SVJedi: genotyping structural variations with long reads

Lolita Lecompte, Pierre Peterlongo, Dominique Lavenier, Claire Lemaitre

Subject Section

SVJedi: Genotyping structural variations with long reads

Lolita Lecompte[1], Pierre Peterlongo[1], Dominique Lavenier[1] and Claire Lemaitre[1]

[1]Univ Rennes, Inria, CNRS, IRISA - UMR 6074, F-35000 Rennes, France.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Studies on structural variants (SV) are expanding rapidly. As a result, and thanks to third generation sequencing technologies, the number of discovered SVs is increasing, especially in the human genome. At the same time, for several applications such as clinical diagnoses, it is important to genotype newly sequenced individuals on well defined and characterized SVs. Whereas several SV genotypers have been developed for short read data, there is a lack of such dedicated tool to assess whether known SVs are present or not in a new long read sequenced sample, such as the one produced by Pacific Biosciences or Oxford Nanopore Technologies.

Results: We present a novel method to genotype known SVs from long read sequencing data. The method is based on the generation of a set of representative allele sequences that represent the two alleles of each structural variant. Long reads are aligned to these allele sequences. Alignments are then analyzed and filtered out to keep only informative ones, to quantify and estimate the presence of each SV allele and the allele frequencies. We provide an implementation of the method, SVJedi, to genotype SVs with long reads. The tool has been applied to both simulated and real human datasets and achieves high genotyping accuracy. We show that SVJedi obtains better performances than other existing long read genotyping tools and we also demonstrate that SV genotyping is considerably improved with SVJedi compared to other approaches, namely SV discovery and short read SV genotyping approaches.

Availability: https://github.com/llecompte/SVJedi.git

Contact: lolita.lecompte@inria.fr

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

1 Introduction

Structural variations (SVs) are characterized as genomic segments of at least 50 base pair (bp) long, that are rearranged in the genome. There are several types of SVs such as deletions, insertions, duplications, inversions or translocations. With the advent of Next-Generation Sequencing (NGS) technologies and the re-sequencing of many individuals in populations, SVs have been admitted as a key component of human polymorphism (Audano *et al.*, 2019). This kind of polymorphism has been shown involved in many biological processes such as diseases or evolution (Lupski, 2015). Databases referencing such variants grow as new variants are discovered. At this time, dbVar, the reference database of human genomic SVs (Phan *et al.*, 2017) now contains more than 36 million variant calls, illustrating that many SVs have already been discovered and characterized in human populations.

When studying SV in newly sequenced individuals, one can distinguish two distinct problems: discovery and genotyping. In the SV discovery problem, the aim is to identify all the variants that differentiate the given re-sequenced individual with respect usually to a reference genome. In the SV genotyping problem, the aim is to evaluate if a given known SV (or set of SVs) is present or absent in the re-sequenced individual, and assess, if it is present, with which ploidy (heterozygous or homozygous). At first glance, the genotyping problem may seem included in the discovery problem, since present SVs should be discovered by discovery methods. However, in discovery algorithms, SV evidence is only investigated for present variants (*i.e.* incorrect mappings) and not for absent ones. If a SV has not been called, we cannot know if the caller missed it (false negative) or if the variant is truly absent in this individual and this could be validated by a significant amount of correctly mapped reads in this region. Moreover, in the genotyping problem, knowing what we are looking for should make the problem simpler and the genotyping result hopefully

more precise. With the fine characterization of a growing number of SVs in populations of many organisms, genotyping newly sequenced individuals becomes very interesting and informative, in particular in human medical diagnosis contexts or more generally in any association or population genomics studies.

In this work, we focus on the second problem: genotyping already known SVs in a newly sequenced sample. Such genotyping methods already exist for short reads data (Alkan *et al.*, 2011; Chander *et al.*, 2019): for instance, SVtyper (Chiang *et al.*, 2015), SV^2 (Antaki *et al.*, 2017), Nebula (https://www.biorxiv.org/content/10.1101/566620v1). Though short reads are often used to discover and genotype SVs, this is well known that their short size makes them ill-adapted for predicting large SVs or SVs located in repeated regions. SVs are often located alongside repeated sequences such as mobile elements (Kidd *et al.*, 2010), resulting in mappability issues that make the genotyping problem harder when using short read data.

Third generation sequencing technology, such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), can produce much longer reads compared to NGS technologies. Despite their high error rate, long reads are crucial in the study of SVs and have enabled new SV discoveries (Norris *et al.*, 2016; Huddleston *et al.*, 2017; Stancu *et al.*, 2017; Jain *et al.*, 2018). Indeed, the size range of these sequences can reach a few kilobases (kb) to megabases, thus long reads can extend over rearranged sequence portions as well as over the repeated sequences often present at SV's breakpoint regions.

Following long read technology's development, many SV discovery tools have emerged, such as Sniffles (Sedlazeck *et al.*, 2018), NanoSV (Stancu *et al.*, 2017), pbsv (unpublished) or SVIM (Heller and Vingron, 2019). Among these tools, some implement a genotyping module that gives the frequency of alleles after calling SVs of the sequenced samples. Nonetheless, discovery tools require post-processing to evaluate if a set of SVs is present or not in the sample and to compare the SV calls between different samples. To our knowledge, there exist only two tools, that can perform genotyping with long read data for a given set of SVs, the discovery tool Sniffles with a specific option and the SV visualisation tool svviz2, but both are not purely dedicated to the genotyping task.

The main contribution of this work is a novel method to genotype known SVs using long read data. We also provide an implementation of this method in the tool named SVJedi. SVJedi accuracy and robustness were evaluated on simulated data of real deletions in a human chromosome. It was also applied to a real human dataset and compared to a gold standard call set provided by the Genome in a Bottle (GiaB) Consortium, containing both deletions and insertions. High genotyping accuracy was achieved on both simulated and real data. We also demonstrated the improvement of such a dedicated method over other long read genotyping tools and other approaches, namely SV discovery with long reads and SV genotyping with short reads.

