GOLDEN HELIX
Enabling Precision Medicine

The Sentieon Genomic Tools Improved Best Practices Pipelines for Analysis of Germline and Tumor-Normal Samples

Andreas Scherer, Ph.D. President and CEO

Dr. Donald Freed, Bioinformatics Scientist, Sentieon

CIOReview
20 most promising Biotech Technology Providers

pharma TECH OUTLOOK
Top 10 Analytics Solution Providers

Gartner
Hype Cycle for Life sciences

Golden Helix

Golden Helix is a global bioinformatics company founded in 1998.

GlaxoSmithKline

VarSeq

- Variant Calling
- Filtering and Annotation
- Clinical Reports
- CNV Analysis
- Pipeline: Run Workflows

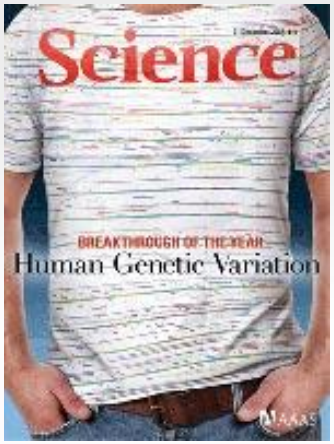
VariantWarehouse

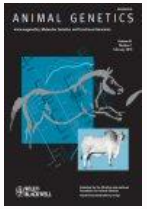
- Variant Warehouse
- Centralized Annotations
- Hosted Reports
- Sharing and Integration

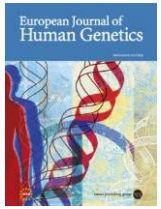
SNP & Variation Suite

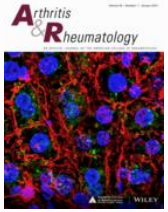
- GWAS
- Genomic Prediction
- Large-N-Population Studies
- RNA-Seq
- Large-N CNV-Analysis

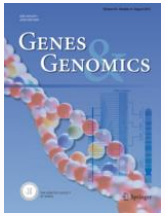
Cited in over 1,100 peer-reviewed publications











Over 350 customers globally

SickKids
THE HOSPITAL FOR SICK CHILDREN

Cincinnati Children's
Hospital Medical Center

The Children's Hospital of Philadelphia

Stanford University

UBC

Ucla

Yale

The Feinstein Institute for Medical Research
North Shore-Long Island Jewish Health System
North Shore LIJ

