Genotyping Structural Variations using Long Read data

Lolita Lecompte, Pierre Peterlongo, Dominique Lavenier, Claire Lemaitre

Genotyping Structural Variations using Long Read data

Lolita LECOMPTE[1], Pierre PETERLONGO[1], Dominique LAVENIER[1] and Claire LEMAITRE[1]
Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes, France

Corresponding author: lolita.lecompte@inria.fr

Abstract *Studies on structural variants (SV) are expanding rapidly. As a result, and thanks to third generation sequencing technologies, more and more SVs are discovered, especially in the human genome. At the same time, for several applications such as clinical diagnoses, it becomes important to genotype newly sequenced individuals on well defined and characterized SVs. Whereas many SV genotypers have been developed for short read data, there still have no approaches to assess whether some SVs are present or not in a new sequenced sample of long reads, from third generation sequencing technologies, such as Pacific Biosciences or Nanopore.*

In this work, we present a method to genotype known SVs from long read sequencing. The principle is based on the generation of a set of reference sequences that represent the two alleles of each structural variant. Alignments are built from mapping the long reads to these reference sequences. They are then analyzed and filtered out to keep only informative ones, in order to quantify and estimate the presence of each allele. Currently, the genotyping of large deletions have been investigated. Tests on simulated long reads based on 1000 deletions from dbVAR show a precision of 95.8%. We also applied the method to the whole NA12878 human genome.

Keywords Structural Variations, Genotyping, Long Reads

1 Introduction

Structural variations (SV) are characterized as genomic segments of a least 50 base pairs (bp) long, that are rearranged in the genome. There are several types of SV such as deletions, insertions, duplications, inversions, translocations. With the advent of Next Generation Sequencing (NGS) and the re-sequencing of many individuals, SVs have been shown to be a key component of polymorphism [1]. This kind of polymorphism are involved in many biological processes such as diseases or evolution [2]. Databases referencing such variants grow as new variants are discovered. At this time, dbVar, the reference database of human genomic SVs [3], contains 35,428,724 variant calls, illustrating that many SVs have already been discovered and characterized in the human population.

When studying the SVs of 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 resequenced individuals with respect 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 evidences are only investigated for present variants (ie. 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 probably more precise. With the fine characterization of a growing number of SVs in the human populations, genotyping newly sequenced individuals becomes very interesting and informative, in particular in medical diagnosis contexts.

In this work, we focus on this second problem: genotyping already known SVs in a newly sequenced sample. Such genotyping methods already exist for short reads data: for instance, SVtyper [4], SV^2 [5], Nebula [6], Malva [7]. Though short reads are often used to discover and genotype SVs, this is

well known that their short size make them ill-adapted for predicting large SVs or SVs located in repeated regions. As a matter of fact, SVs are often located alongside repeated sequences such as mobile elements, resulting in mappability issues that make the genotyping problem harder when using short read data.

Third generation sequencing technology, such as Pacific Biosciences (PB) and Oxford Nanopore Technologies (ONT), can produce long read data compared to NGS technologies. Long read sequences have enabled many applications, including new SVs discoveries. Despite their high error rate, long reads are crucial in the study of SVs. Indeed, the size range of this data can reach a few kilobases (kb) to megabases, thus long reads can extend over rearranged SV sequences as well as over repeated sequences often present at SV's breakpoint regions.

Following long reads technology's development, many SV discovery tools have emerged, such as Sniffles [8] and NanoSV [9]. Among these tools, some have a genotyping module that gives the frequency of alleles after calling SVs of the sequenced samples, nonetheless their required post-processing to evaluate if a set of SVs is present or not in the sample. To our knowledge there is currently no tool that can perform genotyping from a set of known SVs with long read data. Thus, there is a need to develop accurate and efficient methods to genotype SVs with such data, especially in the context of clinical diagnoses.

The main contribution of this work is a new method to genotype known SVs using long reads data. We also provide an implementation of this method in the tool named SVJedi. SVJedi was applied on simulated data of the human genome and on real data of the individual NA12878. High precision was achieved on both simulated and real data.

