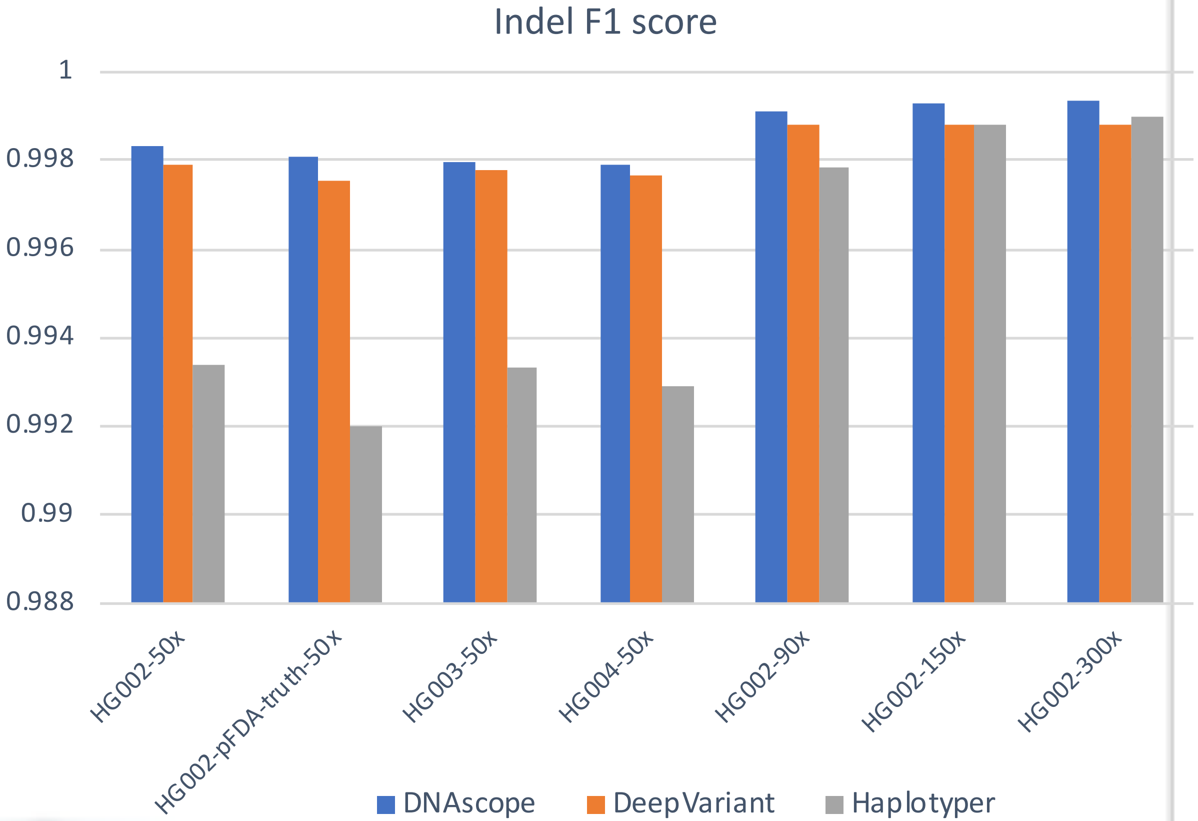
DNAscope and Machine Learning

- DNAscope sensitivity benefits from intelligent filtering
- High sensitivity + machine learning filtering => better results
- Machine leaning model improves candidate filtering:
  - Train model based on NIST truth-sets to capture sequencing artifacts
  - Apply model to enhance F1 score by reducing FP without sacrificing TP
  - In addition, model can enhance TP by changing incorrect genotype in output
  - Approach leverages first principles math and statistics from existing methodologies

DNAscope vs Haplotyper vs DeepVariant

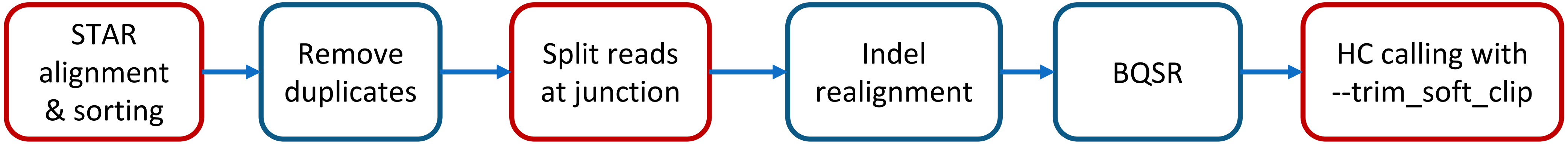
- Evaluate precisionFDA Truth Challenge and subset lanes from GiAB against NIST Truth-sets v3.3.2
  - DNAscope significantly improves over Haplotyper
  - DNAscope give similar or better result than DeepVariant





Sentieon RNAseq Pipeline Example

- Supports RNAseq with minor changes from typical DNAseq pipeline:
  - Alignment with RNA-aligner, eg. STAR.
  - Split reads at junction after removing duplicates.
  - Add --trim_soft_clip option in Haplotype caller.



