High Accuracy Somatic Variant Detection with Sentieon TNscope

Sentieon

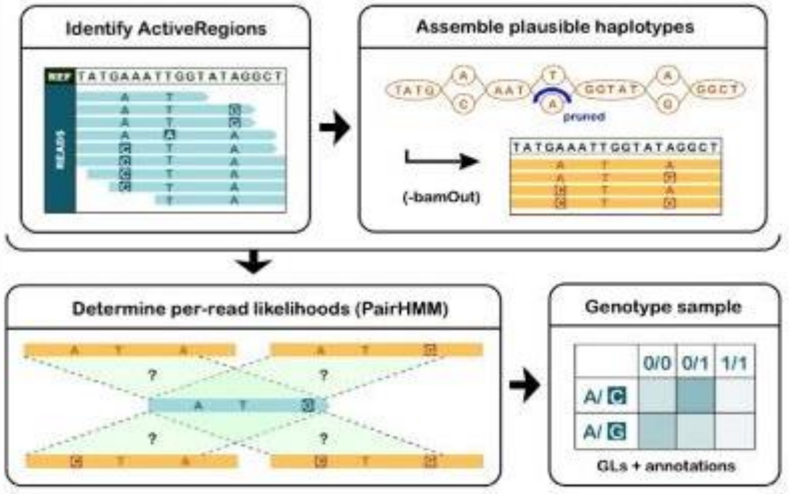
WWW.SENTIEON.COM

Somatic Variant Calling in Cancer

- Applications
  - Understanding cancer biology (TCGA, ICGC, etc.)
  - Improved diagnosis
  - NGS-guided therapy
  - Pharmacogenomics
  - Neoantigen discovery for cancer immunotherapy
- High accuracy is crucial
  - False-positives may result in ineffective treatments
  - False-negatives may results in missed treatment opportunities

Best Practices in Somatic Variant Calling

- Haplotype-aware variant calling
- Rigorous statistical model of errors in the NGS data



$$TLOD = log_{10}\left(\frac{L(M_f^m)P(m,f)}{L(M_0)(1\ -P(m,f)}\right)$$

$$NLOD = log_{10}\left(\frac{L(M_0)P(m,f)}{L(M_{0.5}^m)P(germline)}\right)$$

R. Poplin, *et al*. Scaling accurate genetic variant discovery to tens of thousands of samples. (https://doi.org/10.1101/201178)
K. Cibulskis, *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.

MuTect and MuTect2

- Rigorous mathematical model
- MuTect2 has haplotype-based variant calling
- Over 1,200 citations
- "…we recommend joint tumor-normal calling with MuTect, EBCall or Strelka…" R. Bohnert, *et al.* (2017)