2 Materials and Methods

2.1 Methods

Pipeline We propose a method that aims at assigning a genotype for a set of already known SVs in a given individual sample sequenced with long read data. In other words, the method assesses if each SV is present in the given individual, and if so, how many variant alleles it holds, ie. whether the individual is heterozygous or homozygous for the particular variant. The method is described and implemented here for only one type of SV, the deletions, but the principle can be easily generalized to other types of SVs. The method takes as input a variant file with deletion coordinates, a reference genome and the sample long read sequences. It outputs a variant file complemented with the individual genotype information for each input variant.

The principle of the method is based first on generating reference sequences that represent the two alleles of each SV. Then the sample long reads are aligned on the whole set of reference alleles. An important step of our method consists in selecting and counting only informative alignments to finally estimate the genotype for each known variant. The main steps are illustrated in Fig. 1.

Generating references 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 alleles. Deletions are sequences 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. The reference allele (allele 0) is therefore the sequence of the deletion with adjacent sequences at each side, and the alternative allele (allele 1) consists in the joining of the two previous adjacent sequences. Given that reads of several kb will be mapped on these references, the size of the adjacent sequences was set to 5,000 bp at each side, giving a 10 kb sequence for allele 1 and 10 plus the deletion size kb for allele 0.

Mapping Sequenced long reads are aligned on all previously generated references. We use Minimap2 [10] (version 2.16-r922), with default parameters, as it is a fast and accurate mapper, specifically designed for long erroneous reads.

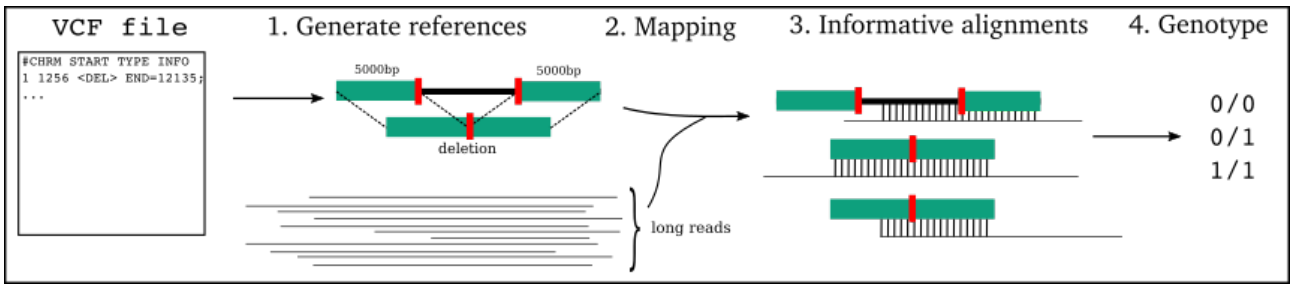

Fig. 1. SVJedi steps. 1. Two corresponding to sequences are generated for each selected SV, one correspond to the orginal sequence and the other the sequence with the deletion. 2. Long read sequenced data are aligned on these references using Minimap2. 3. Informative alignments are selected. 4. Genotypes are estimated.

Selecting informative alignments Minimap2 raw alignment results have to be carefully filtered out in order to remove i) uninformative alignments, that is 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 shorter one, contains one such breakpoint at the junction of the two adjacent sequences (see the red thickmarks 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 distances of the breakpoint to respectively the start and end coordinates of the alignment on the reference sequence (see Fig. 2), they must satisfy the following condition in eq 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 quality score. To focus on uniquely mapped reads, the quality 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 reference and the read sequences. To maximize the probability that the aligned read really originate from the reference locus, we therefore require that the two sequences are aligned in a semi-global manner, where each alignment extremity must correspond to an extremity of at least one of the two aligned sequences. This criteria gathers four types of situation, namely the read is included in the reference, or *vice-versa*, or the read left end aligns on the reference right end, or *vice-versa*. Indeed this criteria is not strictly applied and a distance of d_{end} of the alignment to an extremity is tolerated (d_{end} is set by default to 100 bp). More formally, if a and b (resp. c and d) are the distances of the alignment to the, respectively, left and right extremities of the reference sequence (resp. read sequence), then the alignment must fulfill the following condition in Eq. 2 to be kept:

$$(a < d_{end} \parallel c < d_{end}) \ \& \ (b < d_{end} \parallel d < d_{end}) \tag{2}$$

Estimating genotypes For each variant, the genotype is estimated based on the ratio of amounts of reads informatively aligned to each reference allele. Each variant has two references of different sizes, so even if both alleles are covered with the same read depth, there would be fewer reads that align on the shortest reference. To prevent a bias towards the larger allele, reported read counts for the larger alleles are normalized according to the reference sequence length ratio, assuming that read count is proportional to the sequence length.

