Towards a better understanding of the low discovery rate of short-read based insertion variant callers

Wesley Delage, Julien Thevenon, Claire Lemaitre

Paper 135

Towards a better understanding of the low discovery rate of short-read based insertion variant callers

Wesley Delage[1], Julien Thevenon[2] and Claire Lemaitre[1]
[1] Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000, Rennes, France
[2] Unité de Génétique Clinique,Pôle Couple Enfant, CHU de Grenoble Site Nord-Hôpital Couple-Enfant, 38043, Grenoble, France

Corresponding author: wesley.delage@inria.fr

Abstract *Since 2009, numerous tools have been developed to detect structural variants using short read technologies. Insertions are one of the hardest type to discover and are drastically underrepresented in gold standard variant call sets. The advent of long read technologies has completely changed the situation. In 2019, two independent cross technologies studies have published the most complete variant call sets with sequence resolved insertions in human individuals. Among the reported insertions, only 17% could be discovered with short-read based tools. In this work, we performed an in-depth analysis on one of these unprecedented insertion call sets, in order to investigate the causes of such failures. We have first established a precise classification of insertion variants according to three different layers of characterization: the nature of the inserted sequence, the genomics context of the insertion site and the breakpoints junction complexity. Because these levels are intertwined, we used simulations to characterize the impact of each complexity factor. Most reported insertions exhibited characteristics that may interfere with their discovery: 56% were tandem repeat expansions, 25% contained homology larger than 20 bp within their breakpoints junctions and 64% were located in simple repeats. Consequently, the recall of short-read based variant callers was significantly lower for such insertions (6% vs 48% for mobile element and novel insertions). Simulations showed that the most impacting factor on the discovery rate was the insertion type rather than the genomics context, and that the different factors of insertion complexities were handled differently depending on the chosen tool.*

Keywords short reads, variant calling, structural variants, insertions

1 Introduction

Structural variants (SVs) are defined as a fragment of DNA of at least 50 bp that differs between two individuals[1]. SV are categorized by type : deletion (DEL) for a loss of a fragment, insertion (INS) for a gain of a fragment, inversion for a reversion of a fragment (INV) and translocation (TRANS) for moving a fragment to another position in the genome. Such variations in the genome sequence may have important functional impacts on the organism and SVs are commonly associated to human genetic diseases or disorders [2].

The classical approach to call SVs from Whole Genome sequencing (WGS) with short reads relies on a first mapping step to a reference genome. Then SV callers look for atypical mapping signals, such as discordant read pairs, clipped reads or abnormal read depth, to identify putative SV breakpoints along the reference genome [3,4]. More than 70 SV callers have been developed up to date and several benchmarks have highlighted the low level of agreement between the different methods, demonstrating that SV detection using short reads sequencing remains challenging [5]. Indeed the size of the reads is small compared to the target event size and the detection is mainly based on alignments which may produce artefacts[6]. In particular, insertions are one of the most difficult SV types to call. Because the inserted sequence is absent from the reference genome, or at least at the given locus of insertion, calling such variants and resolving the exact inserted sequence require trickier approaches such as de novo or local assembly [7,8]. This increased difficulty is well exemplified by the dramatic under-representation of such SV type in usual reference databases or standard variant call sets such as dbVar.

Recently, the commercialization of novel long reads technologies has completely changed the situation, and insertion variants are finally being discovered and referenced in human populations[9]. Thanks to several international efforts, some gold standard call sets have been produced in 2019, referencing tens of thousands of insertions in a given human individual [10,11]. Among the reported insertions by Chaisson et al, a great majority (83 %) could not be discovered by any of the tested short-read based tools. This result of discovery rate below 17 % is drastically different from the announced performances of insertion callers when evaluated on simulated datasets [12]. Indeed, Chaisson et al showed that 59 % of insertion variants are found in a tandem repeat context, highlighting the fact that most real insertion variants in human individuals are probably not "simple" sequences inserted in "easy" genomic contexts. However, their analysis went no further in order to precisely identify the actual features of insertion events that make them so difficult to be discovered by short read data.

