Somatic small variant and copy number alteration calling with the Genome Analysis Toolkit

LEE LICHTENSTEIN[1], JONN SMITH[1], DAVID BENJAMIN[1], AARON CHEVALIER[1], KRISTIAN CIBULSKIS[1], JULIAN HESS[1], SAMUEL K. LEE[1], IGNATY LESHCHINER[1], DIMITRI LIVITZ[1], DANIEL ROSEBROCK[1], VALENTIN RUANO-RUBIO[1], TAKUTO SATO[1], ANDREY SMIRNOV[1], CHIP STEWART[1], GAD GETZ[1,2], ERIC BANKS[1].

[1] Broad Institute, Cambridge, MA, USA [2] Massachusetts General Hospital, Boston, MA USA

#5108 – 16
Booth: #3048

Presenter at booth:
Sun, March 31, 2019: 3:30PM
Mon, April 1, 2019: 2PM
Tue, April 2, 2019: 10AM

Abstract

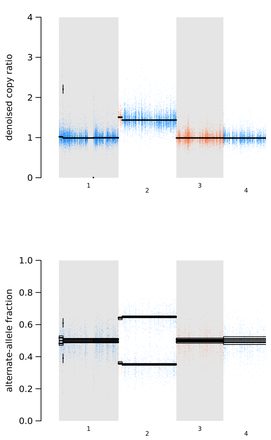
Somatic small mutations (SNVs or Indels) and copy number alterations are the two categories of mutations with the largest impact on cancer tumors. The Broad Institute has released somatic variant calling workflows for small mutations (Mutect2) and copy number alterations (ModelSegments) based on the Genome Analysis Toolkit (GATK[17]). These workflows are the result of extending and/or redesigning previous methods, such as those in the tools MuTect[15], Tangent Normalization[14], and Oncotator[10]. The new suite of workflows can call variants in capture or whole-genome sequencing data and will include functional annotations (Funcotator), such as protein change and impacted gene for small variants. Common artifacts in sequencing data, such as those arising from oxidative DNA damage[8,16], FFPE/deamination, or mapping errors[1,19], are corrected automatically. Evaluation of the workflows is standardized and repeatable, which allows tracking of performance across versions, both detection performance (e.g. sensitivity, precision), as well as runtime performance (e.g. CPU and RAM usage). The workflows are freely available, are portable (i.e. can be run on local, on-prem, or cloud compute), are optimized for cost reduction, and can be tuned to optimally leverage available compute.

Results: The measured sensitivity of M2 was at least 0.93 for small somatic nucleotide variants (SNVs) and 0.83 for small insertions/deletions (Indels) on synthetic data and on a titrated mixture of germline samples (>=100x depth, AF = 0.2). The measured precision of Mutect2 ranged from 0.94 to 0.98 on the synthetic data for both SNVs and Indels. The false positive count of Mutect2 was between 0 and 3 for SNVs, and between 0 and 2 for indels, on twenty pairs of replicate normal-normal samples. The cost of the M2 workflow is about USD$1.50 for a pair of 60/30x WGS matched tumor-normal samples, using Google Cloud Compute, and required about 32 hours of CPU time on a single core with 3GB RAM.

The measured sensitivity of ModelSegments was at least 0.91 for deletions and amplifications across three cohorts of TCGA whole-exome samples (Stomach adenocarcinoma N=39, Thyroid carcinoma N=50, and Lung adenocarcinoma N=60). The measured specificity for the same set of cohorts was at least 0.96 for both deletions and amplifications. All results reported here were using the corresponding SNP Array results as a truth set.

GATK MS cost was approximately USD$0.74 for a 60/30x WGS pair using Google Cloud Compute and required about 6 hours of CPU time with a single core. The RAM usage was varied automatically in the workflow to minimize cost, but was in the range of 2-13GB.

ModelSegments Method[1]



Denoising of copy ratio estimates, for each exome target, was performed with two steps, described in previous literature[1,2,14]. The first was to create a ratio of the proportional coverage for each target against the corresponding median in a panel of normals (PoN). The second was to perform a singular value decomposition (SVD) and projection to remove typical noise based on estimates from the PoN[14].

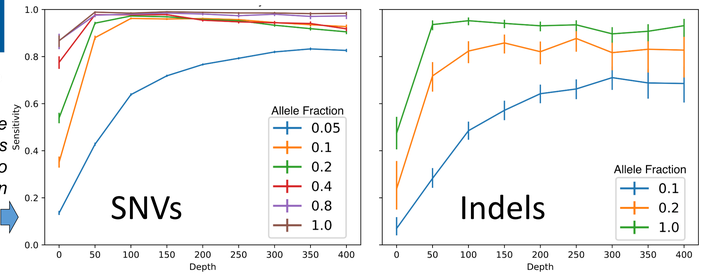
We infer allelic fraction of each segment from germline heterozygous SNP sites. A Bayesian model estimates the allelic fraction per segment and a global reference bias[1].

ModelSegments uses a nonlinear kernel segmentation algorithm[3], which enables a **single, joint segmentation of allelic fraction and copy ratio**. This is a difference from previous approaches[4].

