Structural Variant Simulator Improves Variant Calling Accuracy in NGS

Gilad Mishne, Alex Robertson, Nish Bhat, Annette Leon, Anjali Zimmer, Jeroen Van Den Akker

color

Introduction

While structural variants (SV) are believed to account for at least 10% of pathogenic variants[1,2], historically their detection has been challenging. Therefore, it can be difficult to obtain sufficient clinical cases with known SVs for use in the development and validation of next generation sequencing (NGS) assays[3]. One potential approach to overcome this sample shortage would be to synthesize DNA molecules that contain SVs and run them through the NGS assay under development. While this approach accurately incorporates factors present in an actual sample, it is extremely costly and time-consuming. An alternative approach is to simulate SVs by creating reads from a modified reference genome[4]. This approach has been useful for the development of whole genome sequencing methodologies[5], but it is not as applicable to target enrichment based-assays in which read coverage patterns differ from whole genome sequencing. Additionally, it does not account for key factors that affect SV detection, including sample quality, target enrichment chemistry, and platform-dependent artifacts.

Here, we present the Color Structural Variant Simulator that directly modifies aligned reads of a sequenced sample, resulting in a realistic noise model of both the capture and sequencing stages. We demonstrate its application for evaluating SV calling methods in our target enrichment-based assay and its application to other NGS workflows to incorporate laboratory-specific effects.

Methods

Simulator specifications

The *in silico* simulator uses a sequence alignment file (BAM or SAM) as input along with properties of the SV to be simulated, including variant type, position, and parameters specific to the variant type being simulated (see Table 1). SV types supported are copy number variations (CNVs), insertions, and inversions. Typical simulations take a few seconds to complete. Aligned reads in the alignment file are then modified as follows:

Copy-number variations (deletions and duplications)

To simulate changes in copy-number, reads in the affected region are stochastically removed or duplicated from the alignment file, depending on a user-provided rate. For an example, see Figure 1. All components of a read are modified, including hard/soft clipping, CIGAR strings, mapped sequence, or other user-supplied parameters. By controlling the unaligned sequence, the user can simulate different types of variants, i.e. tandem duplications, processed pseudogenes, etc.

Insertions

To simulate an insertion, a user provides a position and a sequence to be inserted. Reads that span the breakpoint are clipped in a way similar to that for CNVs; the inserted sequence is used as the alternative mapped sequence. Reads fully within the inserted sequence are not simulated, as they would not be captured with the original assay.

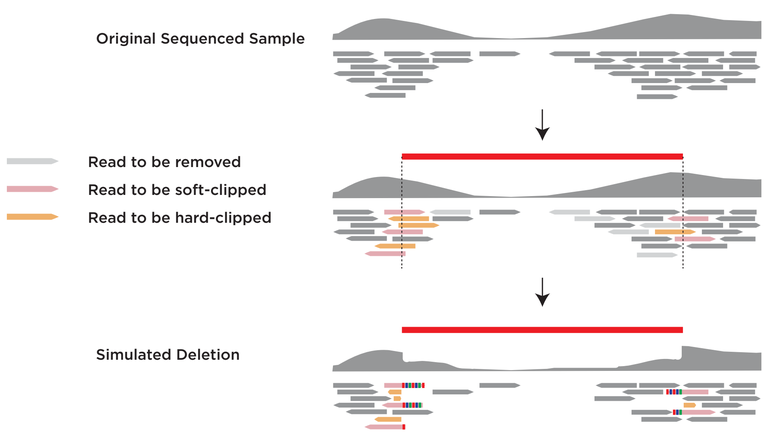
Inversions

To simulate an inversion, a user provides start and end genomic coordinates, and the simulator constructs the inverted sequence -- the reverse complement of the inverted region. The sequence to be inverted is taken from the aligned sequence (BAM) file. Subsequently, the sequence and CIGAR of every read overlapping the inverted region are modified to match the new sequence.

Additional operations

All simulated variants can be further modified by adding noise such as additional base substitutions and indels, at a user-selected rate. A final optional step realigns modified reads to the reference genome, to correct any mismatches. Currently, for paired-read data, reads are modified independently of their mates, which may impact paired-end-based variant callers. Additionally, inserted sequences are simulated only through the clipped read, and not through the paired mate. Improved paired-read support is planned for future versions.