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

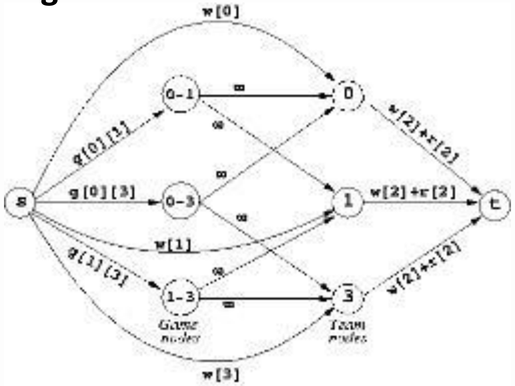
-compute algorithms

-software implementation

-Same mathematical models as the Broad Institute

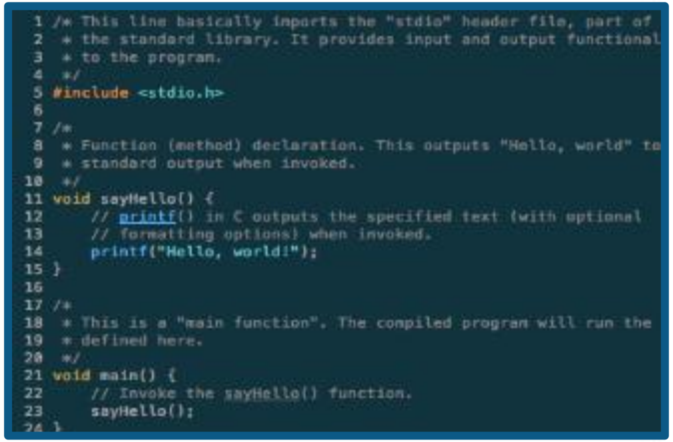
$$
\begin{aligned}
&\frac{D}{Dt}\overline{w'^i w'^j} + \overline{w'^i w'^\alpha}\nabla_\alpha \bar{u}^j + \overline{w'^j w'^\alpha}\nabla_\alpha \bar{u}^i - \alpha\left(g^{i\alpha}\overline{w'^j \frac{T'}{\bar{T}}} + g^{j\alpha}\overline{w'^i \frac{T'}{\bar{T}}}\right)\left(\nabla_\alpha \Phi + \frac{D\bar{u}_\alpha}{Dt}\right) \\
&\quad + \frac{1}{\bar{\rho}}\nabla_\alpha\left[\overline{\bar{\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')}\right] \\
&\quad + \frac{1}{\bar{\rho}}\overline{w'^i w'^j \nabla_\alpha(\bar{\rho} u'^\alpha)} - \overline{P'(g^{i\alpha}\nabla_\alpha w'^j + g^{j\alpha}\nabla_\alpha w'^i)} = -\frac{1}{\bar{\rho}}\left[\overline{\sigma^{i\alpha}(u')\nabla_\alpha w'^j} + \overline{\sigma^{j\alpha}(u')\nabla_\alpha w'^i}\right] = -\epsilon_2^{ij}\,, \qquad (30) \\
&(1+e_4)\frac{D}{Dt}\overline{\left(\frac{T'}{\bar{T}}\right)^2} - 2f(t)\overline{\left(\frac{T'}{\bar{T}}\right)^2} - 2\overline{w'^\alpha \frac{T'}{\bar{T}}}D_\alpha + \frac{1}{(1+e_4)\bar{\rho}C_p^2}\nabla_\alpha\left[(1+e_4)^2 C_p^2 \bar{\rho}\overline{w'^\alpha\left(\frac{T'}{\bar{T}}\right)^2}\right] + \frac{1+e_4}{\bar{\rho}}\overline{\left(\frac{T'}{\bar{T}}\right)^2 \nabla_\alpha(\rho u'^\alpha)} \\
&\quad + \frac{2}{\bar{\rho}\bar{T}C_p}\overline{\frac{T'}{\bar{T}}\left[P'\nabla_\alpha w'^\alpha - \nabla_\alpha(P'_g w'^\alpha) - \frac{DP'_g}{Dt}\right]} = \frac{2}{\bar{\rho}\bar{T}C_p}\overline{\frac{T'}{\bar{T}}\left[\sigma^{\alpha\beta}(u')\nabla_\alpha u'_\beta - \nabla_\alpha F'^\alpha_r\right]} = -\epsilon_2\,, \qquad (31) \\
&(1+e_4)\left[\frac{D}{Dt}\left(\overline{w'^i \frac{T'}{\bar{T}}}\right) + \overline{w'^\alpha \frac{T'}{\bar{T}}}\nabla_\alpha \bar{u}^i - \alpha\overline{\left(\frac{T'}{\bar{T}}\right)^2} g^{i\alpha}\left(\nabla_\alpha \Phi + \frac{D\bar{u}_\alpha}{Dt}\right)\right] - f(t)\overline{w'^i \frac{T'}{\bar{T}}} - \overline{w'^i w'^\alpha}D_\alpha \\
&\quad + \frac{1}{\bar{\rho}C_p}\nabla_\alpha\left[(1+e_4)C_p \bar{\rho}\overline{w'^i w'^\alpha \frac{T'}{\bar{T}}}\right] + \frac{1+e_4}{\bar{\rho}}\overline{w'^i \frac{T'}{\bar{T}}\nabla_\alpha(\rho u'^\alpha)} + \frac{1}{\bar{\rho}\bar{T}C_p}\overline{w'^i\left[P'\nabla_\alpha w'^\alpha - \nabla_\alpha(P'_g w'^\alpha) - \frac{DP'_g}{Dt}\right]} \\
&\quad = \frac{1+e_4}{\bar{\rho}}\overline{\frac{T'}{\bar{T}}\nabla_\alpha \sigma^{i\alpha}(u')} + \frac{1}{\bar{\rho}\bar{T}C_p}\overline{w'^i\left[\sigma^{\alpha\beta}(u')\nabla_\alpha u'_\beta - \nabla_\alpha F'^\alpha_r\right]} = -\epsilon_2^i\,, \qquad (32)
\end{aligned}
$$