Sentieon TNseq: Overview

- Product description:
  - Somatic variant discovery pipeline
  - Identical* result as Broad Institute's "Somatic Variant Discovery Workflow" MuTect/MuTect2(v3.8)/Mutect2(v4)
  - Identical math, but with more efficient algorithm and enterprise-strength engineering
  - Currently, match version MuTect/MuTect2 of version 3.8 and Mutect2 version 4.0.2.1
  - Features:
  - Rigorous and faithful implementation of the mathematics of MuTect and MuTect2/Mutect2
  - 10x faster with high concordance
- Key algorithms:
  - TNsnv: MuTect
  - TNhaplotyper: MuTect2 (v3.8)
  - TNhaplotyper2: Mutect2 (v4.0.2.1)

Sentieon TNseq: Performance Benchmark

- Benchmark setup:
  - See our white paper for complete details on the benchmark
  - 32 core 2.4 GHz Intel Xeon server with 64 GB memory and 2TB dual stripped SSDs
  - Actual runtime may vary from sample to sample, may be significantly impacted by IO throughput
  - Sentieon recommends high throughput storage devices, such as SSD, to take advantage of the highly streamlined computation and CPU capability
  - Below shows runtime of three samples after DNAseq processing and co-realignment
  - Note: MuTect does not have multithreading implementation, while MuTect2 does
- Result:

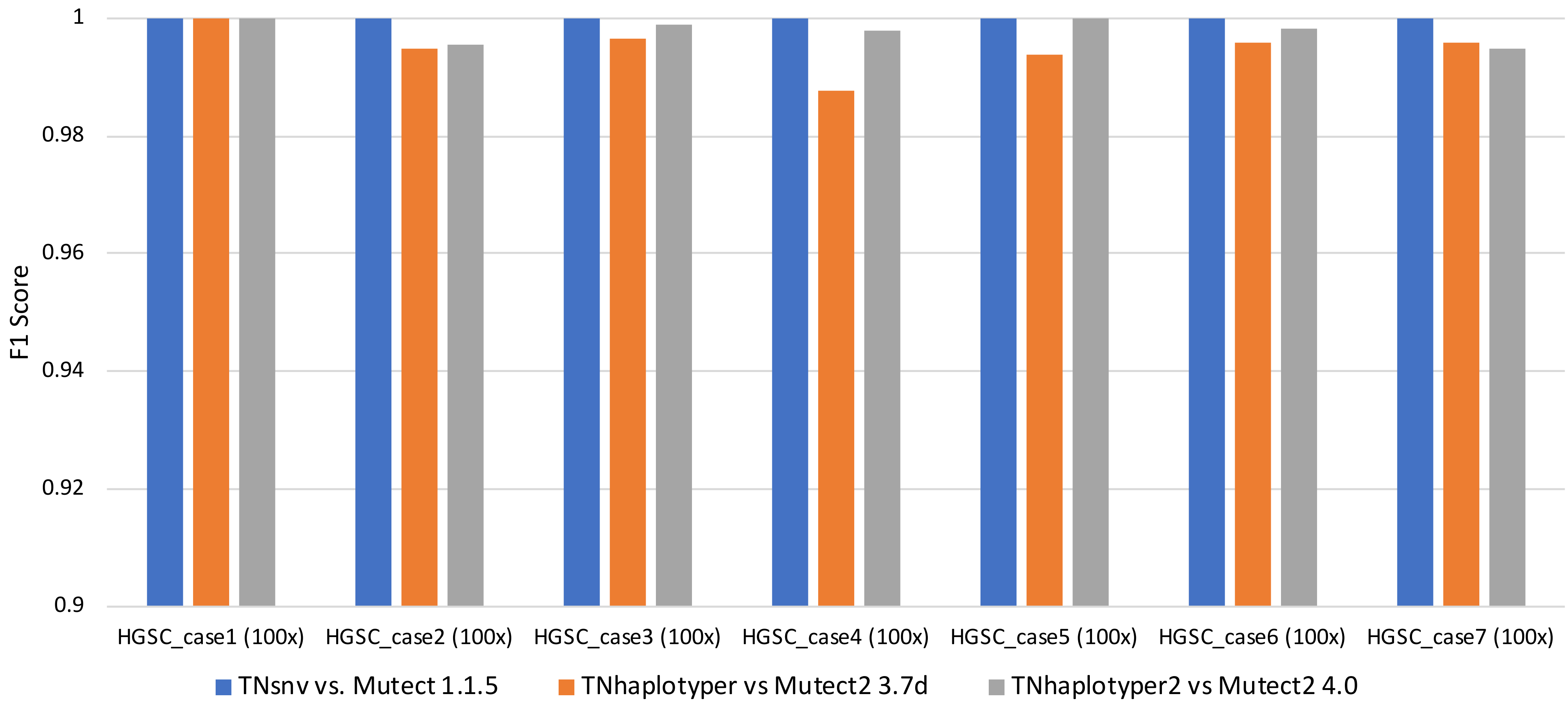
Sample	MuTect	TNsnv	MuTect2 (v3.8)	TNHaplotyper
HGSC_case6_WGS	59 hours	29 min (123x)	24 hours	2.2 hours (11x)
HGSC_case7_WGS	42 hours	7.8 min (325x)	220 hours	29 hours (7.5x)
HGSC_case7	6.4 hours	1.9 min (193x)	4.2 hours	15 min (16x)