Figure 1: Schematic representation of a simulated deletion



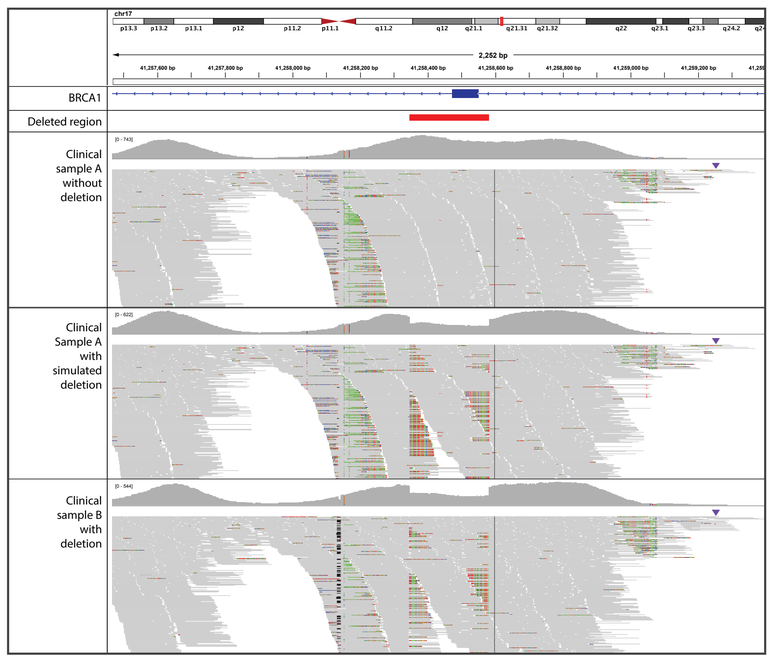
Simplified representation of some of the simulator's functionality. A real sample is used as input. The user defines the location and size of the deletion to be simulated (shown as a red bar) and percentage of reads to be deleted (here reads in the selected region are chosen to be removed at a probability of 50%, independent of other reads in the region, indicated in light grey). Reads that span the breakpoint will be soft- or hard-clipped (indicated in red and orange, respectively). Soft-clipped reads are replaced with a user provided sequence (indicated as multi-colored segment).

Table 1: Parameters of the Color Structural Variant Simulator

Parameter	Value	Applies to	Note
Input, output	BAM/SAM file	All	The aligned sequence file to simulate a SV in, and the output to generate.
Position	Chromosome, start, end	All	The genomic coordinates in which the SV should be simulated.
SNP rate	0-100%	All	Rate of bases to randomly substitute, in the affected region.
Indel rate	0-100%	All	Rate of bases to randomly insert or delete, in the affected region.
Insertion rate	0-100%	Insertions	The rate of reads spanning an insertion breakpoint to be modified.
Insertion sequence	FASTA sequence	Insertions	The sequence to be inserted.
Replacement sequence	FASTA sequence	CNVs	The sequence to be used for soft-clipped reads spanning breakpoints.
Max clipping	0-500	All	The maximum number of bases that can be soft-clipped within a modified read. If more bases in read would be need to be clipped, use hard-clips.
CNV rate change	0-500%	CNVs	The rate by which to change the read depth within the affected region.
Breakpoint rate	0-100%	CNVs	The fraction of reads spanning a breakpoint that will be modified.
Seed	Integer	All	A random seed used for reproducibility of the results.

