Icahn School
of Medicine at
**Mount
Sinai**

# Validating a new somatic mutation caller using TCGA data

**Arun Ahuja, Ryan Williams, Tim O'Donnell, Jeff Hammerbacher**
Icahn School of Medicine at Mount Sinai

## Overview

- ❖ Tuning a new somatic variant caller is difficult due to lack of readily available ground truth data. Most benchmarks use synthetic data containing known mutations or compare to an ensemble of existing callers.
- ❖ Some TCGA submissions include variants validated using targeted sequencing. These submissions enable an assessment of the sensitivity of a mutation caller on real data.
- ❖ Since the validated variants must first have been identified by a standard caller, they are subject to ascertainment bias. Therefore, TCGA validated variants cannot rigorously be used to compare callers or measure the absolute sensitivity of a caller, but they can find variants missed by a particular caller to diagnose errors.
- ❖ Using a set of 24,629 validated calls across 16 TCGA tumor/normal pairs, we evaluated the performance of two popular mutation callers, Strelka and Mutect, as well as an experimental caller we are developing called Guacamole.
- ❖ We intend to extend this preliminary work into a collection of curated calls to the aid the development of new variant callers.

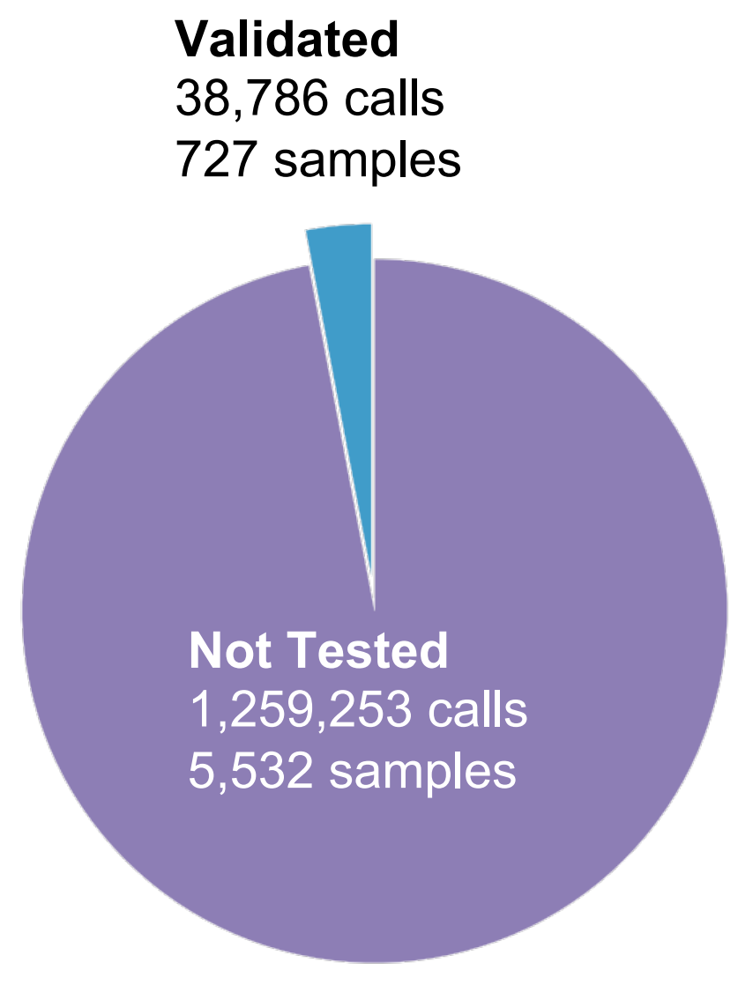
## Previous Work

The ICGC-TCGA DREAM Mutation Calling challenge is a competition to accurately call somatic variants. The first five phases have used synthetic data. The final phase will involve 10 patient datasets and validation of prioritized calls, but is not yet available.