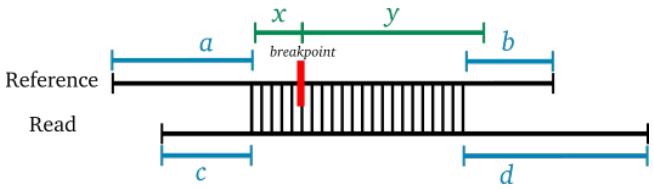

Fig. 2. Definition of the different distances of the alignment with respect to the breakpoint (x and y) and to the sequence extremities (a, b, c and d) used to select informative alignments.

Finally, a genotype is estimated if the variant presence or absence is supported by at least *min_cov* different reads after normalization (sum of read counts for each variant). The allele frequency is defined as the proportion of reads supporting the reference allele 0. A SV is called heterozygous, 0/1, if the allele frequency is within [0.2; 0.8]. Otherwise, if the frequency is > 0.8, the SV is called homozygous reference, 0/0, and if the frequency is < 0.2, the SV is called homozygous alternative, 1/1.

Implementation and availability We provide an implementation of this method named SVJedi. SVJedi is written in Python 3, it requires as input a set of deletion (vcf format), a genome (fasta format), a reads' file (fastq or fasta format). Also the genotyping program can be runned independently from the whole pipeline, then the user must provide a vcf file and a paf file. SVJedi is available at `https://data-access.cesgo.org/index.php/s/JhD0TNgJocVew0E`, under GNU Affero GPL licence.

2.2 Evaluation

Long read simulated datasets SVJedi was assessed on simulated datasets of the human genome GRCh37 based on real characterized deletions for the human genome. From the dbVar database [3], we selected 1000 existing deletions on the chromosome 1, which are separated by at least 10,000 bp. The size of the deletions varies from 50 bp to 10 kbp. In this experiment, deletions were distributed into the three different genotypes: 333 deletions are considered as 0/0 genotype, 334 deletions as 0/1 genotype and the 333 remaining deletions as 1/1 genotype. We consider the homozygous 1/1 genotype, as the genotype where the deletion is present in both alleles. So, deletions were simulated on two different reference sequences, corresponding to the two haplotypes of the human genome. 1/1 genotype deletions were simulated on both reference sequences, while deletions of 0/1 genotype were randomly simulated on one of the reference sequences. Then we simulate PB long reads using SimLoRD[11] (version v1.0.2) with 16% error rate (`-pi 0.11 -pd 0.04 -ps 0.01 --max-passes 1`), at 20x depth of coverage.

Real data SVJedi was assessed on a human genome real dataset. As sequenced reads for the individual NA12878, we used ONT MinION data rel 5, from the ONT whole genome sequencing nanopore consortium data [12] (European Nucleotide Archive : PRJEB23027). ONT data were called with Guppy 0.3 (`https://s3.amazonaws.com/nanopore-human-wgs/rel5-guppy-0.3.0-chunk10k.fastq.gz`). This sequencing dataset contains 15,891,898 reads, totaling 123 Gbp, which correspond to a 39x depth of coverage.

As the set of deletions to genotype, we use the call set of variants provided by the Genome in a Bottle consortium (GiAB), for the NA12878 individual from PB data [13], (`ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/NA12878.sorted.vcf.gz`). From this input set, only deletion sizes greater than 50 bp were used. In a second experiment we selected high confidence deletion calls from the GiAB call set.

Evaluation In order to evaluate our method, for simulated data, we compute contingency table, giving us a clear view of the number of correctly predicted genotypes as well as the number of incorrectly predicted genotypes, for each category. Also, we can assess the number of corrected predicted genotypes over all predicted genotypes, which gives the precision of the method, as shown in equation 3.

$$Precision = \frac{\text{number of correctly predicted genotypes}}{\text{number of predicted genotypes}} \tag{3}$$

3 Results

3.1 Simulated data

SVJedi was applied on PB simulated long reads for the human chromosome 1, with 1000 real characterized deletions found in dbVar, ranging from 50 to 10,000 bp. Deletions are equally distributed among the 3 different genotypes. On simulated data, SVJedi achieves 95.846% precision, it correctly predicts 942 over 987 predicted deletions among the 1,000 assessed deletions, while 13 are not estimated at all. Results are described in Tab. 1. This simulation was repeated 10 times, and gives similar results regarding the number of correctly predicted deletions and precision.

