SVJedi: Genotyping structural variations with long reads
Supplementary material

Lolita Lecompte, Pierre Peterlongo, Dominique Lavenier and Claire Lemaitre

Contents

1 Data accessibility and generation

The gold standard call set for individual HG002, provided by Genome in a Bottle (GiaB) Consortium, is available at the following link:

- `ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/HG002_SVs_Tier1_v0.6.vcf.gz`

In our study, we focus only on variants of the Tier 1 call set which are isolated and sequence-resolved SVs, corresponding to the `PASS` filter of this VCF file. This call set represents 5,464 deletions and 7,281 insertions.

The Pacific Biosciences (PacBio) sequence datasets for Ashkenazi trio individuals provided by GiaB are available at the following links:

- ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/PacBio_MtSinai_NIST/
- ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149_father/PacBio_MtSinai_NIST/
- ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143_mother/PacBio_MtSinai_NIST/

The PacBio data of the individual HG002 has a sequencing depth of 63x. The sequence data were sub-sampled using SAMtools Li *et al.* (2009) to a depth of 30x.

The ONT Promethion dataset for HG002 is available at the following link:

- ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/UCSC_Ultralong_OxfordNanopore_Promethion/

The Illumina dataset for HG002 is available at the following link:

- ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/

As with the PacBio data sample of the individual HG002, the Illumina data sample was sub-sampled to a sequencing depth of 30x as well.

Simulated PacBio reads with SimLoRD Long reads are simulated on two distinct haplotypes where the deletions from dbVar are added according to the simulated genotype (0/0, 0/1, 1/1). The SimLoRD commands are shown below for a dataset with a sequencing depth of 30x and a sequencing error rate of 16 %. The same distribution of errors between insertions, deletions, and substitutions, as LRCstats La *et al.* (2017) has been defined.

```
simlord --read-reference haploA_with_1000_simulated_DEL.fasta --coverage 15
    --max-passes 1 -pi 0.11 -pd 0.04 -ps 0.01 simulated_reads_haploA

simlord --read-reference haploB_with_1000_simulated_DEL.fasta --coverage 15
    --max-passes 1 -pi 0.11 -pd 0.04 -ps 0.01 simulated_reads_haploB
```

2 Tools command lines

SVJedi:

```
svjedi.py -t 40 -r GRCh37.p13.fasta \
    -v HG002_SVs_Tier1_v0.6.PASS.vcf \
    -i 30x_pacbio.fastq.gz
```

Sniffles:

```
ngmlr -t 40 -r GRCh37.p13.fasta -q 30x_pacbio.fastq.gz -o HG002_PB30x.sam
samtools view -Shb -@ 40 HG002_PB30x.sam | \
samtools sort -@ 40 -o HG002_PB30x_ngmlr.sorted.bam -

#Sniffles genotyping
sniffles -t 40 -m HG002_PB30x_ngmlr.sorted.bam \
    -v HG002_PB30x_sniffles_gtcall.vcf \
    --Ivcf HG002_SVs_Tier1_v0.6.PASS.vcf

#Sniffles discovery
sniffles -t 40 -m HG002_PB30x_ngmlr.sorted.bam \
    -v HG002_PB30x_sniffles_svcall.vcf --genotype
```

svviz2:

```
minimap2 --MD -ax map-pb -t 40 \
    -R "@RG\\tID:{ID}\\tSM:{SM}" \
    GRCh37.p13.fasta 30x_pacbio.fastq.gz | \
    samtools sort -@ 40 -o HG002_PB30x_minimap2.sorted.bam

samtools index HG002_PB30x_minimap2.sorted.bam
bwa index GRCh37.p13.fasta

svviz2 --ref GRCh37.p13.fasta --variants HG002_SVs_Tier1_v0.6.PASS.vcf \
    -o outDIR/ HG002_PB30x_minimap2.sorted.bam
```

pbsv:

```
pbmm2 align GRCh37.p13.fasta 30x_pacbio.fastq.gz HG002_PB30x_pbmm2.bam \
    --sort --median-filter --sample sample1

pbsv discover HG002_PB30x_pbmm2.bam HG002_PB30x.svsig.gz
pbsv call GRCh37.p13.fasta HG002_PB30x.svsig.gz HG002_PB30x_pbsv.vcf
```

