More bang for your buck! Universal Sequence Detection: A novel full spectrum variant genotyper for NGS data

Jeroen Van Den Akker, Karen H.Y. Wong, Asha Rostamianfar, Zachary Langley, Michael Cusack, Anju Ondov, Anjali D. Zimmer, Alicia Y. Zhou, Scott Topper
Color Health, Burlingame, CA

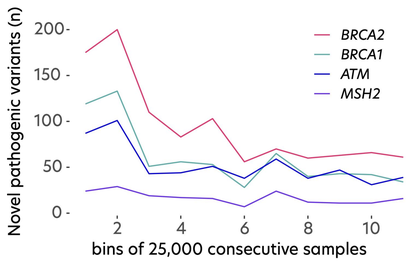
Introduction

The continuously decreasing cost of NGS sequencing has driven the uptake of NGS in the field of clinical genetic testing, rapidly replacing the once widely-used genotyping array. NGS offers the critical advantage of allowing for novel variant detection, however this opportunity comes at the cost of increased bioinformatic complexity. As NGS adoption continues to grow, with many large-scale programs implementing exome and whole genome sequencing, the ability for researchers and clinicians to implement and fine-tune complex variant calling pipelines can be prohibitive. However, genotyping of known variants improves sensitivity and operational efficiency while reducing the complexity of bioinformatics pipelines and clinical workflows.

While NGS methods allow for novel variant discovery, in practice most variants that are identified are not novel. A noted exception to this would be sequencing of individuals from races/ethnicities that have historically not been included in genomics studies. However, even with these populations, as more samples are sequenced, the majority of variants identified will be recurrent. Indeed, we have found that many novel variants were identified in the first 100,000 sequenced samples, but over time the number of novel variants has leveled off (Figure 1). To improve the detection of recurring or otherwise known variants, we created a novel genotyping algorithm.

Here, we describe the development of Unique Sequence Detection (USD), a genotyping-based variant caller that utilizes a library of predefined unique sequences to accurately and efficiently genotype known variants in short-read data. We also show the ability of USD to genotype the entire spectrum of variant types, from single nucleotide variants (SNVs) to complex structural variants (SVs).

Figure 1. Number of novel pathogenic and likely pathogenic variants detected over time at Color.



Methods

The USD workflow operates in two distinct phases: 1) unique sequence generation: a one-time unique sequence generation step where unique sequences are constructed based on previously well-characterized variants, and 2) unique sequence detection: a pattern matching step where an alignment file is scanned to identify reads supporting the constructed unique sequences. Each of these phases are broken down in Figure 2.

Each unique sequence has three parts: 1) a "core" sequence that spans the variant and needs to be matched exactly and 2) left and 3) right "flanking" sequences wherein some variation in nucleotide mismatching and base calling quality scores are tolerated to provide resilience against co-variants and sequencing errors. The length of the unique sequences can vary, and in some cases having a flanking sequence may be optional. An example of a specific unique sequence is shown in the case study (Figure 4). Here, the core sequence is 10 nucleotides (nt), and the left flanking sequence is 11 nt. The right flanking sequence is shorter, only 3 nt, due to the presence of a neighboring homopolymer.

Here we present an application of USD in the Color Diagnostics clinical laboratory running a targeted NGS panel for hereditary disease risk. The library of unique sequences used for the initial validation of the genotyper is shown in Table 1. This set was enriched with hard-to-call variants that are especially suited for variant identification by a genotyper. Once validated, USD was employed in our clinical bioinformatics pipeline. The results section shows USD genotyping performance in an initial cohort of 24,783 samples.

Figure 2. The USD workflow. Nt = nucleotides, indels = insertions and deletions.

