Integrating DeepVariant into a Clinical Bioinformatics Pipeline

Lawrence Hon[1], Jeroen Van den Akker[1], Ziga Mahkovec[1], Gilad Mishne[1], Anju Ondov[1],
Lizzie Dorfman[2], Pi-Chuan Chang[2], Ryan Poplin[2], Mark DePristo[2], Jeremy Ginsberg[1]
[1]Color Genomics, Burlingame, CA; [2]Google, Inc., Mountain View, CA

Introduction

Bioinformatics pipelines are a necessary and integral component in next generation sequencing based approaches for genetic testing. In a clinical setting, these pipelines utilize a number of different variant callers to maximize sensitivity, add redundancy, and increase system robustness. Many variant callers such as GATK[1] have been developed to detect small-to-mid sized variants (Table 1). However, some variants are difficult to call because they are near challenging sequence context (e.g. near homopolymer regions or in extreme GC regions) or they are complex variants that can be obscured in the sequence alignment (e.g. complex indels). This creates opportunities to add variant callers to supplement performance. Here, we describe the integration of DeepVariant, a state-of-the-art learning-based variant caller developed by Google[2], into a bioinformatics pipeline at a clinical laboratory, Color Genomics. We show how DeepVariant adds redundancy and maximizes sensitivity in clinical tests. Furthermore, we demonstrate the increased sensitivity of DeepVariant for calling single nucleotide variants (SNVs, 1 bp) and insertion and deletions (indels, <50 bp) in high-GC regions in 59 genes that can cause serious and preventable clinical conditions such as breast and ovarian cancer, familial hypercholesterolemia, Lynch syndrome, and others (hereafter referred to the ACMG59 genes)[3].

Table 1. Parameters of different small variant callers

Variant Caller	Notes
GATK	v3.4, following the Broad's "best practices", with variant confidence model[4] used to filter variants
DeepVariant	v0.6, training using WES model, using default quality filter
MNV	in-house implemented multiple nucleotide variant calling algorithm to phase and merge nearby variants called by GATK/DeepVariant
Scalpel	v0.5.3, limited to indels >10bp
Samtools	custom implementation of samtools/bcftools applied to homopolymer regions

Methods

To bolster bioinformatics pipeline performance for difficult-to-call variants and regions, Color has incorporated multiple variant callers (Figure 1). The small variant callers are run following best practices, generally tuned to increase sensitivity while balancing the false positive rate (Table 1). Variant calls are merged and annotated with Color's variant confidence model[4]. Low confidence and novel variants are confirmed using Sanger sequencing. After an initial analysis showed DeepVariant having promising gains in sensitivity in high GC regions, DeepVariant (v0.6) was integrated into the Color clinical pipeline to increase redundancy and strengthen variant detection capabilities.

To explore the contribution of the callers in the bioinformatics pipeline, 478 samples were sequenced for 30 genes associated with hereditary cancer[5] (hereafter referred to as the HC30 genes). Collectively, these samples had 487 pathogenic or likely pathogenic variants that were subsequently confirmed by Sanger sequencing (Figure 2). To assess the additive value of DeepVariant in regions of high-GC content (Figure 3), several Coriell cell lines and 7000 research consented samples were sequenced for the ACMG59 genes (Figure 4, Table 2, and Figure 5).

Maximizing Performance through an Ensemble of Variant Callers

Figure 1. Variant detection methods in the bioinformatics pipeline

The bioinformatics pipeline utilizes a number of different variant callers to detect different variant types and sizes.