In this work, we performed an in-depth analysis of this unprecedented insertion call set, in order to investigate the causes of short read based caller failures. We have first established a precise classification of insertion variants according to three different layers of characterization: the nature of the inserted sequence, the genomic context of the insertion site and the breakpoint junction complexity. Because these levels are intertwined, we used simulations to characterize the impact of each complexity factor on the discovery rate of several SV callers, accounting for the different types of methodological approaches.

2 Results

2.1 In-depth analysis of an exhaustive insertion variant call set

In this work, we first aimed at precisely characterizing an exhaustive set of insertion variants present in a given human individual. We based our study on a recently published SV call set published by Chaisson and colleagues in 2019[10]. Using an extensive sequencing dataset, combining several different sequencing technologies and methodological approaches (short, linked and long reads, mapping-based and assembly-based SV calling), three human trios were thoroughly analysed to establish exhaustive and gold standard SV call sets. We focused our study on the individual NA19240, son of the so-called Yoruban (YRI) Nigerian trio, whose SV call set contains 15,693 insertions greater than 50 bp.

We have established a precise classification of these insertion variants according to three different layers of characterization: the nature of the inserted sequence, the genomic context of the insertion site and the breakpoint junction complexity.

Insertion sub-types. Insertion variants can be classified in different sub-types according to the nature of the inserted sequence. Whereas only 3 insertion categories were distinguished in the original publication, namely tandem repeats, mobile element (ME) insertions and complex ones for all the other types, we chose to refine this classification in six insertion sub-types, illustrated in Figure 1. A classical subdivision consists in opposing novel sequence to duplicative insertions. In the first case, the inserted sequence is completely absent in the reference genome, whereas in the second, the inserted sequence has one or several homologous copies elsewhere in the genome. Among duplications, *mobile element* insertions are a very specific sub-type and are defined based on the homology of the inserted sequence with an already known mobile element. Then, several sub-types of duplicative insertions are then defined according to the location or amount of the inserted sequence copies in the reference genome. We therefore distinguish *tandem duplications*, for which at least one copy of the inserted sequence is adjacent to the insertion site, from *dispersed duplications*, for which its copies can be located anywhere else in the genome. Among tandem duplications, we defined a specific sub-type called *tandem repeats*, where the inserted sequence itself is composed of multiple tandem repetitions of a seed motif. Mobile elements (ME) are characterized by very high copy numbers in the genome (typically greater than 500), other dispersed duplication types were then required to have a copy number lower than 50, in order not to be confounded with potential MEs. Finally, a sub-type of dispersed duplications is a segmental duplication, that must be larger than 1kb and share more than 90 % of sequence identity with at least one copy, following previous definitions [2].