Sentieon TNseq: Accuracy and Consistency

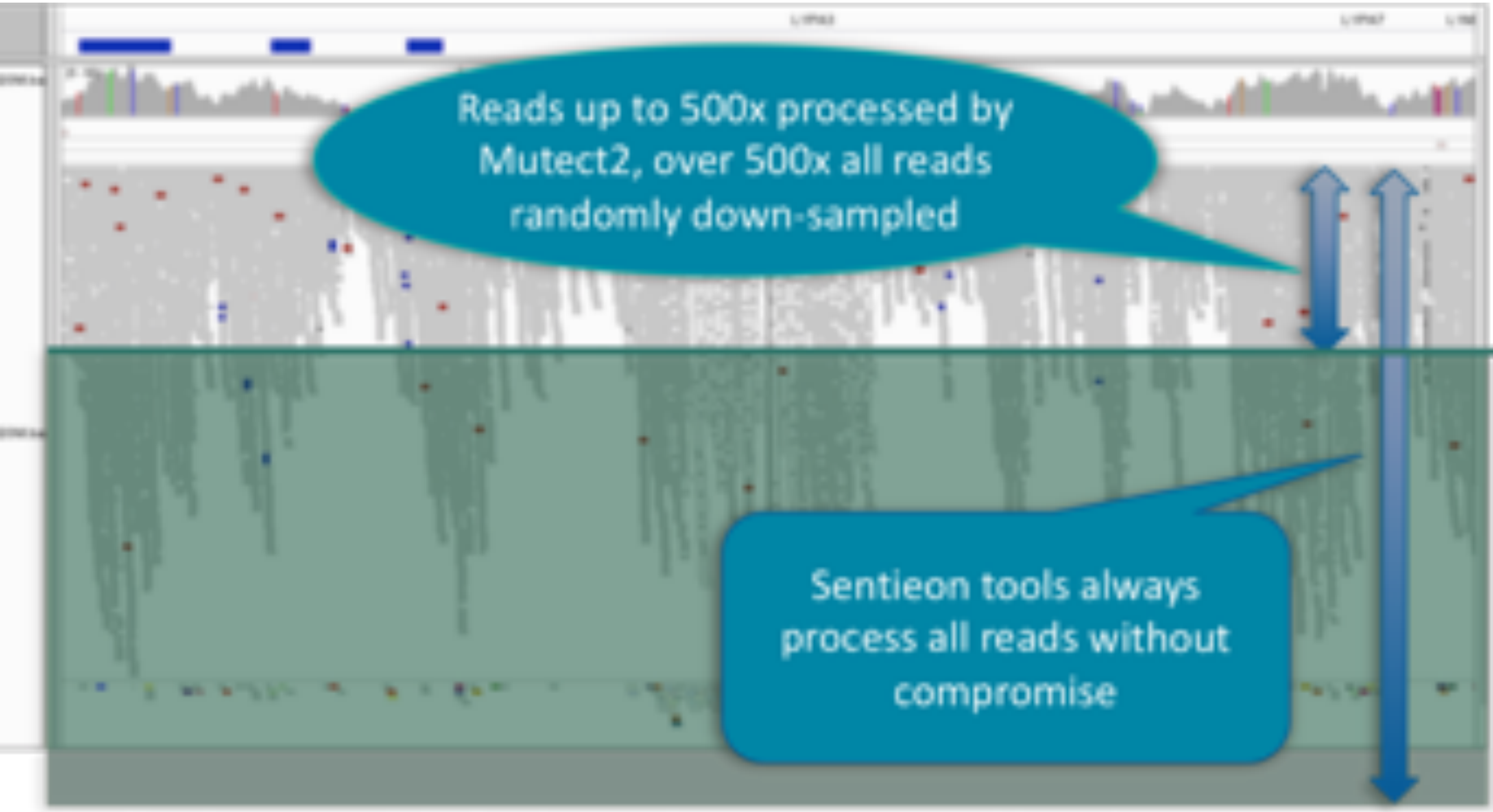
- High concordance with MuTect and MuTect2/Mutect2:
  - See table below for details
- Difference comes from:
  - Rounding error
  - Run-to-run difference due to thread dependency of MuTect2, down-sampling in high-depth region in MuTect/MuTect2/Mutect2
  - Down-samples reads in high-depth (>500) region, while Sentieon DNAseq is capable of handling arbitrary read depth without down-sampling