2 Methods

The method assigns a genotype for a set of already known SVs in a given individual sample sequenced with long read data. It assesses for each SV if it is present in the given individual, and if so, how many variant alleles it holds, *i.e.* whether the individual is heterozygous or homozygous for the particular variant.

For clarity purposes, we describe here the method for deletion genotyping only. The genotyping of insertions is perfectly symmetrical, the genotyping of inversions and translocations differs only by the number of breakpoints to examine and follows the same strategy. The method takes as input a variant file with SV coordinates in VCF format, a reference genome and the sample of long read sequences. It outputs a variant file complemented with the individual genotype information for each input variant in VCF format.

The method consists of four different steps, that are illustrated in Fig. 1. The fundamentals of the method lie in its first step, which generates *representative allele sequences* that represent the two alleles of each SV. Long reads are then aligned on the whole set of representative allele sequences. An important step consists in selecting and counting only informative alignments to finally estimate the genotype for each input variant.

2.1 Representative allele sequence generation

Starting from a known variant file in VCF format and the corresponding reference genome, the first step consists in generating two sequences for each SV, corresponding to the two possible allele sequences. These representative allele sequences are hereafter simply called *allele sequences*. In the case of deletions, these are parts of the reference genome that may be absent in a given individual. They are characterized in the VCF file by a starting position on the reference genome and a length. We define the reference allele sequence (allele 0) as the sequence of the deletion with adjacent sequences at each side, and the alternative allele sequence (allele 1) consists in the joining of the two previous adjacent sequences. Given that reads of several kb will be mapped on these allele sequences, the size of the adjacent sequences, denoted by L_{adj}, was set to 5,000 bp at each side, giving a 10 kb sequence for allele 1 and 10 kb plus the deletion size for allele 0. For deletions larger than $2 \times L_{adj}$, that is here larger than 10 kb, two representative sequences are generated for allele 0, one for each breakpoint. The same adjacent sequence size is used, *i.e.* 5,000 bp, on each side of the breakpoints, giving then three 10 kb sequences: one for allele 1, and two for allele 0 (left and right breakpoints).

2.2 Mapping

Sequenced long reads are aligned on all previously generated allele sequences, using Minimap2 (Li, 2018) (version 2.17-r941). Option `-c` is specified to generate a CIGAR for each alignment. Alignments are then output in a PAF file.

2.3 Informative alignment selection

Minimap2 raw alignment results have to be carefully filtered out to remove i) uninformative alignments, which are those not discriminating between the two possible alleles, and ii) spurious false positive alignments, that are mainly due to repeated sequences.

Informative alignments for the genotyping problem are those that overlap the SV breakpoints, that is the sequence adjacencies that are specific to one or the other allele. In the case of a deletion, the reference allele contains two such breakpoints, the start and end positions of the deletion sequence; the alternative sequence, the shortest one, contains one such breakpoint at the junction of the two adjacent sequences (see the red thick marks of Fig. 1).

To be considered as overlapping a breakpoint, an alignment must cover at least d_{over} bp from each side of the breakpoint (d_{over} is set by default to 100 bp). In other words, if x and y are the sizes of the aligned parts on the allele sequence at the respectively left and right sides of the breakpoint, they must satisfy the following condition in equation (1) for the alignment to be kept :

$$(x > d_{over}) \;\&\; (y > d_{over}) \tag{1}$$

Concerning the filtering of spurious false positive alignments, Minimap2 alignments are first filtered based on the mapping quality (MAPQ) score. To focus on uniquely mapped reads, the MAPQ score of the alignments must be greater than 10. This is not sufficient to filter out alignments due to repetitive sequences since mapping is performed on a small subset of the reference genome and these alignments may appear as uniquely mapped on this subset.

As Minimap2 is a sensitive local aligner, many of the spurious alignments only cover subsequences of both the allele and the read sequences. To maximize the probability that the aligned read originates from the genomic region holding the SV, we, therefore, require the read to be aligned with the allele sequence in a semi-global manner. Each alignment extremity must correspond to an extremity of at least one of the two aligned sequences. This criterion gathers four types of situations, namely the read is included in the allele sequence, or *vice-versa*, or the read left end aligns on the allele sequence right end or *vice-versa*.

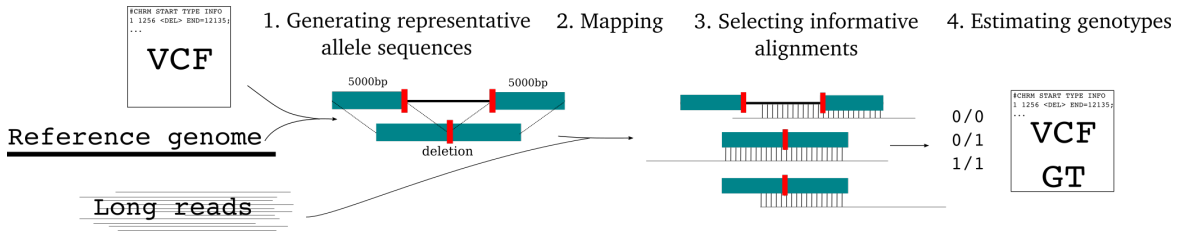

Fig. 1. SVJedi steps for deletion genotyping. Steps for insertion genotyping are symmetrical and are not shown on the figure for clarity purposes. 1. Two corresponding representative allele sequences are generated for each selected SV, one corresponds to the original sequence and the other to the sequence with the deletion. 2. Long read sequenced data are aligned on these allele sequences using Minimap2. 3. Informative alignments are selected. 4. Genotypes are estimated.