-more efficient compute algorithms



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

-Enterprise strength software implementation



```
/* 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

The Sentieon Genomic Tools

- Identical* results to BWA-MEM, Picard, BQSR, GATK HaplotypeCaller, MuTect (TNsnv), MuTect2 (TNhaplotyper)
- Consistency
  - Winner of pFDA consistency challenge
  - No random seed
- 10x faster fastq to VCF in core-hours
- Processes all the data (no downsampling)
- An enterprise-strength implementation
  - Rigorous testing and architecture
  - Easy parallelization

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

Sentieon TNscope

Improving upon the mathematical model of MuTect2

- Uses the same general mathematical model used in MuTect and MuTect2
- Haplotype-based variant detection, including joint genotyping of haplotypes in the tumor and normal samples
- Improvements
  - Improved active regions
    - Use statistics for active region detection
    - More accurate detection of active regions
  - Improved local assembly
    - Assembles through “blindspots”
    - More accurate identification of the correct haplotype
  - A novel variant quality score combining NLOD and TLOD
  - Additional nonparametric variant annotations

ICGC-TCGA DREAM Mutation Calling Challenge 6

Final Leaderboard (8/19/2016)

DREAM CHALLENGES

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

Further Improvements of TNscope

Machine learning model for variant filtration

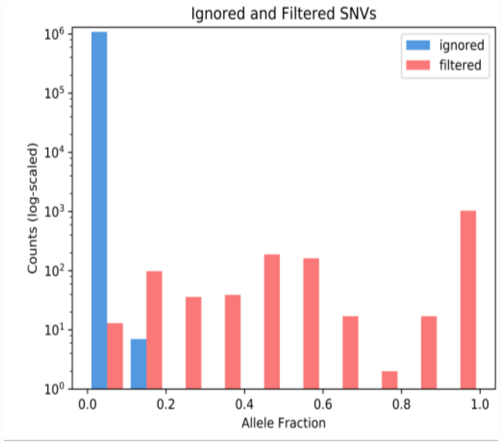
- Additional variant annotations allow for improved filtering
- Constructed a random forest model for variant filtration