Sample	MuTect vs. TNsnv	MuTect2 (v3.8) vs. TNhaplotyper	Mutect2 (v4) vs TNhaplotyper2
HGSC_case6_WGS	0.9998	0.9929	0.9974
HGSC_case7_WGS	0.9995	0.9965	0.9706
HGSC_case7	0.9992	0.9867	0.9948

Sentieon TNseq: High concordance with MuTect/Mutect2



No Downsampling in any Sentieon tools



Reads up to 500x processed by Mutect2, over 500x all reads randomly down-sampled
Sentieon tools always process all reads without compromise

http://bioinformatics-ca.github.io/bioinformatics_for_cancer_genomics_IGV_lab_2016/

MuTect2 random down-samples reads below this line, causing run-to-run error, sacrificing accuracy

All Sentieon tools processes *all* reads with no compromise.

Two sources of error w/Mutect2:

- dcov: fixed at 500. Cannot be changed
- maxReadsInRegionPerSample: Default 1000. Can be changed but will blow up memory

Study on the effect of down-sampling in variant calling

- Input:
  - NextSeq 550: Comprehensive v3 DNA Panel (Coriell and Horizon Samples) from Basespace
    - Coriell (NA12878) and Horizon samples were prepared using the AmpliSeq Comprehensive Panel v3 for Illumina kit. Samples were sequenced on the NextSeq system at 2x151 bp read length with dual indexing.
  - Amplicon regions: Comprehensive-v3.dna_manifest.20180509.bed
  - References: ucsc.hg19.fasta with GATK hg19 bundle
- Variant calling:
  - Germline: Haplotyper(Sentieon 201808.01) vs HaplotypeCaller(GATK 3.7)
  - Somatic: TNhaplotyper(Sentieon 201808.01) vs MuTect2 (GATK 3.7)

- Fastq2bam preprocessing:
  - BWA -> Realign -> BQSR (->ReadWriter*)

  *ReadWriter (PrintReads in GATK) is not required for Sentieon tools.

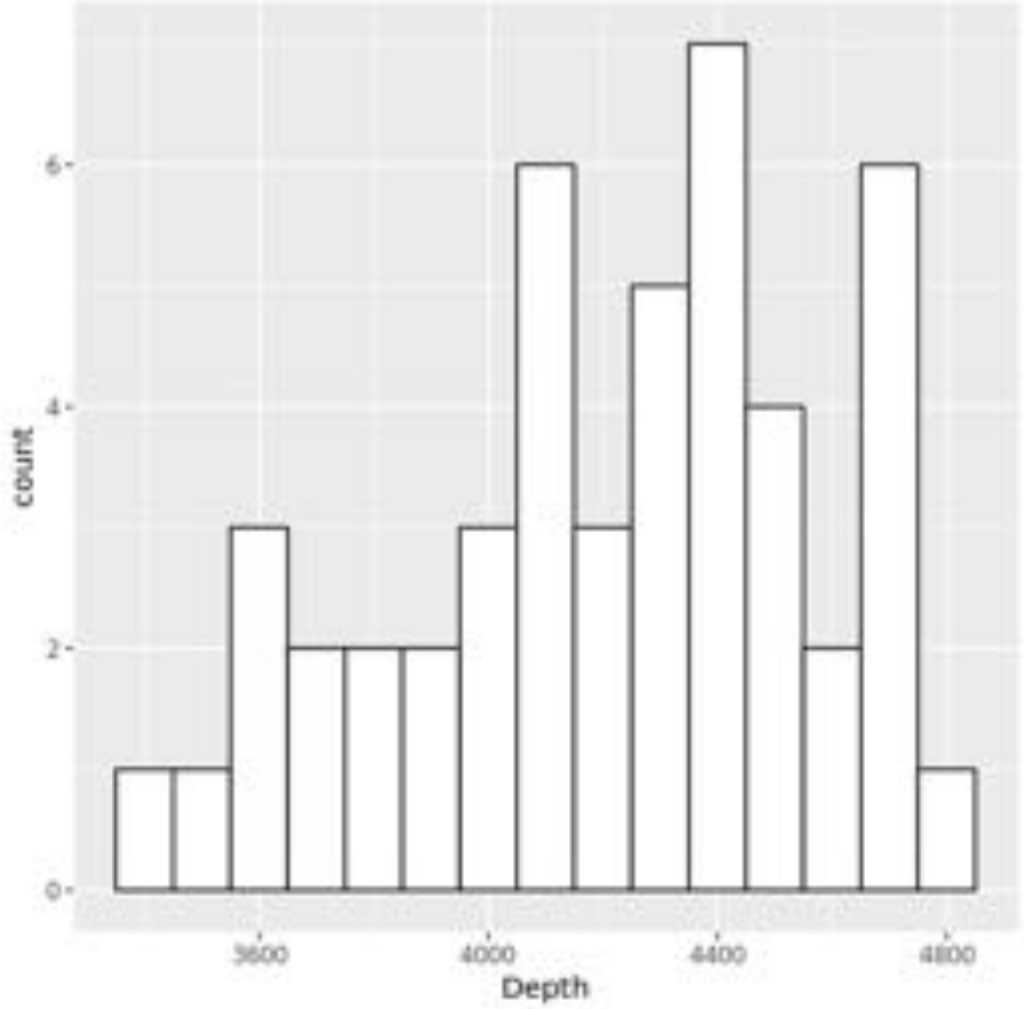
- Parameters:
  - Germline: call_conf=30; emit_conf=30
  - To disable downsampling in GATK: -dcov 100000 --maxReadsInRegionPerSample 100000
- Matching evaluation:
  - RTGtools (3.9.1)