Mutect2 (SNVs/Indels) Evaluation

All evaluations were run using default parameters, except where noted otherwise.

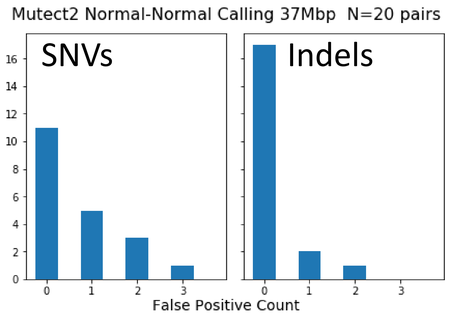
Titration mixtures drawn from ten HapMap[5] samples were sequenced at different ratios. This simulated somatic variants at different allelic fractions using known germline variants. No panel of normals was used. Cross-individual contamination correction was disabled.



Sample	SNV sens	SNV prec	Indel sens	Indel prec
DREAM 1	0.975	0.954	Not in DREAM1	
DREAM 2	0.979	0.944	Not in DREAM2	
DREAM 3	0.929	0.966	0.897	0.976
DREAM 4	0.843	0.959	0.814	0.984

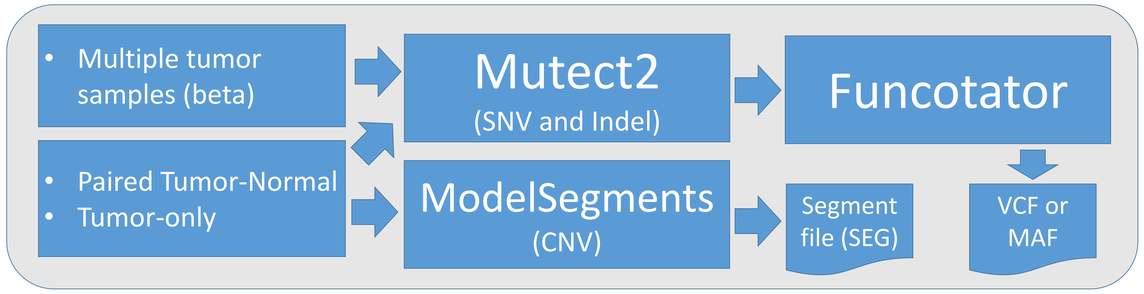
We ran M2 calls on the synthetic pairs from the somatic ICGC-TCGA DREAM8.5 challenge[6]. These samples generated a known truth set of with known events, introduced in silico.

Five WES replicates of a blood-normal (NA12878) were called as tumor-normal pairs in M2, for a total of twenty pairs. Any variant should be considered a false positive.



Percentage (Ground truth)	N	RMSE
0 - 1%	10	0.013
0 – 5%	15	0.013

CalculateContamination was run on mixtures of two non-replicate blood normal samples. The mixtures were made at known proportions. The RMSE was calculated based on the predicted cross-individual contamination and the mixture ground truth.



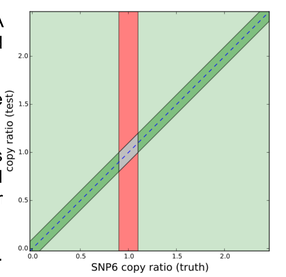
All evaluations run with GATK 4.1.0.0

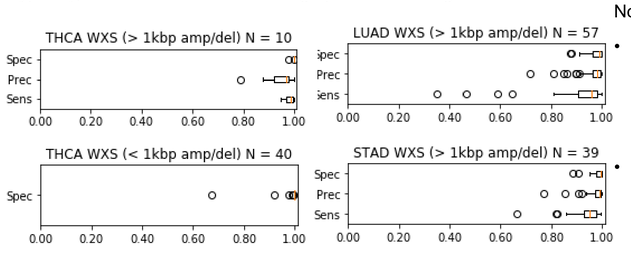
ModelSegments (CNV) Evaluation

We ran a concordance evaluation between ModelSegments and the TCGA SNP array pipeline[14] (as ground truth) to capture sensitivity, precision, and specificity. Default parameters for WES were used in ModelSegments.

- Evaluation was run on WES from three different tumor types, to capture low, medium, and high-mutation rate samples.
- Sensitivity, precision, and specificity were scored by concordant bases with the associated SNP arrays for a sample. To have precision and sensitivity scored, there had to be at least 1kbp of amplifications or deletions in the exome.
- Sensitivity outliers were due to differences in segment mean estimates. Typically, of few events that cover large portions of the genome.

precision = TP / (TP + FP)
sensitivity = TP / (TP + FN)
specificity = TN / (TN + FP)





Not shown:

- STAD WXS samples with less than 1kbp amplified or deleted (in the ground truth), since there were no samples fitting that criteria.
- Two LUAD WXS samples with less than 1kbp amplified or deleted (in the ground truth), both had specificity=1.0

Mutect2 Method[7]

Mutect2 is the next version of MuTect[15] and includes Indel calling and major differences to the core approach (assembly vs. pileup). Mutect2 can run on paired tumor-normal samples, multiple tumor samples (beta feature), or a single tumor sample.