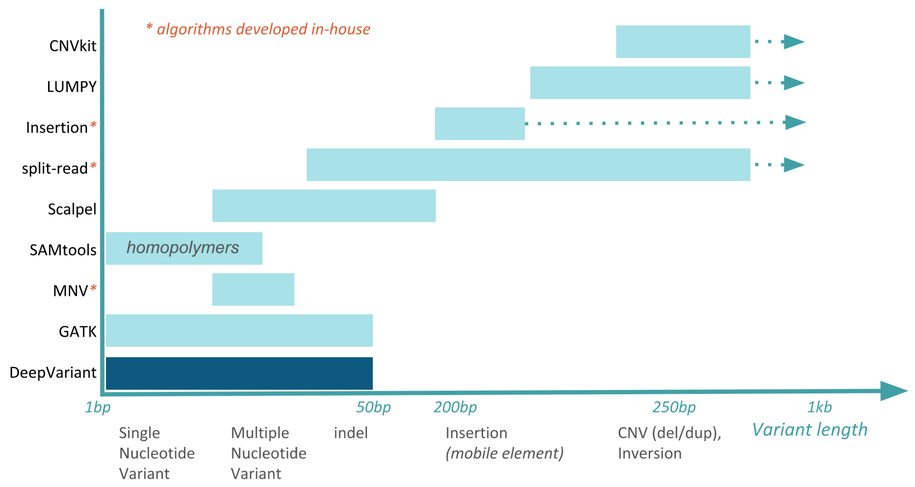
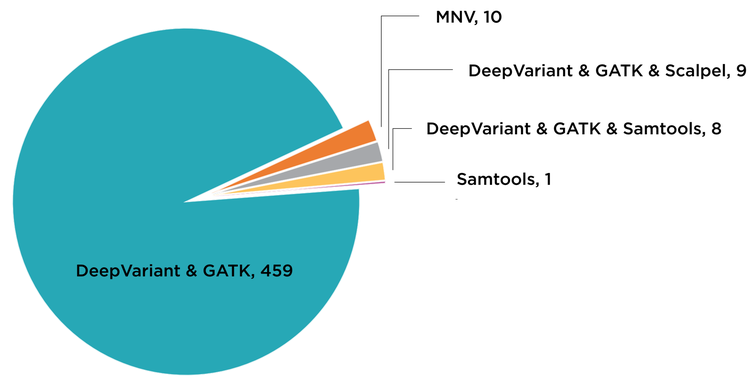


Figure 2. Previously confirmed variants*, by caller

Multiple different small-to-mid size variant callers were required to accurately identify 487 variants including SNVs (1 bp) and indels (<50 bp) in 30 genes associated with hereditary cancer.

*These calls reflect the validation process for variants before inclusion of those unique to DeepVariant, which is why there are no DeepVariant only calls. These are discussed in the section below.



Increased Sensitivity of DeepVariant in High-GC Regions of ACMG59

Figure 3. GC content of the HC30 and ACGM59 panels

The ACMG59 panel has 3.4x the number of high-GC bp compared to the HC30 panel. This enrichment remains true even after normalization for panel size.

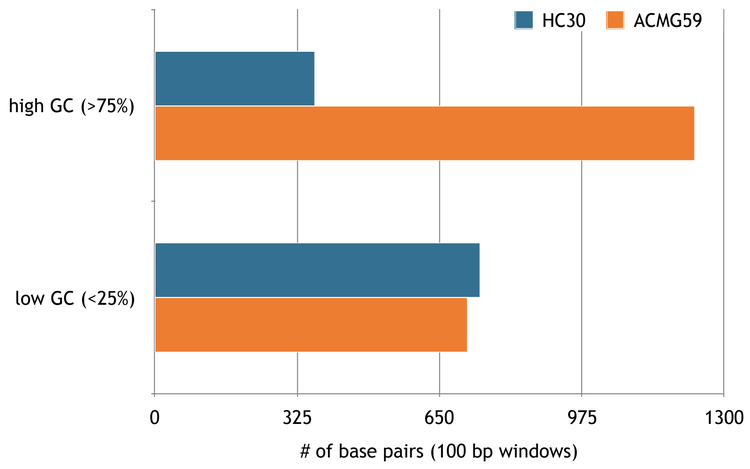


Figure 4. Example of a known pathogenic variant in the ACMG59 panel detected by DeepVariant alone

DeepVariant detected the known pathogenic variant *MEN1* c.206_207insGCCCC (p.Asp70Profs)[6] in Coriell cell line NA11630. This variant was not detected by any other callers.