PREVENTION GENETICS
DISEASE PREVENTION THROUGH GENETIC TESTING

MAYO CLINIC

LINEAGEN

MOFFITT CANCER CENTER

Dartmouth-Hitchcock
NORRIS COTTON CANCER CENTER

BAYER

Lilly

abbvie

TEVA

GOLDEN HELIX
Enabling Precision Medicine

NATIONAL CANCER INSTITUTE

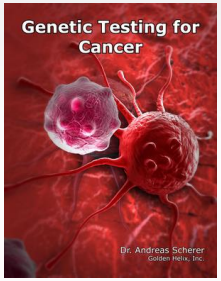
USDA
USDA Agricultural Research Service

FDA

Golden Helix – Who We Are

When you choose a Golden Helix solution, you get more than just software

- REPUTATION
- TRUST
- EXPERIENCE

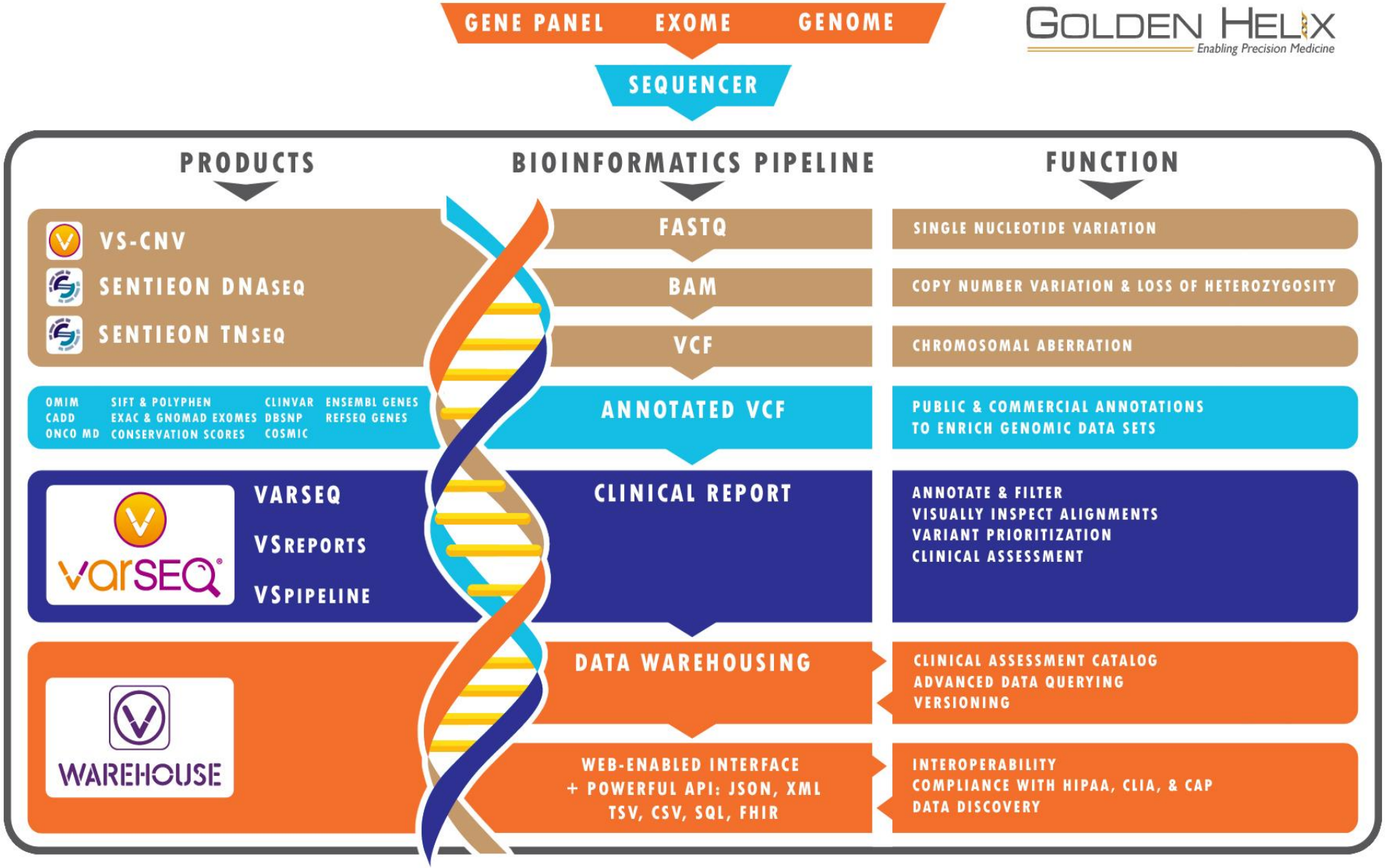


- INDUSTRY FOCUS
- THOUGHT LEADERSHIP
- COMMUNITY

- TRAINING
- SUPPORT
- RESPONSIVENESS

- TRANSPARENCY
- INNOVATION and SPEED
- CUSTOMIZATIONS

End-to-End Architecture for Clinical Testing Labs



The Sentieon Genomic Tools – Enabling Precision Data for Precision Medicine

WWW.SENTIEON.COM

What is Precision Data

- Hottest Word: Big Data
- But, PRECISION is the goal:
  precision recommendation, precision prediction
- Big genomics data for precision medicine:
  -precision diagnostics
  -improve precision treatment for individual

Big + Accurate | “better data in, better results out”

→ **Precision Data**

Sentieon's Mission

Enable Precision Genomics Data for Precision Medicine

- Ability to process big data at affordable cost and time
- With confidence
  - The highest accuracy
  - Consistent results

