Optimizing Delivered Sequencing Data with a Theoretical Sensitivity to Heterozygous SNPs

Kylee Degatano, David Benjamin, Jonathan M. Bloom, Maura Costello, Jason Rose, Kathleen Tibbetts, Charlotte Tolonen, Yossi Farjoun

75 Ames St, Cambridge, MA 02142

Introducing Theoretical Het Sensitivity

Traditional sequencing metrics do not optimize delivered data. Using only coverage as a deliverable fails to consider the potential for poor data quality.

Here we derive and improve upon a theoretical measure of sensitivity to heterozygous single nucleotide polymorphisms (SNPs), which we propose as a new sequencing deliverable, called Theoretical Het Sensitivity (THS). Specifically, we focus on estimating and improving THS based on distributions of read depth and base quality and comparing THS to NIST het sensitivities.

This metric has the potential to deliver better quality sequencing data by translating the researchers' desired sensitivity to genetic variants to sequencing attributes more easily controlled in the lab.

Estimating Theoretical Het Sensitivity

Calculating the Estimate

The generative model of het detection, a Bernoulli random variable, is as follows:

1.Draw depth *n* from the coverage distribution *P(n)*.

2.Draw the number of true alternate alleles m ~binomial(*n*, 0.5) covering the HET site.

3.Draw i.i.d. base qualities $q_1,...q_n$ from the base quality distribution *P(q)*, where the first *m* are assigned to the alternates.

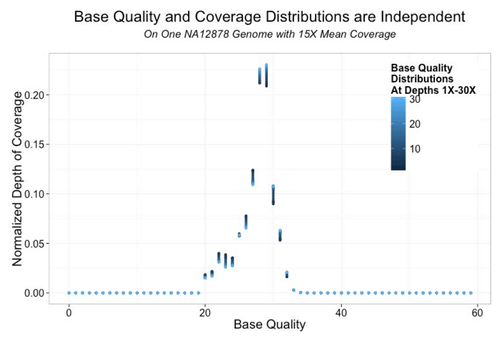
4.Compare the likelihood ratio of het vs hom ref to the log odds threshold *T*:

$$\frac{\binom{n}{m}\left(\frac{1}{2}\right)^n}{\binom{n}{m}\prod_{j=1}^{m} e_j \prod_{j=m+1}^{n}(1-e_j)} > T$$

where $e_j = 10^{-q_j/10}$ is the probability of error.

Simplifying Assumptions

- The sequence data is derived from a diploid sample (no cancer, copy number variation, or contamination)
- At a het site with genotype A/B, the only possible alleles are A and B.
- There is no reference bias: at an A/B site it is equally probable to sequence A as B, regardless of the reference.
- The coverage distribution and base quality distribution are known and statistically independent.

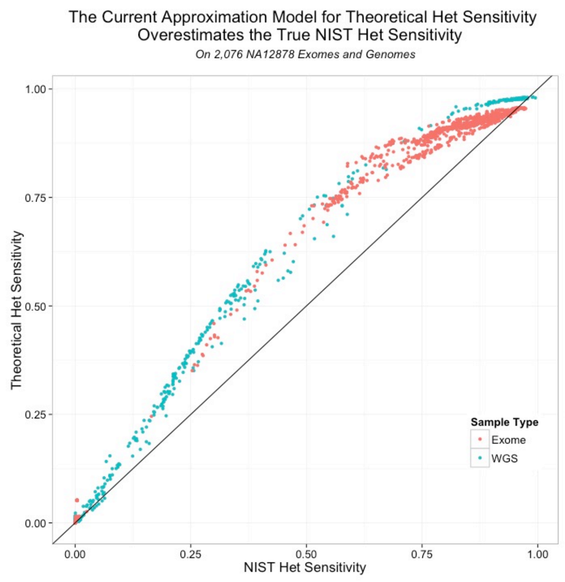


Base quality distributions for each level of coverage from 1X to 30X are graphed on the same axis

Accuracy of Theoretical Het Sensitivity

THS has been implemented in the Broad Institute's[1] publically available software suite, Picard[2], for whole genome, exome, and targeted PCR samples in their respective metrics collectors.

The current implementation over-estimates the actual sensitivity when samples have coverage below ~12X or low base quality. Here, THS and NIST het sensitivity are calculated for over 2000 exomes and whole genomes with varying coverage and quality.



Citations

(1) genomics.broadinstitute.org

(2) broadinstitute.github.io/picard

(3) *Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls.* Zook et al. 2014.