  VCF files were split into SNP and INDEL-only vcf files by bcftools (1.9) before evaluation.

Sequencing depth

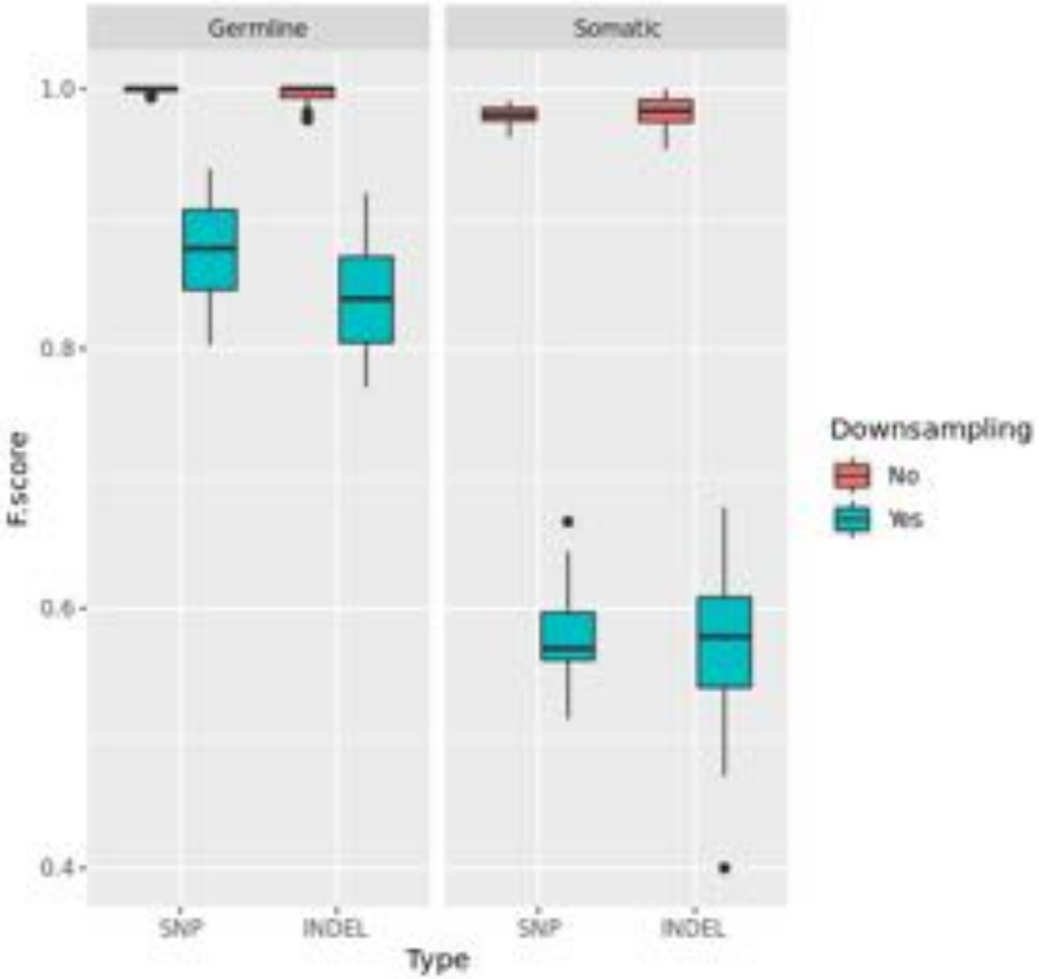
- Total number of samples: 48
- Average depth statistics:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3381	4002	4266	4212	4465	4785