Three components of analytical software

-mathematical methods → **-compute algorithms** → **-software implementation**

-Same mathematical models as the Broad Institute

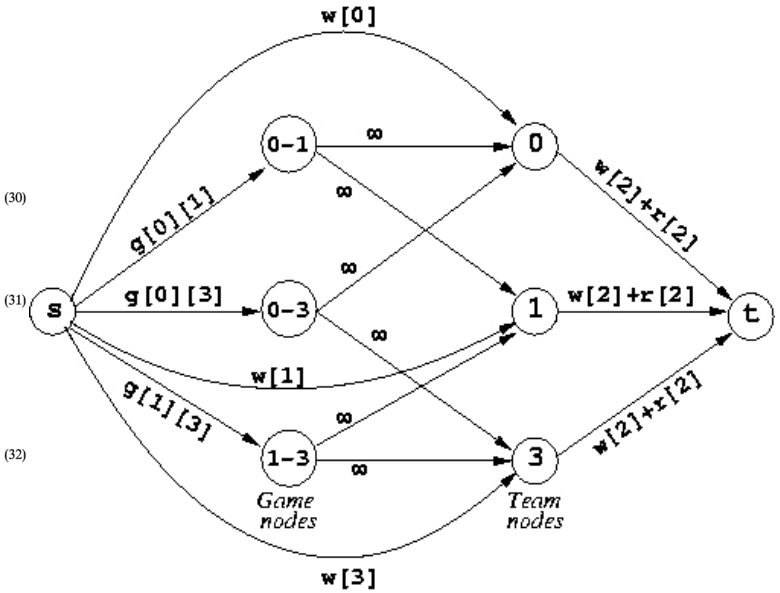
-more efficient compute algorithms

-Enterprise strength software implementation

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

$$(1+e_4)\frac{D}{Dt}\overline{\left(\frac{T'}{\tilde{T}}\right)^2} - 2\overline{f(t)\left(\frac{T'}{\tilde{T}}\right)^2} - 2\overline{w'^\alpha\frac{T'}{\tilde{T}}}D_\alpha + \frac{1}{(1+e_4)\tilde{\rho}C_p^2}\nabla_\alpha\left[(1+e_4)^2C_p^2\tilde{\rho}\overline{w'^\alpha\left(\frac{T'}{\tilde{T}}\right)^2}\right] + \frac{1+e_4}{\tilde{\rho}}\overline{\left(\frac{T'}{\tilde{T}}\right)^2\nabla_\alpha(\rho u'^\alpha)}$$
$$+\frac{2}{\tilde{\rho}\tilde{T}C_p}\overline{\frac{T'}{\tilde{T}}\left[P'\nabla_\alpha w'^\alpha - \nabla_\alpha(P'_g w'^\alpha) - \frac{DP'_g}{Dt}\right]} = \frac{2}{\tilde{\rho}\tilde{T}C_p}\overline{\frac{T'}{\tilde{T}}[\sigma'^{\alpha\beta}(u')\nabla_\alpha u'_\beta - \nabla_\alpha F'^\alpha_r]} = -\epsilon_2, \quad (31)$$

$$(1+e_4)\left[\frac{D}{Dt}\overline{\left(w'^i\frac{T'}{\tilde{T}}\right)} + \overline{w'^\alpha\frac{T'}{\tilde{T}}}\nabla_\alpha\tilde{u}^i - \alpha\overline{\left(\frac{T'}{\tilde{T}}\right)^2}g^{i\beta}\left(\nabla_\alpha\tilde{\Phi} + \frac{D\tilde{u}_\beta}{Dt}\right)\right] - f(t)\overline{w'^i\frac{T'}{\tilde{T}}} - \overline{w'^i w'^\alpha}D_\alpha$$
$$+\frac{1}{\tilde{\rho}C_p}\nabla_\alpha\left[(1+e_4)C_p\tilde{\rho}\overline{w'^i w'^\alpha\frac{T'}{\tilde{T}}}\right] + \frac{1+e_4}{\tilde{\rho}}\overline{w'^i\frac{T'}{\tilde{T}}\nabla_\alpha(\rho u'^\alpha)} + \frac{1}{\tilde{\rho}\tilde{T}C_p}\overline{w'^i\left[P'\nabla_\alpha w'^\alpha - \nabla_\alpha(P'_g w'^\alpha) - \frac{DP'_g}{Dt}\right]}$$
$$= \frac{1+e_4}{\tilde{\rho}}\overline{\frac{T'}{\tilde{T}}\nabla_\alpha\sigma'^{i\alpha}(u')} + \frac{1}{\tilde{\rho}\tilde{T}C_p}\overline{w'^i[\sigma'^{\alpha\beta}(u')\nabla_\alpha u'_\beta - \nabla_\alpha F'^\alpha_r]} = -\epsilon_2^i, \quad (32)$$