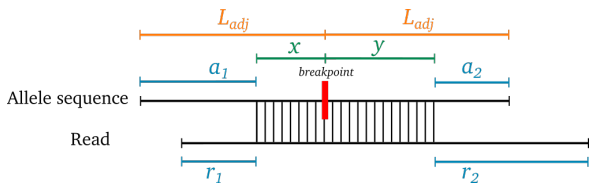

Fig. 2. Definition of the different distances used to select informative alignments between an allele sequence and a read. The aligned parts of the sequences are illustrated by vertical bars. The allele sequence is composed of two adjacent sequences of size L_{adj} on either side of the breakpoint, which is represented by a red vertical thick bar. x and y are the distances of the breakpoint to respectively the start and end coordinates of the alignment on the allele sequence. a_1 and a_2 (resp. r_1 and r_2) are the distances of the alignment left and right extremities to the, respectively, left and right extremities of the allele sequence (resp. read sequence). It follows here, that $a_1 + x = L_{adj}$ and $a_2 + y = L_{adj}$.

Indeed this criteria is not strictly applied and a distance of d_{end} of the alignment to an extremity of at least one of the two aligned sequences is tolerated (d_{end} is set by default to 100 bp). More formally, if a_1 and a_2 (resp. r_1 and r_2) are the sizes of the unaligned parts at the, respectively, left and right sides of the alignment on the allele sequence (resp. read sequence) (see Fig. 2), then the alignment must fulfill the following condition in equation (2) to be kept:

$$(a_1 < d_{end} \parallel r_1 < d_{end}) \;\&\; (a_2 < d_{end} \parallel r_2 < d_{end}) \tag{2}$$

The left member of equation (2) imposes that the unaligned part at the left of the alignment is small in at least one of the two aligned sequences; the right member imposes the same condition at the right side of the alignment.

2.4 Genotype estimation

For each variant, the genotype is estimated based on the ratio of amounts of reads informatively aligned to each allele sequence. In the case of deletions and insertions, the allele sequences of a given variant are of different sizes and contain a different number of breakpoints (for a deletion for instance, the reference allele contains 2 breakpoints, whereas the alternative allele contains only 1), so even if both alleles are covered with the same read depth, there would be fewer reads that align on the shortest allele sequence and that overlap at least one breakpoint. To prevent a bias towards the longest allele, reported read counts for the longest allele are normalized according to the allele sequence length ratio, assuming that read count is proportional to the sequence length. More precisely, in the case of a deletion, the reference allele is the longest allele. Its allele sequence size is the deletion size plus $2 \times L_{adj}$ (cumulative size of the adjacent sequences) if the deletion is smaller than $2 \times L_{adj}$. Otherwise it is composed of two sequences of size $2 \times L_{adj}$ each centered on each breakpoint. We therefore apply the following formula to compute the normalized read count for the reference allele, c_0^*, as a function of the observed read count for the reference allele, c_0, and the deletion size, del_{size}:

$$c_0^* = \begin{cases} c_0 \times \frac{2 \times L_{adj}}{(2 \times L_{adj} + del_{size})} & \text{if} \quad del_{size} < 2 \times L_{adj} \\ c_0 \times \frac{1}{2} & \text{otherwise} \end{cases} \tag{3}$$

Finally, a genotype is estimated if the variant presence or absence is supported by at least min_cov different reads after normalization (sum of the read counts for each allele). By default, this parameter is set to 3.

Genotypes are estimated according to a maximum likelihood strategy. The likelihoods of the three possible genotypes given the observed normalized read counts (c_0^* and c_1) are computed based on a simple binomial model, assuming a diploid individual, as described in Nielsen *et al.* (2011) (see also (Li, 2011)):

$$\mathcal{L}(0/0) = (1-err)^{c_0^*} \times err^{c_1} \times C^{c_0^*}_{c_0^*+c_1} \tag{4}$$

$$\mathcal{L}(1/1) = err^{c_0^*} \times (1-err)^{c_1} \times C^{c_0^*}_{c_0^*+c_1} \tag{5}$$

$$\mathcal{L}(0/1) = \left(\frac{1}{2}\right)^{c_0^*+c_1} \times C^{c_0^*}_{c_0^*+c_1} \tag{6}$$

where err is the probability that a read maps to a given allele erroneously, assuming it is constant and independent between all observations. err was fixed to 5.10^{-5}, after empirical experiments on a simulated dataset (see Supplementary Figure 1).

Finally, the genotype with the largest likelihood is assigned and all three likelihoods are also output (-log10 transformed) as additional information in the VCF file.

2.5 Implementation and availability

We provide an implementation of this method named SVJedi, freely available at https://github.com/llecompte/SVJedi, under the GNU Affero GPL license. SVJedi can also be installed from Bioconda. SVJedi is written in Python 3, it requires as input a set of SVs (VCF format), a reference genome (fasta format)

and a sequencing read file (fastq or fasta format). Notably, the main steps are implemented in a modular way, allowing the user to start or re-run the program from previous intermediate results. As an example, the first step is not to be repeated if there are several long read datasets to be genotyped on the same SV set. Results shown here were obtained with release version 1.1.0.

3 Material

3.1 Long read simulated dataset

SVJedi was assessed on simulated datasets on the human chromosome 1 (assembly GRCh37) based on real characterized deletions for the human genome. From the dbVar database (Phan *et al.*, 2017), we selected 1,000 existing deletions on chromosome 1 (defined as `<DEL>` SV type), which are separated by at least 10,000 bp. The sizes of the deletions vary from 50 bp to 10 kb (with median and average sizes of 950 bp and 2,044 bp respectively) . In this experiment, deletions were distributed into the three different genotypes: 333 deletions are simulated with 0/0 genotype, 334 deletions with 0/1 genotype and the 333 remaining deletions with 1/1 genotype. Two different sequences were simulated containing each overlapping sets of deletions, representing the two haplotypes of the simulated individual. 1/1 genotype deletions were simulated on both haplotype sequences, whereas deletions of 0/1 genotype were simulated each on one randomly chosen of the two haplotype sequences. Then PacBio data were simulated on both haplotypes, using SimLoRD (Stöcker *et al.*, 2016) (version v1.0.2) with varying sequencing error rates (6 %, 10 %, 16 % and 20 %), and at varying total sequencing depths (6x, 10x, 16x, 20x, 30x, 40x, 50x and 60x). Most results presented in the main text are for 16 % error rate and 30x sequencing depth. Ten such datasets were simulated to assess the reproducibility of results.