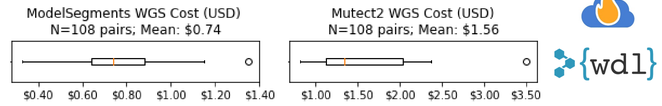
Steps 1 and 2) Local sequence assembly performed on reads near possible somatic variants to obtain set of candidate haplotypes. Then calculate the likelihood that each haplotype generated each read. These steps are described elsewhere[13].

Step 3) Use Bayesian somatic genotyping model to determine probability that candidate haplotypes are real.

Step 4) Apply filters to account for learned features of the tumor and for detecting artifacts, such as orientation bias artifacts (incl. deamination and Oxo-G[8,16]), cross-individual contamination (similar to ContEst[12]), and mapping errors[1,19].

Cost

We ran the M2 and MS workflows on 108 TCGA WGS tumor-normal (THCA: 50, LUAD: 58). For the runs, we ran the freely available workflows in FireCloud[9], which uses Google Compute Engine. By default, workflows are pre-emptible (≤5 preemptions) to reduce cost. The outlier sample in the ModelSegments plot was due to the maximum preemption count being hit for a coverage collection step (an expensive step). The outlier sample in the Mutect2 costs was due to a hyper-mutated LUAD sample, which caused extra processing.



Funcotator[11]

Funcotator is a functional annotation tool in the core GATK toolset. It was built as a successor to Oncotator[10], but to handle both somatic and germline use cases, rather than somatic only. Funcotator reads in a VCF file, labels variants with one of twenty-three distinct variant classifications, produces gene information, and associations to information in datasources. Supported datasources include GENCODE (gene information and protein change prediction), dbSNP, gnomAD, and COSMIC. The corpus of datasources are extensible and user-configurable. Cloud-based datasources are supported for Google Cloud Storage. Funcotator produces either a Variant Call Format (VCF) file (with annotations in the INFO field) or a Mutation Annotation Format (MAF) file. Annotation of seg files can still be performed using Oncotator[10].

References and acknowledgments

1. https://github.com/broadinstitute/gatk/blob/4.1.0.0/docs/CNVs/CNV-methods.pdf
2. http://gatkforums.broadinstitute.org/gatk/categories/recapseg-documentation
3. Celisse A, et al. New efficient algorithms for multiple change-point detection with kernels. **arXiv:1710.04556v1** [math.ST]
4. Olshen AB, et al. “Circular binary segmentation for the analysis of array-based DNA copy number data.” Biostatistics. 2004 Oct;5(4):557-72.
5. International HapMap 3 Consortium "Integrating common and rare genetic variation in diverse human populations". Nature. 467 (7311): 52–58. doi:10.1038/nature09298.
6. https://www.synapse.org/#!Synapse:syn312572/wiki/58893
7. https://github.com/broadinstitute/gatk/blob/4.1.0.0/docs/mutect/mutect.pdf
8. Costello M, et al. “Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation” Nucleic Acids Research, 41 (6) 2013, Page e67, https://doi.org/10.1093/nar/gks1443
9. Birger C, et al. 2017. “FireCloud, a cloud-based platform for collaborative genome analysis: Strategies for reducing and controlling costs” bioRxiv doi: 10.1101/209494
10. Ramos AH, et al. “Oncotator: Cancer Variant Annotation Tool. Human Mutation” (2015). http://dx.doi.org/10.1002/humu.22771
11. https://software.broadinstitute.org/gatk/documentation/article?id=11193#1.1
12. Cibulskis K, et al. “ContEst: estimating cross-contamination of human samples in next-generation sequencing data”, Bioinformatics, Volume 27, Issue 18, 15 September 2011, Pages 2601–2602, https://doi.org/10.1093/bioinformatics/btr446
13. Poplin R, et al. “Scaling accurate genetic variant discovery to tens of thousands of samples”, bioRxiv 201178; doi: https://doi.org/10.1101/201178
14. Tabak B, et al. “The Tangent copy-number inference pipeline for cancer genome analyses”, bioRxiv 566505; doi: https://doi.org/10.1101/566505
15. Cibulskis K, et al., “Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples” Nature Biotechnology volume 31, pages 213–219 (2013)
16. Stewart C et al., “Comment on ‘DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification’” Science 28 Sep 2018, doi: 10.1126/science.aas9824
17. McKenna et al., “The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data”, Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. Epub 2010 Jul 19.
18. Giannakis M, et al., “Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma”, Cell Rep. 2016 Apr 26;15(4):857-865. doi: 10.1016/j.celrep.2016.03.075
19. Ellrott K, et al., “Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines“, Cell Syst. 2018 Mar 28;6(3):271-281.e7. doi: 10.1016/j.cels.2018.03.002

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Workflow Definition Language (WDL) files for on-prem and Google Cloud compute are also available in the GATK github repo (https://github.com/broadinstitute/gatk)

Thanks to Chet Birger for editing help.