HAL
open science

SVJedi-graph: genotyping close and overlapping structural variants with a variation graph and long-reads

Sandra Romain, Claire Lemaitre

To cite this version:

Sandra Romain, Claire Lemaitre. SVJedi-graph: genotyping close and overlapping structural variants with a variation graph and long-reads. JOBIM 2022 - Journées Ouvertes en Biologie, Informatique et Mathématiques, Jul 2022, Rennes, France. hal-03885541

HAL Id: hal-03885541

https://inria.hal.science/hal-03885541v1

Submitted on 6 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SVJedi-graph: genotyping close and overlapping structural variants with a variation graph and long-reads

Sandra Romain[1] and Claire Lemaitre[1]
Université de Rennes 1, Inria, IRISA, 35000, Rennes, France

Corresponding author: claire.lemaitre@inria.fr

Abstract

Structural variants (SVs) are genomic segments of more than 50 bp that have been rearranged in the genome. The advent of long-read sequencing technologies has increased and enhanced their study, and a great number of SVs has already been discovered in many species. Complementary to their discovery, the genotyping of known SVs in newly sequenced individuals is of particular interest for several applications such as trait association and clinical diagnosis. Due to SVs' large size range (up to a few megabases), long-reads are more suited for their study than short-reads. As such, our team previously released SVJedi [1], one of the first SV genotypers using long-read data. SVJedi's method of representing independently both SV's allelic sequences reduced reference bias in genotyping and showed improved genotyping performances. However, the method failed to genotype closely located or overlapping SVs due to redundancy in representative allelic sequences.

To overcome this limitation, we present SVJedi-graph, a long-read SV genotyper based on a variation graph to represent SV alleles. The use of sequence graphs to represent SVs for genotyping is fairly recent [2,3,4,5], but existing methods are restricted to short-read data, and SVJedi-graph is the first graph-based SV genotyper using long-reads. In our method, we build the variation graph from a reference genome and a given set of SVs. The genome sequence is split in fragments at each SV's start and end positions, and each fragment becomes a node in the graph. Edges are added between nodes to indicate reference and alternative paths for each SV, and additional nodes are added for insertions. Then, the long reads are mapped on the variation graph using GraphAligner [6] and the resulting alignments are filtered on their quality and mapping localization. Finally, the most likely genotype for each SV is predicted from the ratio between the number of reads supporting each allele.

SVJedi-graph can genotype four SV types as of now, namely deletions, insertions, inversions and translocations. Running SVJedi-graph on simulated sets of deletions showed that the use of a variation graph was able to restore the genotyping quality on close and overlapping SVs. For instance, with a simulated set of deletions that had another close deletion 0 to 50 bp apart, we obtained a genotyping rate (proportion of SVs with a predicted genotype) of 99.9% and an accuracy (proportion of accurate genotype predicted among all predicted genotypes) of 99.0%, compared to a genotyping rate of 78.9% and an accuracy of 97.3% with SVJedi on the same dataset. We also tested our method on the real gold standard dataset of Genome In A Bottle (human individual HG002), and were able to obtain a higher genotyping rate than SVJedi on the same data (97.4% against 90.2%), with a similar or slightly better accuracy (92.9% against 92.2%). SVJedi-graph is distributed under an AGPL license and available on GitHub at https://github.com/SandraLouise/SVJedi-graph.

References

[1] L. Lecompte et al. SVJedi: genotyping structural variations with long reads. *Bioinformatics*, 36(17):4568–4575, 2020.

[2] E. Garrison et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36(9):875–879, 2018.

[3] S. Chen et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, 20(1):291, 2019.

[4] H. P. Eggertsson et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nature Communications*, 10(1):5402, 2019.

[5] G. Hickey et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biology*, 21(1):35, 2020.

[6] M. Rautiainen and T. Marschall. GraphAligner: rapid and versatile sequence-to-graph alignment. *Genome Biology*, 21(1):253, 2020.