Accuracy improvements in somatic whole-genome small-variant calling with the DRAGEN platform

Konrad Scheffler, Sangtae Kim, Varun Jain, Jeffrey Yuan, Westley Sherman, Taylor O'Connell, Eric Ojard, Lisa Murray, Rami Mehio, and Severine Catreux

Illumina Inc., 5200 Illumina Way, San Diego, CA 92122, USA.

INTRODUCTION

- Next-generation whole-genome sequencing promises to enable dramatic expansion of precision oncology research and personalized cancer care.
- We present the DRAGEN somatic small-variant calling pipeline, which achieves improved turn-around-time and accuracy compared to state-of-the-art alternatives.
- DRAGEN somatic small-variant caller
  - Produced 27-52% and 18-97% fewer false single-nucleotide-variant (SNV) calls, and 32-75% and 42-89% fewer false indel calls than state-of-the-art pipelines Strelka2 and Mutect2.
  - Exhibits higher tolerance to tumor-in-normal contamination than Strelka2 and Mutect2.
  - The average end-to-end workflow runtime of the DRAGEN somatic pipeline was 66 minutes, 76% and 615% faster than Strelka2 and Mutect2 taking DRAGEN alignments as input.

METHOD HIGHLIGHTS

DRAGEN somatic caller with Strelka Genotyping

DRAGEN v3.4 | Sept. 2019

Hardware-accelerated (FPGA) implementation of GATK4/Mutect2 [1]

+

nature methods BRIEF COMMUNICATION

Strelka2: fast and accurate calling of germline and somatic variants

Sangtae Kim, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, Yeonbin Kim, Daniel Beyter, Peter Krusche and Christopher T. Saunders

Strelka2 [2]: Best-in-class in 2018 (combination of speed and accuracy)

Methods from DRAGEN v3.4

- Fast hardware-accelerated implementation of key algorithms
  - Smith-Waterman alignment
  - Pair Hidden Markov Models (HMMs)
- Adaptive parameter estimation
  - HMM parameters estimated per sample, separately for insertions and deletions depending on reference contexts

Methods from Strelka2

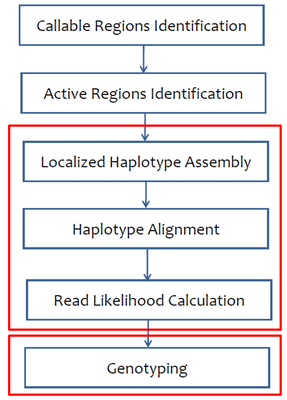
- Joint genotype modeling of tumor and normal samples allows shared systematic error and tumor-in-normal (TiN) contamination to be modeled.

Additional improvements

- Hardware-accelerated local *de novo* assembly
- Foreign Read Detection (FRD) model
  - Account for errors due to mismapped reads
  - Integration of the map quality (MAPQ) into the core probabilistic model
- Strand-Bias (SB) model
  - Account for errors affecting reads on one strand
- Both the FRD and strand-bias models are integrated into the core Bayesian model
- Additional filtering steps
  - Use signals not yet incorporated into the Bayesian model

METHODS

DRAGEN variant caller workflow

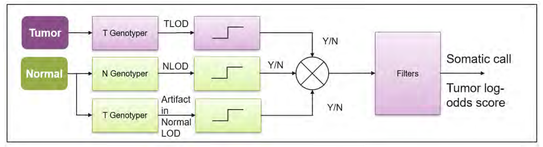


Hardware acceleration

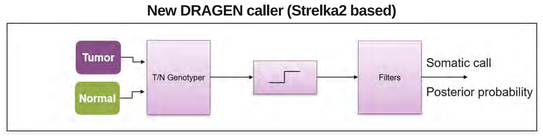
Strelka's joint genotyping with new FRD and SB modeling

Comparison of the legacy pipeline and the new pipeline

Legacy DRAGEN caller (GATK4/Mutect2)

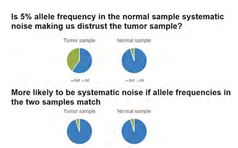


New DRAGEN caller (Strelka2 based)

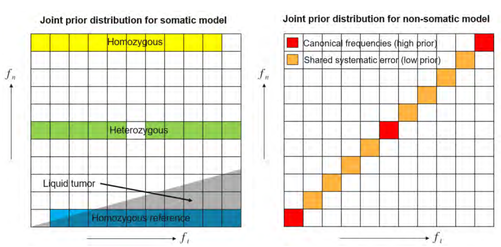


Somatic genotyping

- Core problem: Tumor allele frequencies are unknown
- Strelka solution: Discretize allele frequencies and integrate over all possible allele frequencies



Shared error model



Joint distributions of the frequency of candidate somatic allele for somatic (left) and non-somatic (right) model (f_t: tumor allele frequency, f_n: normal allele frequency)

