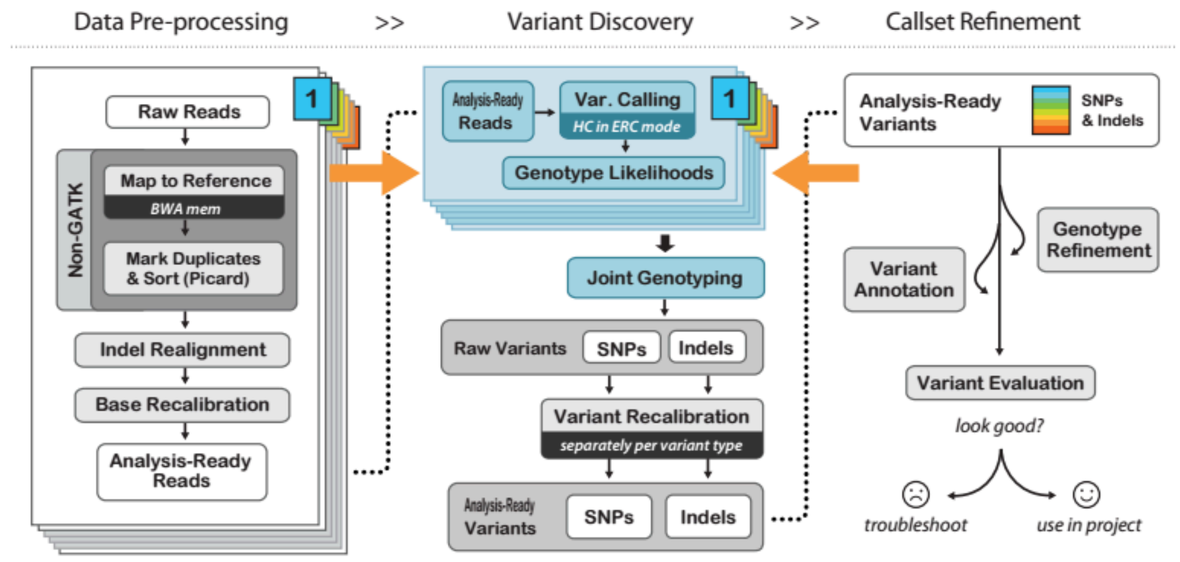
Scaling variant calling up to hundreds of thousands of samples with GATK

Laura D. Gauthier, David Benjamin, Kylee Bergin, Jonathan M. Bloom, Yossi Farjoun, Mark Fleharty, Samuel K. Lee, Monkol Lek, Heng Li, Valentin Ruano-Rubio, Takuto Sato, Megan Shand, Eric Banks, Daniel MacArthur

Motivation

- Cohort size for modern sequencing studies continues to rise into the hundreds of thousands of samples
- Processing needs to be efficient
- Variant calling accuracy needs to be preserved
- Rare variation sensitivity should not be sacrificed

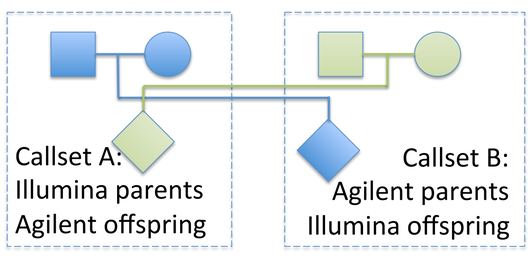
GATK Analysis Pipeline Overview



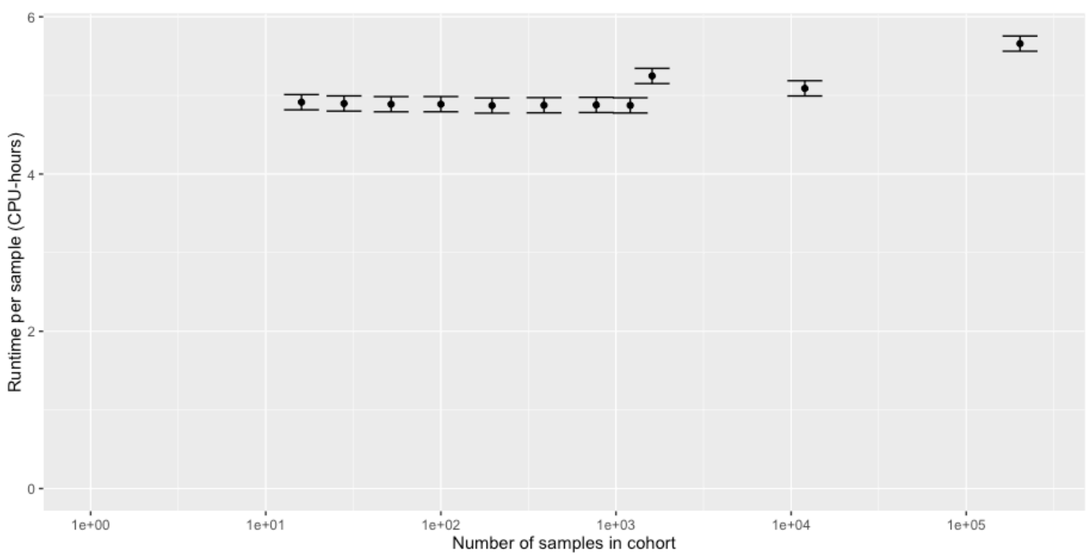
GATK best practices call each sample's variant likelihoods individually and then genotype all samples together for computational efficiency.