http://www.cs.princeton.edu/courses/archive/spr05/cos226/assignments/baseball/

```c
/* 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();
}
```

http://www.wikiwand.com/en/Programming_language

Sentieon

DNAseq

TNseq

DNAseq

Identical* results as
Broad Institute's
"Best Practice Workflow"
BWA-GATK HaplotypeCaller

TNseq

Identical* results as
Broad Institute's
"Somatic Variant Discovery Workflow"
MuTect and MuTect2

- Identical math, much more efficient computing algorithm and enterprise-strength software engineering
- 10+ X faster whole pipeline FASTQ to VCF; 20X-50X faster GATK/MuTect/MuTect2 portion; in core-hours
- Both products kept up-to-date with Broad Institute's releases

*1/1000 vcf differences due to GATK down-sampling, thread dependency, rounding differences

GATK is the gold standard

GATK Haplotype Caller is the most accurate DNA analysis tool

- "As of GATK version 3.3, we recommend using HaplotypeCaller in all cases, with no exceptions."
  - Broad Institute
- "Haplotype Caller is more accurate than the Unified Genotyper"
  - https://hpc.mssm.edu/files/Carneiro_workshop.pdf
- "GATK HaplotypeCaller called a substantial number of indels not called using VarScan-Cons (as well as GATK UnifiedGenotyper)"
  - Warden CD, Adamson AW, Neuhausen SL, Wu X. PeerJ. 2014 Sep 30;2:e600. doi: 10.7717/peerj.600. eCollection 2014.

DNAseq

TNseq

GATK is the gold standard

GATK Haplotype Caller is the most accurate DNA analysis tool

But GATK Haplotype Caller is too slow

Other speedup efforts:

(1) massive parallel on cloud (challenges: workflow, data privacy, cost) – may combine with our solution;

(2) hardware acceleration (challenges: cost, intrusive, inflexibility, scalability);

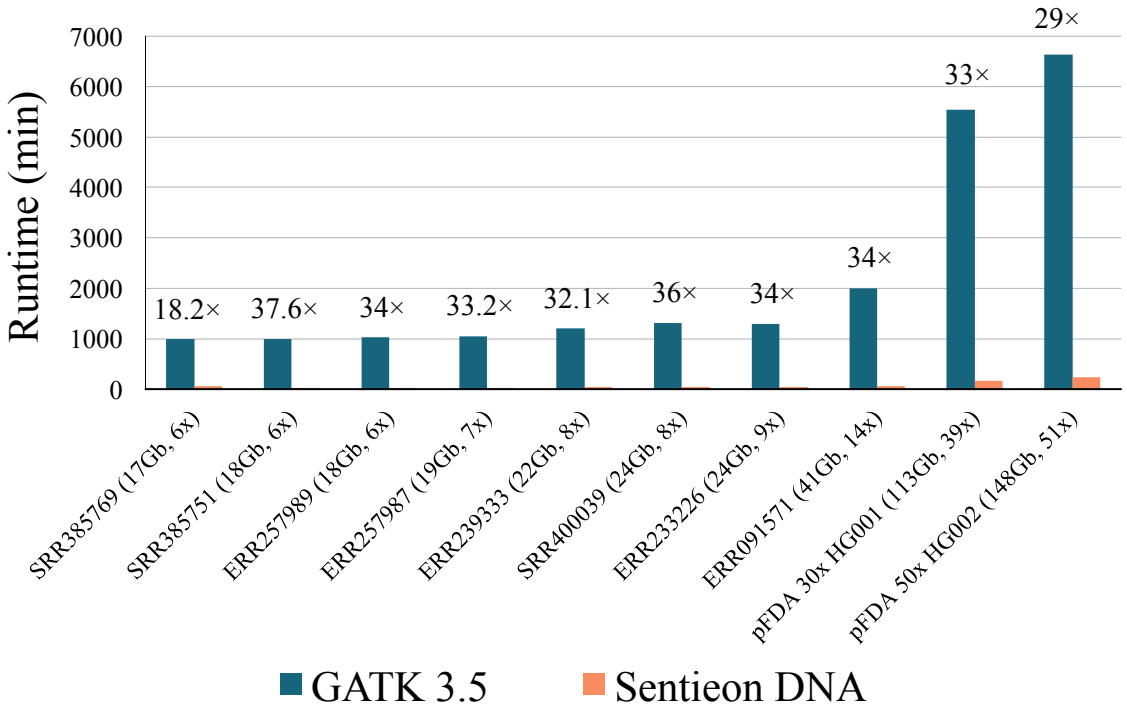
(3) corner-cutting (challenge: lose info)