		Prediction			
		0/0	0/1	1/1	./.
Truth	0/0	330	3	0	0
	0/1	19	305	6	4
	1/1	3	10	311	9

Precision : 95.846%

Tab. 1. Contingency table on simulated data

As we can observe in Tab. 1, the majority of false positive genotypes result from an over-mapping on reference allele 0, rather than on the alternative allele 1 (bottom left corner of the table). Indeed 19 deletions were called as 0/0 whereas they are 1/1, 10 were called as 0/1 instead of 1/1, finally 3 called as 0/0, while they are 1/1. Thus, there is a clear mapping bias towards the longest reference sequence (allele 0 contains the deletion sequence).

Interestingly, among these 32 false positives, we noticed that nearly half of them have a size less than 100 bp. This suggests that the precision of the method may depend on the deletion size. As a matter of fact, the precision is of 85.4 % for deletions smaller than 100 bp versus 97.9 % for deletions greater than 500 bp.

The remaining false positive deletions of size $\geq$ 100 bp, were manually investigated, and most of them occur in regions with a high density of mobile elements.

Comparison with SV discovery approaches One can wonder if these simulated deletions could be easily detected and genotyped by a long read SV discovery tool. We applied here the best to date such tool, Sniffles [8,14] to the chromosome 1 simulated read dataset. As expected, none of the 333 simulated deletions with 0/0 genotypes were assigned a genotype in the Sniffles output call set, since a discovery tool naturally only reports present variants. Surprisingly, among the 667 deletions simulated with either a 0/1 or 1/1 genotype, only 406 were discovered by Sniffles, which gives a recall of only 60.9 %, and with mainly the heterozygous genotypes missing (74% of 0/1 deletions were missed, versus 6 % for the homozygous ones). Interestingly, Sniffles also mis-predicts the genotype of the discovered deletions, assigning most of the 1/1 discovered deletions (n= 254, 81%) as heterozygous. This highlights the fact that Sniffles, a SV discovery tool, is much less precise for the genotyping task than a dedicated genotyping tool.

3.2 Real data

SVJedi was also applied on real ONT data for the whole human genome of NA12878 individual. We try to genotype deletions called by Genome in a Bottle (GiAB). This set of SV, refers as GiAB Mt Sinai VCF, was obtained from PB data using three different SV detection approaches, including PBHoney [15] and SMRT-SV [16]. These approaches also estimated genotypes for present variants (heterozygous 0/1 and alternative homozygous 1/1 only), thus we can compare SVJedi prediction results with the genotype calls predicted by different methods, for this individual.

Full initial GiAB call set The input set of deletions contains 15,616 deletions, with a significant imbalance towards the heterozygous genotype with 14,185 (90%) predicted with a 0/1 genotype (deletion is present in one allele only) and 1,431 with a 1/1 genotype (deletion is present in both alleles). As

expected, as a discovery result, this call set does not contain any 0/0 SV. As a result, SVJedi has assigned a genotype to 9,388 deletions, 4,388 of which were identical to the GiAB ones. As we can observe in Tab. 2, most of the differently predicted genotypes are genotyped as heterozygous in GiAB. Surprisingly, our method did not assign any genotype to 6,228 deletions, again mostly 0/1 genotyped deletions in GiAB, as indicated by the last column. This means that too few reads could be mapped to one or the other allele reference. This could be due to redundancy within the deletion call set (several closely located deletions), resulting in very similar reference sequences between variants preventing the mapper to map reads uniquely. As a matter of fact, almost half of the deletions are less than 1,000 bp apart from the preceding one.

We also predicted 3,325 deletions as 0/0, in other words, absent in NA12878. This might suggest potential False Positive calls of the GiAB call set. Finally, we notice that an important number of differently genotyped deletions, 1,607, were predicted as 1/1, whereas they were called as 0/1 in GiAB. This is in contradiction with results obtained on simulated datasets, where SVJedi errors tend to over-estimate the reference allele. This again may suggest that some deletions of the GiAB call set are mis-genotyped.