- Model provides a single ensemble quality score for variant filtration
  - Tuned for maximum F1-score
  - Encompasses the most important variant annotations
  - Allows the user to set their desired sensitivity-specificity cutoff

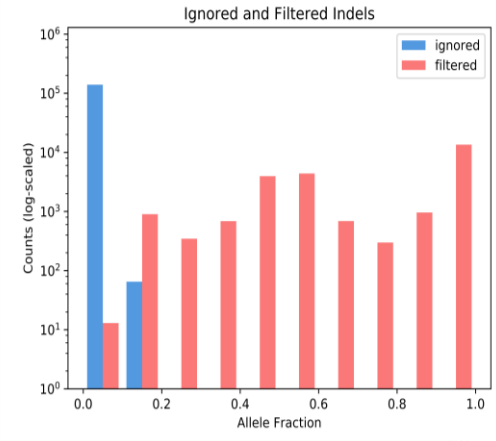
Benchmarking Methodology

- Use real sequence data
- Use samples with a known ground truth (GIAB samples)
- Can use in-silico mixtures of these data to create synthetic tumors
  - Variants will be present and 100% and 50% of the tumor sample purity
- Process these data through our standard variant calling pipelines

Benchmarking Truth Sets

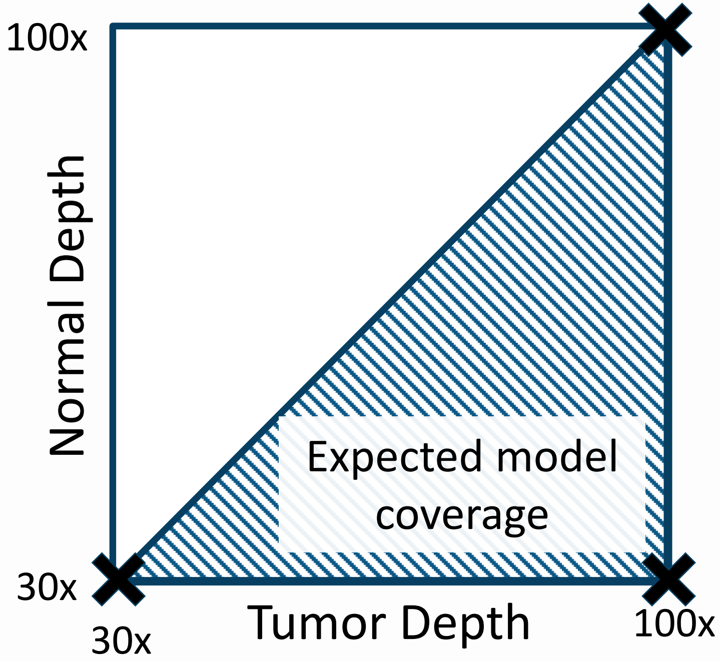
- Subtract variants in the normal sample from the tumor sample
- Intersect the high-confidence BED regions
- Remove unique sites in the tumor with substantial support in the normal sample
  - Mostly removes noisy indels