Data and Methods

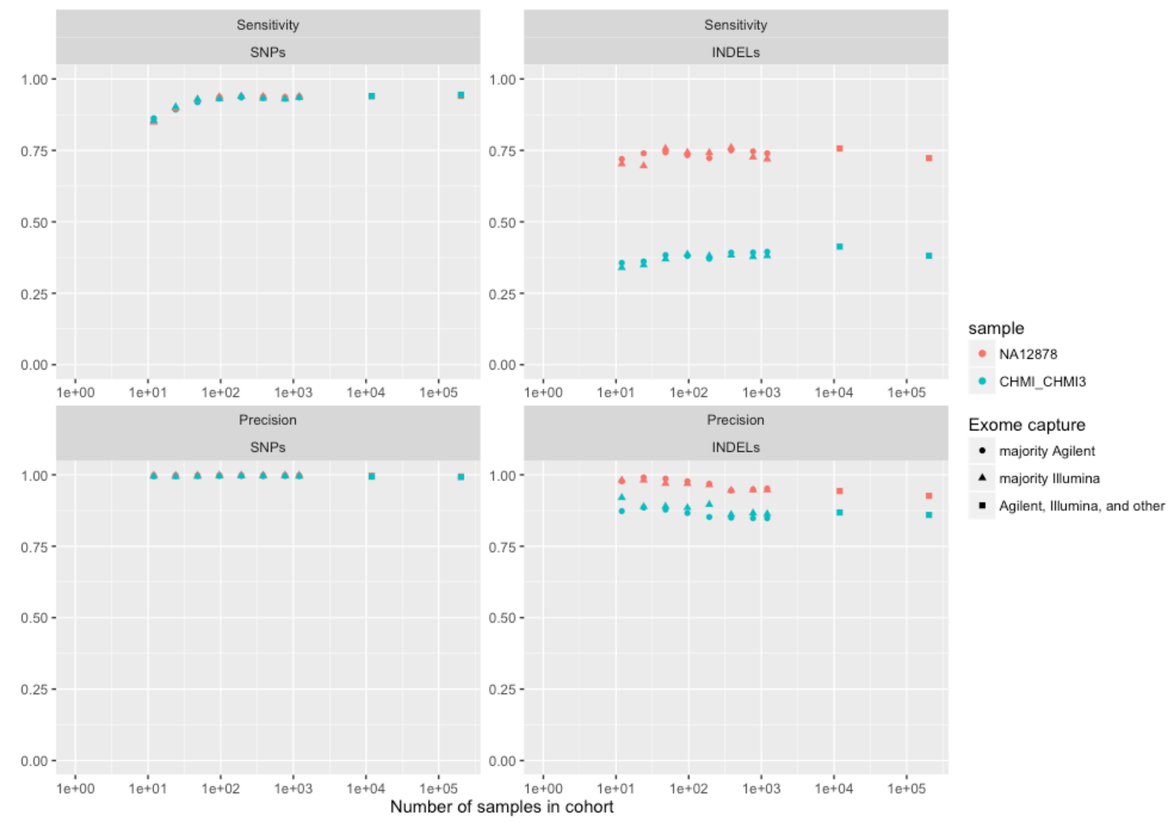
To examine the effects of different sequencing platforms, a pool of 802 trios of European ancestry and equal proportion of Illumina and Agilent captures was analyzed. Callsets of varying sizes were composed of unrelated samples consisting of parents from one capture type and offspring from the other. Data from ExAC[1] version 2 was also analyzed, which contains a wide variety of exome captures and sequencing platforms, along with a subset of approximately 6% of ExAC. Truth samples (NA12878 and the pseudo-diploid sample published in [2]) were included as members of each cohort. "Singletons" were defined with respect to the 1,203 sample callset and evaluated for each callset size. Due to compute constraints, only the 1,200 and 12,000 sample cohorts were run with multiple trials.