SVTyper:

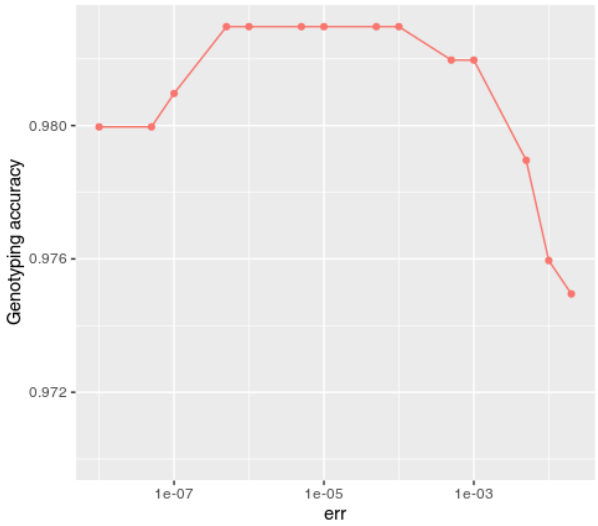
```
speedseq align -t 40 -M 20 \
    -R "@RG\tID:HG002\tSM:Illumina\tLB:2x250" \
    -o HG002_Illumina30x_speedseq \
    GRCh37_latest_genomic.fna reads.end1.fq reads.end2.fq

python -m svtyper.classic -i svtype_format_HG002_SVs_Tier1_v0.6.PASS.vcf \
    -B HG002_Illumina30x_speedseq.bam > HG002_svtyper.vcf
```

Note: Both `CIPOS` and `CIEND` were set to 0,0 for all variants in the VCF before running SVtyper.

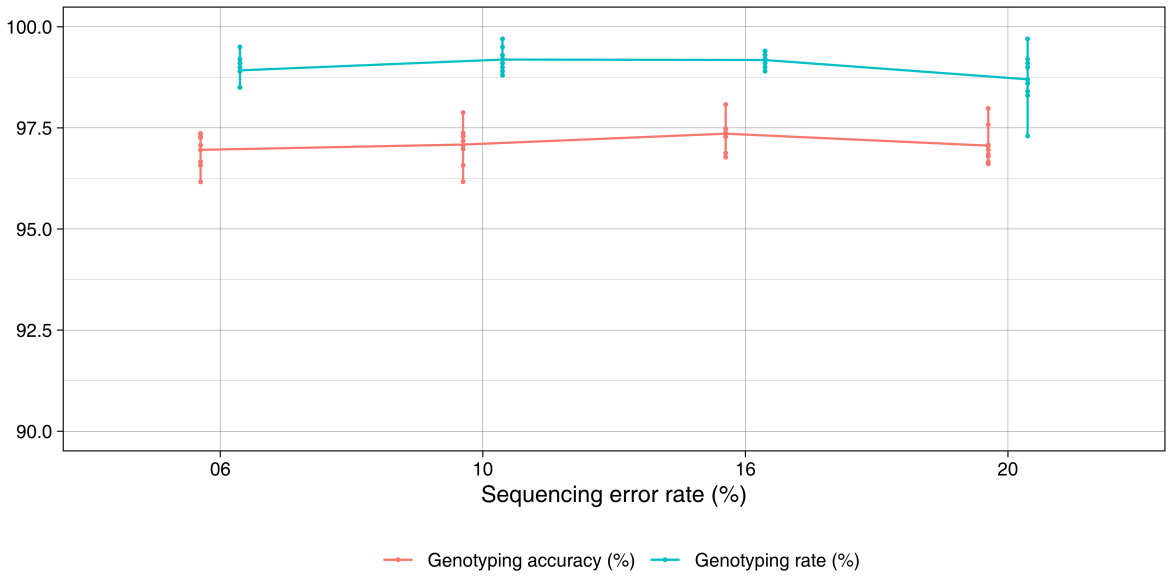
3 Assessment of SVJedi performances and robustness with simulated datasets