Our approach: stay with the most accurate math, but use much more efficient compute algorithm with enterprise-strength software implementation

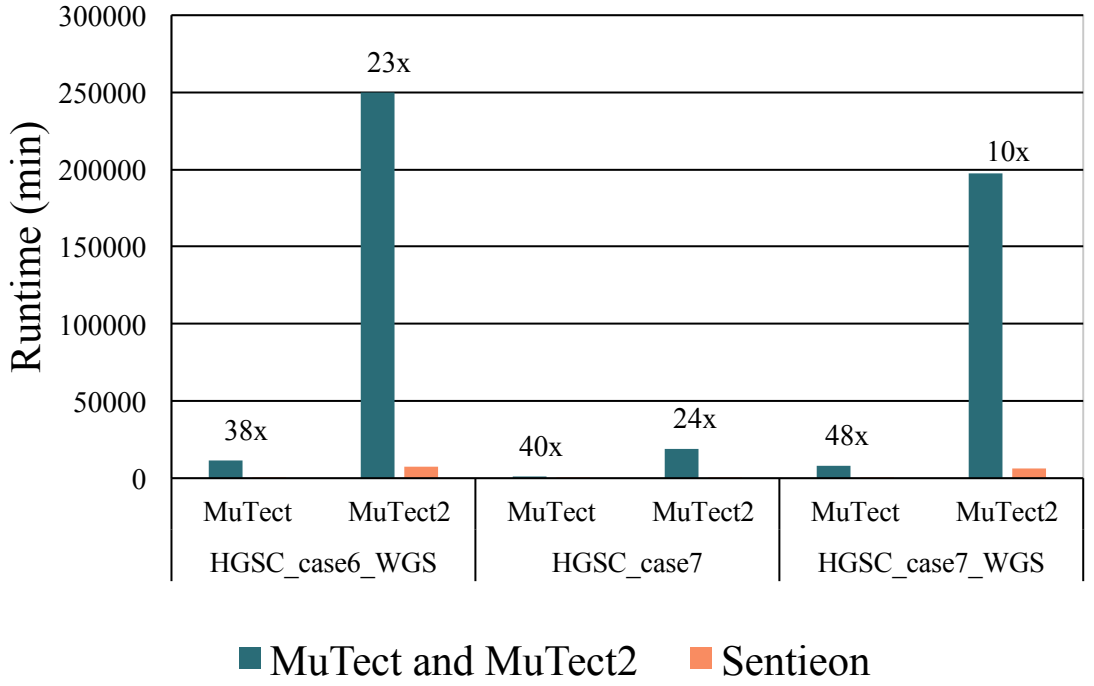
DNAseq

TNseq

DNAseq



TNseq



*Server specs: 32 core 2.4 GHz Intel Xenon server, 64 GB memory

DNAseq

TNseq

Highlights beyond speed:

- 100% consistency, no run-to-run differences

  No down-sampling in high coverage region, no thread dependency

  ➔ Higher accuracy by eliminating software noise
- System robustness

  ➔ Large dataset joint call over 100K samples together (without intermediate merging)

precisionFDA

DREAM

Consistency Challenge

(2/25-4/25/2016)

-Top Overall Performance

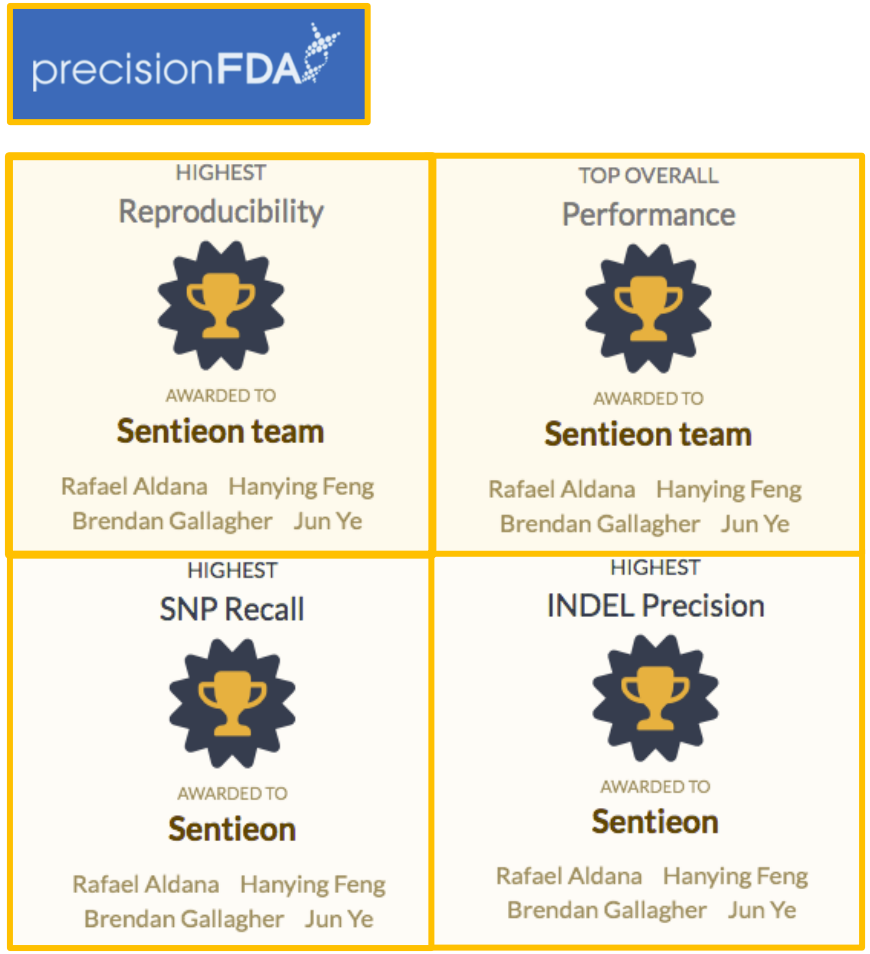
-Highest Reproducibility

Truth Challenge

(4/26-5/26/2016)

-Highest INDEL Precision

-Highest SNP Recall



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

precisionFDA

DREAM

ICGC-TCGA DREAM Mutation Calling challenge

Annual open contest by ICGC-TCGA for somatic variant calling accuracy.