3.2 Real data

SVJedi was applied on a real human dataset, from the individual HG002, son of the so-called *Ashkenazi trio* dataset. A PacBio Continuous Long Read (CLR) sequencing dataset for HG002 was downloaded from the FTP server of GiaB and down-sampled to 30x read depth (FTP links are given in Supplementary Material). We considered the assembly GRCh37.p13 as the human genome reference and as a gold standard call set, we used the SV benchmark set (v0.6) of HG002 individual provided by the GiaB Consortium (https://www.biorxiv.org/content/10.1101/664623v3). This set contains 5,464 high confidence deletions and 7,281 insertions (`PASS` filter tag), whose sizes range from 50 bp to 125 kb (median sizes of 149 bp and 215 bp for deletions and insertions, respectively). We used the `TRgt100=TRUE` tags present in the GiaB VCF file to identify SVs located in Tandem Repeats greater than 100 bp (denoted as TRs, n= 6,469). 48 SVs were found located inside large ($>$ 10 kb) segmental duplications, using the UCSC Segmental Dups feature track (Bailey *et al.*, 2002).

These SVs were also genotyped in PacBio sequencing datasets of the two parents (HG003 and HG004, 30x and 27x, respectively) to assess the level of Mendelian inheritance consistency of the son predicted genotypes.

SVJedi was applied on a real human ONT PromethION 44x dataset for the individual HG002 as well. Finally, we considered a real short read dataset for the HG002 individual, 2 X 250 bp Illumina dataset from GiaB, that was down-sampled to 30x read depth. This short read dataset is used for comparison with a short read based SV genotyping approach. FTP links for all real sequencing datasets are given in Supplementary Material Section 1).

3.3 Evaluation

To evaluate the accuracy of the method, a contingency table between the estimated genotypes and the true (simulated) ones is computed, providing a clear view of the number and type of correctly and incorrectly estimated genotypes. The genotyping accuracy of the method is then assessed as the number of correctly estimated genotypes overall all estimated genotypes, as shown in equation (7). The percentage of SVs for which a genotype could be estimated is also measured, and hereafter called the genotyping rate (equation (8)).

$$Genotyping\ accuracy = \frac{\text{\# of correctly estimated genotypes}}{\text{\# of estimated genotypes}} \quad (7)$$

$$Genotyping\ rate = \frac{\text{\# of estimated genotypes}}{\text{\# of known SVs}} \quad (8)$$

3.4 Comparison with other genotyping approaches

Comparisons with other genotyping approaches were performed on the real PacBio 30x HG002 dataset.

SVJedi was first compared to two tools, Sniffles (Sedlazeck *et al.*, 2018) and svviz2 (Spies *et al.*, 2015). Both tools, although not dedicated to genotyping, have options that allow them to also do SV genotyping from a set of SVs and with a long read sequencing dataset. Following the recommendations of Sniffles[1], reads were first aligned with NGMLR (version 0.2.7) on the human reference genome. Then, we used Sniffles (version 1.0.11) with the `-Ivcf` option to genotype the GiaB call set. For svviz2, reads were aligned on the human reference genome using Minimap2 (version 2.17-r941). Genotyping was then performed from the sorted Minimap2 alignments using svviz2 (version 2.0a3) with default parameters.

We also compared our approach with two SV discovery tools, Sniffles again but in its default mode and Pacific Biosciences SV caller, pbsv [2]. Sniffles was run with the `-genotype` parameter with the previously obtained NGMLR read alignments. For pbsv, reads were aligned to the reference genome using its own mapper pbmm2 (version 1.1.0) with the `-sort`, `-median-filter` and `-sample` parameters. SVs were then discovered and called with pbsv (version 2.2.2) using default parameters. Both Sniffles and pbsv analyses do not always predict SVs at the exact simulated coordinates, so a predicted SV is considered identical as the expected SV if both SVs overlap by at least 70 %.

Finally, SVJedi was also compared to a SV genotyping approach based on short read data. To do this, the short reads are first aligned with SpeedSeq (Chiang *et al.*, 2015) (version 0.1.2), then the known variants are genotyped with SVtyper (version 0.7.0) with the default settings.

All tools were run on a Linux 40-CPU node running at 2.60 GHz, all command lines are given in Supplementary Material Section 2.

4 Results

4.1 Assessing SVJedi accuracy and robustness on simulated deletions

To comprehensively assess the accuracy and robustness of SVJedi, it was first applied to simulated data. Results for SVJedi are shown here only for deletion type SVs, as insertions variants are simply the counterpart of deletions, results for inversions and translocations are shown in Supplementary Table 1. PacBio long reads were simulated on artificial diploid genomes obtained by introducing deletions in the human chromosome 1. Importantly, the sets of introduced and genotyped deletions are made of real characterized deletions in human populations, to reflect the real size distribution and the real complexity of deletion breakpoints and neighboring genomic contexts. To do so, one thousand deletions located on human chromosome 1 were randomly selected from the dbVar database, ranging from 50 to 10,000 bp in size.