Model Training

- Trained with HG002 (tumor) and HG001 (normal) using ~2% of GIAB variants
- Trained with tumor sample purities of 10% and 30% (alternate allele fractions from 5% to 30%)
- Tumor normal depths
  - 30x – 30x
  - 100x – 30x
  - 100x - 100x



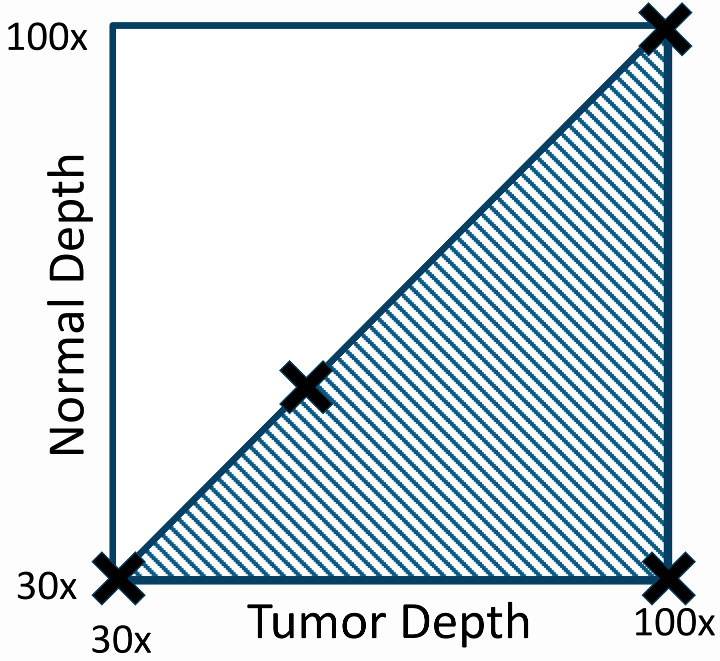
Model Performance

Performance of TNscope after application of the model on held-out data from HG001-HG002

			SNPs			Indels		
Tumor Purity	Tumor Depth	Normal Depth	Precision	Sensitivity	F1-Score	Precision	Sensitivity	F1-Score
0.3	100	100	0.990	0.998	0.994	0.934	0.9860	0.959
	100	30	0.990	0.997	0.994	0.944	0.973	0.959
	30	30	0.975	0.929	0.951	0.9290	0.875	0.901
0.1	100	100	0.989	0.891	0.938	0.932	0.822	0.874
	100	30	0.975	0.897	0.934	0.920	0.815	0.865
	30	30	0.956	0.469	0.630	0.885	0.398	0.550

Accuracy Benchmarking

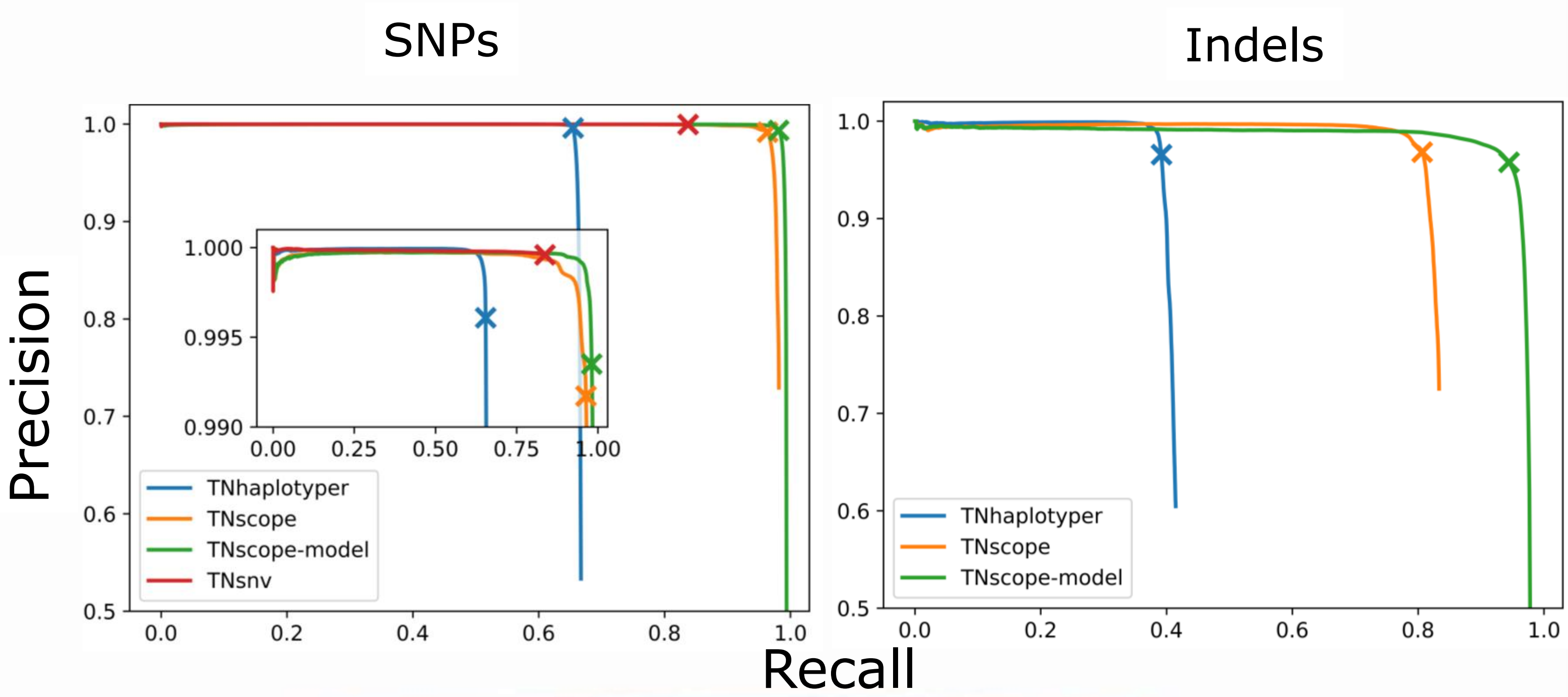