Challenge-6 due 8/19/2016 (extended from 4/22/2016)

precisionFDA

DREAM

ICGC-TCGA DREAM Mutation Calling challenge

DREAM CHALLENGES

Final Leaderboard (8/19/2016)

SNV	INDEL	SV
Sentieon 98.57%	Sentieon 98.14%	Sentieon 100%
on 4/21 98.17%	on 4/21 97.48%	on 4/21 100%
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-

Consistency

precisionFDA Consistency Challenge Reproducibility

F1-score(%) between runs and between samples

	Garvan vs. Garvan rerun			Garvan vs HLI		
	FP	FN	F1-score	FP	FN	F1-score
Sentieon by UNM*	0	0	100	134633	315831	95.07
Sentieon (dual mapping)	0	0	100	107302	295568	95.53
ISAAC**	0	0	100	112255	266952	95.32
Genalice**	3621	3673	99.92	286306	433504	92.08
Edico Dragen**	3147	3216	99.93	161213	315611	94.87

* Sentieon standard pipeline results from Jeremy Edwards (University of New Mexico) submission

** Edico, Genalice (MAP 2.2.0) and Isaac (aligner v01.14.07.14 & variant caller v2.0.13) run results from Changhoon Kim's (Macrogen Clinical Laboratory) submission

Accuracy

precisionFDA Consistency Challenge Accuracy

F1-score(%) to NIST truth set

	Garvan sample			HLI sample		
	All	SNP only	Indel only	All	SNP only	Indel only
Sentieon by UNM*	99.39	99.86	95.85	98.97	99.73	92.94
Sentieon (dual mapping)	99.47	99.88	96.37	99.06	99.77	93.46
ISAAC**	97.29	98.57	86.66	96.34	97.99	82.10
Genalice**	98.04	99.20	89.10	97.25	98.83	84.66
Edico Dragen**	99.25	99.74	95.49	98.85	99.62	92.74

* Sentieon standard pipeline results from Jeremy Edwards (University of New Mexico) submission

** Edico, Genalice (MAP 2.2.0) and Isaac (aligner v01.14.07.14 & variant caller v2.0.13) run results from Changhoon Kim's (Macrogen Clinical Laboratory) submission

Accuracy

precisionFDA Truth Challenge Accuracy

F1-score(%) to NIST truth set

	HG002 sample			HG001 sample		
	All	SNP only	Indel only	All	SNP only	Indel only
Sentieon	99.8950	99.9548	99.3628	99.8971	99.9605	99.3324
GATK (Sanofi)	99.8905	99.9456	99.4009	99.7053	99.7483	99.3229
Consensus (Bina)	99.8805	99.9382	99.3675	99.8776	99.9325	99.3903
Verily	99.8597	99.9587	98.9802	99.8620	99.9554	99.0328
Edico Dragen (Macrogen)	99.7569	99.8268	99.1359	99.7739	99.8454	99.1397
GATK (Garvan)	99.5679	99.5934	99.3424	99.6208	99.6487	99.3742
Genalice (Macrogen)	98.9492	99.5188	93.9183	98.9624	99.5204	94.0347
Isaac (Macrogen)	98.2618	98.5357	95.8099	98.2800	98.5433	95.9257

Observations:

- Much higher F1 score (~99.9%) than Consistency Challenge, due to masking out complex regions in truth set
- Sentieon has excellent consistency between HG001 and HG002, due to no run-to-run differences

Value of Sentieon solutions

- Highest accuracy: most rigorous math, no noise in algorithm and software
- No down-sampling in high-coverage regions ← Critical for clinical samples
- No run-to-run difference ← Critical for medical decision
  --- proven in precisionFDA challenges and DREAM challenge
- Fast turnaround, ability to scale below 20 minute turnaround → Improved productivity, faster medical decision
- >10X reduced core-hours → drastically reduced compute cost
- Enable large dataset joint call → Enable genomics big data analysis

Sentieon tools: fastest, most accurate, zero run-to-run difference, no down-sampling, large cohort joint call, pure software solution, running on any generic computer systems

DNAseq

TNseq

Sentieon Software in Use

- Deployed at >100 sites worldwide (academia and industry)
- Accumulated usage at customer sites:
  >100K WGS/WES, >5e15 bases

Sentieon products are built on the solid foundation of the most rigorous and most extensively validated mathematical models used in Broad Institute's Best Practice Workflow, but with more efficient computing algorithms and enterprise-strength software implementation

Thank You

Contact Don Freed (don.freed@sentieon.com)
if you would like to talk to me in person