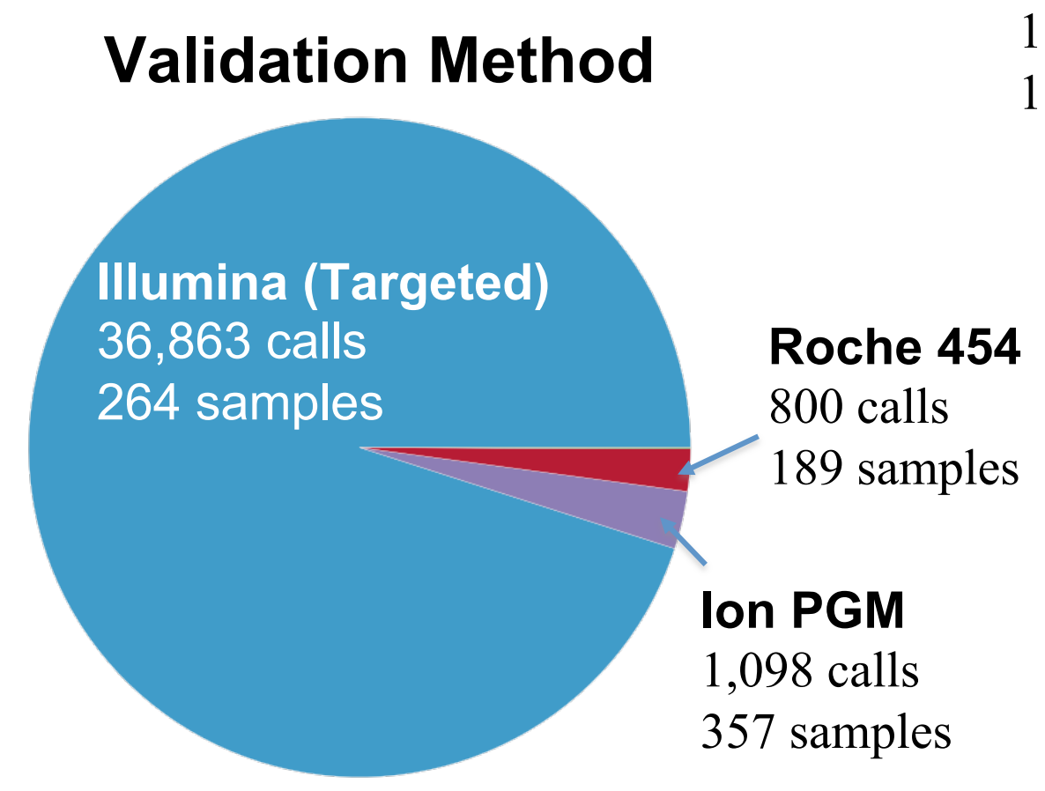
(Kim 2013) use validation data from TCGA to assess the performance of many callers, but their analysis cannot be readily reproduced as sample identifiers are not specified.

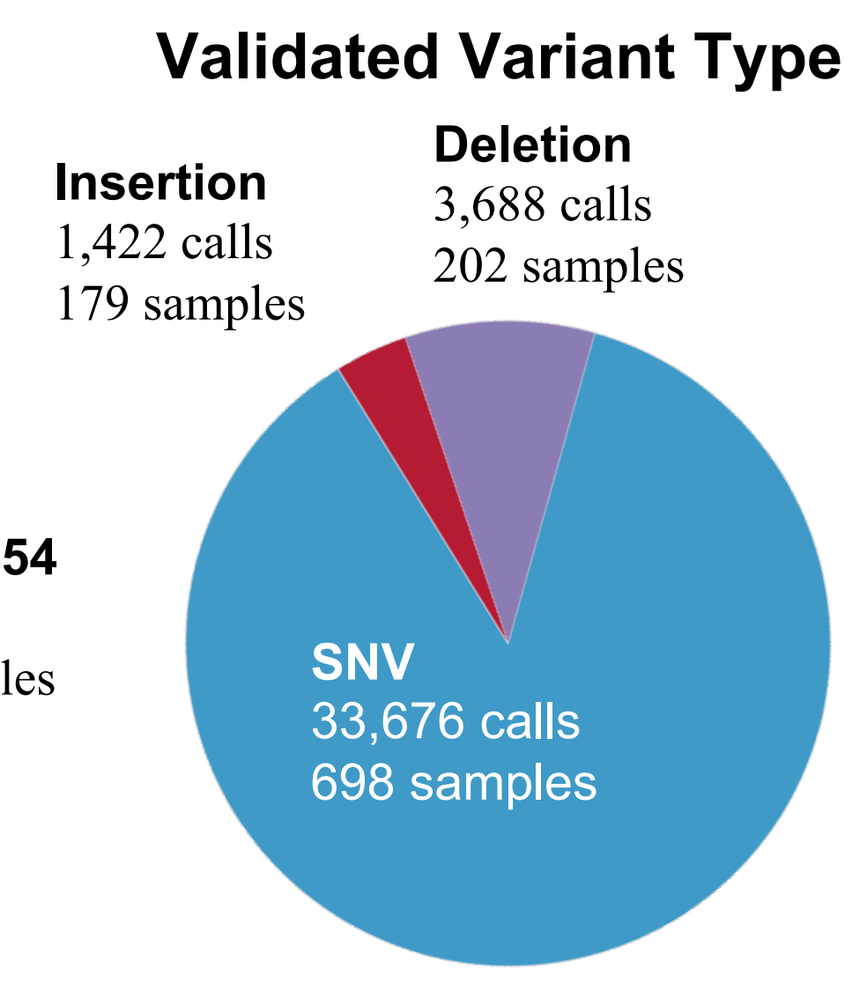
Other work has used mice (Löwer 2012), mixtures of normal cells to simulate tumor samples (Xu 2014), assessed only concordance with other callers (Roberts 2013), or done their own sequencing and validation but have not released raw data (Alito 2014, Goode 2013).