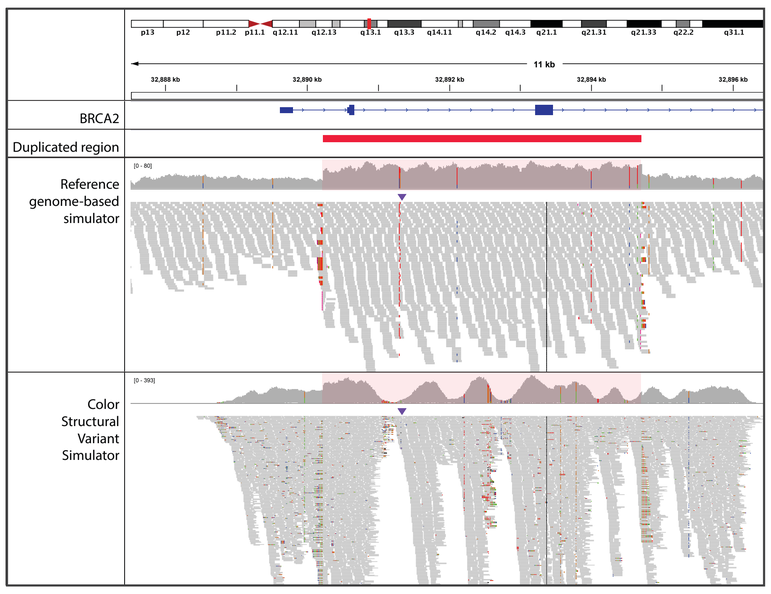
Results

Figure 2: Simulation of a clinically identified *BRCA1* deletion



Using the Color Structural Variant Simulator to simulate a 244-bp deletion of *BRCA1* exon 4, a variant previously reported in the literature [6,7]. *Top*: Sequencing of Clinical Sample A that does not have the deletion, used as reference for the simulation. *Middle*: Simulation of the deletion using the sequencing of Clinical Sample A as input. *Bottom*: Sequencing of a Clinical Sample B that carries the deletion.

Figure 3: Reference genome based simulators cannot realistically simulate SVs in targeted NGS data.

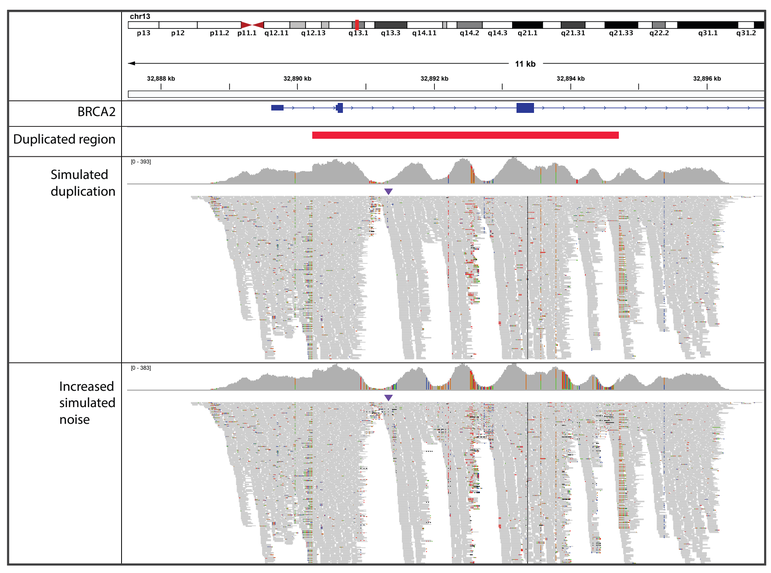


Simulated duplication in *BRCA2*. *Top*: SV simulators that use a whole genome sequencing derived reference genome result in relatively stable read coverage, with an increase in coverage in the duplicated region, and simulated noise different from the noise found in our lab-generated data. *Bottom*: The Color Structural Variant Simulator uses sequenced reads from an actual sample. The data used here is from a target enrichment methodology, note the characteristic probe coverage.

