Subject Section

VISOR: a versatile haplotype-aware structural variant simulator for short and long read sequencing

Davide Bolognini [1,3,*], Ashley Sanders [2], Jan O. Korbel [2], Alberto Magi [4], Vladimir Benes [3], Tobias Rausch [2,3]

[1] Department of Experimental and Clinical Medicine, University of Florence, Florence, 50134, Italy
[2] European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, 69117, Germany
[3] European Molecular Biology Laboratory (EMBL), GeneCore, Heidelberg, 69117, Germany
[4] Department of Information Engineering, University of Florence, Florence, 50134, Italy

[*] To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: VISOR is a tool for haplotype-specific simulations of simple and complex structural variants (SVs). The method is applicable to haploid, diploid or higher ploidy simulations for bulk or single-cell sequencing data. SVs are implanted into FASTA haplotypes at single-basepair resolution, optionally with nearby single-nucleotide variants. Short or long reads are drawn at random from these haplotypes using standard error profiles. Double- or single-stranded data can be simulated and VISOR supports the generation of haplotype-tagged BAM files. The tool further includes methods to interactively visualize simulated variants in single-stranded data. The versatility of VISOR is unmet by comparable tools and it lays the foundation to simulate haplotype-resolved cancer heterogeneity data in bulk or at single cell resolution.

Availability and implementation: VISOR is implemented in python 3.6, open-source and freely available at `https://github.com/davidebolo1993/VISOR`. Documentation is available at `https://davidebolo1993.github.io/visordoc/`

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic structural variants (SVs) are a major source of variation in the human genome (Alkan *et al*., 2011), (Weischenfeldt *et al*., 2013). Over the past years, a wide variety of tools for SV calling from both short and long reads have been developed, including DELLY (Rausch *et al*., 2012), LUMPY (Layer *et al*., 2014), Manta (Chen *et al*., 2016), NanoSV (Stancu *et al*., 2017), Sniffles (Sedlazeck *et al*., 2018) and SVIM (Heller *et al*., 2019). Benchmarking SV callers using ground truth datasets is essential to assess their performance and correctness.

High-coverge whole-genome and whole-exome datasets are publicly available (1000 Genomes Project Consortium *et al*., 2015) but their full SV spectrum is still unknown; moreover, generating new real datasets with specific features requires extensive and cost-prohibitive biological experiments (Zook *et al*., 2014) and can be biased by the experimental design. *In silico* simulations are an inexpensive and unbiased alternative and the available ground truth enables an accurate estimation of precision and recall of SV calling methods.

With the advent of long and linked-read technologies that allow read-backed phasing, a SV simulator should be able to simulate haplotype-specific SVs. In addition, we are entering an era where single-cell sequencing complements bulk sequencing approaches; therefore, a versatile simulation approach that can handle SVs of different size, clonality, haplotype-configuration and single-cell versus bulk sequencing resolution is greatly needed.

Over the last years some SV simulators have been developed, including RSVSim (Bartenhagen *et al*., 2013), SCNVSim (Qin *et al*., 2015), VarSim (Mu *et al*., 2015), BAMSurgeon (Ewing *et al*., 2015) and SVEngine (Xia *et al*., 2018). However, these tools have some limitations, either with respect to simulating certain SV types or to the lack of support for haplotype-resolved SVs. Most tools are also restricted to short reads only and, to the best of our knowledge, VISOR is the only tool capable to simulate strand-specific DNA sequencing (Strand-seq) data (Supplementary Table S1).

VISOR fills an important gap in the field of SV simulations by allowing users to rapidly generate one or more haplotypes with desired SVs and to simulate sequencing reads with features mirroring those from commercial sequencing providers (Illumina, PacBio, 10X Genomics and Oxford Nanopore Technologies).

Once haplotypes have been created, users can decide whether to simulate short (single- or double-stranded) or long reads from selected regions/chromosomes. For bulk sequencing and cancer genomics applications, users can adjust for tumour purity, heterogeneity and ploidy (Liu *et al*., 2018). VISOR also incorporates capture biases for targeted assays such as cancer panels or whole-exome target enrichment. Simulating such biases accurately is important for assessing the performance of genomic variation callers (Wang *et al*., 2018). Moreover, VISOR allows to simulate SVs and small variants jointly which is essential for testing SV detection approaches employing haplotype-resolved local assemblies (Chaisson *et al*., 2019).

2 Materials and methods

An overview of VISOR's worfkfow is available in Supplementary Figure S2 while its main modules and methods are outlined below.

2.1 Generate haplotypes with SVs

VISOR offers a module, called HACk, to generate haplotype-resolved FASTA files with implanted SVs. The module requires a template FASTA file and one or more BED files containing non-overlapping SVs. VISOR scans the template FASTA file by chromosome and inserts the SVs and optionally, additional short variants.

In addition to simple SVs such as deletions, tandem duplications, inversions and insertions, VISOR HACk improves the creation of more complex SVs such as inverted duplications, interspersed duplications or reciprocal translocations. A unique strength of VISOR is the simulation of perfect/approximate tandem repetitions and the contraction/expansion of existing microsatellites, thus being potentially suitable to test performances of tandem repeat calling methods (Dashnow *et al*., 2018), (Gymrek *et al*., 2012).

2.2 Simulate short and long reads

VISOR provides two different modules, called SHORtS and LASeR, for short and long read simulations respectively. These modules require a folder containing one or more FASTA haplotypes generated by VISOR HACk and a BED file describing the desired regions/chromosomes to simulate. VISOR SHORtS and LASeR implement state-of-the-art pipelines to generate a final BAM file modelling current sequencing error rates, fragment sizes and read-lengths. For each region in the BED file, FASTQ files are simulated using either wgsim (SHORtS) or PBSIM (Ono *et al*., 2013) (LASeR), SAM files are computed using BWA (Li *et al*., 2009) (SHORtS) or Minimap2 (Li, 2018) (LASeR), and SAM-to-BAM conversion is carried out by Samtools (Li *et al*., 2009). Each read in the final BAM file is eventually tagged with the haplotype number ("HP"-tag), making the alignments easily haplotype-resolvable, for example with IGV (Thorvaldsdóttir *et al*., 2013). For each region in the BED file, users must also define their capture efficiency and purity values, the defaults assume no capture bias and a fully clonal simulation at 100 percent haploid variant allele frequency.

2.2.1 Simulate heterogeneous data

By giving more than one folder with FASTA haplotypes as inputs, both VISOR SHORtS and LASeR are capable to simulate heterogeneous data. This is particularly useful to simulate tumour sub-clonality, a prominent feature of some cancer types during tumorigenesis (Dagogo-Jack *et al*., 2018). When multiple input folders are given, these are considered as sub-clones of a heterogeneous sample and they will be simulated according to relative user-defined percentages (Supplementary Figure S3). As for reads coming from different haplotypes, reads coming from different sub-clones can be easily discriminated in the final BAM file as they are tagged with different clone numbers ("CL"-tag).

2.2.2 Simulate Strand-seq data

By default, VISOR SHORtS performs double-stranded data simulation; however, it can also simulate Strand-seq data. Strand-seq is a single-cell template strand sequencing technology which generates directional genomic libraries that allow clear distinction between the individual homologs of chromosomes. Homolog resolution allows to identify sister chromatid exchanges (SCE) events (Falconer *et al*., 2013) and to locate haplotype-specific SVs, including inversions, translocations and SVs with copy-number changes (Sanders *et al*., 2017). VISOR's *strand-seq* mode considers each input folder as a single-cell and, for each haplotype, VISOR SHORtS generates strand-specific BAM files (see Supplementary Information, S4).

3 Discussion

VISOR is a comprehensive SV simulator: it generates haplotype-resolved SVs and implements state-of-the-art pipelines to generate tagged BAM files, which can in turn be used to evaluate SVs callers performances (Supplementary Figure S5). VISOR supports simulations of reads coherent with short read, linked-read and long read technologies as well as with double-stranded and single-stranded data, where it enables visualization of haplotype-resolved SVs. The applications of VISOR range from simulating bulk, aneuploid, and heterogeneous cancer genomics samples to single-cell studies of germline mosaicisms.

4 Acknowledgements

The authors thank the Human Genome Structural Variation Consortium for data access, and EMBL GeneCore and IT for technical support.

Funding

JOK is supported by ERC grant 773026. AM is supported by Associazione Italiana per la Ricerca sul Cancro (AIRC Investigator Grant 20307, 'Third Generation Cancer Genomics'). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68-74.

Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011 May;12(5):363-76.

Bartenhagen C, Dugas M. RSVSim: an R/Bioconductor package for the simulation of structural variations. Bioinformatics. 2013 Jul 1;29(13):1679-81.

Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019 Apr 16;10(1):1784.

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications.Bioinformatics. 2016 Apr 15;32(8):1220-2.

Dagogo-Jack I, Shaw AT.Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018 Feb;15(2):81-94.

Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, Davis M, Lamont P, Clayton JS, Laing NG, MacArthur DG, Oshlack A. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol. 2018 Aug 21;19(1):121.

Ewing AD, Houlahan KE, Hu Y, Ellrott K, Caloian C, Yamaguchi TN, Bare JC, P'ng C, Waggott D, Sabelnykova VY; ICGC-TCGA DREAM Somatic Mutation Calling Challenge participants, Kellen MR, Norman TC, Haussler D, Friend SH, Stolovitzky G, Margolin AA, Stuart JM, Boutros PC. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. 2015 Jul;12(7):623-30.

Falconer E, Lansdorp PM. Strand-seq: a unifying tool for studies of chromosome segregation. Semin Cell Dev Biol. 2013 Aug-Sep;24(8-9):643-52.

Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res. 2012 Jun;22(6):1154-62.

Heller D, Vingron M. SVIM: Structural Variant Identification using Mapped Long Reads. Bioinformatics. 2019 Jan 21.

Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery.Genome Biol. 2014 Jun 26;15(6):R84.

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 Jul 15; 25(14): 1754–1760.

Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094-3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9.

Liu J, Dang H, Wang XW. The significance of intertumor and intratumor heterogeneity in liver cancer. Exp Mol Med. 2018 Jan; 50(1): e416.

Mu JC, Mohiyuddin M, Li J, Bani Asadi N, Gerstein MB, Abyzov A, Wong WH, Lam HY. VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. Bioinformatics. 2015 May 1;31(9):1469-71.

Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator–toward accurate genome assembly. Bioinformatics. 2013 Jan 1;29(1):119-21.

Qin M, Liu B, Conroy JM, Morrison CD, Hu Q, Cheng Y, Murakami M, Odunsi AO, Johnson CS, Wei L, Liu S, Wang J. SCNVSim: somatic copy number variation and structure variation simulator. BMC Bioinformatics. 2015 Feb 28;16:66.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012 Sep 15;28(18):i333-i339.

Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nat Protoc. 2017 Jun;12(6):1151-1176.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single molecule sequencing. Nat Methods. 2018 Jun; 15(6): 461–468.

Stancu MC, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Giorgia M, Valle-Inclan JE, Korzelius J, de Bruijn E, Cuppen E, Talkowski ME, Marschall T, de Ridder J, Kloosterman WP. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017; 8: 1326.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar;14(2):178-92.

Wang VG, Kim H, Jeffrey H. Chuang. Whole-exome sequencing capture kit biases yield false negative mutation calls in TCGA cohorts. PLoS One. 2018; 13(10): e0204912.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013 Feb;14(2):125-38.

Xia LC, Ai D, Lee H, Andor N, Li C, Zhang NR, Ji HP. SVEngine: an efficient and versatile simulator of genome structural variations with features of cancer clonal evolution. Gigascience. 2018 Jul 1;7(7).

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014 Mar;32(3):246-51.