Expanding the diagnostic yield of germline genetic testing in cancer patients using deep learning

(AlDubayan et al. JAMA. 2020 Nov 17)

Saud H. AlDubayan, M.D., Jake R. Conway, B.Sc, Leora Witkowski, Ph.D., Eric Kofman, B.Sc, Brendan Reardon, B.Sc, Sabrina Y. Camp, B.Sc, Seunghun Han, B.A, Haitham Elmarakeby, Ph.D., Abdullah M. Al-Rubaish, M.D., Abdulsalam A. Al-Sulaiman, M.D., Amein K. Al-Ali, Ph.D., Amaro Taylor-Weiner, Ph.D., Eliezer M. Van Allen, M.D

Broad Institute of MIT and Harvard, Cambridge, MA; Dana-Farber Cancer Institute, Boston, MA; Harvard Medical School, Boston, MA, **E-mail:** SAUD_ALDUBAYAN@DFCI.HARVARD.EDU

Introduction

Despite the established clinical utility of germline genomic analysis, thus far, only a small fraction of cancer patients are found to carry germline pathogenic variants, leaving the majority of the cases with unidentified genetic drivers of cancer risk. Moreover, even when the clinical presentation is highly suggestive of a particular cancer predisposition syndrome, most patients have a negative genetic analysis of all protein-coding genes. In addition to highlighting non-coding or epigenetic variants as a possible explanation, these observations also raise concern about the possibility of incomplete detection of known or expected pathogenic disease-causing variants by the currently-used "gold standard" germline variant detection methodologies.

Recently, advances in computational methods that utilize deep neural networks demonstrated enhanced germline variant detection in "The Genome in a Bottle" ground-truth set. We hypothesized that these methods may also improve detection of rare clinically relevant variants in germline data from cancer patients and fundamentally change clinical genetics practice.

Methods

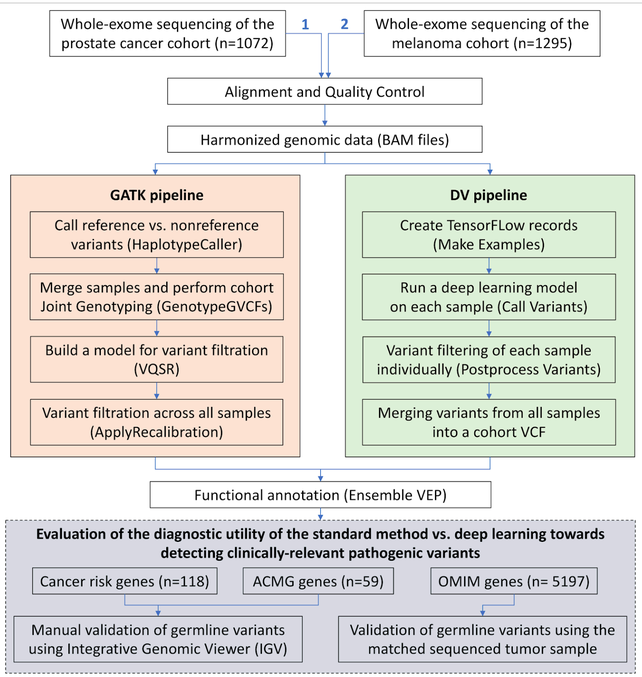

Figure 1: An overview of the study design and the used germline variant detection pipelines. Germline WES data of two independent cohorts (1072 PC and 1295 melanoma patients) were analyzed using GATK, the "gold-standard" method, and DV, a deep-learning variant calling approach. (OMIM: Online Mendelian Inheritance in Man).