RESULTS

Datasets for benchmarking

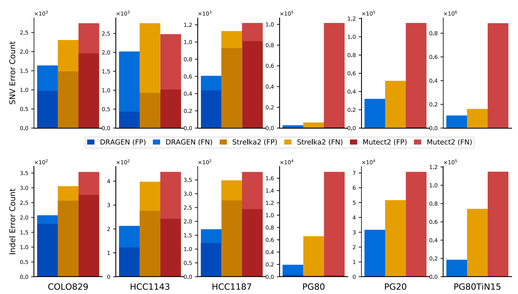
Dataset	Type	Depth (T:N)	#Variants in the truth set	Note
COLO829	Cancer Cell Line [3]	80:40	38k	COLO-829, Skin Melanoma
HCC1143		80:40	25k	HCC-1143, Ductal breast carcinoma
HCC1187		80:40	13.5k	HCC-1187, Ductal breast carcinoma
PG80	In silico NA12877/ NA12878 Mixture	110:40	1,158k	Tumor purity 80%, TiN 0%
PG20		110:40	1,158k	Tumor purity 20%, TiN 0%
PG80TiN15		110:40	1,158k	Tumor purity 80%, TiN 15%

All sequencing data were generated with Illumina NovaSeq using the TruSeq DNA PCR-free library prep kit. The 3 cancer cell line datasets are obtained from NYGC [3]. The PG datasets were generated in house by mixing NA12878 and NA12877 data.

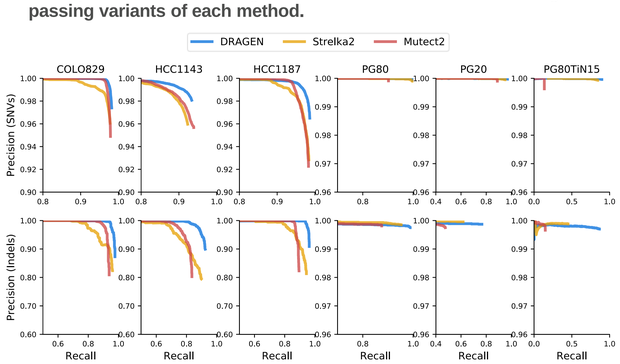
Benchmarking

- Tools: DRAGEN (v3.6) vs Strelka2 (v2.9.9) vs Mutect2 (included in GATK v4.1.2)
- Accuracy measurement: RTG vcfeval v3.9.1
- Truth sets: NYGC high confidence callset (COLO829, HCC1143, HCC1187), NA12878 variant calls where NA12877 genotype is homref (PG80, PG20, PG80TiN15)
- For the 3 cancer cell-line datasets for which the truth set is not comprehensive, we created "normal-normal" datasets by randomly assigning reads to two normal datasets with the depth profile matching the tumor-normal datasets. We used the normal-normal data for false positive counting and the tumor-normal data for false negative counting.

Accuracy comparison



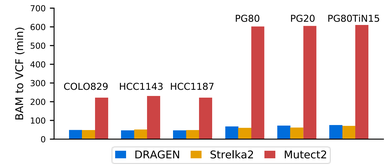
False positive and false negative error counts when considering passing variants of each method.



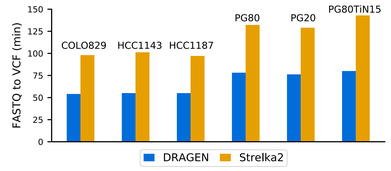
Precision/recall curves generated using the following scoring metrics: SQ for DRAGEN, TLOD for Mutect2, SomaticEVS for Strelka2.

RESULTS

Runtime comparison



Somatic-variant calling runtime. Each method took DRAGEN alignments (BAM format) as an input. The runtime was measured on DRAGEN servers with 2 Intel Xeon Gold 6126 CPUs (total 24 cores, 2.66 GHz), a Xilinx U200 FPGA board and 384 GB memory.



End-to-end workflow runtime for DRAGEN and Strelka2. DRAGEN is highly optimized for end-to-end workflows where the input consists of raw read (e.g. FASTQ), with alignment and variant calling performed in a single workflow. DRAGEN took raw reads (compressed FASTQ format) as an input. Strelka2's runtime is the sum of DRAGEN mapping/alignment time and Strelka2 runtime.

CONCLUSIONS

- DRAGEN outperforms other state-of-the-art solutions in terms of both accuracy and speed.
- DRAGEN is robust against variations in coverage, sequencing platform, sample preparation chemistry, and tumor purity (not shown here).
- The DRAGEN pipeline enables reliable whole-genome analysis that can be scaled to large numbers of samples for both research and clinical use.
- The presented algorithm will be available as part of DRAGEN version 3.6.
- Users who don't own a DRAGEN server can run it on the cloud: https://basespace.illumina.com

REFERENCES

1. Cibulskis, K et al., Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* (2013)
2. Kim, S. et al., Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* (2018)
3. Arora, K. et al., Deep Whole-Genome Sequencing of 3 Cancer Cell Lines on 2 Sequencing Platforms, *Sci Rep.* (2019)

illumina®

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