Overall matching accuracy

- Boxplots of concordance F-scores grouped by calling mode (germline/somatic), with down-sampling turned on/off in GATK and variant type (SNP/INDEL).
- Both germline and somatic callers achieve excellent matching accuracy when down-sampling in GATK was disabled.
- Down-sampling in high-depth samples is detrimental to its accuracy



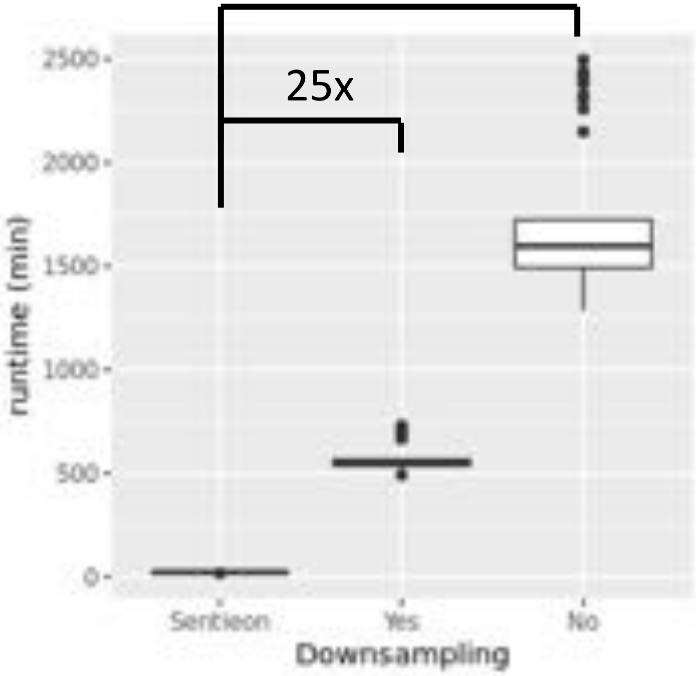
Sentieon

GATK runs slower when downsampling is disabled

- Runtime statistics:

Somatic

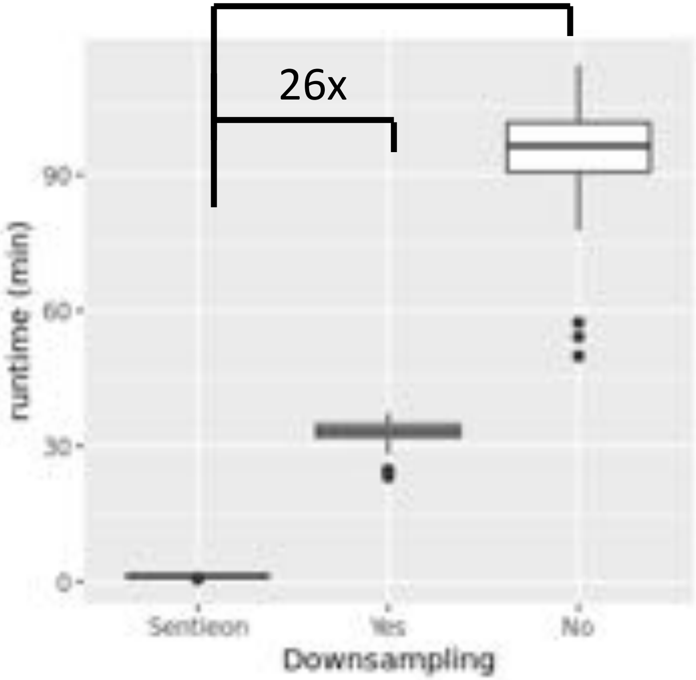
76x



25x

Germline

78x



26x

* Jobs were run on a 32-core machine.

Sentieon TNscope: Overview

- Product description:
  - Sentieon's own algorithm for somatic variant discovery
  - Finished first in latest open somatic variant calling challenge ICGC-TCGA Dream Challenge in all categories: SNV, INDEL, and Structural Variants
  - Supports tumor-normal, tumor-only, with or without panel of normals
- Features:
  - Improved mathematical model
  - Better local assembly to handle complex region
  - More annotations for variant assessment and filtering: QUAL, PV, etc.
  - Structural variant calling
  - Tumor-only and Tumor/Normal functionality for SNV, InDel, SV
  - Efficient algorithm and enterprise-strength engineering for efficient and stable computing
- Introduced machine-learning model-based variant filtration (Beta)

Sentieon TNscope: ICGC-TCGA DREAM challenge

ICGC-TCGA DREAM Mutation Calling challenge

Final Leaderboard (8/19/2016)