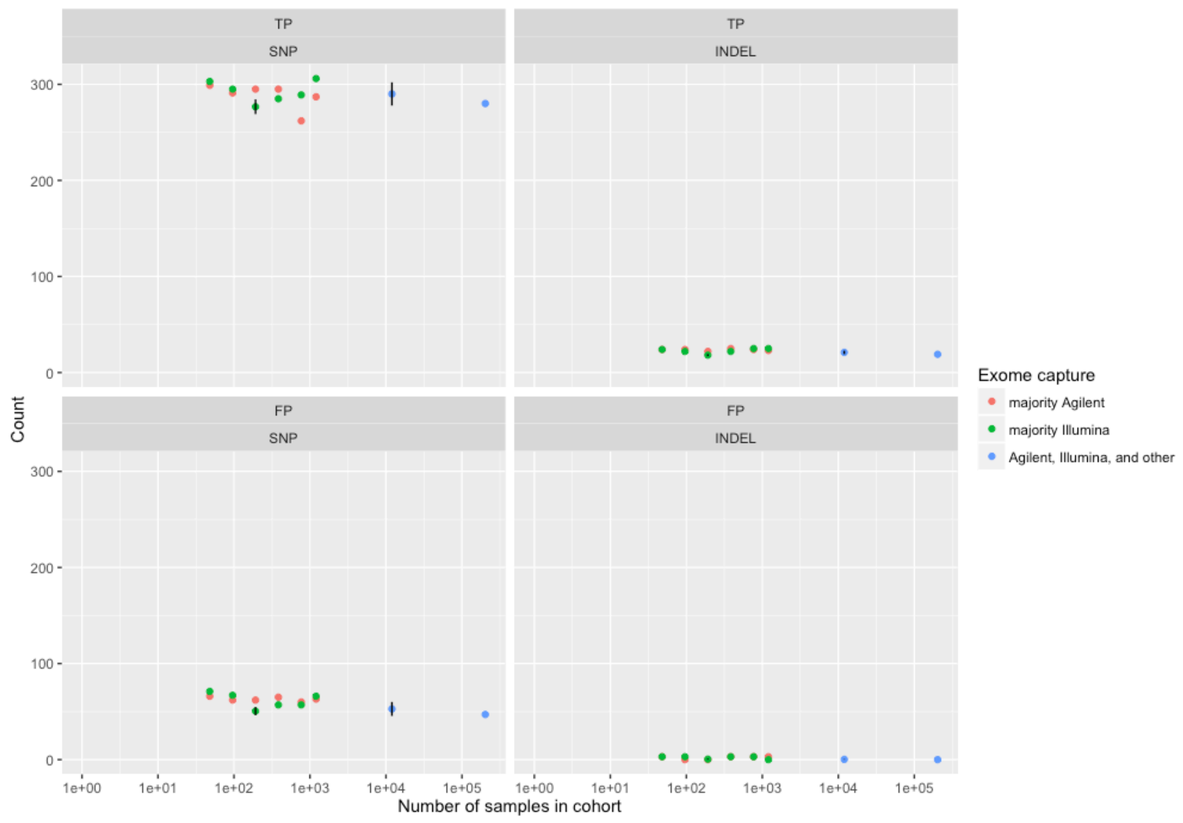
Compute time per sample is constant



Increasing joint-calling cohort size improves variant sensitivity



Joint-calling cohort size scales without loss of singleton accuracy



The set of singleton truth variants was compared across cohort sizes. At the scale of the largest cohort, many of the variants are no longer singletons, but the comparison of a constant set of variants is informative.

Conclusions and Recommendations

- Hundreds of thousands of samples can be joint-called at constant time per sample
- Sensitivity gains from larger cohorts saturate near 600 samples
- Precision may be decreased by heterogeneous capture types and/or differing read lengths
- Sensitivity to singletons is maintained as more samples are added

References

[1] Lek et al. "Analysis of protein-coding genetic variation in 60,706 humans" Nature: 536, 285-291.
[2] Schneider et al. "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly" http://biorxiv.org/content/early/2016/08/30/072116

Acknowledgements

This work was funded by U54DK105566 and R01GM104371