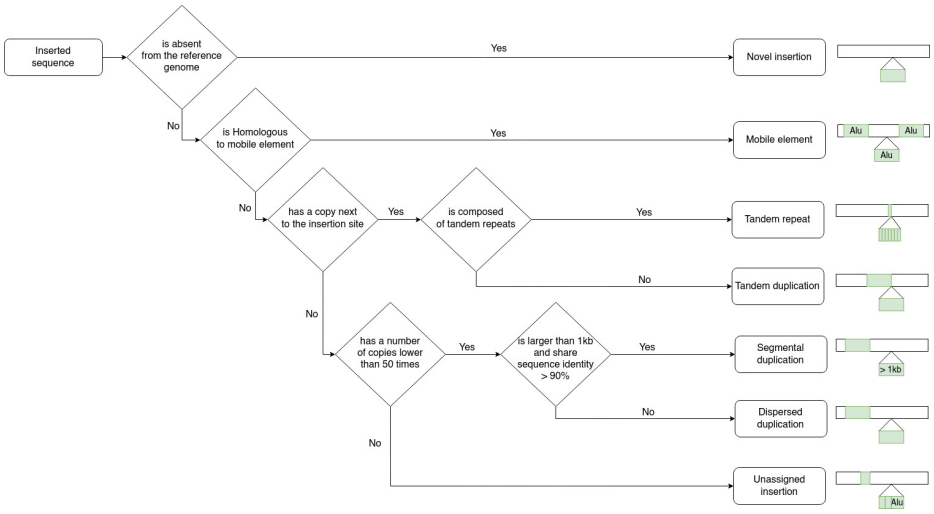


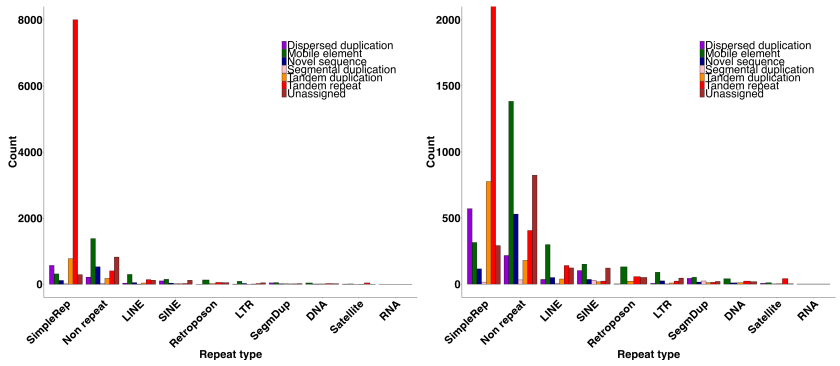
Fig. 1. Decision tree used to classify insertion variants in six insertion sub-types.

In order to classify the insertion call set, all inserted sequences were aligned against the human reference genome, a mobile element database and were scanned for tandem repeats (see Material and Methods). We used a minimal sequence coverage threshold to annotate each insertion to an insertion sub-type according to the decision tree described in Figure 1. We set the threshold to 80% for our analysis to ensure a good compromise between specificity and quantity of annotated insertions in all sub-types. For instance, an insertion is classified as a novel sequence insertion if more than 80 % of its inserted sequence is not covered by any alignment with the reference genome nor with the ME reference sequences, nor contains tandem repeats. Insertions that does not meet the 80 % coverage requirement to be annotated as one of the previous sub-types are qualified as *unassigned* insertions.

With a threshold set at 80%, 90% of insertions could be assigned to a given type. Among the 15,693 insertions, 56% were annotated as tandem repeats, 16 % as mobile elements, 7 % as tandem duplications, 5 % as novel sequences, 6 % as dispersed duplications and 1% as segmental duplications (Figure 2). Compared to the classification of Chaisson et al, the proportions of tandem repeats (57% vs 56%) and mobile elements (23% vs 16%) were very similar. The difference in mobile element proportions represent mainly insertions that are unassigned in our annotation, suggesting that our classification is more conservative. Interestingly, 77 % of their complex insertions were more precisely classified in one of our six sub-types, with mainly 3 sub-types being roughly equally involved: novel sequences, tandem and dispersed duplications. Short read based SV callers used in the original study were able to detect 17 % of these insertions, mainly represented by MEs. This short-read recall was highly variable with respect to the insertion type: ME and novel sequence insertions showed the best recalls (49 and 45 % respectively), whereas other types were all below 11 %. In particular, tandem repeats appeared to be a very hard insertion type to discover (recall of 5 %), although it represents most of the insertion variation in a human genome.

Characterization of insertion locations in the genome. We then characterized the insertions based on the genomic context of their insertion site. We investigated in particular genomic features that can make read mapping and SV calling difficult, such as the repetitive content. A strong over-representation was found in regions annotated as simple repeats, with 64% of the insertions located in these regions that only represent 1.2 % of the genome. As expected, 93 % of tandem repeats were found in simple repeat regions, revealing expansions of already known sites to be highly repeated. We also observed most of the duplications, tandem (72 %) or dispersed (58 %) in these regions. Conversely, 68 % of novel sequence insertions and 56 % of mobile element insertions were located in non repeated regions (Figure 2). We did not find a higher rate of insertions among exon, intron or intergenic regions compared to their distribution along the genome. Compared to GC content

variation along the genome, insertions showed an under-representation in regions with GC content lower than 41% (20 % vs 29% of the genome content) and an over-representation in regions with GC content higher than 46% (17 % vs 7% of the genome content). Novel sequence and mobile element insertions showed to be located in lower GC content regions (median lower than 40 %) than tandem and dispersed duplications, and tandem repeats (median greater than 43%).



(a) (b)

Fig. 2. Dispersion of insertion sub-types according to the repeat content of their insertion site. (a) overview of the dispersion, (b) zoom-in for class counts below 2,000.