3.1 Choosing the *err* value



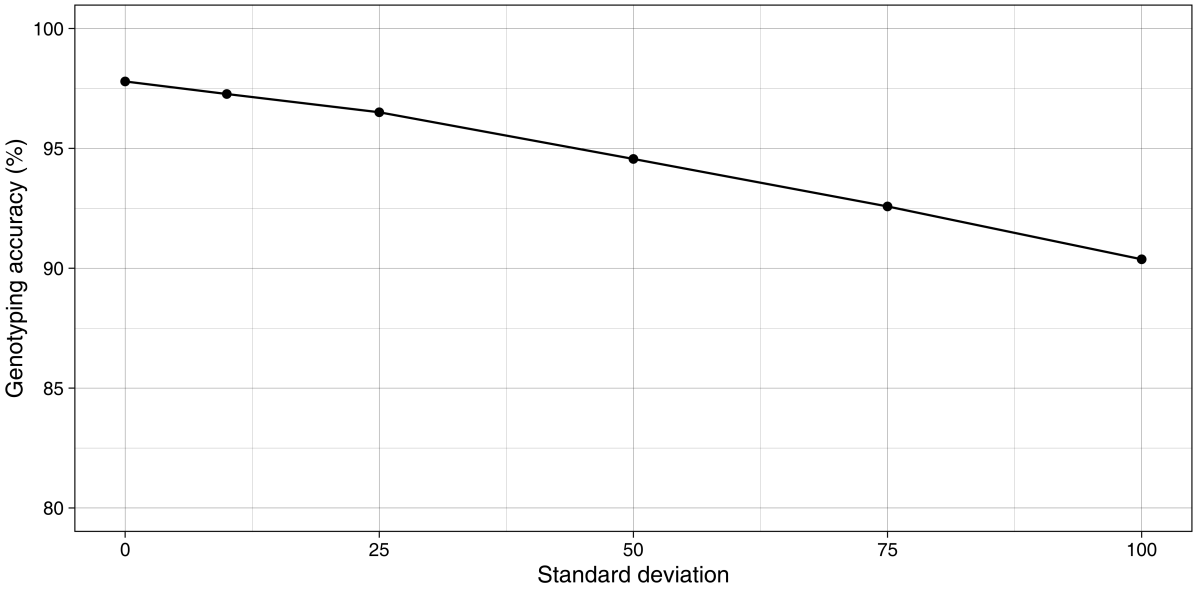
Supplementary Figure 1: Genotyping accuracy according to the *err* parameter on a simulated dataset, with 1,000 existing (in dbVar) deletions on human chromosome 1. *err* is the probability that a read maps to a given allele erroneously, it is used to compute each genotype likelihood. The default value for *err* was fixed to 5.10^{-5}.

3.2 SVJedi results on simulated PacBio datasets for several error rates



Supplementary Figure 2: SVJedi results on simulated 30x datasets at several sequencing error rates : 6 %, 10 %, 16 %, 20 %.

3.3 SVJedi results on a simulated PacBio dataset with shifted breakpoints



Supplementary Figure 3: SVJedi genotyping accuracy on a simulated 30x dataset with respect to the precision of breakpoint positions in the input deletion file. All breakpoint positions have been randomly shifted according to a Normal distribution centered on the exact breakpoint position with several standard deviations values ranging from 10 to 100 bp.

3.4 SVJedi results for inversions and translocations

Inversions

		SVJedi estimations			
		0/0	0/1	1/1	./.
Truth	0/0	150	0	0	0
	0/1	0	150	0	0
	1/1	0	10	140	0

Genotyping accuracy: 97.8 %

Genotyping rate: 100 %

Translocations

		SVJedi estimations			
		0/0	0/1	1/1	./.
Truth	0/0	150	0	0	0
	0/1	0	150	0	0
	1/1	1	7	142	0

Genotyping accuracy: 98.2 %

Genotyping rate: 100 %

Supplementary Table 1: Contingency tables of SVJedi genotyping results on two 30x PacBio simulated dataset: one dataset with 450 inversions (left) and one dataset with 450 translocations (right). Inversions were randomly simulated on human chromosome 1, by sampling the first breakpoint location in a uniform distribution and choosing the inversion size uniformly between 50 bp and 15 kb. Translocations were successively simulated between human chromosome 1 and chromosome 2, by sampling uniformly breakpoint locations on both chromosomes. Only variants simulated outside the gap regions according to UCSC, and at a distance of at least 10 kb from one another, were kept. Grey labelled boxes correspond to identical predictions between the two methods. The number of genotypes that SVJedi fails to assess is indicated by the "./." column.