## Validated Calls in TCGA



Summary of Illumina HiSeq and MiSeq calls from the TCGA public MAF files.





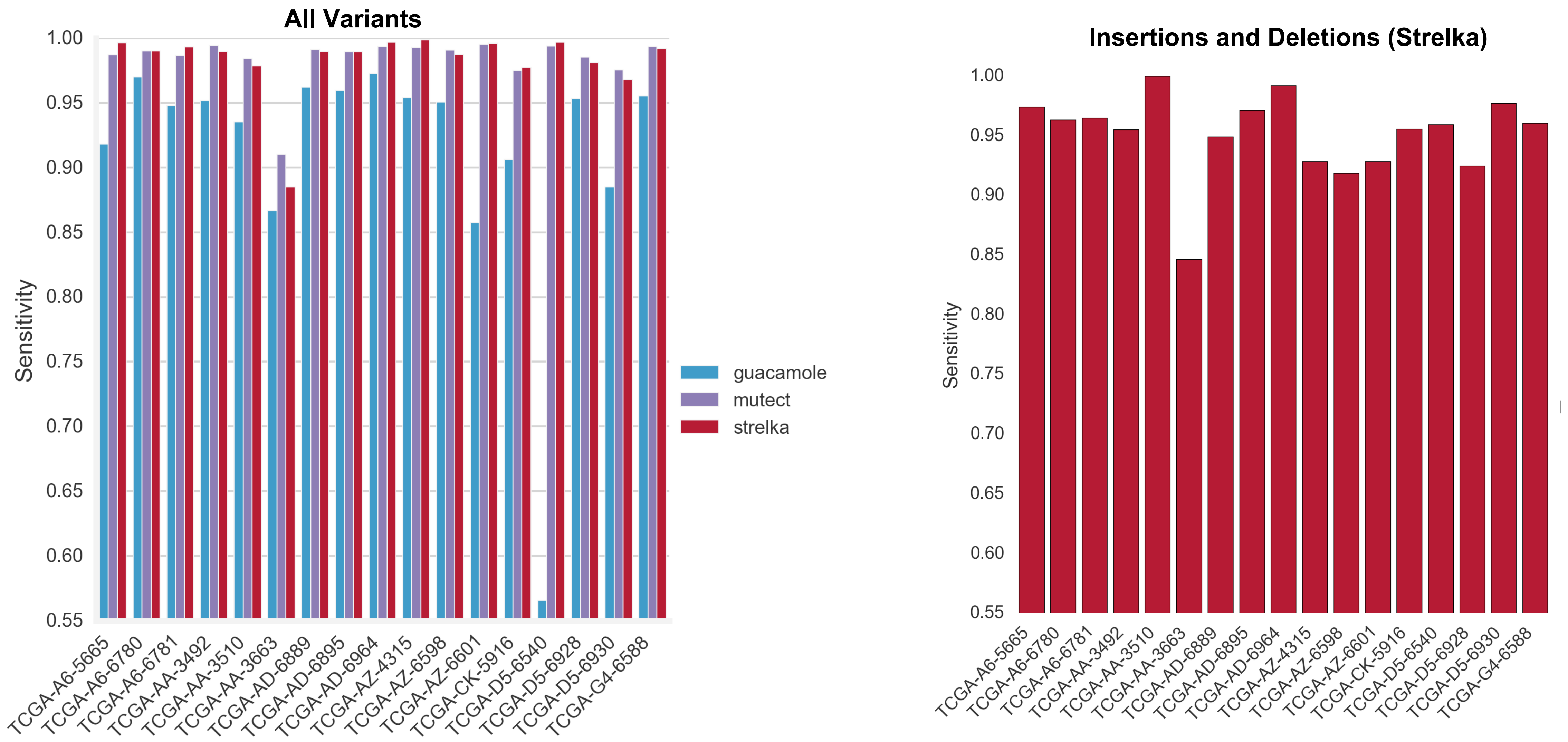
**Samples with the most validated variants**
Calls generated from Illumina sequencing and validated with targeted Illumina.