- HG005 (tumor) and HG004 (normal)
- TNsnv (MuTect), TNhaplotyper (MuTect2), TNscope, TNscope + model
- Tumor sample purities
  - 10%, 15%, 20%
- Tumor normal depths
  - 30x – 30x
  - 50x – 50x
  - 100x – 30x
  - 100x - 100x



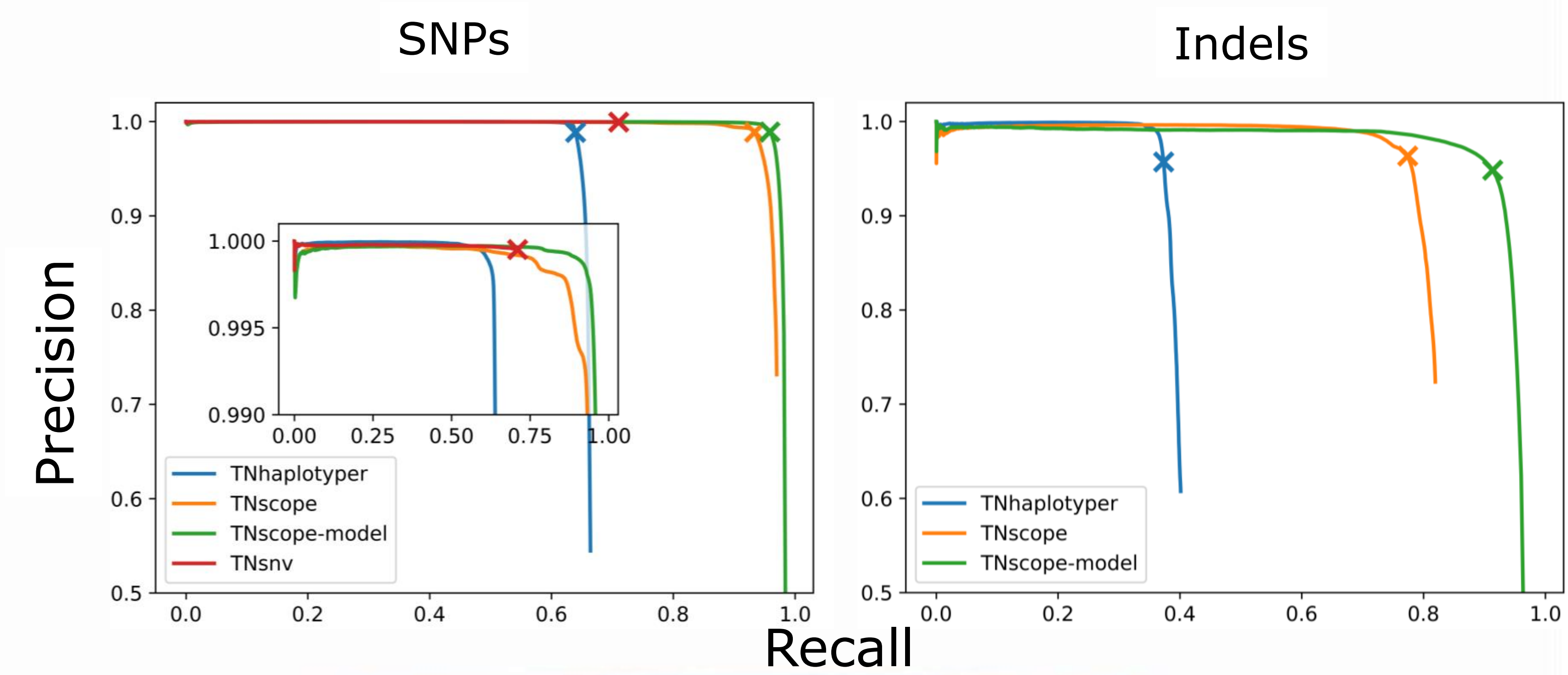
Benchmarking Results – F1-score

			SNPs				Indels		
Tumor Purity	Tumor Depth	Normal Depth	TNsnv	TNhap	TNscope	TNscope Model	TNhap	TNscope	TNscope Model
0.2	100	100	0.911	0.770	0.960	0.987	0.523	0.860	0.952
	100	30	0.912	0.773	0.963	0.985	0.522	0.881	0.946
	30	30	0.499	0.403	0.529	0.822	0.273	0.501	0.761
0.1	100	100	0.609	0.598	0.771	0.917	0.397	0.695	0.869
	100	30	0.597	0.592	0.760	0.914	0.395	0.699	0.856
	30	30	0.332	0.266	0.360	0.707	0.183	0.350	0.645

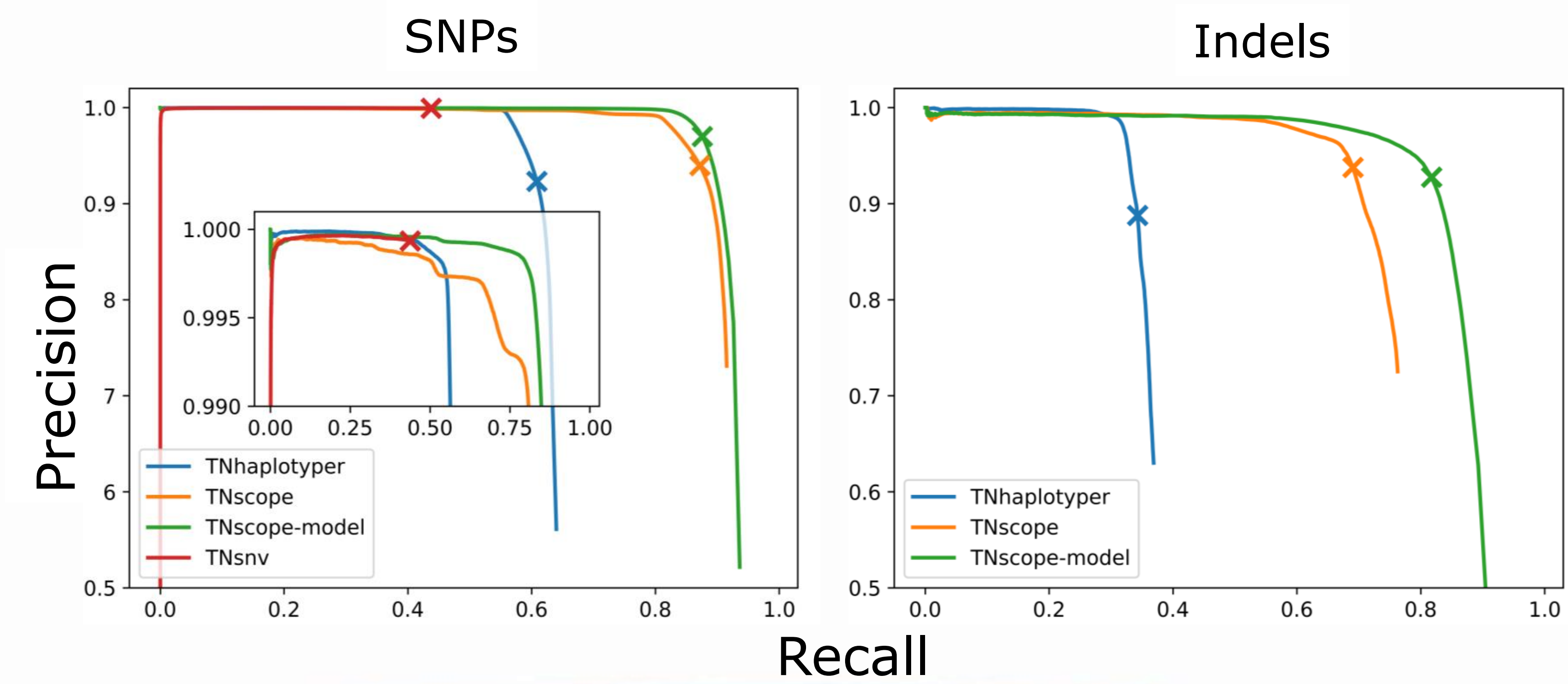
Tumor Purity – 20% 100x/100x Depth



Tumor Purity – 15% 100x/100x Depth