Table 1 shows the obtained genotypes compared with expected ones for one simulated dataset at 30x read depth. On this dataset, SVJedi achieves 97.8 % genotyping accuracy, with 974 deletions correctly predicted over 996 with

[1] `https://github.com/fritzsedlazeck/Sniffles/wiki/SV-calling-for-a-population`

[2] `https://github.com/PacificBiosciences/pbsv`

an assigned genotype. Among the 1,000 assessed deletions, only 4 could not be assigned a genotype due to insufficient coverage of informative reads, the genotyping rate being thus 99.6 %. Among the few genotyping errors, most concern 1/1 genotypes that were incorrectly predicted as 0/1.

		SVJedi predictions			
		0/0	0/1	1/1	./.
Truth	0/0	331	1	0	1
	0/1	3	330	0	1
	1/1	0	18	313	2

Genotyping accuracy : 97.8 %

Table 1. Contingency table of SVJedi results on PacBio simulated data (30x) of human chromosome 1 with 1,000 deletions from dbVar. SVJedi genotype predictions are indicated by column and the expected genotypes are shown by row. The genotype "./." column corresponds to deletions for which SVJedi could not assess the genotype.

SVJedi genotyping accuracy results were evaluated in terms of varying sequencing depths, ranging from 6x to 60x (see Fig. 3). As expected, the accuracy of SVJedi increases with the read depth, but interestingly, even at low coverage (6x) the accuracy is on average above 94 % and a plateau is quickly reached between 20x and 30x, with already 97 % of genotyping accuracy at 20x. The genotyping rate reaches its plateau at a sequencing depth of 16x.

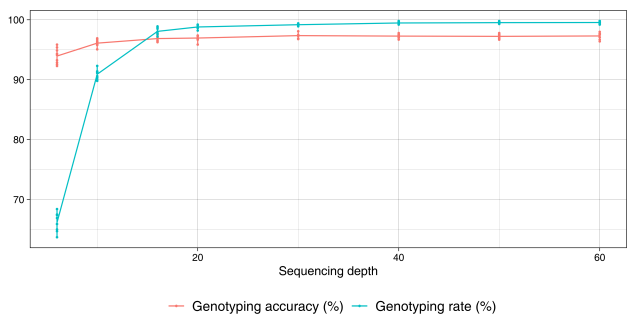

Fig. 3. SVJedi genotyping accuracy results as a function of the sequencing depth for nine simulated PacBio datasets of human chromosome 1, containing 1,000 deletions from the dbVar database. The red dots correspond to the average genotyping accuracy and the red segments represent the standard deviations, at each sequencing depth.

Similarly, SVJedi results were evaluated in terms of varying sequencing error rates. In this case, both genotyping accuracy and genotyping rate were not impacted by a lower or higher sequencing error rate as long as it stays realistic (see Supplementary Figure 2).

The breakpoint coordinates of SVs detected by SV discovery methods are not always defined at the base pair resolution. To assess to what extent this potential imprecision can impact the genotyping accuracy of SVJedi, we performed experiments with altered breakpoint positions in the input variant VCF file. All breakpoint positions have been randomly shifted according to a Normal distribution centered on the exact breakpoint position with several standard deviations (σ) values ranging from 10 to 100 bp. We show that the genotyping accuracy with σ equals 50 bp does not fall below 94 % (see Supplementary Figure 3), indicating that SVJedi is not much impacted by the exact definition of the positions of the reference breakpoints.

4.2 Application of SVJedi to a real human dataset

To get closer to the reality of biological data, we applied our tool to a real human dataset, the HG002 individual, son of the so-called *Ashkenazi trio* dataset, which has been highly sequenced and analyzed in various benchmarks and especially by the GiaB Consortium (Zook *et al.*, 2016). The latter, precisely, provides a set of high confidence SV calls together with their genotype in the individual HG002. SV discovery and genotyping were based on several sequencing technologies, SV callers and careful call set merging. Their work estimated genotypes for 5,464 deletions and 7,281 insertions, which can then be considered as the ground truth. It should be noted that we can focus only on heterozygous (0/1) and homozygous for the alternative allele (1/1) genotypes. Indeed, the SV call set was obtained from SV discovery methods, which can only detect variations between the individual and the reference genome. SVJedi was applied on a 30x PacBio long read dataset from individual HG002, to assess the genotypes of both deletions and insertions of this high confidence set.

4.2.1 SVJedi results on the HGOO2 individual

We observe a good overlap of 92.2 % between the estimated genotypes of SVJedi and the GiaB call set. More precisely, among the assigned genotypes, there are 91.7 % of deletions and 92.5 % of insertions that are genotyped by SVJedi identically as the GiaB call set (the detailed contingency tables are provided in Supplementary Table 2).

Among the SVs differently genotyped between SVJedi and GiaB, a large part is represented by small variants (57 % are smaller than 100 bp). The genomic context of the SVs seems also to impact the genotyping accuracy: 75 % of the differently genotyped variants are located in Tandem Repeats greater than 100 bp (TRs), compared to 51 % for the whole SV set. Both features, size and location in TR, have similar impacts on the genotyping accuracy, with a difference of 9 and 11 % for small-vs-large and TR-vs-nonTR located SVs, respectively. Combining the two factors leads to a larger difference, with the small SVs that are located in TRs having the lowest genotyping accuracy of 81.3 % compared to near perfect accuracy of 97.9 % for the larger ones outside TRs (see the cross table in Supplementary Table 3).

Compared to previous results on simulated data, SVJedi shows a lower genotyping rate on this real dataset, for both deletions and insertions (85.8 % and 93.6 %, respectively). As in the case of accuracy, we notice that the great majority of the non-genotyped variants are either small or located in TRs: 87 % are of size less than 100 bp and 84 % are located in TRs. The factor impacting most the genotyping rate is the SV size (genotyping rates of 74.6 % and 98.1 % for small and large SVs respectively, see Supplementary Table 3). The presence of a TR at the breakpoint of small SVs worsens the genotyping task, with only 68.2 % of such SVs that could be genotyped. Notably, the GiaB deletion set contains more in proportion of such small SVs (39 vs 29 % for deletions and insertions respectively), explaining the observed difference in genotyping rate between the two SV types. Interestingly, these kinds of variants seem to be more impacted by the heterogeneity of PacBio sequencing depth since when using the full 63x dataset, the overall genotyping rate increases to 96.6 %.

