Simulating pathogenic and likely pathogenic variants for bioinformatics pipeline testing

Qian Zeng, Binbin Huang, Pingsha Hu, Neil Russell, Stan Letovsky, Natalia Leach & Angela Kenyon
Center of Excellence for Data Sciences, AI and Bioinformatics, Labcorp, Westborough, MA

Introduction

Next Generation Sequencing (NGS) is widely used to detect clinically significant variants in clinical samples. The analytical performance (sensitivity and specificity) of NGS tests is dependent on many factors, including the selection of an analysis pipeline, targeted genomic regions, variant types and length, library quality, sequencing technology, read length and quality, and coverage. It is critical to understand the impact of these factors in order to optimize testing performance.

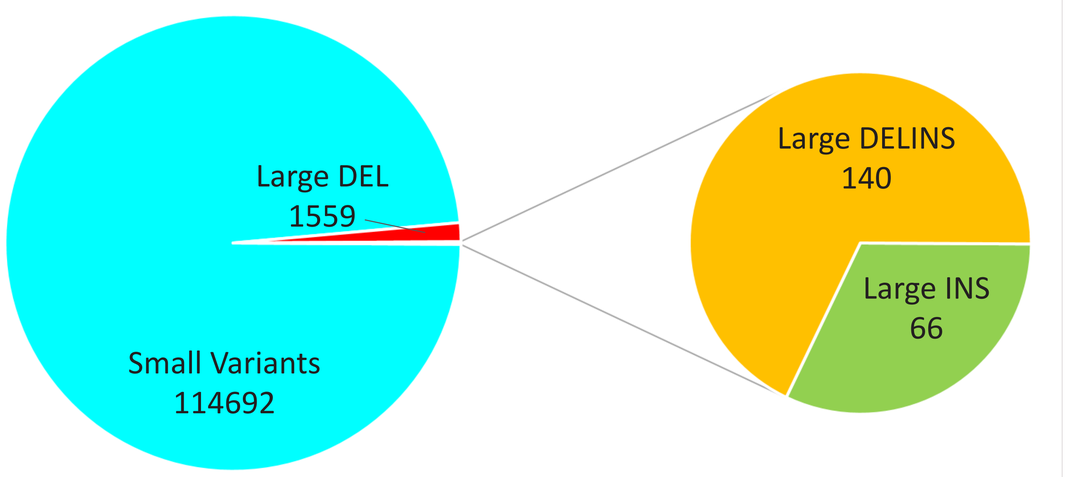
Current methods for assessing NGS pipeline performance are limited in the number of pathogenic and likely pathogenic variants (PLPs):

- High quality reference data set (e.g., GIAB "gold standard") [Wagner 2021]
- Public repositories (e.g., from Coriell Institute)
- Previously characterized clinical samples with reported PLPs

Methods

Simulation

- "Can we detect a variant with typical NGS coverage?"
- Generate simulated Illumina reads for >116,000 ClinVar PLPs (100X coverage, 50% variant allele frequency, 150bp paired ends, 300bp fragment size) [Huang 2012]



Decimation

- "What is the minimum coverage needed for variant detection?"
- Down-sample real Illumina data for GIAB reference genomes to 90%,80%,…, 1% of original fastq reads

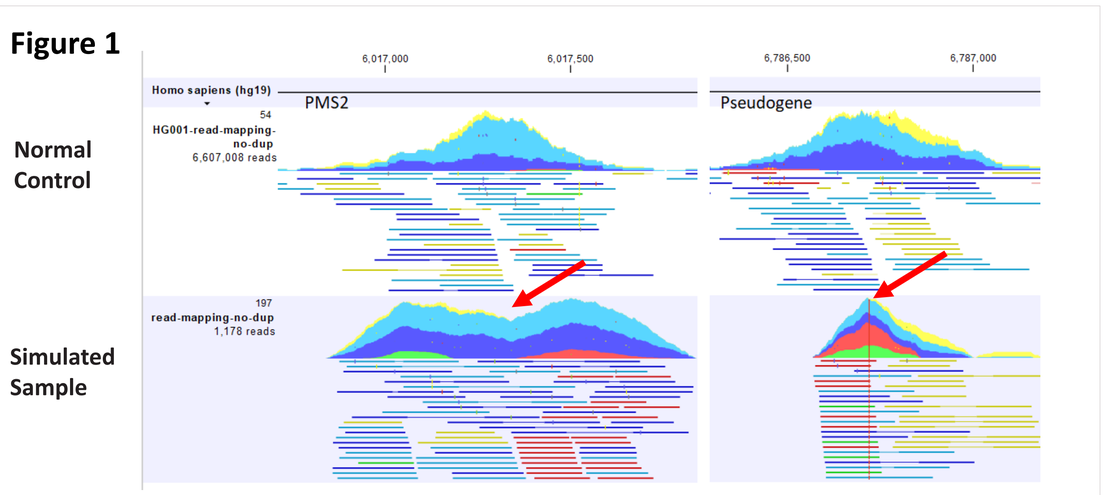
NIST/GIAB ID	Variants	Acc#	Description
HG001	26,741	NA12878	European trio daughter
HG002	27,426	NA24385	Ashkenazim trio son
Total	42,822		