Results: Exome-wide detection of common and rare variants

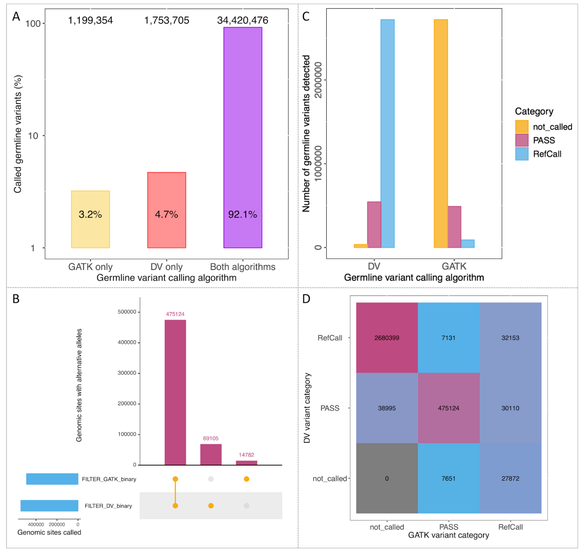

Figure 2: Exome-wide germline variant detection in 1072 in PC patients.

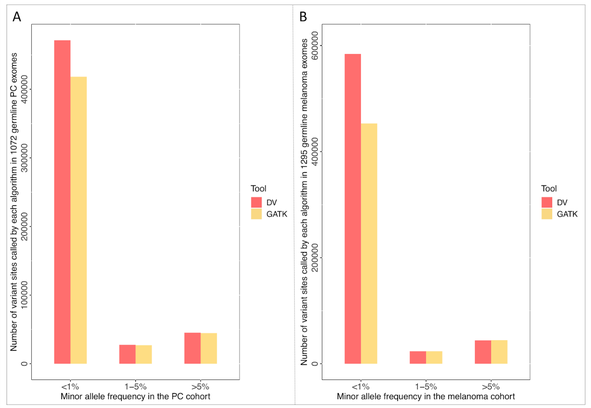

Figure 3: Although deep learning only called 1.3% and 2.1% more common (MAF>5%) and uncommon variants (MAF:1-5%) than the gold-standard method in 1072 germline PC exomes, it identified 12.7% (n=53,161) more rare variants (MAF<1%) than the standard approach, suggesting a substantially higher performance of deep learning for detecting this variant subset which are highly enriched for Mendelian disease-causing variants.

Results: Detection of pathogenic cancer risk variants

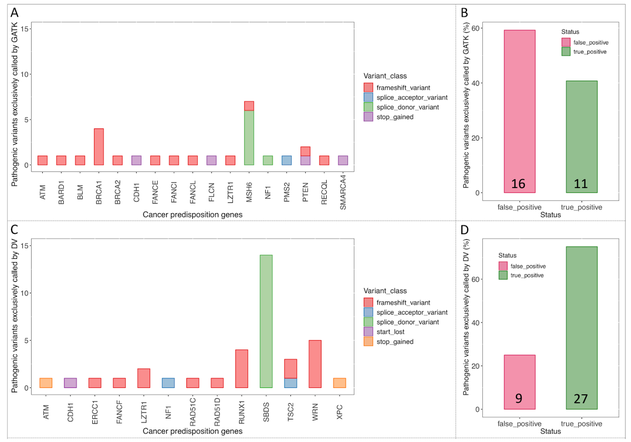

Figure 4: The standard method, GATK, exclusively identified 27 cancer predisposition variants in 1072 PC patients (A) 60% of which were found to be computational artifacts (B). In contrast, deep learning exclusively identified 37 germline pathogenic variants (C), 75% of which were validated true variants (D).