4 SVJedi results on the real HG002 long read data

4.1 Detailed SVJedi results on the PacBio 30x dataset

DELETIONS

		SVJedi predictions			
		0/0	0/1	1/1	./.
GiaB	0/1	227	2,773	38	395
	1/1	2	124	1,522	383

INSERTIONS

		SVJedi predictions			
		0/0	0/1	1/1	./.
GiaB	0/1	18	2,870	290	327
	1/1	1	199	3,436	140

Supplementary Table 2: Contingency tables of SVJedi genotyping results on the real 30x PacBio dataset of human individual HG002 with respect to the high confidence GiaB call set. Results for the 5,464 deletions (left) and 7,281 insertions (right) are indicated in two separated tables, where columns indicate SVJedi genotypes and rows GiaB ones. Grey labelled boxes correspond to identical predictions between the two methods. The number of genotypes that SVJedi fails to assess is indicated by the "./." column.

4.2 Detailed analysis of SVJedi results with respect to SV genomic context and SV size

SV count

	len(SV) < 100bp	len(SV) > 100bp	Total
Inside TR	2,873	3,596	6,469
Outside TR	1,399	4,877	6,276
Total	4,272	8,473	12,745

Genotyping accuracy (%)

	len(SV) < 100bp	len(SV) > 100bp	Total
Inside TR	81.3	91.2	87.6
Outside TR	89.7	97.9	96.3
Total	84.5	95.1	92.2

Genotyping rate (%)

	len(SV) < 100bp	len(SV) > 100bp	Total
Inside TR	68.2	96.4	83.9
Outside TR	87.8	99.4	96.8
Total	74.6	98.1	90.2

Supplementary Table 3: Comparison of SVJedi results on the HG002 real PacBio dataset between several classes of SVs. SVs are classified according to two binary variables: the SV size is larger or not than 100 bp and the SV is located inside a Tandem Repeat of size greater than 100 bp (TR) or not. The top, middle and bottom tables indicate respectively the SV count, the average genotyping accuracy and the average genotyping rate in each class. Both features, SV size and location in TR, are not independently distributed among the SVs, as confirmed by a Chi-squared test applied on the top table (SV counts) (Chi-squared statistics of 698.5, p-value $< 2.10^{-16}$).

4.3 SVJedi results on the ONT dataset

DELETIONS

		SVJedi predictions			
		0/0	0/1	1/1	./.
GiaB	0/1	401	2,483	57	492
	1/1	1	42	1,431	557

Genotyping accuracy: 88.7 %

Genotyping rate: 80.8 %

INSERTIONS

		SVJedi predictions			
		0/0	0/1	1/1	./.
GiaB	0/1	43	2,704	306	452
	1/1	4	167	3,351	254

Genotyping accuracy: 92.1 %

Genotyping rate: 90.3 %

Supplementary Table 4: Contingency tables of SVJedi genotyping results on the real 44x Oxford Nanopore (PromethION) dataset of human individual HG002 with respect to the high confidence GiaB call set. Results for the 5,464 deletions (left) and 7,281 insertions (right) are indicated in two separated tables, where columns indicate SVJedi genotypes and rows GiaB ones. Grey labelled boxes correspond to identical predictions between the two methods. The number of genotypes that SVJedi fails to assess is indicated by the "./." column.

5 Comparisons with other approaches on the HG002 GiaB call set

5.1 Detailed genotyping results

long read SV genotyping tool predictions

		SVJedi				Sniffles –Ivcf				svviz2			
		0/0	0/1	1/1	./.	0/0	0/1	1/1	./.	0/0	0/1	1/1	./.
GiaB	0/1	245	5,643	328	722	404	5,447	1,078	9	301	5,947	690	0
	1/1	3	323	4,958	523	47	757	4,996	7	1,446	1,903	2,458	0