(4) software.broadinstitute.org/gatk

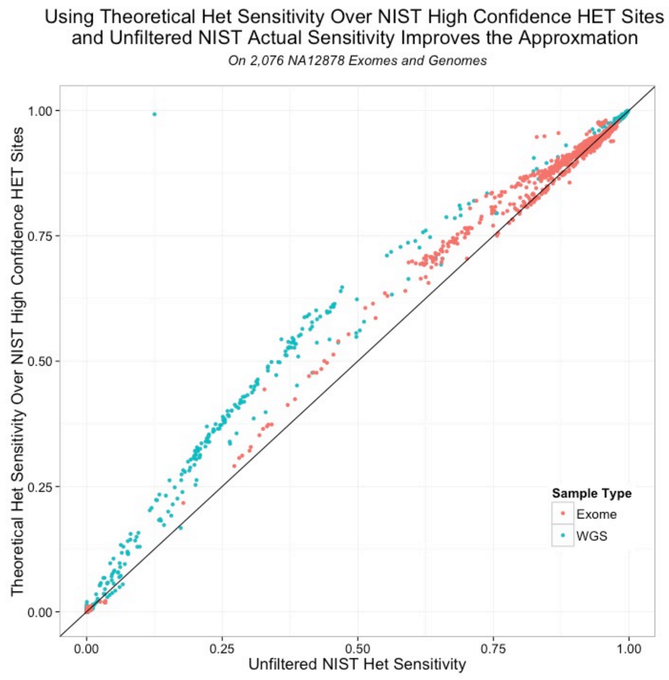
Overall Impact: Optimized Delivered Sequencing Data

Analyses on the effects of THS as a deliverable were done on a data set of 9,616 exomes and 6,643 genomes aggregated in production. Exomes are delivered at 85% of target bases at 20X coverage (85%@20X) and genomes are delivered 80%@20X.

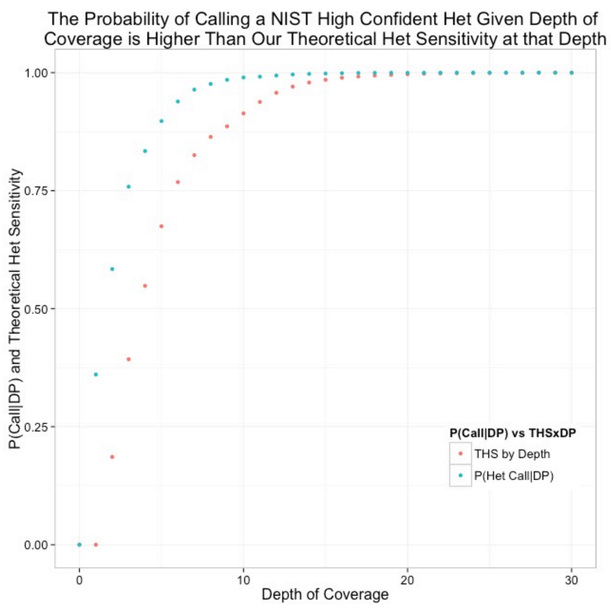
*A transform was used to calculate the average quality from the target base quality distribution. Because of the log scale of Q scores, low quality bases will drastically bring down the average.

Using a deliverable of 96% THS, 5.3% more genomes would be ready for delivery and 4.11% fewer exomes. We are currently delivering more genome data than necessary to achieve at least 96% sensitivity to calling het SNPs. The new deliverable requirement would be more stringent for exomes. Currently, the average quality* of delivered bases is inconsistent. THS would contribute to optimizing delivered data.

Limiting to NIST High Confidence Het SNPs Increases Accuracy



On the left, we calculate THS across the known het sites in the NIST[3] high confidence region within the WGS and exome evaluation intervals, instead of only the evaluation intervals on the Y axis. When we calculate NIST het sensitivity, we filter to increase our specificity, sacrificing sensitivity. THS is meant to calculate how many variants the caller will discover, not how many will be in the final call set after filtering. Therefore, in the top panel we compare THS to the unfiltered NIST het sensitivity.



In the bottom graph, we show that our ability to make a het call given the depth of coverage at that variant for NIST het high confidence SNPs in the WGS evaluation intervals correlates with our ability to calculate THS at that depth over the same intervals. However, the likelihood of calling a het using HaplotypeCaller from GATK[4] is higher than even our improved theoretical sensitivity to calling a het.

Future Improvements

We are working to improve THS by:

- Altering parameters and the log odds threshold used to calculate THS over NIST high confidence het SNP sites for more accuracy.
- Modifying the THS calculation to work for variable allele fractions, to look at Low AF cancer samples.

Acknowledgements

Thank you to Megan Shand, Mark Fleharty, Laura Gauthier, Eric Banks for their help with these analyses and this poster.