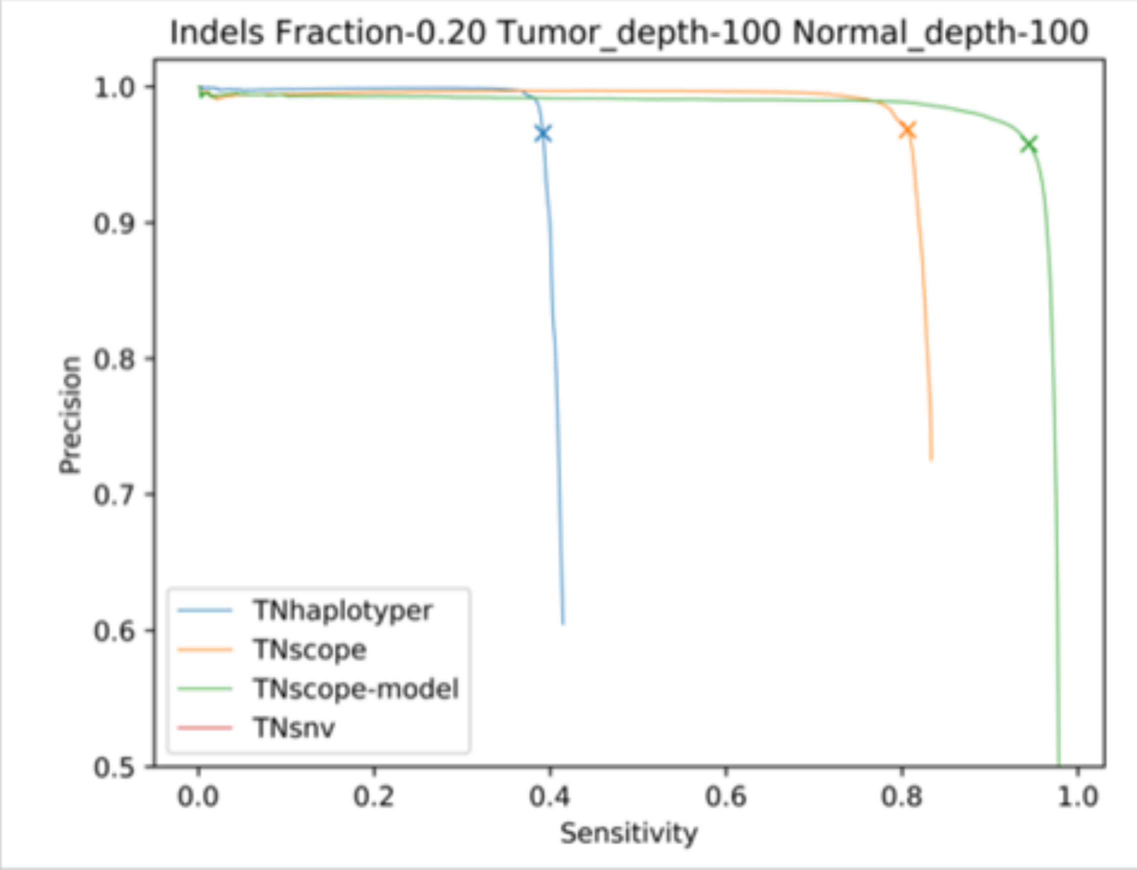
SNV	INDEL	SV
Sentieon 98.57%	Sentieon 98.14%	Sentieon 100%
Bina/Roche 97.57%	Bina/Roche 97.01%	Genowis 99.82%
Genowis 96.92%	OICR-GSI 86.99%	Gridss 99.63%