Results and Conclusion

Simulation

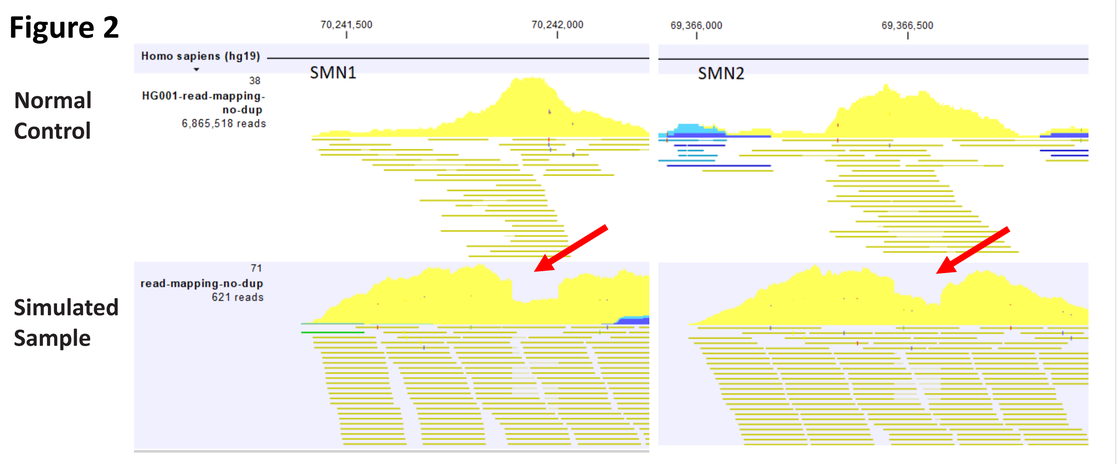
- All ClinVar PLPs are detected except for the three noted below:
- Challenging variants: 3
  - Pseudogene match: 2
    - PMS2 5bp DELINS (see **Figure 1** below) and STRC SNV
    - Similar to CYP21A2 variants and GBA 55bp DEL [Zeng 2020]
    - Coverage drop detected by CNV caller (see **Figure 1** below)

Figure 1



- Need more investigation: SMN1 112bp DEL (see **Figure 2** below)

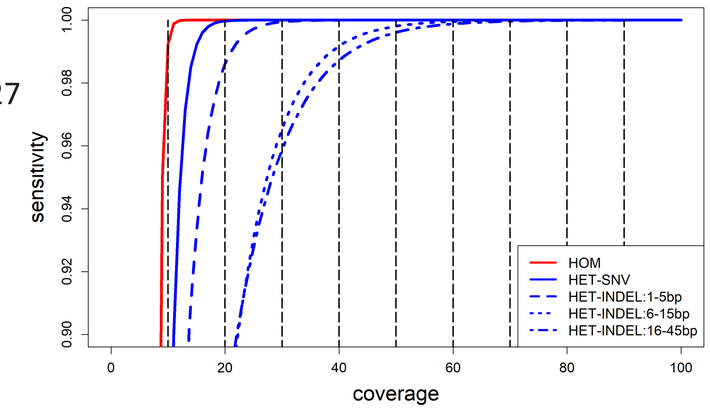
Figure 2



Decimation

Figure 3

Challenging variants: 37
- Pseudogene match: 27
- Repeat region: 3
- Borderline allele frequency (~20%): 7



Variant Group	Count	Sensitivity at 30X
All	42822	0.9999
HOM	15077	0.9999
HET-SNV	27103	0.9999
HET-INDEL:1-5bp	522	0.9995
HET-INDEL:6-15bp	93	0.9655
HET-INDEL:16-45bp	27	0.9589

References

Wagner J. et al., *BioRxiv*, 2021.
Huang W. et al., *Bioinformatics*, 2012, 28, 593-594.
Zeng Q. et al., *Sci. Rep.*, 2020, 10, 15060.

ASHG Annual Meeting, Los Angeles, CA, October 25-29, 2022

labcorp