4.2.2 Mendelian inheritance analysis

Since sequencing data are available for the parents of the studied individual (HG003 for the father and HG004 for the mother), we can check, as an alternative validation approach, if the predicted genotypes for the son are consistent with his parent genotypes, assuming perfect Mendelian inheritance and a very low de novo mutation rate. To do so, from the same set of deletions and insertions, which is the GiaB call set, SVJedi was applied to three PacBio sequence datasets, one per individual, with a sequencing depth of about 30x for each dataset. Overall, the Mendelian inheritance consistency of SVJedi on this trio dataset is high, with 96.9 % of the son genotypes that are consistent with his parent genotypes. As expected, most inconsistent genotypes concern SVs that were genotyped differently between SVJedi and GiaB (48.7 %, n = 154), confirming for those that they are probably wrongly assessed by SVJedi. However, these confirmed errors represent only 1.2 % of the dataset.

4.2.3 SVJedi results on ONT data

SVJedi was applied on the same SV call set and for the same HG002 individual, but with sequencing data obtained by a different long read technology, namely Oxford Nanopore. With a 44x PromethION dataset, SVJedi shows very similar genotyping performances as with the PacBio dataset (90.7 % accuracy and 86.2 % rate, see Supplementary Table 4), highlighting its versatility with respect to long read sequencing technologies.

4.3 Comparison with other approaches

4.3.1 Comparison with other genotyping tools

SVJedi was compared on the PacBio HG002 dataset to two other tools that can genotype a set of SVs with long read sequencing data, Sniffles (Sedlazeck *et al.*, 2018) and svviz2 (Spies *et al.*, 2015). Both tools are not purely dedicated to the genotyping problem. Sniffles is a SV discovery tool that has an option (`-Ivcf`) enabling a genotyping mode instead of a discovery mode, we will thereafter refer to this tool usage as Sniffles`-Ivcf`. svviz2 is a visualisation tool enabling to visualize how reads align to the reference and alternative alleles of a given SV, as a byproduct it can estimate a genotype based on the aligned read counts.

As shown in Table 2, both Sniffles`-Ivcf` and svviz2 have genotyping rates close to 100 % but at the expense of lower genotyping accuracies (detailed results for all genotypes are given in Supplementary Table 5). Sniffles`-Ivcf` is 10 % less accurate than SVJedi (82.0 % vs 92.2 %). svviz2 obtained the lowest genotyping accuracy (65.9 %, with a 10 % difference between deletions and insertions).

A stratified analysis of the genotyping performances of all three tools with respect to the SV size is presented in Fig. 4. We can observe that SVJedi has a better accuracy than Sniffles`-Ivcf` for all SV size classes. As mentioned previously, the lowest accuracy of SVJedi is observed for small SVs (<100 bp), but, apart from this size class, its accuracy is quite robust with respect to the size of SVs. On the opposite, svviz2 obtained its best genotyping accuracy for the smallest SVs (<100 bp) and it rapidly drops for SVs larger than 250 bp, falling below 30 % for SV sizes between 1 kb and 10 kb. When comparing between SV types, svviz2 genotyping accuracy is significantly lower for insertions than deletions, with, in particular, less than 10 % of the insertions larger than 1 kb that are correctly genotyped (see Supplementary Figure 4). This inability for genotyping large SVs can probably be explained by the way svviz2 identifies informative reads for a given SV: only the reads mapped initially to the reference genome are selected before re-aligning them against both the reference and alternative alleles. Consequently, most reads coming from large insertion alternative alleles could probably not be used for estimating these genotypes. To a lesser extend, Sniffles`-Ivcf` genotyping accuracy is also lower for large insertions than large deletions (69.6 % for insertions vs 85.7 % for deletions, $\geq$ 1 kb), whereas SVJedi genotyping accuracy is unaffected by SV type for all size classes.

We then compared the genotyping performances with respect to the genomic context of the SVs (Fig. 5). As shown previously, SVs falling in a Tandem Repeat greater than 100 bp are harder to genotype with SVJedi, with a 9 % decrease of accuracy for these SVs, compared to those outside TRs. A larger decrease of genotyping accuracy (14 %) is observed with Sniffles`-Ivcf` in these regions. Although less frequent (here, only 48 concerned SVs), large segmental duplications, typically larger than 10 kb, are also likely to affect long read mapping accuracy and thus genotyping accuracy. SVJedi accuracy seemed not to be affected by these duplications, contrary to Sniffles`-Ivcf` (Fig. 5).

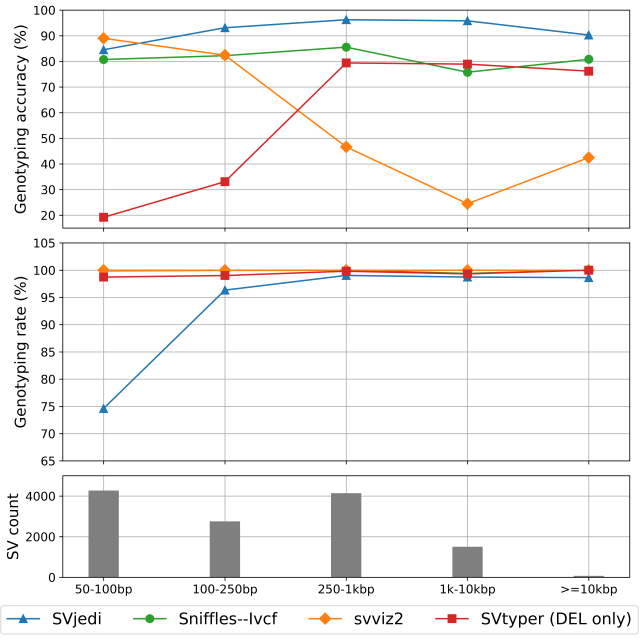

Fig. 4. Results of genotyping tools for the 12,745 deletions and insertions from the GiaB call set in the HG002 individual according to different SV size classes: 50 to 100 bp, 100 to 250 bp, 250 bp to 1 kb, 1 to 10 kb and $\geq$ to 10 kb. The two figures on top represent the genotyping accuracies and the genotyping rates of SVJedi, Sniffles`-Ivcf` and svviz2 on a 30x PacBio dataset, and of SVtyper for deletions only on a 30x Illumina dataset. The bottom figure represents the SV count of each SV size class.