| Tumor Barcode | Disease | Type | Calls<br>Validated / Total |
|---|---|---|---|
| TCGA-CA-6717-01A-11D-1835-10 | COAD | WGS | **4289** / 7007 |
| TCGA-AZ-4315-01A-01D-1408-10 | COAD | WXS | **2798** / 6086 |
| TCGA-AA-3510-01A-01D-1408-10 | COAD | WXS | **1549** / 2963 |

**Samples with the most validated variants from an orthogonal platform**
Calls generated from Illumina sequencing and validated with targeted Ion PGM or 454.

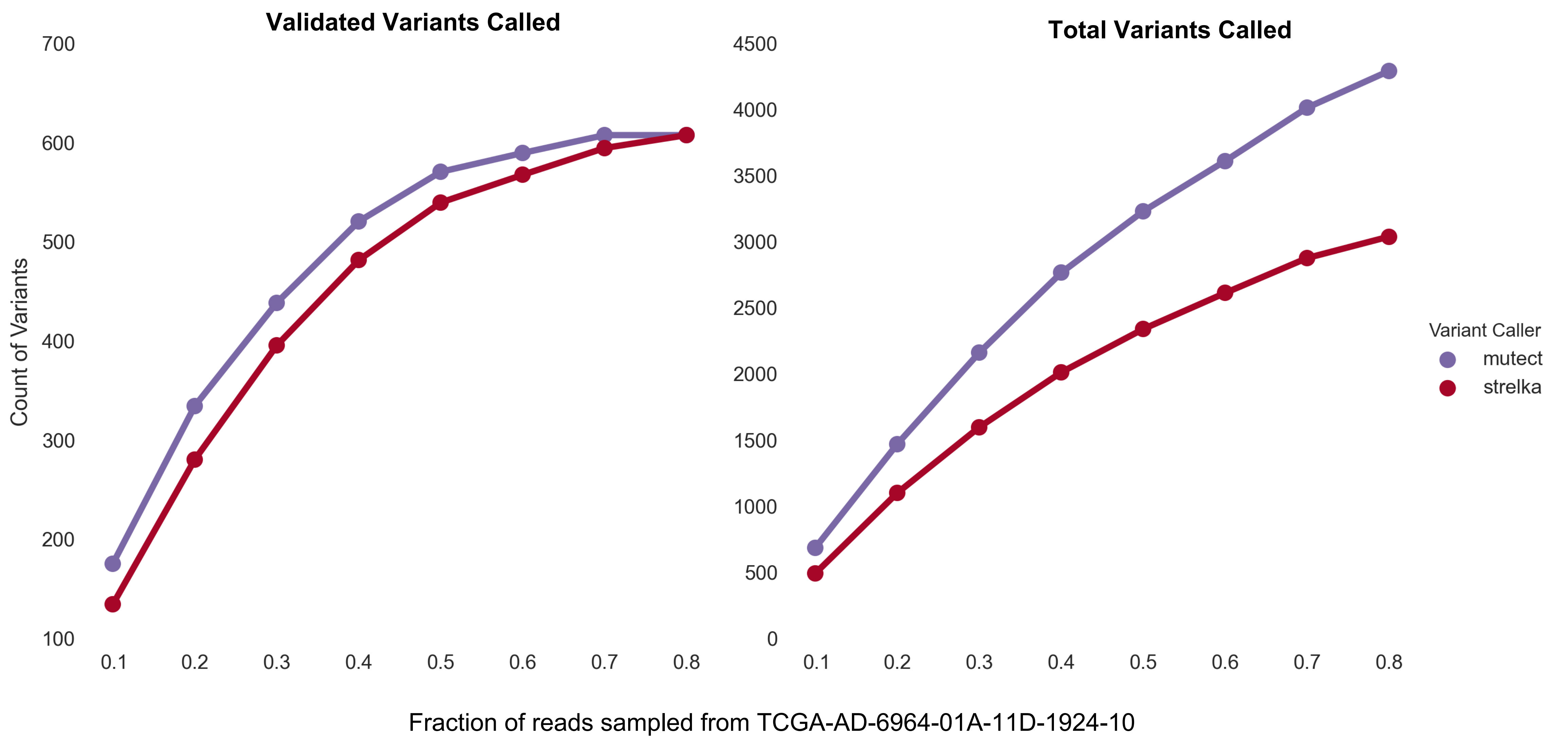
| Tumor Barcode | Disease | Type | Calls<br>Validated / Total |
|---|---|---|---|
| TCGA-D5-6931-01A-11D-1924-10 | COAD | WXS | **21** / 320 |
| TCGA-CA-6716-01A-11D-1835-10 | COAD | WXS | **13** / 208 |
| TCGA-CK-4950-01A-01D-1719-10 | COAD | WXS | **13** / 436 |
| TCGA-A3-3308-01A-01D-0966-08 | KIRC | WXS | **13** / 90 |