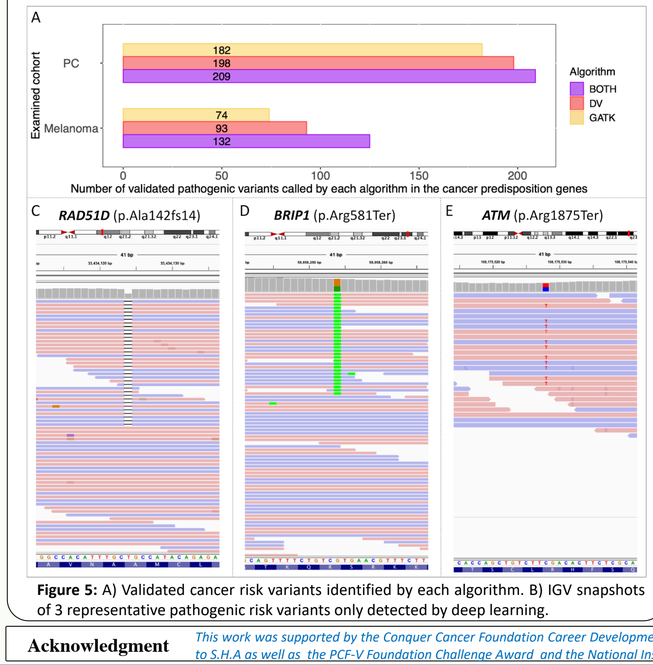

Figure 5: A) Validated cancer risk variants identified by each algorithm. B) IGV snapshots of 3 representative pathogenic risk variants only detected by deep learning.

Results: Exome-wide detection of pathogenic variants

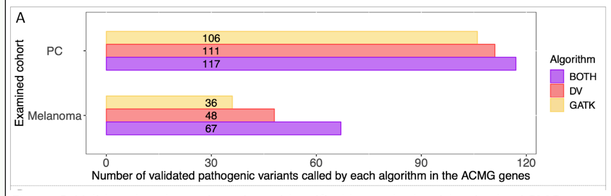

Figure 6: Validated pathogenic variants identified by each algorithm in the ACMG genes.

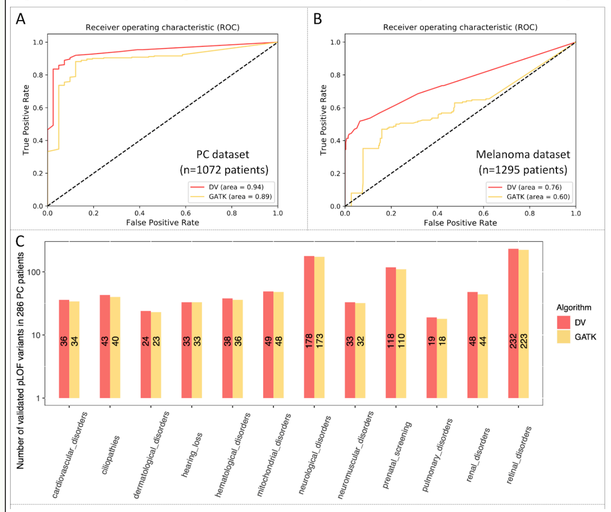

Figure 7: A & B; Evaluation of the model performance of the standard method, GATK, and deep learning, DV, towards detecting clinically relevant pathogenic variants in the ACMG and cancer predisposition genes (n=151). C; Germline exome-wide analysis of 286 PC patients, using deep learning, identified more putative LOF variants in 11 (91.7%) of the 12 phenotype-targeted multi-gene panels.

Conclusions

Our analysis showed that on average, deep learning identified one additional patient with a pathogenic cancer-risk variant for every 52 to 67 cancer patients undergoing expanded germline testing and one additional patient with a pathogenic variant in the highly-actionable ACMG genes for every 182 patients undergoing this test, underscoring a nontrivial missing rate of the currently used method. Collectively, our analysis consistently showed a clinically meaningful increase in the diagnostic yield of deep learning variant detection over the current gold-standard method regardless of the examined gene set. Deep learning higher sensitivity for rare pathogenic variant detection may also lead to an improved ability to uncover novel gene-disease associations.

References

- Poplin R, Chang P-C, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983-987.
- DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491-498.

Acknowledgment *This work was supported by the Conquer Cancer Foundation Career Development Award (CDA#13167) and the Prostate Cancer Foundation Young Investigator Award (YIA#18YOUN02) to S.H.A as well as the PCF-V Foundation Challenge Award and the National Institutes of Health R01CA222574 to E.M.V.*