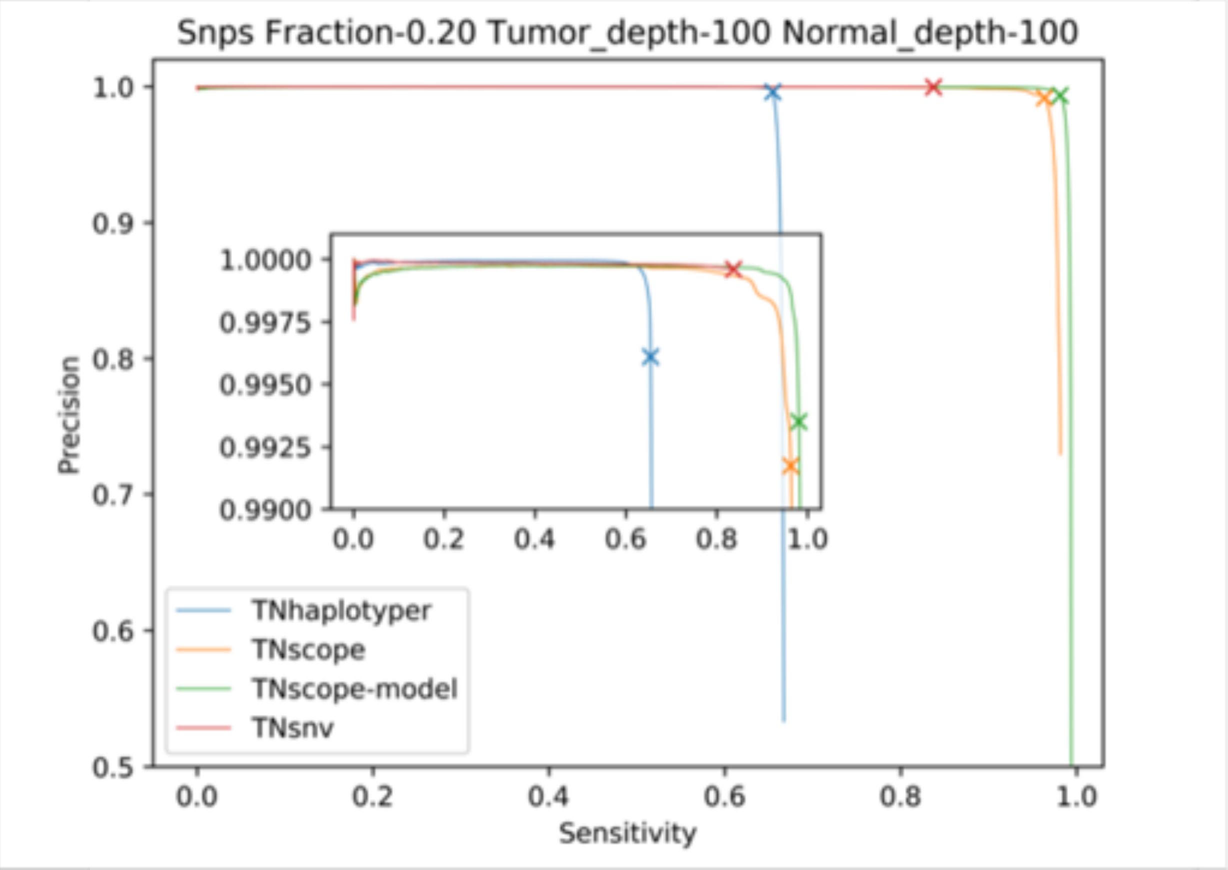
-Sentieon leads in all categories-

https://www.synapse.org/#!Synapse:syn312572/wiki/247695

TNscope Performance White Paper: Result

- Result:
  - TNscope outperforms both TNsnv and TNhaplotyper in SNP and Indel
  - Model-based variant filtration works well with TNscope in the cases tested





Distributed mode of Sentieon pipeline

- Sentieon provides native support for fine-grained splitting of chromosomes for efficient distributed processing
- Sentieon provides a cross-platform implementation of the pipeline using
  - Common Workflow Language (CWL)
  - Toil workflow engine
- Case study:
  - Turnaround time:
    - < 30 mins on 8 64-core servers (~250 core-hours)
    - 1 hour with 3 servers (~190 core-hours)
  - All tests used GCE n1-highcpu-64 instances (64 cores, 57.6 GB RAM) with the precisionFDA consistency challenge 30x Garvan sample
- For comparison, GATK4 on FireCloud takes approximately 800 core-hours to process this same sample.

Other Products

- CNV:
  - Match GATK 4.0 somatic CNV tool
  - Somatic CNV calling with normalization based on Panel of Normals
- Python API engine:
  - Allows user-defined BAM processing logic with Python scripts
  - Example case:
    - Worked with St Jude Children’s Hospital to improve speed of CREST by 10x
    - Remove reads of non-unique mapping for specialized applications

Summary

- Highest accuracy:
  - Most rigorous math and no down-sampling in high-coverage regions
  - No run-to-run difference
  - Proven in precisionFDA challenges and DREAM challenge
- High concordance with BWA/GATK/Mutect/MuTect2(v3.8)/Mutect2(v4): DNAseq and TNseq for snv and indel calling
- Improved accuracy with DNAscope/TNscope and DNAscope + Structural Variant calling
- Fast turnaround: 10X reduced core-hours
- Easy deployment:
  - Generic commodity hardware
  - Supports distributed processing
- Ease of use, drop-in replacement
- Strong technical support

References

- Sentieon website: http://www.sentieon.com
  - Latest product announcement
  - News and full publication list
- Sentieon product manuals application notes:
  - Manuals: https://support.sentieon.com/manual/
  - App notes: https://support.sentieon.com/appnotes/
- Publications:
  - Biorxiv Benchmark Paper: https://www.biorxiv.org/content/early/2017/05/12/115717
  - TNscope paper: https://www.biorxiv.org/content/early/2018/01/19/250647
  - PeerJ GATK compatibility paper: https://peerj.com/preprints/1672/
- More references upon request