Tumor Purity – 10% 100x/100x Depth



Summary

- TNscope has substantially improved accuracy
    - TNscope significantly higher accuracy over MuTect and MuTect2
    - Accuracy is further improved using machine learning for variant filtration
- Published on bioRxiv - https://www.biorxiv.org/content/early/2018/01/19/250647
- Results generalize to other tumor-normal samples at similar depths

Thank You

Contact info@goldenhelix.com
for more information

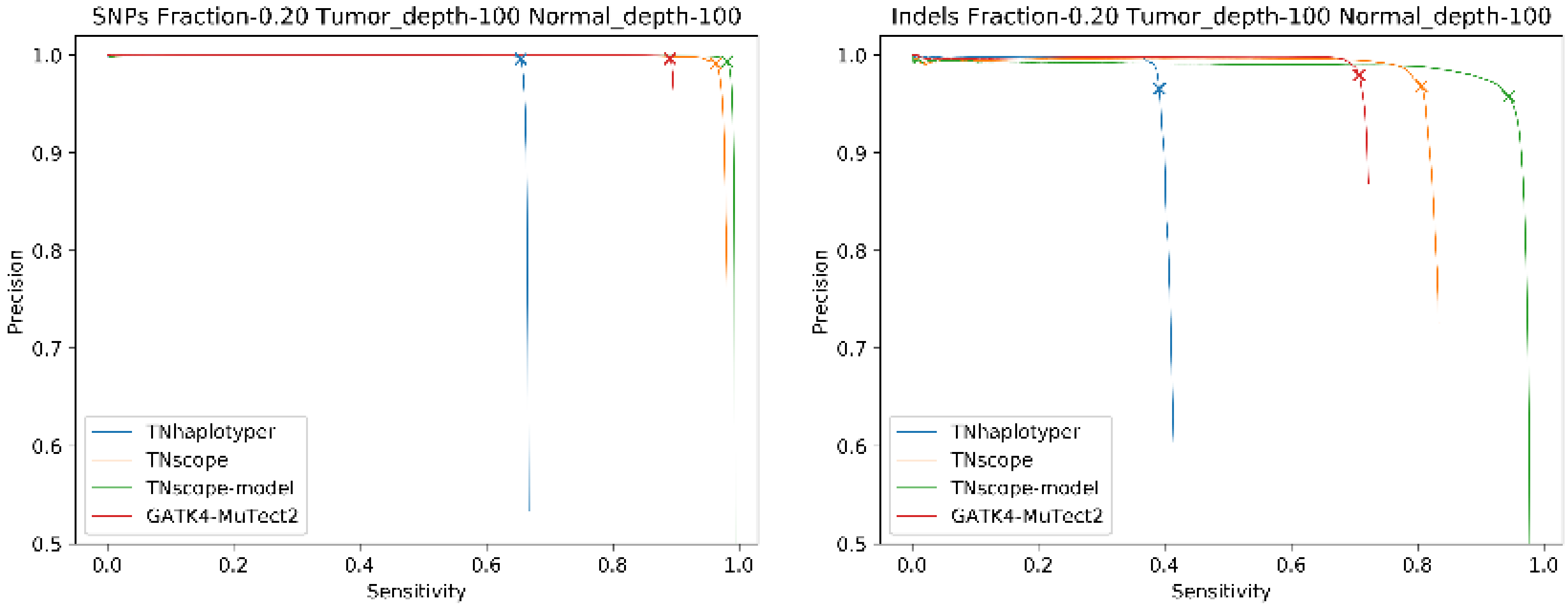
Email me at don.freed@setieon.com

Golden Helix – Special Pricing

- VarSeq PowerPack (2 seats) $17,500
  - VarSeq
  - VS-CNV
  - VSReports
  - Sentieon – Tier One

Contact info@goldenhelix.com | 406-999-0176

GATK4 MuTect2 – Preliminary Benchmarks



TNhaplotyper will match the GATK4 MuTect2 in the near future (with improved performance)