## Sensitivity on Validated Calls



Three callers were evaluated for sensitivity on 16 TCGA tumor/normal pairs. Mutect and Strelka consistently found most of the validated SNVs. Strelka, the only caller supporting indels, performed similarly on insertions and deletions. The validated variants missed by our caller (Guacamole) can be used to improve it in a future version.

## Sensitivity on downsampled data from a single tumor/normal pair



Variant call counts for a single tumor/normal pair, in which the tumor reads have been randomly downsampled across a range of sampling fractions. Many of the validated variants are still called at lower depths, suggesting the validated variants are biased toward being the easiest to identify variants.

## References

Kim, S. Y., & Speed, T. P. (2013). Comparing somatic mutation-callers: beyond Venn diagrams. BMC Bioinformatics, 14(1), 189. http://doi.org/10.1186/1471-2105-14-189

Löwer, M., Renard, B. Y., de Graaf, J., Wagner, M., Paret, C., Kneip, C., … Sahin, U. (2012). Confidence-based somatic mutation evaluation and prioritization. PLoS Computational Biology, 8(9), e1002714. http://doi.org/10.1371/journal.pcbi.1002714

Roberts, N. D., Kortschak, R. D., Parker, W. T., Schreiber, A. W., Branford, S., Scott, H. S., … Adelson, D. L. (2013). A comparative analysis of algorithms for somatic SNV detection in cancer. Bioinformatics (Oxford, England), 29(18), 2223–30. http://doi.org/10.1093/bioinformatics/btt375

Xu, H., DiCarlo, J., Satya, R. V., Peng, Q., & Wang, Y. (2014). Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics, 15, 244. http://doi.org/10.1186/1471-2164-15-2445

Alioto, T. S., Derdak, S., Beck, T. A., Boutros, P. C., Bower, L., Eldridge, M. D., … Hennings-yeomans, P. (2014). A Comprehensive Assessment of Somatic Mutation Calling in Cancer Genomes.

Goode, D. L., Hunter, S. M., Doyle, M. A., Ma, T., Rowley, S. M., Choong, D., … Campbell, I. G. (2013). A simple consensus approach improves somatic mutation prediction accuracy. Genome Medicine, 5(9), 90. http://doi.org/10.1186/gm494