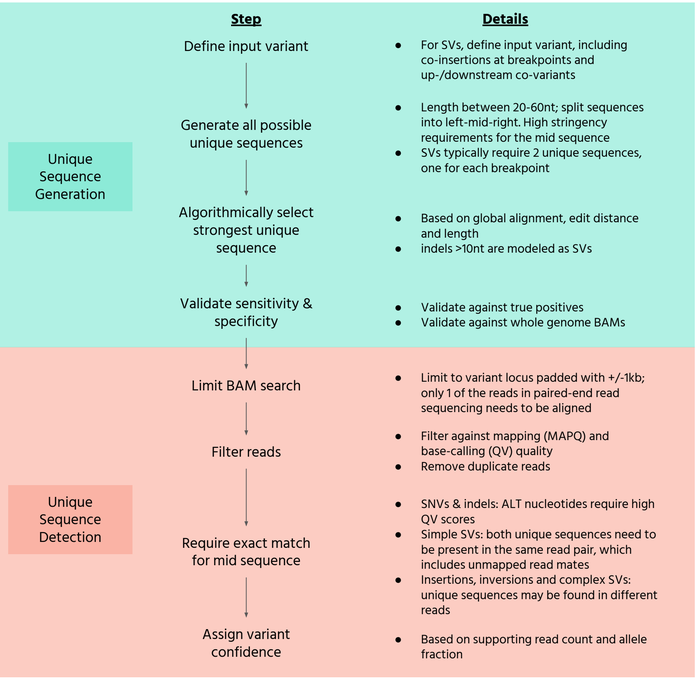


Table 1. Unique sequences in validation set by variant type.

Variant Type		Count
Deletion (including deletion with co-insertions)		510
Duplication		141
Mobile Element Insertion		27
Inversion (including inversion with co-insertions)		6
SNV/indel*	Variant length 5-20nt Variant length ≥ 20nt GC content ≤ 25% GC content ≥ 75% Neighboring homopolymer ≥ 10nt *PMS2* exons 12-15	282 55 75 75 68 15
Total		**1530**

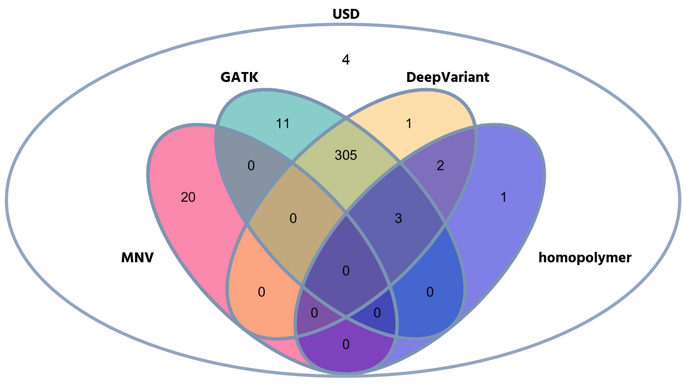
**GC content and maximum homopolymer length were assessed in the +/-50 base pairs (bp) sequence flanking the variant.*

Results

Comparative performance of small variant callers

While most small variants (SNVs and small indels <5nt) can be called by out-of-the-box implementations of common variant callers (for example, GATK[1] or DeepVariant[2]), such variant callers have known difficulties with some challenging variants, including those with neighboring homopolymers, low or high GC content, and high homology. USD can reliably detect these small variants, reducing the need for complex bioinformatic pipelines that incorporate multiple callers specialized to address specific challenges. To illustrate this, Figure 3 shows the performance of a suite of variant callers on a set of 351 challenging small variants genotyped by USD in patient samples. While a majority (305/351, 86.9%) could be called by common variant callers, using any single caller would miss variants. USD was able to call all variants, including four variants that were missed by all callers.

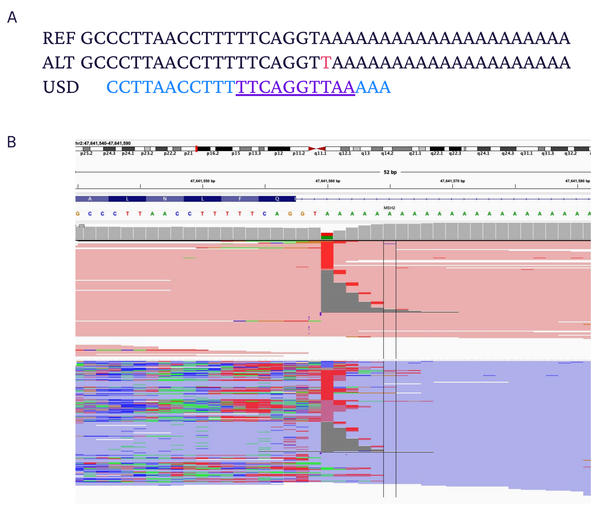
Figure 3. Performance of small variant callers. All 347 variants, of which 300 classified as pathogenic or likely pathogenic, were detected by USD. The following callers were assessed: GATK v3.4, DeepVariant v0.10, "MNV" (Color-developed algorithm for phasing and merging neighboring variants, referred to as Multiple Nucleotide Variant (MNV)), and "homopolymer" (Color-developed algorithm for the detection of variants in the vicinity of long homopolymers). Note that all 11 variants that were only called by GATK were located in *PMS2* exons 12-15, where GATK was run with ploidy = 4 due to the high sequence similarity with *PMS2CL* exons 3-6.



Case study: *MSH2* c.942+3A>T

MSH2 c.942+3A>T is a pathogenic splice variant with a neighboring homopolymer, which we have reported 39 times to date. Standard variant callers, such as GATK and DeepVariant, do not reliably call variants with adjacent homopolymers, but they can be consistently detected using USD.

Figure 4. (A) Generation of the "unique sequence" used for genotyping by USD. Ref = reference sequence, ALT = alternate sequence with the single nucleotide variant (highlighted in red), and USD = the unique sequence generated for USD. The left and right "flanking" sequences which allow for some matching variability are colored blue, and the "core" sequence which requires an exact match is underlined and colored purple.
(B) IGV plot showing alignment of reads surrounding the variant. Note, many reads sequenced from right-to-left (shown in purple) suffer from base-calling errors at the locus of interest due to the 27nt homopolymer. This strand bias presents a complex signal that is challenging for variant callers, and both GATK and DeepVariant were unable to call this variant in this sample.



USD increases operational efficiency

Variant calling in *PMS2* exons 12-15 is complicated by a segmental duplication shared with *PMS2CL*, which results in unreliable alignment of short (2*150bp) sequencing reads. Consequently, germline variant calling algorithms assuming a diploid genome are typically not usable because a heterozygous variant is present in only ~25% of the sequencing reads aligned to *PMS2*. Tables 2 and 3 show 274 SNVs/indels and 58 SVs genotyped by USD in a consecutive cohort of 24,783 samples, which has greatly reduced the number of required operational interventions.

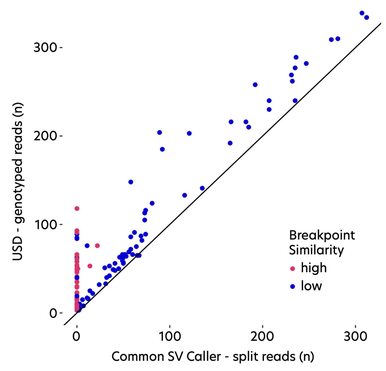
Table 2. Comparison of performance of USD to other common variant callers for small variants in *PMS2*. Note, candidate variants require disambiguation to *PMS2* or *PMS2CL* using long-range PCR.

Candidate SNV/indel in *PMS2*	Samples (n = 247)	Called by DeepVariant v0.10	Called by GATK v3.4 with default settings	Called by GATK v3.4 with ploidy = 4	Called by USD
c.2113G>A	1	x	x	✓	✓
c.2174+1G>A	4	x	x	✓	✓
c.2182_2184delinsG	130	x	x	2 variants	✓
c.2192_2196del	2	x	x	✓	✓
c.2243_2246del	113	x	x	✓	✓
c.2404C>T	2	x	x	✓	✓
c.2444C>T	2	x	x	✓	✓
c.2444_2445insTT	13	x	x	✓	✓
c.2500_2501delinsG	7	x	x	2 variants	✓

Improved structural variant calling with USD

SV calling using short read sequencing poses a challenge. Reads that span breakpoints (so-called split reads) frequently have poor mappability and lack explicit soft-clipping, especially when the SV results from recombination between common repeat elements (e.g., Alu-Alu). Consequently, such reads are typically not utilized by common SV-calling algorithms, with SVs sharing high sequence similarity between their breakpoints (Figure 5), defined by a Levenshtein edit distance < 40 based on the +/- 50bp flanking sequence (details in van den Akker *et al,* 2021[3]). In addition, co-insertions between SV breakpoints obscure split read signals. USD, which does not rely on alignment metadata such as insert size, read pair orientation and read clipping, utilizes such reads to improve SV identification.