Junctional homology. Junctional homology is defined as a DNA sequence that has two identical or nearly identical copies at the junctions of the two genomic segments involved in the rearrangement, when the sequence is short (<70 bp) this is often called a micro-homology [13]. These homologies and micro-homologies have been found involved in several molecular mechanisms generating rearrangements (NAHR for homologies, and MMEJ or MMBIR for microhomologies) [14,15]. In the case of an insertion, a junctional homology is a sequence segment at the left (resp. right) side of the insertion site which is nearly identical to the end (resp. beginning) of the inserted sequence. From a detection point of view, these homologies can have an impact on SV calling performance, since the concerned region at the inserted site is no longer specific to the reference allele and it is no longer possible to identify the exact location of the insertion site. Therefore, we systematically compared the insertion site junction sequences with the inserted sequence extremities to identify stretches of identical or nearly identical sequences. Half of the insertions contained junctional homologies larger than 5 bp, and still 25 % larger than 20 bp, mainly represented by dispersed duplications. The size distribution of the homologies varied between insertion types, novel sequences had small microhomologies (median of 5bp), mobile elements a medium size (median of 15 bp) and dispersed duplications showed a higher homology size (median of 86 bp). Interestingly, insertions called by long reads only had larger junctional homologies than insertions that could be discovered by short reads also (median size of 64 bp vs 12 bp resp.), pointing towards junctional homologies being a potential difficulty factor for short-read based callers.

2.2 Using simulations to investigate the factors impacting the insertion calling recall

In real insertion call sets, most of the previously identified factors impacting SV discovery are correlated. In order to quantify the impact of each factor independently, we produced various simulated datasets of 2x150 bp reads at 40x coverage, containing each 200 homozygous insertion variants on the human chromosome 3. As a baseline, we simulated 250 bp novel sequences taken from yeast exonic sequences inserted inside human exons. This is meant to represent the easiest type of insertions to detect, where inserted sequences contain very few repeats and are novel in the genome, the genomic context of insertion is also simple and repeat-free, and breakpoint junctions do not have any homology. Then, we considered 3 scenarii of simulations, where only one of the three factors, among insertion type (complexity of the inserted sequence), insertion site location and homology at the breakpoints, is changed at a time with respect to the baseline simulation. Four insertion variant callers were evaluated

on these datasets. They were chosen according to their good performances in recent benchmarks [5] and to maximise the methodological diversity. GRIDSS[8], Manta[12] and SVaba[4] are based on a first mapping step to the reference genome, contrary to MindTheGap[7] which uses solely an assembly data structure (the De Bruijn graph).

Discovery rates for all four methods are presented for the different simulated datasets in Table 1. On the baseline simulation, all tools had a close to perfect discovery rate. However, it should be noted that the tools were evaluated solely on their ability to detect an insertion event at a given site regardless of the predicted genotype and the resolution of the inserted sequence. As a matter of fact, only MindTheGap was able to return sequence resolved insertions. The other tools returned either only the insertion site or the insertion site with a partial inserted sequence.

Impact of the insertion type. When simulating various insertion types, GRIDSS was the only tool whose discovery rate was not impacted. Manta could not find any dispersed duplications and very few mobile elements, MindTheGap was unable to detect any type of tandem duplications and SVaba was not able to detect any tandem repeat with a small motif and almost half of the mobile element insertions (Table 1).

Impact of microhomology. Concerning junctional homology, GRIDSS and SVaba were both the less impacted tools. Only the scenario with 50 bp size microhomology impacted them, reducing by 30 to 40 % their discovery rate. Manta discovery rate decreased with the size of microhomology, starting at 50 bp size, reaching 0 % with 150 bp homologies. MindTheGap was the most impacted by microhomology, being unable to detect insertions with microhomology at any tested size.