Simulator evaluation

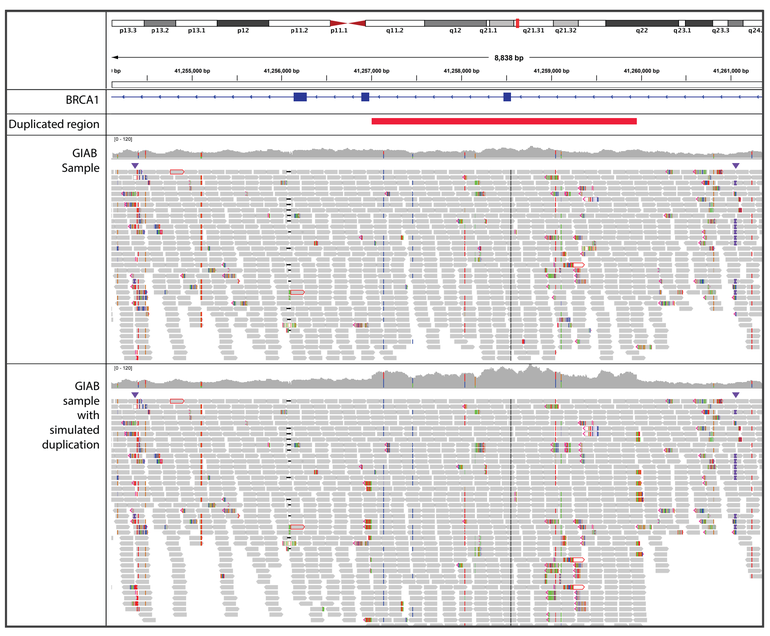
To evaluate the simulator, we tested a depth-based CNV caller on a high-homology region. The same caller was able to achieve 100% sensitivity on other genomic regions, but due to the small number of samples in the high-homology region (less than 10 clinical samples with known CNVs were available), the caller was over-fitted and did not achieve the expected sensitivity. We achieved further improvement by performing a literature review of CNVs in this region and simulating additional samples with CNVs using published breakpoints. The simulated samples increased the amount of data for development and validation significantly, enabling more robust parameter tuning and evaluation, and resulting in 100% sensitivity.

Figure 4: Simulated noise is tunable



A simulated duplication in *BRCA2*, with default and increased levels of simulated noise. Simulating increased noise could help evaluate a variant caller's sensitivity in various scenarios, such as amplification and sequencing artifacts. *Top*: Simulated duplication with the default levels of simulated noise: SNP rate of 0.05% and indel rate of 0.02%. *Bottom*: Increased simulated noise: SNP rate of 0.3% and indel rate of 0.1%.

Figure 5: Using Color Structural Variant Simulator for other workflows



A duplication in *BRCA1* simulated by the Color Structural Variant Simulator using whole genome sequencing data from Genome in a Bottle (GIAB)[8]. *Top*: Original whole-genome sequencing reads of reference sample NA12878 from GIAB. *Bottom*: Simulated duplication, note the incorporation of noise from original sample.

Conclusions

- Simulating structural variants directly on aligned reads is a high-quality, yet cost-effective method for generating realistic data to develop structural variant detection algorithms.
- The simulator was effective for our capture-based assay, which is difficult to simulate with currently available reference genome based simulators, and can be applied to other NGS workflows.
- Using realistic simulated variants, we were able to significantly improve structural variant calling on homologous regions and other areas for which biological samples are rare.
- The Color Structural Variant Simulator is available at **github.com/color/clrsvsim**.

References

1 Judkins, T. *et al.* Clinical significance of large rearrangements in BRCA1 and BRCA2. *Cancer* **118**, 5210–5216 (2012).

2 Ewald, I. P. *et al.* Genomic rearrangements in BRCA1 and BRCA2: A literature review. *Genet. Mol. Biol.* **32**, 437–446 (2009).

3 Lupski, J. R. Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environ. Mol. Mutagen.* **56**, 419–436 (2015).

4 Mu, J. C. *et al.* VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics* **31**, 1469–1471 (2015).

5 Abyzov, A. *et al.* Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nat. Commun.* **6**, 7256 (2015).

6 Preisler-Adams, S. *et al*. Gross rearrangements in BRCA1 but not BRCA2 play a notable role in predisposition to breast and ovarian cancer in high-risk families of German origin. *Cancer Genet. Cytogenet.* **168**, 44–49 (2006).

7 Engert, S. *et al.* MLPA screening in the BRCA1 gene from 1,506 German hereditary breast cancer cases: novel deletions, frequent involvement of exon 17, and occurrence in single early-onset cases. *Hum. Mut*at. **29**, 948–958 (2008).

8 Zook, J. M. *et al*. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).

Figure 5: Using Color Structural Variant Simulator for other workflows

A duplication in *BRCA1* simulated by the Color Structural Variant Simulator using whole genome sequencing data from Genome in a Bottle (GIAB)[8]. *Top*: Original whole-genome sequencing reads of reference sample NA12878 from GIAB. *Bottom*: Simulated duplication, note the incorporation of noise from original sample.