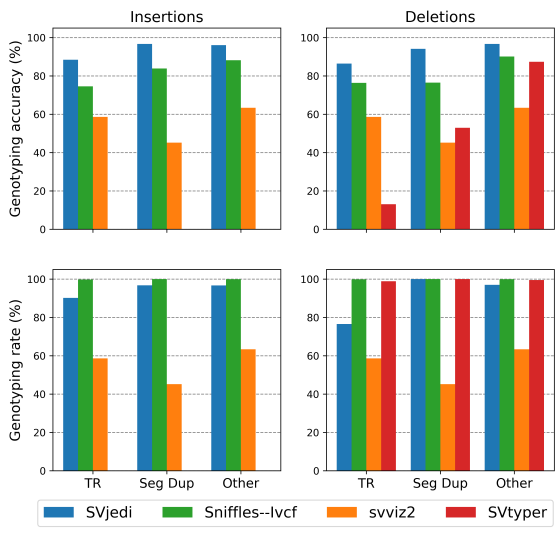

Fig. 5. Stratified analysis of genotyping accuracy and rate for several genotyping tools with respect to the genomic context of the SVs. Three categories of genomic context are considered: tandem repeats greater than 100 bp (TR), segmental duplications larger than 10 kb (SeqDup) and all other regions (Other).

4.3.2 Comparison with a short read based genotyping approach

For this same individual (HG002), some short read datasets are also available. We, therefore, can compare SV genotyping performances between two approaches and data types, namely long versus short reads. SVJedi predictions were compared to a SV genotyping tool for short reads, SVtyper, known as a reference tool in the state of the art (Chiang *et al.*, 2015; Chander *et al.*, 2019). Since SVtyper does not support insertion variants, we focus here only on deletions, and the 5,464 deletions from the GiaB call set were genotyped with SVtyper using a 2x250 bp 30x Illumina read dataset of HG002.

Table 2 shows that more than half of the deletions are genotyped differently by SVtyper than in the high confidence GiaB call set, resulting in a genotyping accuracy of only 46.5 %, while this percentage rises to 91.7 % for SVJedi with long reads. Remarkably, many of the discrepancies of SVtyper with

	Deletions		Insertions		
Tool	Genotyping accuracy	Genotyping rate	Genotyping accuracy	Genotyping rate	Time
SVJedi	**91.7**	85.8	**92.5**	93.6	**2h25m**
Sniffles-`Ivcf`	82.5	**99.9**	81.7	**99.8**	17h16m
svviz2	72.5	**100**	61.0	**100**	5days*
SVtyper (Illumina dataset)	46.5	99.2	-	-	5h32m
Sniffles (discovery mode)	48.7	52.4	39.8	44.8	18h04m
pbsv	90.1	72.7	68.8	59.8	5h29m

Table 2. Comparison of several tools and approaches for genotyping the 12,745 deletions and insertions of the GiAB call set in the HG002 individual. Three approaches are compared: using long read genotyping tools (first three tools), using a short read genotyping tool (SVTyper), and using long read discovery tools (last two tools). Except for the short read genotyping tool (SVtyper) that uses a 30X Illumina sequencing dataset, all other tools were run with a 30X PacBio long read dataset. Runtimes were measured on a 40-CPU computing node. * svviz2 is not parallelized.

GiaB are totally contradictory with 0/0 genotypes instead of 1/1 ones (see Supplementary Table 5). We can clearly see, in Fig. 5, that short read based genotyping is much more impacted by the presence of TRs at the breakpoint. As expected, mapping reads in these regions is much more challenging for short than long reads. This demonstrates the higher benefit of using long reads and a dedicated genotyping tool such as SVJedi rather than short reads.

4.3.3 Comparison with SV discovery approaches

One can wonder if these SVs could be easily detected and genotyped by long read SV discovery tools. We applied here two such tools, among the bests to date, Sniffles and the Pacific Biosciences SV caller, pbsv (Sedlazeck *et al.*, 2018; De Coster *et al.*, 2019). As a result, both tools obtained the lowest genotyping rates over all genotyping approaches: among the 12,745 SVs, only 6,127 were discovered by Sniffles, and 8,326 by pbsv (genotyping rates of 48.1 % and 65.3 % respectively, see Table 2 and details in Supplementary Table 5). As expected, most of the missed SVs have an heterozygous genotype in the GiaB call set. More surprisingly, for the discovered SVs, their genotyping accuracy is overall smaller than with other approaches, with 43.9 % and 78.9 % for Sniffles and pbsv respectively. In particular, Sniffles misassigns 85 % of the discovered SVs with a 1/1 genotype in GiaB as heterozygous. pbsv shows the same type of errors but mainly for insertions, resulting in an important difference of genotyping accuracy between deletions and insertions (90.1 vs 68.8 %). These results highlight the fact that SV discovery tools, are much less precise for the genotyping task than a dedicated genotyping tool.