		Recall (insertion site only)			
		GRIDSS	**Manta**	**MindTheGap**	**SVaba**
Baseline simulation: 250 bp novel sequences in exons		100	95	99	97
# False positive		33	0	14	184
Scenario 1 Insertion type	Dispersed duplication	97	0	97	91
	Tandem duplication	98	98	0	100
	Mobile element	100	5	70	58
	Tandem repeat (6 bp pattern)	100	92	0	0
	Tandem repeat (25 bp pattern)	100	71	5	99
# False positive		33-533	1-22	14-20	6-592
Scenario 2 Microhomology	20 bp	100	99	0	96
	50 bp	70	45	0	59
	100 bp	100	14	0	100
	150 bp	100	0	0	100
# False positive		33-200	2-56	15	2-595
Scenario 3 Genomic location	Low GC content	84	100	72	99
	medium GC content	85	100	69	99
	high GC content	86	100	75	99
	Non repeat	83	99	76	99
	Simple repeat	86	100	71	98
	SINE	86	100	53	99
	LINE	82	99	91	100
	Real locations	84	80	38	71
# False positive		106-144	3-9	16-21	6-25
Scenario 4: real insertions at real locations		45	35	6	44
# False positive		513	107	19	523

Tab. 1. Discovery rate of several short-read insertion callers according to different simulation scenarii. Cells of the table are colored according to the variation of the recall value of the given tool with respect to the recall obtained with the baseline simulation (first line, colored in blue): cells in red show a loss of recall >10%, cells in grey show no difference compared to baseline recall at +/- 10%. For each scenario, the last line indicates the range of the number of false positive predictions.

Impact of the genomic location. Concerning the impact of the genomic context of insertion, the tools showed two distinct behaviors. On the one hand, Manta and SVaba were not affected by

the repeat or GC content of the regions hosting the insertion site. On the other hand, both GRIDSS and MindTheGap showed a loss of recall even in repeat-free and medium GC contexts with respect to exonic locations simulated in the baseline simulation. Interestingly, when using the locations of the real insertions of NA19240 to simulate simple insertions, all tools underwent a loss in their recall compared to the same inserted sequences but in exonic locations in the baseline simulation.

Finally, when simulating the real insertions at their real location as described in the NA19240 variant calling file for the chromosome 3, the discovery rate of all tools dropped to less than 45 %, reaching for many tools their lowest values among the different simulated datasets. This suggested that several levels of difficulties might be combined in real insertions. GRIDSS reached the largest discovery rate (45 %), but it produced the largest amount of false positive discoveries. Surprisingly, the amount of false positives was not constant for most tools, it increased when the simulated insertions are less well discovered or with particular insertion types.

3 Discussion

We have presented here one of the most detailed and comprehensive analyses of factors impacting the detection of insertion variants in the human genome with short read re-sequencing data. This could be possible thanks to the publication of an exceptional SV call set by Chaisson et al[10]. Not only, this catalog of insertion variants is considered as the most exhaustive for a given human individual, but this is also the first set with sequence-resolved events for any size and type of insertions. This resolution of sequence enabled us to propose a refined classification of insertion variants and to quantify the presence of sequence homologies at the breakpoint junctions. Our results showed a strong over-representation of insertion types and contexts towards the most difficult ones to detect with short-read data, for instance tandem repeats inserted in simple repeat contexts. Moreover, most insertions and even the simplest types, such as novel sequence insertions, showed junctional homologies of substantial size that affect SV calling with short reads.

Our simulation protocol enabled to study each difficulty factor independently and highlighted the larger impact of insertion type compared to insertion location. However, all studied factors taken independently could not explain the whole loss of discovery rate and there is probably an important synergetic effect of combining in a single insertion event several of the studied factors. Surprisingly, the different evaluated tools showed very contrasted sensitivities to the different simulated difficulties. This result suggests that combining the calls of several SV callers could improve substantially the overall discovery rate. Currently, Structural Variation studies are based on intersection selections of a combination of SV callers, selecting only the calls that are discovered concordantly between different tools to increase the precision. [5]. Our results suggest an utterly different way of combining tools by taking a careful union of calls based for instance on the type or location of insertions. The main shortcoming of this strategy would then be to control the false positive rate. Our results on simulated data showed that except for MindTheGap, short-read based tools can not provide sequence-resolved variants. We argue that systematically assembling the inserted sequence, such as what is performed with MindTheGap using the whole read set instead of a sub-sample, could help in controlling the false discovery rate.

4 Methods

Data origin. The SV call set of individual NA12940 was downloaded from the following link: `ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/Homo_sapiens/by_study/genotype/nstd152/NA19240.BIP-unified.vcf.gz`. Out of the 17,026 described insertions, only insertions that were sequence resolved (ie. with an inserted sequence entirely defined) and that were also present in at least one of the parent were kept, resulting in a set of 15,693 insertions. The human reference genome version for this study was Hg38.

Insertion annotation. TandemRepeatFinder (TRF) was used to annotate tandem repeats within each inserted sequence [16]. Recommended parameters were used, except for the maximum expected TR length (-l) which was set to 6 millions. In order to annotate Mobile Elements (MEs) in inserted sequences, we used one of the annotation tools of RepeatMasker, namely dfam [17]. Each

inserted sequence was scanned by dfam with the standard hmm profile database of human MEs provided by the tool. For the detection of dispersed duplications and the occurrence count of their copies in the reference genome, each inserted sequence was locally aligned against the human genome using Blat with default parameters [18]. Only the alignments with at least 80 % identity were kept. For the detection of tandem duplications, both breakpoint junction sequences were aligned against the inserted sequence using Blat.

Junctional homology detection. From the previous obtained alignments between the breakpoint junctions and the inserted sequence, only the alignments with at least 90 % identity and occurring as close as 10 bp from extremities of the inserted sequence and from the insertion site were kept. Only alignments between the left (resp. right) side of the insertion site and the end (resp. beginning) of the inserted sequence were kept. In case of multiple candidates hits at one side of the junction, the one located at the closest position from extremities was kept. If homologies were found at both sides of the junction, the homology size was obtained by summing both alignment sizes.

Genomic context characterization. To study the genomic context of insertions, we used the repeat content annotations of RepeatMasker from the UCSC genome browser for the Hg38 genome and the gene annotations from the Gencode v32. To study the GC content, we segmented the genome into isochores with isoSegmenter [19], giving the following five families of isochores: <37 %, 37-41%, 41-46%, 46-53% and >53% GC content.

Simulations. 18 sequencing datasets were simulated to characterize the impact of potential difficulties for variant calling. Each dataset was obtained by altering the human chromosome 3 with 200 insertions. Reads were generated using ART with the following parameters : 2x150 bp reads, at 40 X coverage, with insert size of 300 bp on average and 20 bp standard deviation [20].
Baseline simulation. We simulated 250 bp novel sequence insertions located in exons without any microhomology at the breakpoint junctions. Novel sequences were extracted from random exonic regions of the *Saccharomyces cerevisae* genome.
Scenario 1: Insertion type impact. Insertion locations were identical to the baseline simulation, but the 250 bp inserted sequences were alternatively replaced by dispersed duplications, tandem repeats, tandem duplications or mobile elements. Two types of tandem repeats were simulated, with a pattern size of 6 bp or 25 bp, the pattern originating from the left breakpoint junction. 200 Alu mobile element sequences with a size ranging between 200 and 300 bp were randomly extracted from the human genome based on the RepeatMasker annotation. Tandem duplications were generated by duplicating the 250 bp left breakpoint sequence. The inserted sequences of simulated dispersed duplications were extracted from exons of the chromosome 3.
Scenario 2 : Microhomology impact. The 250 bp insertion sequences produced in the baseline simulation were altered with microhomology. To simulate microhomologies, we replaced the X first bases of each insertion with the same size sequence originating from the right breakpoint sequence. We simulated four microhomology sizes : 20, 50, 100 and 150 bp.
Scenario 3 : Location impact. The 250 bp insertions from the baseline simulation were inserted in specific genomic contexts : either inside different types of mobile elements, namely SINEs and LINEs, in simple repeats or in non-repeated regions with different GC contents. We defined three families of GC content : low (<41%), middle (41-46%) and high (>46 %).
Scenario 5 : Real insertions at real locations. The 889 insertions located on the chromosome 3 from the NA19240 call set were simulated as described in the vcf file.

Insertion calling. Simulated reads were aligned with bwa against the hg38 reference genome, and read duplicates were marked. GRIDSS v2.8.0, Manta v1.6.0, MindTheGap v2.2.1 and SVaba v1.1.0 were all run using recommended, or otherwise default, parameters [8,12,7,4]. Only "PASS" insertions, that were larger than 50 bp, were kept for the recall calculation. Since most of the tools are not able to output sequence resolved variants, the discovery rate was assessed solely based on the insertion site location prediction with a 10 bp margin around the expected location (after left-normalization).

Acknowledgements

We acknowledge the GenOuest bioinformatics core facility (https://www.genouest.org) for providing the computing infrastructure.

References

[1] Monya Baker. Structural variation: the genome's hidden architecture. *Nature methods*, 9(2):133–137, 2012.

[2] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, 2006.

[3] Mehdi Pirooznia, Melissa Kramer, Jennifer Parla, Fernando S Goes, James B Potash, W Richard McCombie, and Peter P Zandi. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8(1):14, 2014.

[4] Jeremiah A. Wala, Pratiti Bandopadhayay, Noah Greenwald, Ryan O Rourke, Ted Sharpe, Chip Stewart, Steve Schumacher, Yilong Li, Joachim Weischenfeldt, Xiaotong Yao, Chad Nusbaum, Peter Campbell, Gad Getz, Matthew Meyerson, Cheng-Zhong Zhang, Marcin Imielinski, and Rameen Beroukhim. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Research*, mar 2018.

[5] Shunichi Kosugi, Yukihide Momozawa, Xiaoxi Liu, Chikashi Terao, Michiaki Kubo, and Yoichiro Kamatani. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*, 20(1), jun 2019.

[6] Irina Abnizova, Rene te Boekhorst, and Y Orlov. Computational errors and biases of short read next generation sequencing. *J Proteomics Bioinform*, 10(1):1–17, 2017.

[7] Guillaume Rizk, Anaïs Gouin, Rayan Chikhi, and Claire Lemaitre. Mindthegap : integrated detection and assembly of short and long insertions. *Bioinformatics*, 30(24):3451–3457, Dec 2014.

[8] Daniel L. Cameron, Jan Schröder, Jocelyn Sietsma Penington, Hongdo Do, Ramyar Molania, Alexander Dobrovic, Terence P. Speed, and Anthony T. Papenfuss. Gridss: sensitive and specific genomic rearrangement detection using positional de bruijn graph assembly. *Genome Research*, 2017.

[9] 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. *Nature methods*, 15(6):461–468, 2018.

[10] Mark J.P. Chaisson, Ashley D. Sanders, ..., Tobias Marschall, Jan Korbel, Evan E. Eichler, and Charles Lee. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10:1784, April 2019.

[11] Justin M. Zook, Nancy F. Hansen, ..., Mark JP Chaisson, Noah Spies, Fritz J. Sedlazeck, Marc Salit, and the Genome in a Bottle Consortium. A robust benchmark for germline structural variant detection. *bioRxiv*, jun 2019.

[12] Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J Cox, Semyon Kruglyak, and Christopher T Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics (Oxford, England)*, 32:1220–1222, April 2016.

[13] Diego Ottaviani, Magdalena LeCain, and Denise Sheer. The role of microhomology in genomic structural variation. *Trends in Genetics*, 30(3):85–94, 2014.

[14] Donald F Conrad, Christine Bird, Ben Blackburne, Sarah Lindsay, Lira Mamanova, Charles Lee, Daniel J Turner, and Matthew E Hurles. Mutation spectrum revealed by breakpoint sequencing of human germline cnvs. *Nature genetics*, 42(5):385, 2010.

[15] Jeffrey M Kidd, Tina Graves, Tera L Newman, Robert Fulton, Hillary S Hayden, Maika Malig, Joelle Kallicki, Rajinder Kaul, Richard K Wilson, and Evan E Eichler. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*, 143(5):837–847, 2010.

[16] Gary Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research*, 27(2):573–580, 1999.

[17] Robert Hubley, Robert D Finn, Jody Clements, Sean R Eddy, Thomas A Jones, Weidong Bao, Arian FA Smit, and Travis J Wheeler. The dfam database of repetitive dna families. *Nucleic acids research*, 44(D1):D81–D89, 2016.

[18] W James Kent. Blat—the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[19] Paolo Cozzi, Luciano Milanesi, and Giorgio Bernardi. Segmenting the human genome into isochores. *Evolutionary Bioinformatics*, 11:EBO–S27693, 2015.

[20] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. Art: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, 2012.