SVJedi-graph: Structural Variant genotyping with long-reads using a variation graph

Sandra Romain, Claire Lemaitre

SVJedi-graph: Structural Variant genotyping with long-reads using a variation graph

Sandra Romain[1] and Claire Lemaitre[1]
Univ Rennes, Inria, CNRS, 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 third generation sequencing technologies has increased and enhanced their study, and a great number of SVs has already been discovered in the human genome. 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. Most of the SV genotypers currently available are designed for second generation sequencing data, although third generation sequencing data is more suited to study SVs due to their large range of sizes (up to few mega bases). As such, our team previously released SVJedi, the first SV genotyper dedicated to long read data[1]. The method is based on linear representations of the allelic sequences of each SV and each SV is represented and genotyped independently of the other ones. While this is very efficient for distant SVs, the method fails to genotype some closely located or overlapping SVs due to redundancy in representative allelic sequences.

To overcome this limitation, we present a novel approach, SVJedi-graph, which uses sequence graphs instead of linear sequences to represent the SVs. The use of sequence graphs to represent SVs for genotyping is fairly recent [2,3,4] and only designed for short-reads as for now. Here, we chose to represent only the SV sequences and that of the SV flanking regions in our graph, in order to reduce the long-read mapping time. This results in a variation graph composed of multiple connected components, each representing the possible alleles for a region of one, or several SVs in case of close SVs (less than 10 kb apart). In SVJedi-graph, the variation graph is built using VG toolkit[5] after a pre-processing step performed on the data. The long reads are then mapped on the graph using GraphAligner[6], and the mapping results are filtered to keep only the informative alignments. Finally, the genotype for each SV of the dataset is predicted using the estimation method implemented in SVJedi[1].

Tests on simulated long-reads on the human chromosome 1, with 1,000 deletions from the dbVar database, show a similar precision compared to SVJedi (98.1 %, against 97.8 %). Importantly, when additional deletions are added progressively closer to the original 1,000 in the dataset, SVJedi-graph maintains a 100 % genotyping rate with a high precision, when SVJedi is not able to assign a genotype to 21 % of the deletions when they are too close to each other (0-50 bp apart). SVJedi-graph also supports other SV types such as insertions and inversions, for which similar performances were obtained. We are planning to apply SVJedi-graph and to compare it to other approaches on real human re-sequencing data from the Genome In a Bottle consortium.

References

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

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

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

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

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

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