		Prediction			
		0/0	0/1	1/1	./.
GiAB call	0/0	0	0	0	0
	0/1	3,317	3,210	1,607	6,051
	1/1	8	68	1,178	177

Tab. 2. Contingency table comparing deletion genotypes of NA12878 between the GiAB full initial call set (n= 15,616) and SVJedi predictions with ONT data.

Higher confidence call set We filtered the GiAB Mt Sinai VCF in order to focus on deletions with a higher confidence call. The initial call set is the union of the deletions calls obtained with seven different discovery pipelines. Here, as a higher confidence call set, we selected the intersection set, keeping only the deletions that were detected by all seven different pipelines, which corresponds to the 'NS=1111111' flag in the INFO column. This new set of deletions contains 1,685 deletions, of which 922 are 0/1 genotypes and 763 are 1/1 genotypes. Compared to the previous experiment, we note that the ratio is more balanced between heterozygous and homozygous genotypes in this set of filtered deletions.

SVJedi was run on these selected deletions. Only one deletion could not be assigned a genotype by the method. For the 1,684 predicted deletions, 1,514 were genotyped exactly as in GiAB. This results in a much higher overlap, 89.9 %, than with the full call set. As we can observe in Tab. 3, the majority of differently predicted genotypes are 1/1 whereas they are 0/1 in GiAB call set. Again, these results are in contradiction with the evaluation of the method on simulated datasets, where the reference allele was overestimated, and therefore question the veracity of GiAB genotypes. Also previously obtained results with the SV discovery tool Sniffles, suggest a similar trend of discovery tools to mis-predict homozygous variants as heterozygous.

		Prediction			
		0/0	0/1	1/1	./.
GiAB call	0/0	0	0	0	0
	0/1	23	780	118	1
	1/1	6	23	734	0

Tab. 3. Contingency table comparing deletion genotypes of NA12878 between the GiAB higher confidence call set (n= 1,685) and SVJedi predictions with ONT data.

Performances On higher confidence GiAB call set for the human genome, with a 39 x coverage, SVJedi took 1h46m to genotype 1,687 deletions, including 1h42 for the alignment with Minimap2 parallelized on 40 cpu. SVJedi reached 6.5 Gbytes as the maximum resident set size, corresponding in fact to the memory usage of Minimap2. On the inital GiAB call set for the human genome, SVJedi took 4h02m, to genotype 15,616 deletions, including 3h42m of mapping, and it reached a maximum memory peak of 14.7 Gbytes during mapping.

4 Discussion and Conclusion

In this work, we provide a novel SV genotyping approach for long read data, that showed good results on both simulated and real datasets. The approach is implemented for the moment only for deletion variants in the SVJedi software. However, this proof of principle on deletion variants is a first step before generalizing the approach for all types of SVs. Insertion variants are simply the counterpart of deletions, and inversions and translocations are SVs even more balanced than insertion/deletion regarding the number of breakpoints (with exactly two breakpoints per allele). Therefore, for all these types of SVs, the method will be easily generalized, to be used in the context of clinical diagnoses or for population genomics analyses.

In the presented analyses, SVJedi ran fast within a few hours on a whole human genome dataset. Our tests show that most of the running time is dedicated to the mapping with Minimap2. Minimap2 is a fast mapper, but it spends time to compute full alignments with optimized similarity scores (ie. optimizing the locations of matches and gaps) whereas only the approximate similarity regions could be used in our approach. Thus, in order to reduce our execution time, we could use other similarity estimation strategies, such as fast alignment-free approaches [17,18].

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 with the Sniffles experiment. Indeed, without a priori SV discovery is a much harder task than genotyping SVs with well characterized alleles, but when the aim is strictly to genotype or compare individuals on already known variants, we have shown that using as much as possible the known features of variants is much more efficient.

As a matter of fact, the efficiency of this approach depends on the quality and precision of the input variants to genotype. Although this issue is inherent to any genotyping approach, our analysis on the full GiAB call set demonstrated that our approach is probably less efficient if there is redundancy in the set of SVs to genotype. This can be frequent when the SV set is obtained from SV calling in several individuals, or with several methods as this was the case here. In these cases, this is still a difficult problem to correctly merge several call sets [19,20], and this can result in a single SV event being described by several SV entries with overlapping coordinates. This is currently not well supported by SVJedi which discards non uniquely mapped reads. The precise or rather imprecise definition of the breakpoints may also impact the genotyping performances and this remains to be assessed for this particular approach. Finally, in the perspective of applying our method for instance on the full SV catalog referenced in the dbVar database, both issues of precision and redundancy of the initial SV call set will be critical issues that may monopolize most of the efforts.