short read SV genotyping tool predictions

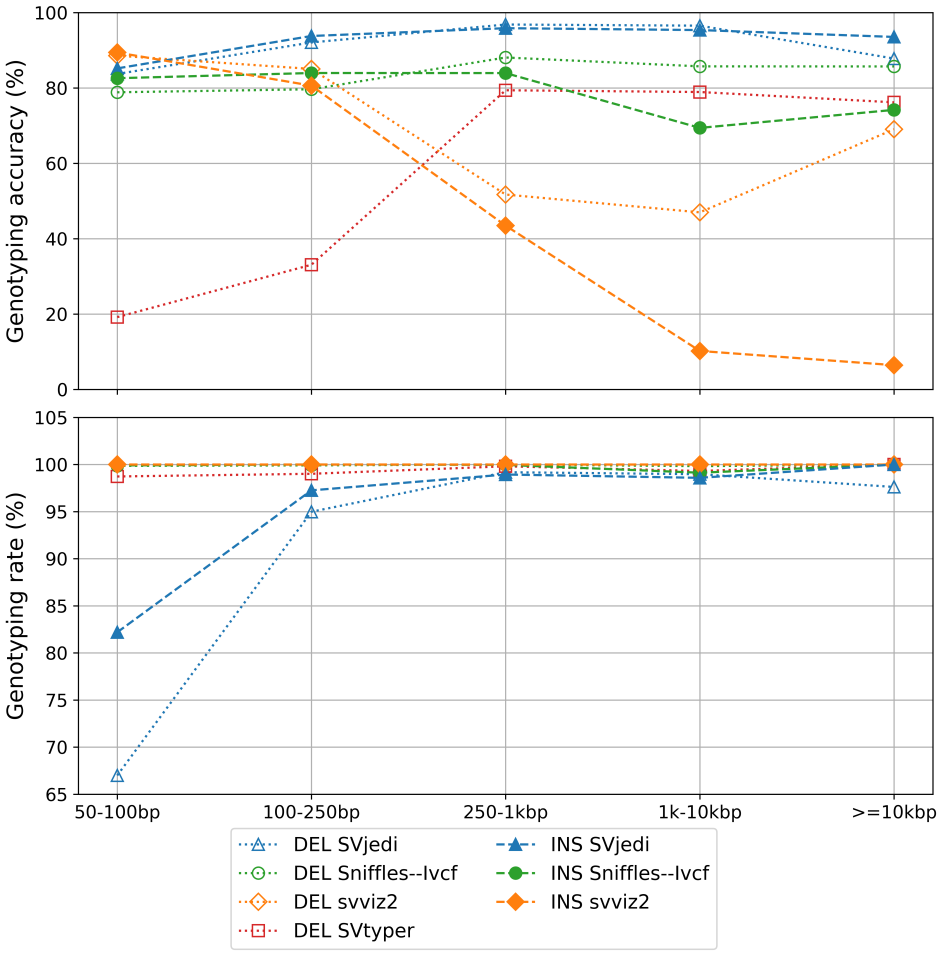
		SVtyper (del only)			
		0/0	0/1	1/1	./.
GiaB	0/1	1,852	1,561	12	8
	1/1	800	233	961	37

long read SV discovery tool predictions

		Sniffles (discovery)				pbsv			
		0/0	0/1	1/1	./.	0/0	0/1	1/1	./.
GiaB	0/1	349	2,189	146	4,254	0	4,110	149	2,679
	1/1	5	2,936	502	2,364	0	1,606	2,461	1,740

Supplementary Table 5: 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 (top), using a short read genotyping tool (middle), and using long read discovery tools (bottom). 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. For each tool, a contingency table of the estimated genotypes with respect to the high confidence GiaB call set is given, where columns indicate the tool estimated genotypes and rows GiaB ones. Grey labeled boxes correspond to identical predictions between the tool and the call set. The number of genotypes that each tool fails to assess is indicated by the "./." column.

5.2 Genotyping results with respect to SV size



Supplementary Figure 4: Results of genotyping tools for the 5,464 deletions (dotted lines) and the 7,182 insertions (dashed lines) of 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.

References

La, S., Haghshenas, E., *et al.* (2017). LRCstats, a tool for evaluating long reads correction methods. *Bioinformatics*, **33**(22), 3652–3654.

Li, H., Handsaker, B., *et al.* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.