Figure 5. USD utilizes more reads to call SVs than a common SV caller (LUMPY v0.2.13[4]). Consequently, USD is capable of identifying SVs that share high sequence similarity between their breakpoints.



USD can detect challenging structural variants

Structural variants overlapping *PMS2* exons 12-15 are challenging to detect due to both the segmental duplication shared with *PMS2CL* and the high density of Alu elements, which result in SVs with high breakpoint similarity (Figure 5). In a consecutive cohort of 24,783 samples, USD identified 10 distinct SVs in a total of 58 samples; 9 of these SVs exhibit very little/no split read signal due to the high sequence similarity shared between their breakpoints. SV genotyping enabled subsequent stratification to *PMS2* or *PMS2CL* using long-range PCR and variant-specific nested PCR.

Table 3. Detection of challenging structural variants in *PMS2*.

Candidate Structural Variant in *PMS2*	Variant Length	Samples (n = 58)	Mechanism of Recombination	Breakpoint Similarity*
duplication of exons 1-12	39,678	1	none-Tigger2	59
deletion of exons 1-15 (whole gene)	152,231	1	AluSx-AluSq2	23
deletion of exons 2-15	73,294	1	AluSx1-AluSx1	22
deletion of exon 12	1,637	5	AluSx-AluSz	17
deletion of exons 13-14	2,968	43	AluSz6-AluSc	21
deletion of exons 13-15	9,173	1	AluSx-AluSx	19
deletion of exon 14	1,885	1	AluSc-AluY	19
deletion of exon 14	1,890	3	AluSc-AluY	33
deletion of exons 14-15	5,429	1	AluSx-AluSq	31
deletion of exons 14-15	9,429	1	AluSg-AluSq	15

**Levenshtein distance for +/-50bp. High similarity is defined as < 40.*

Conclusions

- USD genotyping is capable of identifying the entire spectrum of variants, including SNVs/indels, variants neighboring long homopolymers, MNVs, CNVs, mobile element insertions, inversions, and complex rearrangements.
- USD increases sensitivity: common germline variant calling algorithms missed, or incorrectly called, more than 10% of the challenging SNVs and indels genotyped by USD. Additionally, USD utilizes signals in sequencing reads that are missed by common structural variant detection algorithms due to sequence similarity and co-insertions.
- USD increases operational efficiency. Genotyping of challenging variants such as MNVs and complex SVs can eliminate the requirement for human review/intervention for accurate variant classification and reporting.
- USD reduces the complexity of a bioinformatics pipeline. Calling MNVs in *PMS2* exons 12-15 requires combining two specialized algorithms: one to call variants with an expected ploidy of 4, and another to phase and merge neighboring variant calls. USD identifies such MNVs in the same way as any regular indel, with no strict expectations for variant allele fraction.

Future Directions

- For large sequencing panels and whole genome sequencing, it is frequently not feasible to identify all variant types using an ensemble of variant calling algorithms due to the high false positive rate and the difficulty of consolidating slightly different calls representing the same variant. For this application, USD could provide a sensitivity boost, with high specificity, to a basic pipeline using GATK/DeepVariant and a read-depth CNV caller.
- USD could be used for pharmacogenomics applications to enable accurate reporting of star alleles such as *CYP2C19**37 based on recurrent deletions (e.g., exons 1-5 and exons 2-5).

References

1. DePristo MA, Banks E, Poplin R, et al. *Nat Genet.* 2011.
2. Poplin R, Chang PC, Alexander D, et al. *Nat Biotechnol.* 2018.
3. van den Akker J, Hon L, Ondov A, et al. *J Mol Diagn.* 2021.
4. Layer RM, Chiang C, Quinlan AR, Hall IM. *Genome Biol.* 2014.

To learn more about research at Color, please visit color.com/research-platform