Acknowledgements

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

References

[1] Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E Welch, Max L Dougherty, Bradley J Nelson, Ankeeta Shah, Susan K Dutcher, et al. Characterizing the major structural variant alleles of the human genome. *Cell*, 2019.

[2] James R Lupski. Structural variation mutagenesis of the human genome: impact on disease and evolution. *Environmental and molecular mutagenesis*, 56(5):419–436, 2015.

[3] Lon Phan, Jeffrey Hsu, Michaela Willi Le Quang Minh Tri, Tamer Mansour, Yan Kai, John Garner, John Lopez, and Ben Busby. dbvar structural variant cluster set for data analysis and variant comparison. *F1000Research*, 5, 2016.

[4] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. Speedseq: ultra-fast personal genome analysis and interpretation. *Nature methods*, 12(10):966, 2015.

[5] Danny Antaki, William M Brandler, and Jonathan Sebat. Sv2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics*, 34(10):1774–1777, 2017.

[6] Parsoa Khorsand and Fereydoun Hormozdiari. Nebula: Ultra-efficient mapping-free structural variant genotyper. *bioRxiv*, page 566620, 2019.

[7] Giulia Bernardini, Paola Bonizzoni, Luca Denti, Marco Previtali, and Alexander Schönhuth. Malva: genotyping by mapping-free allele detection of known variants. *BioRxiv*, page 575126, 2019.

[8] Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, and Michael C Schatz. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*, 15(6):461–468, 2018.

[9] Mircea Cretu Stancu, Markus J Van Roosmalen, Ivo Renkens, Marleen M Nieboer, Sjors Middelkamp, Joep De Ligt, Giulia Pregno, Daniela Giachino, Giorgia Mandrile, Jose Espejo Valle-Inclan, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature communications*, 8(1):1326, 2017.

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

[11] Bianca K Stöcker, Johannes Köster, and Sven Rahmann. Simlord: simulation of long read data. *Bioinformatics*, 32(17):2704–2706, 2016.

[12] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338, 2018.

[13] Matthew Pendleton, Robert Sebra, Andy Wing Chun Pang, Ajay Ummat, Oscar Franzen, Tobias Rausch, Adrian M Stütz, William Stedman, Thomas Anantharaman, Alex Hastie, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature methods*, 12(8):780, 2015.

[14] Wouter De Coster, Arne De Roeck, Tim De Pooter, Svenn D'hert, Peter De Rijk, Mojca Strazisar, Kristel Sleegers, and Christine Van Broeckhoven. Structural variants identified by oxford nanopore promethion sequencing of the human genome. *BioRxiv*, page 434118, 2018.

[15] Adam C English, William J Salerno, and Jeffrey G Reid. Pbhoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC bioinformatics*, 15(1):180, 2014.

[16] Mark JP Chaisson, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608, 2015.

[17] Nicolas Maillet, Claire Lemaitre, Rayan Chikhi, Dominique Lavenier, and Pierre Peterlongo. Compareads: comparing huge metagenomic experiments. *BMC Bioinformatics*, 13(Suppl 19):S10, 2012.

[18] Camille Marchet, Lolita Lecompte, Antoine Limasset, Lucie Bittner, and Pierre Peterlongo. A resource-frugal probabilistic dictionary and applications in bioinformatics. *Discrete Applied Mathematics*, pages 1–11, April 2018.

[19] Hemang Parikh, Marghoob Mohiyuddin, Hugo YK Lam, Hariharan Iyer, Desu Chen, Mark Pratt, Gabor Bartha, Noah Spies, Wolfgang Losert, Justin M Zook, et al. svclassify: a method to establish benchmark structural variant calls. *BMC genomics*, 17(1):64, 2016.

[20] Daniel C Jeffares, Clemency Jolly, Mimoza Hoti, Doug Speed, Liam Shaw, Charalampos Rallis, Francois Balloux, Christophe Dessimoz, Jürg Bähler, and Fritz J Sedlazeck. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature communications*, 8:14061, 2017.