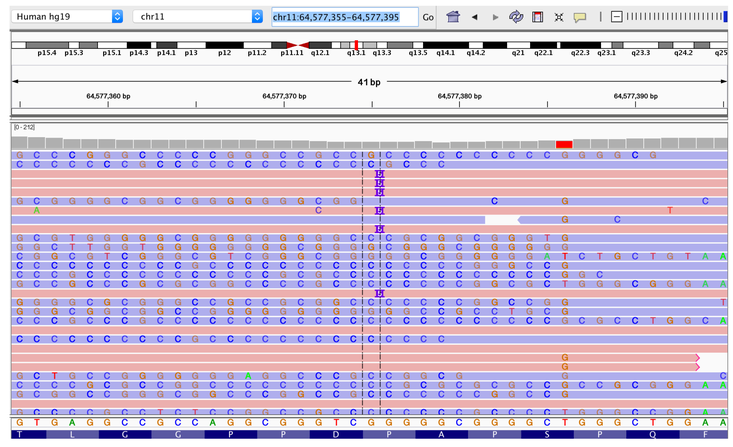


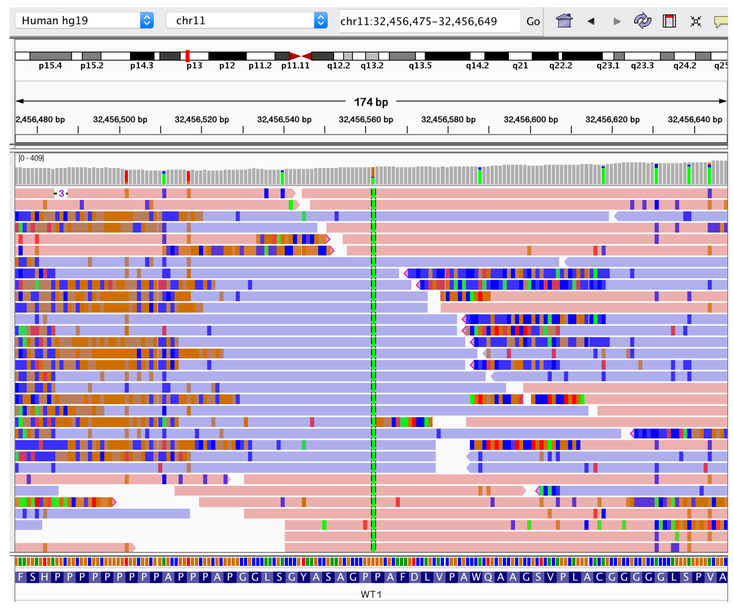
Table 2. Putative variants in the ACMG59 panel detected by DeepVariant alone

DeepVariant identified 15 variants in 7000 research consented samples sequenced for the ACMG59 genes that were not detected by any other callers (after removing commonly recurrent variants). These 15 variants were reviewed and deemed present based on IGV. To note, these variants are enriched for high-GC regions and indels. The limited number of novel variant calls detected only by DeepVariant suggests a limited impact on our downstream workload.

Variant Name	Count	Quality	AF	Gene	Notes
chr1:g.55524323G>A	1	45.9	44.1%	*PCSK9*	high GC
chr2:g.48033885A>AAAAACTTTTTTTTTTTTTTTT	2	8.3	19.1%	*MSH6*	low GC, insertion
chr6:g.7585039C>A	1	31.2	38.4%	*DSP*	high GC, part of complex indel
chr11:g.32456562G>A	3	31.5	37.9%	*WT1*	high GC
chr16:g.2137924T>TCCCTGCAGTGCAGGAAAGGTAGGGCCGGGTGGGG	1	36.1	64.6%	*TSC2*	insertion
chr16:g.15826416A>C	2	37.9	43.2%	*MYH11*	part of complex indel
chr19:g.38956803G>A	1	60.2	48.6%	*RYR1*	high GC
chr19:g.38993372A>G	4	37.2	41.6%	*RYR1*	high GC

Figure 5. Example of a putative benign variant in the ACMG59 panel detected by DeepVariant alone

DeepVariant detected the putative benign variant *WT1* c.330C>T (p.Pro110=) in a high-GC region of exon 1. This variant has been previously reported in ClinVar[7]. The sequencing noise that makes variant calling difficult in this region is visible in this IGV.



Conclusions

- To ensure high sensitivity and redundancy in clinical bioinformatics pipelines, it is critical to include callers optimized for different types of variants.
- Gene panels can have different compositions of difficult to sequence regions and variants types, further highlighting the importance of including multiple callers.
- DeepVariant was shown to have increased sensitivity in regions of high-GC content in the ACMG59 panel.
- Further increases in performance may be possible through customizing DeepVariant's deep learning model by training on a large data set of known rare variants.

References

1. DePristo MA, Banks E, Poplin R, et al. *Nat Genet*. 2011.
2. Poplin R, Chang P-C, Alexander D, et al. *Nat Biotechnol.* 2018.
3. Green RC, Berg JS, Grody WW, et al. *Genet Med.* 2013.
4. van den Akker J, Mishne G, Zimmer AD, Zhou AY. *BMC Genomics. 2*018.
5. Hereditary Cancer Test https://color.com/whitepaper
6. ClinVar. NM_130799.2(MEN1):c.206_207insGCCCC (p.Asp70Profs). https://www.ncbi.nlm.nih.gov/clinvar/variation/183079/. Accessed October 9, 2018.
7. ClinVar. NM_024426.4(WT1):c.330C>T (p.Pro110=). https://www.ncbi.nlm.nih.gov/clinvar/variation/193454/. Accessed October 9, 2018.