4.3.4 Runtime comparison

Importantly, SVJedi does not come with a high computational cost. On a 40-CPU computing node, genotyping the 12,745 SVs with the 30x PacBio HG002 dataset took only 2h25m. The alignment step is actually the most time-consuming step and took 2h15m. Compared to other tools, SVJedi was the fastest among the tested ones (Table 2). Among the long read genotypers, SVJedi was 7 times faster than Sniffles-`Ivcf` and 50 times faster than svviz2. The large runtime of svviz2 (more than 5 days) can be explained by the fact that it is not natively parallelized, when manually parallelized on 20 CPU (only 20 due to memory limits), it took roughly 11h.

5 Discussion and conclusion

In conclusion, we provide a novel SV genotyping approach for long read data, that showed good results on simulated and real datasets. The approach is implemented in the SVJedi software for most SV types (insertions, deletions, inversions and translocations). The robustness of our tool, SVJedi, was highlighted in this work, for several sequencing depths and error rates, but also related to the precision of the breakpoint positions. On a real human dataset with more than 12,000 insertions and deletions, SVJedi obtained a better genotyping accuracy than other tested genotyping and discovery tools. SVJedi, like the other tools, had more difficulties to accurately genotype small variants (<100 bp) and those located in large tandem repeat regions. However, SVJedi showed a more conservative behavior than other tools, with a lower genotyping rate for these most difficult SVs: instead of estimating an incorrect genotype, it favored not assigning any genotype at all.

This work also demonstrated that this is crucial to develop dedicated SV genotyping methods, as well as SV discovery methods. Firstly, because this is the only way to get evidence for the absence of SVs in a given individual. Secondly, and more surprisingly, because SV discovery tools are not as efficient and precise to genotype variants once they have been discovered, at least with long read data as was shown here. Indeed, without a priori SV discovery is a much harder task than genotyping. Because the alternative allele is not known in discovery, discovery methods rely on fewer or noisier signals to identify the SVs than genotyping methods. Consequently, both approaches would likely benefit from different optimal parameter settings, with for instance discovery methods requiring a more stringent set of parameters to limit the false discovery rate. However, in this paper, we have no intention to oppose both approaches but rather argue that they are complementary and are intended to different purposes: when the aim is strictly to genotype or compare individuals on a set of already known variants, we have shown that using as much as possible the known features of variants is much more efficient.

Also, on real human data, we were able to quantify the impact of the sequencing technology on SV genotyping. Although this was expected that long read data would perform better than short read ones, the observed difference is considerable with a twofold increase of the genotyping accuracy with long reads. This is in particular due to the very poor performances obtained with short reads, that are ill-adapted to deal with the complex and repeat-rich regions often present at SV junctions. On the opposite, this work shows that the long-distance information contained in long reads can be efficiently used to discriminate between breakpoints, despite relatively high sequencing error rates and variability in sequencing coverage. This result underlines the relevance of such a method dedicated to genotyping from long read data.

Although long read sequencing technology remains to date more expensive than short read ones, to be used for instance in routine in the clinical setting (Merker *et al.*, 2018), we can hope that this situation will improve in the next few years. The high genotyping accuracy and low computational requirements of SVJedi make it ready for such happening and to be integrated into routine pipelines to screen for instance disease-related SVs and therefore improve medical diagnosis or disease understanding.

Acknowledgements

We are thankful to the Genouest bioinformatics platform, computations have been made possible thanks to the resources of the Genouest infrastructure.

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

References

Alkan, C., Coe, B. P., *et al.* (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**(5), 363.

Antaki, D., Brandler, W. M., *et al.* (2017). SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics*, **34**(10), 1774–1777.

Audano, P. A., Sulovari, A., *et al.* (2019). Characterizing the major structural variant alleles of the human genome. *Cell*, **176**(3), 663–675.

Bailey, J. A., Gu, Z., *et al.* (2002). Recent segmental duplications in the human genome. *Science*, **297**(5583), 1003–1007.

Chander, V., Gibbs, R. A., *et al.* (2019). Evaluation of computational genotyping of structural variation for clinical diagnoses. *GigaScience*, **8**(9).

Chiang, C., Layer, R. M., *et al.* (2015). SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods*, **12**(10), 966–968.

De Coster, W., De Rijk, P., *et al.* (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Res.*, **29**(7), 1178–1187.

Heller, D. and Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics*, **35**(17), 2907–2915.

Huddleston, J., Chaisson, M. J., *et al.* (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.*, **27**(5), 677–685.

Jain, M., Koren, S., *et al.* (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.*, **36**(4), 338.

Kidd, J. M., Graves, T., *et al.* (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, **143**(5), 837–847.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**(21), 2987–2993.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.

Lupski, J. R. (2015). Structural variation mutagenesis of the human genome: Impact on disease and evolution. *Environ. Mol. Mutagen.*, **56**(5), 419–436.

Merker, J. D., Wenger, A. M., *et al.* (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.*, **20**(1), 159.

Nielsen, R., Paul, J. S., *et al.* (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**(6), 443.

Norris, A. L., Workman, R. E., *et al.* (2016). Nanopore sequencing detects structural variants in cancer. *Cancer Biol. Ther.*, **17**(3), 246–253.

Phan, L., Hsu, J., *et al.* (2017). dbVar structural variant cluster set for data analysis and variant comparison. *F1000Research*, **5**.

Sedlazeck, F. J., Rescheneder, P., *et al.* (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, **15**(6), 461–468.

Spies, N., Zook, J. M., *et al.* (2015). svviz: a read viewer for validating structural variants. *Bioinformatics*, **31**(24), 3994–3996.

Stancu, M. C., Van Roosmalen, M. J., *et al.* (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.*, **8**(1), 1326.

Stöcker, B. K., Köster, J., *et al.* (2016). SimLoRD: Simulation of Long Read Data. *Bioinformatics*, **32**(17), 2704–2706.

Zook, J. M., Catoe, D., *et al.* (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*, **3**, 160025.

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture