# SurVIndel2: improving local CNVs calling from next-generation sequencing using novel hidden information

Ramesh Rajaby[1,2] and Wing-Kin Sung[1,2,3,4,5,6,7,*]

[1]Hong Kong Genome Institute, Hong Kong Science Park, Shatin, Hong Kong, China
[2]A*STAR Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore
[3]National Key Laboratory of Crop Genetic Improvement, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[4]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[5]Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong, China
[6]Laboratory of Computational Genomics, Li Ka Shing Institute of Health Sciences, The Chinese University of Hong Kong, Hong Kong, China
[7]School of Computing, National University of Singapore, 13 Computing Drive, Singapore 117417, Singapore
[*]Corresponding author: Wing-Kin Sung, kwksung@cuhk.edu.hk

**Abstract**

Deletions and tandem duplications are two major classes of structural variations. They frequently occur in highly repetitive regions, where existing methods fail to detect most of them. We previously introduced SurVIndel, an algorithm that tackled this issue by employing novel statistical methods and had improved sensitivity in repetitive regions. However, its precision was low. Here, we introduce SurVIndel2, an algorithm that uses a novel type of evidence (which we called hidden split reads) and also borrows and adapts the statistical approach of its predecessor. By combining these two approaches, SurVIndel2 is both more sensitive and more precise than any other method we tested.

# 1 Introduction

Structural Variations (SVs) are mutations in a genome that involve 50 nucleotides or more. Although they are less frequent than point mutations, SVs account for more heritable differences in populations due to their large scale and play a major role in many genetic diseases [15]. Deletions and tandem duplications are two important classes of SVs. Deletions occur when a contiguous sequence of bases is lost, while tandem duplications occur when a source sequence of DNA is duplicated in-place; they are called unbalanced SVs because they alter the copy number of the genome. A recent study of the SVs in 14,891 human genomes [6] identified deletions as the most frequent SV and tandem duplications as the third most common, after insertions. However, the SVs were detected using short reads, and existing SV callers miss most duplications, as demonstrated in the following sections. When detecting SVs using high-quality long reads, tandem duplications are as common as deletions (Section 2.4). Aside from being an important source of polymorphism, deletions and tandem duplications are implicated in many genetic diseases, such as Smith-Magenis syndrome, Potocki-Lupski syndrome and Williams-Beuren syndrome [15]. In this manuscript, we will often refer to deletions and tandem duplications collectively as *local CNV*s, or simply CNVs for short.

Nowadays, local CNVs can be detected using long read sequencing technologies, like PacBio HiFi reads and ONT nanopore reads. However, these technologies are still expensive and this limits their use in large-scale population studies. A less costly solution is to use short-reads sequencing technologies such as Illumina, which is often used for sequencing large populations [6, 1, 3]. Short read datasets are rapidly increasing in number, with studies on tens of thousands of genomes already published. As the Hong Kong Genome Project, Singapore SG100K, the European '1+ Million Genomes' Initiative, and the All of Us Research Project (among others) continue to generate large amounts of sequencing data, short-read whole genome sequencing will remain at the forefront of genomics for the foreseeable future.

However, detecting deletions and especially tandem duplications from short reads is very challenging. Despite a large number of callers published in the last decade, the sensitivity of even the best methods is still low. Manta, a widely used SV caller employed in many large-scale population studies [6, 3], as well as the best performing tool according to recent benchmarking efforts [13, 4], detects around 40% of the existing deletions and less than 10% of the existing tandem duplications when tested on a recent benchmark dataset (Section 2.4). In a previous study [11], we found that most of the CNVs missed by existing callers were located in tandem repetitive regions. This is because CNVs in tandem repetitive regions often do not generate split reads; this result was confirmed in this study (Section 2.1). Furthermore, shorter CNVs generate few or no discordant pairs, and read depth methods are known to only work for large CNVs [14]. Thus, naively counting the number of discordant pairs and split reads will miss many CNVs in repetitive regions. Finally, we showed that the problem can be mitigated by considering the insert sizes of all the read pairs containing the putative CNVs, and performing statistical tests to determine whether they deviate from expectation. Using this technique, SurVIndel improved the detection rate in repetitive regions over existing SV callers.

However, SurVIndel suffered from two problems. First, it generated

many false positives. Because noise is often randomly distributed while CNVs are recurrent, false positives can easily outnumber true positives as the number of samples increases. Second, because split reads are often unavailable in repetitive regions, many of the CNVs detected had imprecise boundaries. Subsequent steps in populations studies, such as clustering of the CNVs across samples, regenotyping and downstream analysis, all rely and greatly benefit from precise breakpoints.

For these reasons we developed a new caller, SurVIndel2, that significantly reduces the number of called false positives, while retaining or even improving the sensitivity of the original SurVIndel, and generates precise breakpoints for most of the called CNVs. SurVIndel2 detects candidate CNVs using split reads, discordant pairs and new type of evidence, that we called hidden split reads. These are reads that can determine the existence and the precise breakpoints of CNVs in repetitive regions. Recognising and using them, SurVIndel2 generates both a more precise and more comprehensive set of candidate CNVs, and then it adapts the statistical approach of SurVIndel, along with depth-based filters, to produce a set of confident CNVs. In Section 2.1, we introduce the notion of hidden split reads, and why they can improve CNV detection. Section 2.2 gives an overview of the algorithm of SurVIndel2. Section 2.3 and 2.4 use publicly available datasets to compare the performance of SurVIndel2 to the best methods in the literature, according to a recent benchmark [4]. Finally, in Section 2.5, we use SurVIndel2 to generate a dataset of deletions and duplications in the 1000 Genomes Project. We show that our dataset is significantly more complete than that produced by a state of the art pipeline: it contains more than twice as much deletions and thrice as much duplications, with high validation rates according to long reads.

SurVIndel2 is fully open source and available at `https://github.com/kensung-lab/SurVIndel2`. It is easy to run and only requires a BAM or a CRAM file and a reference genome, and it outputs a fully standard VCF file.

# 2 Results

## 2.1 Hidden split reads

In the presence of a deletion, two regions that are far apart in the reference genome are adjacent in the sample genome. More precisely, let $S$ be a region in the reference, and assume $D = S[a..b]$ is a length-$d$ region ($d = b-a+1$) that is deleted in the sample. Let $A = S[a-L..a-1]$ and $B = S[b+1..b+L]$ be the two regions flanking the deletion to the left (upstream) and right (downstream), respectively: in the sample genome, the two regions will be adjacent. Therefore, a read $R[1..r]$ may be sequenced such that part of it is equal to a suffix of $A$ while the remaining is equal to a prefix of $B$, i.e., $R = A[L - r_1 + 1..L]B[1..r_2]$, where $r_1 + r_2 = r$. Since $A$ and $B$ are not adjacent in the reference genome, $R$ may not fully align to one location. Instead, only one part of it will be aligned (Fig. 1a, and Supplementary Fig. S1 for a more technical illustration), and we call such a read *clipped*, or *split*. Similarly, tandem duplications may also produce clipped reads (Fig. 1b).

Some deletions will not produce clipped reads. In particular, when a sufficiently long prefix of the deleted region, $D[1..p]$, is homologous to the prefix of the right-flanking region, $B[1..p]$, no right-clipped reads will support the deletion (Fig. 1c). Similarly, when the suffix of the deleted

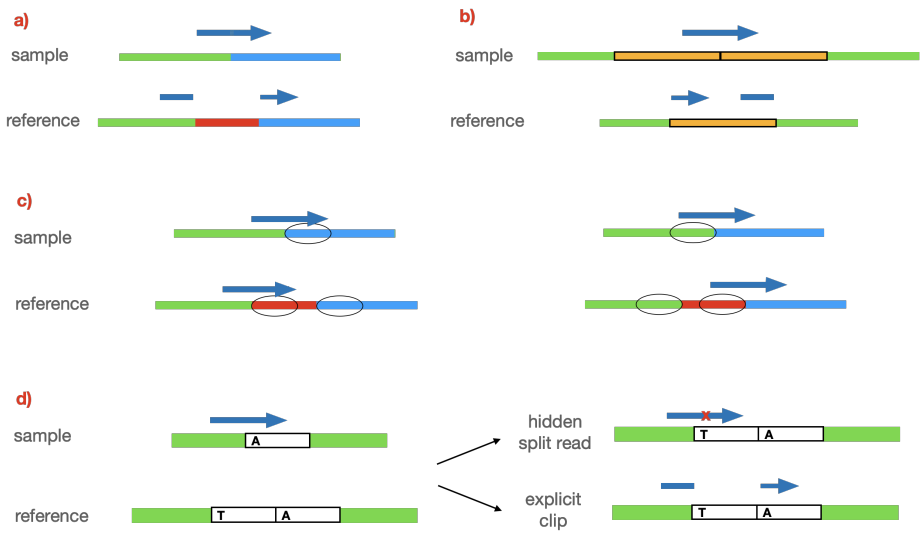


Figure 1: (a) The red segment of the reference is deleted in the sample, causing the green and the blue regions to be adjacent. Reads sequenced partly from the green region and partly from the blue region may not fully align to the reference, causing the read to be clipped. (b) Similarly, tandem duplications may also create clipped reads. (c) When the prefixes (circled) of the deleted (red) region and the right-flanking (blue) region are homologous, many reads sequenced across the left breakpoint of the deletion will fully align to the reference, instead of being clipped. If the length of the prefixes is longer than the read length, no clipped reads will be produced. Similarly, when the suffixes of the deleted (red) region and the left-flanking (green) region are homologous, little or no clipped reads will be produced. (d) Deletions of full copies in tandem repeats are a prime example of (c), which is why they often do not cause clipped reads. However, in practice, different copies of a tandem repeat are often not identical, and we may get hidden split reads. In this example, a sequence (white) is repeated twice in the reference, but only once in the sample. The two copies of the sequence are identical except for one position (first copy has a T, the second has an A). The single copy in the sample has an A, therefore we want to report that the first copy was deleted. Reads spanning both the green left-flanking region and the A character can be aligned to the reference in two ways: fully with one mismatch, or explicitly clipped. Aligners will normally choose the former. While this is usually a reasonable choice, in this case it will hide the existence of the deletion. We call such reads hidden split reads.

region $D[d-p+1..d]$ is homologous to the suffix of the left-flanking region $A[L-p+1..L]$, no left-clipped reads will support the deletion (Fig. 1d). Tandem repeats are common examples of this.

We have a tandem repeat when a sequence $U$ (*repeat unit*) appears consecutively $n$ times in the genome, $n > 1$. Deletions and duplications of entire repeat units often do not create clipped reads [11]. Let $T = U_1..U_n$ be a tandem repeat, and assume $U_i$ is deleted: $U_{i-1}$ and $U_{i+1}$ will be adjacent in the sample. Because $U_{i-1}$, $U_i$ and $U_{i+1}$ are identical to each other, the deletion of $U_i$ will not generate any new sequences, since $U_{i-1}U_{i+1} = U_{i-1}U_i = U_iU_{i+1}$. A similar situation arises from tandem duplications of whole repeat units in tandem repeats (Fig. S2a).

However, in practice, the copies in a tandem repeat may present some differences between each other. For briefness, when two strings $S_1$ and $S_2$ are similar but not identical, we say $S_1 \sim S_2$. If $U_i \sim U_{i-1}$ and $U_i \sim U_{i+1}$, then the sequence $U_{i-1}U_{i+1}$ may be novel, i.e., not exactly present in the reference genome. Let $R$ be a read sequenced such that $R[1..r] = U_{i-1}[u-r_1+1..u]U_{i+1}[1..r_2]$, where $r_1 + r_2 = r$ is the length of $R$ and $u$ is the length of $U_{i-1}$. Since $U_i \sim U_{i+1}$, then $U_{i-1}[u-r_1+1..u]U_{i+1}[1..r_2] \sim U_{i-1}[u-r_1+1..u]U_i[1..r_2]$. Furthermore, since $U_i \sim U_{i-1}$, then $U_{i-1}[u-r_1+1..u]U_{i+1}[1..r_2] \sim U_i[u-r_1+1..u]U_{i+1}[1..r_2]$. The algorithm that maps $R$ to the reference is presented with two options:

1. Align the full read to either $U_{i-1}[p-r_1+1..p]U_i[1..r_2]$ or $U_i[p-r_1+1..p]U_{i+1}[1..r_2]$, accepting that the read does not perfectly align; or
2. Split the read, i.e., map part of the read to $U_{i-1}[p-r_1+1..p]$ and part of the read to $U_{i+1}[1..r_2]$

Which option the aligner chooses depends on many factors, including how similar $U_{i-1}$, $U_i$ and $U_{i+1}$ are, the mapping algorithm used and its parameters. When option 1 is chosen, we call the read a *hidden split read* (Fig. 1e). Similarly, tandem duplications may also produce hidden split reads (Fig. S2b).

Next, we provide a formal definition of hidden split read. Given a genomic fragment $s$, let $score(s)$ be the score of the alignment of $s$ to the reference genome. Given a read $R[1..r]$, we say that $R$ is a hidden split read if there exists $i \in [1..r-1]$ such that $score(R) < score(R[1..i]) + score(R[i+1..r])$; in other words, if $R$ is allowed to split into two without penalty and the two parts can align independently, the score of the alignment will increase. $score(R[1..i]) + score(R[i+1..r]) - score(R)$ is the *HSR-score* of $R$. See Supplementary Fig. S3 for an example of HSR-score calculation.

In order to confirm the impact of deletions deleting full repeat units and the importance of hidden split reads in detecting such deletions, we use two publicly available datasets: the benchmark generated by Zook et al. [16] on HG002 (which we call HG002-GIAB), and the benchmark generated by HGSVC on HG00512, a sample part of the 1000g project [7] (which we call HG00512-HGSVC). A detailed explanation of the experiments mentioned below is provided in Supplementary Section 4. Fig. 2a, Supplementary Fig. S4a show that tandem repeats contain a large portion of the deletions (half of the deletions for HG002-GIAB, two thirds for HG00512-HGSVC), and Fig. 2b, Supplementary Fig. S4b demonstrates that that indeed most deletions delete an integer number of repeat units. For this reason, the number of deletions in tandem repeats supported by split reads is low: 37.5% and 25.6% for both HG002-GIAB and HG00512-HGSVC, compared to 89.2% and 82.5%, respectively, for deletions not

overlapping tandem repeats (Fig. 2c, Fig. 2d, Supplementary Fig. S4c, Supplementary Fig. S4d). However, hidden split reads can help us uncover a significant number of deletions not supported by split reads. Using both split reads and hidden split reads, the number of deletions that we are potentially able to discover increases from 37.5% to 63.3% for HG002-GIAB, and from 25.6% to 54.6% for HG00512-HGSVC.

Despite the potential to detect a large portion of repetitive deletions missed by split reads, using hidden split reads is challenging. Fig. 2e and Supplementary Fig. S4e show the HSR-score distribution for hidden split reads that support existing deletions, while Fig. 2f and Supplementary Fig. S4f show the HSR-score distribution for hidden split reads detected in 10,000 randomly selected tandem repeat regions. Less than 2% of these reads supported an existing deletion. The two distributions look similar, and distinguishing hidden split reads that support existing deletions is not trivial, and novel and sophisticated techniques may need to be developed. In the remainder of this paper, we propose an algorithm that uses hidden split reads to detect deletions and duplications in a sample genome, and significantly outperforms the current state of the art. It should be noted that, unfortunately, the benchmark datasets do not explicitly provide duplications and their coordinates, so we could not obtain similar statistics for them. However, from a theoretical standpoint, the presented notion of hidden split reads stands for tandem duplications as well.

## 2.2 Overview of the method

SurVIndel2 requires (1) a reference genome (in FASTA format) and (2) a BAM or CRAM alignment file containing paired reads sequenced from a single sample aligned to the reference genome. It outputs a VCF file containing a list of predicted deletions and tandem duplications. Fig. 3 shows a high-level overview of the full algorithm.

As a first step, SurVIndel2 iterates through the alignment file and gathers the evidence that will be later used to determine the candidate local CNVs: split reads, hidden split reads and discordant pairs. Split reads and hidden split reads were described in Section 2.1. SurVIndel2 considers potential hidden split reads all the reads that are not clipped and present at least three differences with the reference genome (defined as number of mismatches plus number of indels). Discordant read pairs are pairs of reads that map inconsistently to the library preparation parameters. A deletion may produce pairs where the two reads map very far apart (Fig. S7a). Let $\mu$ and $\sigma$ be the mean and the standard deviation of the distribution of insert sizes for the library: we define as discordant any pair that has insert size greater than $\mu + 3\sigma$ (we will refer to this quantity as $max_{IS}$ throughout the paper). On the other hand, a duplication may produce pairs that have the wrong relative orientation, i.e., the forward-facing read is mapped downstream of the reverse-complemented read (Fig. S7b). We define such pairs as discordant.

Given the split reads, hidden splide reads and discordant pairs, candidate local CNVs are generated by two different modules: the *split module* uses clipped reads and hidden split reads, while the *discordant module* uses discordant pairs.

The split module first generates consensus sequences by clustering clipped reads and hidden split reads (Fig 3a, 3b, Supp. Section 5). The consensus sequences are partitioned into left and right-clipped. Each potentially represents either the left or the right breakpoint of a local CNV:

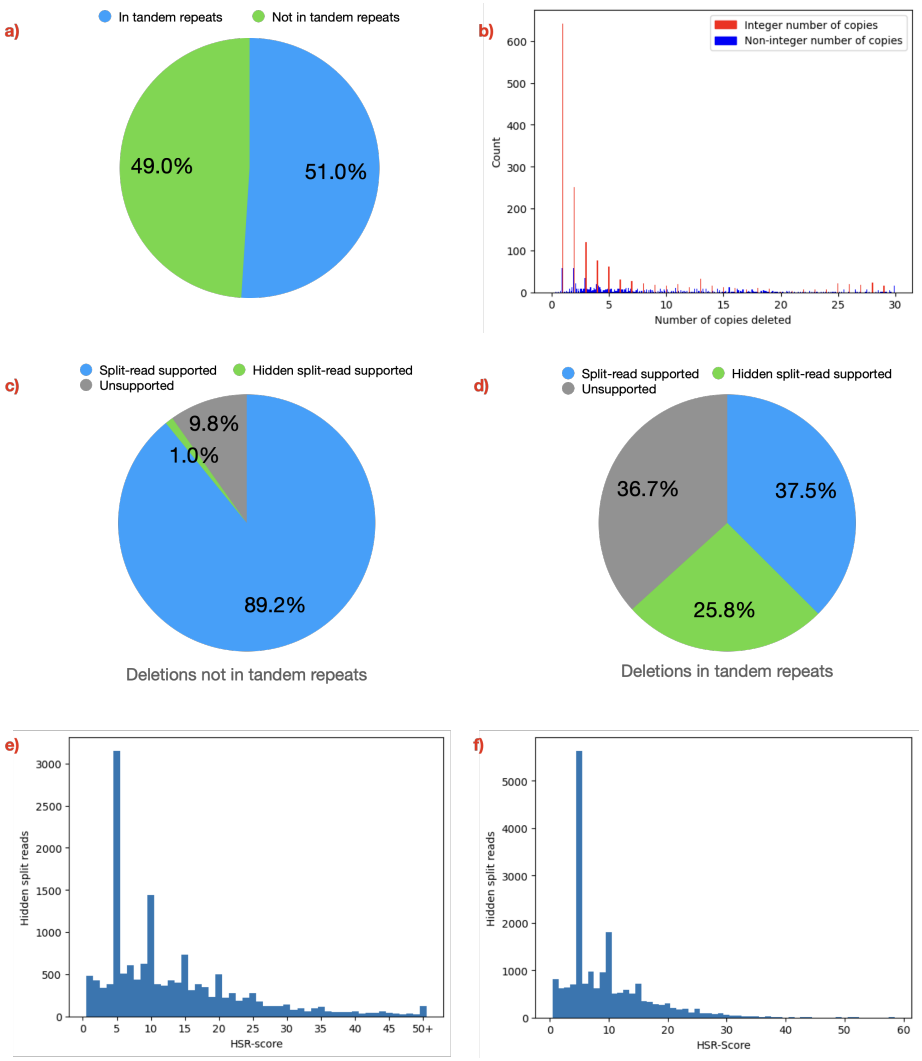


Figure 2: The impact of deletions deleting full repeat units and the importance of hidden split reads in detecting such deletions, studied on HG002-GIAB. (a) We consider a deletion to be contained in a tandem repeat if at least 90% of the deletion is covered by a repetitive region. Around half of the deletions in this dataset are contained in a tandem repeat. (b) For every deletion that overlaps with a tandem repeat, we calculated the number of repeat units that it deletes by dividing the length of the deletion by the length of the repeat unit. Bars representing an integer number of copies (i.e., the length of the deletion is multiple of the length of the repeat unit) are coloured red. Most deletions delete an integer number of repeat units. (c,d) Percentages of deletions outside (c) and contained (d) in tandem repeats that are supported by split reads, by hidden split reads, and by neither. Hidden split reads have the potential to recover a significant number of deletions missed by regular split reads. (e) HSR-score distribution of hidden split reads supporting existing deletions, while (f) shows the distribution of hidden split reads randomly sampled; only 0.4% of such reads supported an existing deletion. The two distributions look similar. This shows that using hidden split reads is very challenging, as it is not trivial to distinguish the ones supporting actual deletions.

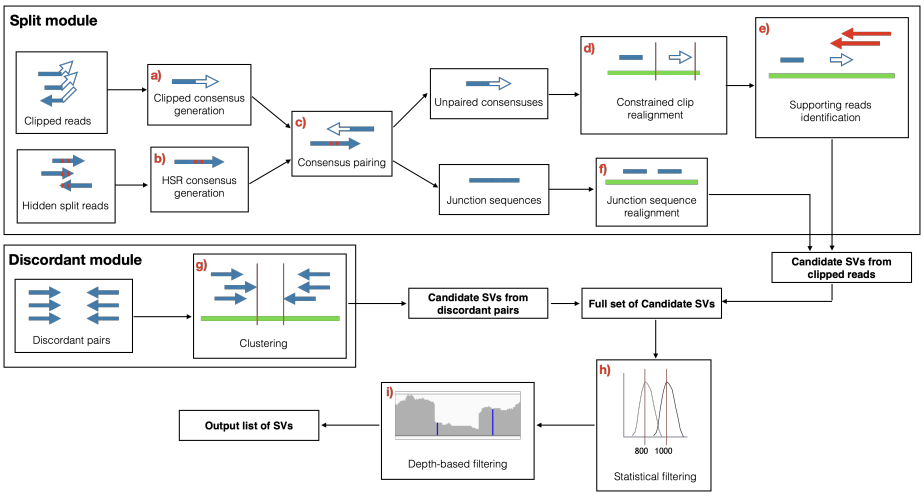


Figure 3: Overview of the SurVIndel2 method. It primarily identifies candidate local CNVs by using split reads and hidden split reads. Candidate local CNVs that do not generate clipped reads are identified using discordant pairs. Finally, the candidate local CNVs are tested by a series of filters, mainly (i) statistically testing the distribution of the read pairs surrounding the local CNV and (ii) testing changes in read depth in the local CNV region as well as its flanking regions.

furthermore, using the mates of the clustered reads, we can calculate a putative range for the location of the second breakpoint (Supp. Section 6, Fig. S8): we call this range *opposite breakpoint range*.

We pair the consensuses so that each pair contains a left-clipped and a right-clipped consensus, and the consensuses are in each other's opposite breakpoint range (Fig 3c). For each pair we generate a *junction sequence* (Supp. Section 7), which is a concatenation of two regions that are adjacent in the sample genome but not in the reference genome (the left and right-flanking regions for deletions, the two copies of the duplicated region for tandem duplications). Finally we realign the junction sequence to the reference genome to obtain the precise breakpoints of the CNV (Fig 3f). For consensuses that are not paired, the clipped portion is realigned within its opposite breakpoint range (Fig 3d) in order to detect the second breakpoint of the CNV (for clusters of hidden split reads, where there is no explicit clip, the read is forcefully clipped in the best location, see Supp. Section 4). Finally, we search for reads supporting the second breakpoint (Supp. Section 8, Fig 3e), and if less than three of them are found, the CNV is discarded.

The discordant module is considerably simpler, and it clusters discordant pairs in order to detect CNVs using the clustering algorithm described in [11] (Fig 3g). One drawback is that, unlike split reads, discordant pairs often provides an approximate position for the breakpoints, rather than a precise one. When the same CNV is predicted by both the split module and the discordant module, the coordinates from the split module are retained, since they are deemed precise.

Finally, the candidate CNVs discovered by the two modules are passed through a series of filters. Two major categories of filters are applied:

*insert size-based* (Fig 3h) and *depth-based* (Fig 3i). Insert size-based filters were first used by SurVindel. Statistical analysis is used to check whether the distribution of the insert sizes of the pairs surrounding a putative CNV is compatible with the presence of the CNV (Supp. Section 9). Depth-based filters check whether the deletion or duplication correspond to a decrease or increase in read depth. For duplications, a specialised algorithm is applied to remove false alignments before computing the read depth (Supp. Section 10).

## 2.3 HG002 benchmark

The Genome in a Bottle Consortium created a benchmark catalogue of 7,281 sequence-resolved insertions and 5,464 deletions, based on the human genome version 19 (hg19) [16]. We classified 3,273 out of the 7,281 insertions as tandem duplications (Supplementary Section 11). Therefore, the final benchmark catalogue contained 5,464 deletions and 3,273 tandem duplications. In the rest of the manuscript, we will refer to it as GIAB-SV. The authors in [16] also defined a set of regions in hg19 (called "Tier 1" regions) where the benchmark is guaranteed to be reasonably complete, and any extra deletion and insertion predicted is very likely to be a false positive. Precision was computed using only calls in Tier 1 regions. We obtained a BAM file of approximately 50× coverage by downloading runs from SRR1766442 to SRR1766486, and mapping them to hg19 using BWA-MEM [9].

We tested three highly cited methods that performed well in a recent benchmarking analysis [4]: Delly [12], Lumpy [8] (which we ran using the Smoove package, as recommended by the authors) and Manta [5]. These SV callers were used in some of the major studies on SVs in human populations [6, 1, 3]. We also included the original SurVIndel [11] in the comparison.

SurVIndel2 shows a 13.6% increase in sensitivity compared to Manta (0.59 vs 0.67, Fig. 4a), while matching it in precision (0.89 for both). SurVIndel is nearly as sensitive as SurVindel2 (0.63 vs 0.67), but falls short on precision (0.56 vs 0.89). Delly and Smoove perform sensibly worse than SurVIndel2 on both precision and sensitivity.

When it comes to tandem duplications, SurVIndel2 outperforms the existing callers by a much wider margin. It shows an increase in sensitivity of nearly 58% compared to Manta (0.19 vs 0.3), while also being more precise (0.62 for Manta, 0.68 for SurVIndel2). SurVIndel nearly matches SurVIndel2 in sensitivity, but its precision is nearly half (0.35 for SurVIndel, 0.68 for SurVIndel2). Both Delly and Lumpy perform much worse: SurVIndel2 is 10 and 15 times more sensitive, respectively, while also being several times more precise. In general, tandem duplications are extremely challenging to detect using short reads.

## 2.4 HGSVC2 benchmark

The Human Genome Structural Variation Consortium (HGSVC) has published a catalogue (called HGSVC2) of insertions and deletions in 34 human genomes from the 1000 genomes project, identified from long-read PacBio whole-genome sequencing and Strand-seq data [7]. Furthermore, the New York Genome Center (NYGC) has recently performed high-coverage, PCR-free sequencing for the 2,504 original samples in the 1000g genome project, along with 698 additional samples. We ran the callers on

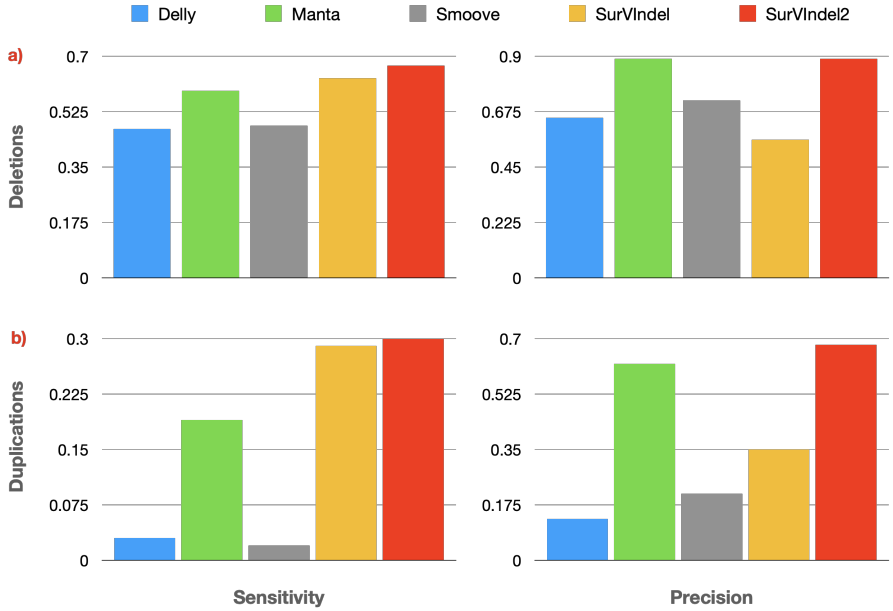


Figure 4: Sensitivity and precision for deletions (a) and tandem duplications (b) on the HG002 benchmark for the five tested methods. SurVIndel2 has the highest sensitivity and the highest precision for both deletions and tandem duplications (for deletions, its precision is tied with Manta at 0.89).

the CRAM files provided by the NYGC for the 34 samples in HGSVC2. Note that HGSVC2 reports a much higher number of SVs per sample when compared to GIAB-SV: 9,872 deletions and 15,136 insertions (9,343 tandem duplications) on average. Therefore, we expect this benchmark to be a much harsher judge of the sensitivity of the methods.

Similarly to Section 2.3, SurVIndel2 outperforms the other methods (Fig. 5). It has both the highest sensitivity and the highest precision for deletions, and the increase in sensitivity over Manta is more marked (29%, 0.41 average sensitivity for Manta, 0.53 average sensitivity for SurVIndel2). SurVIndel is sensitive but not precise, while Delly and Lumpy are both less precise and less sensitive.

For tandem duplications, SurVIndel is the most sensitive method, but once again it suffers in precision. SurVIndel2 is significantly more sensitive and precise than the other methods. Once again, tandem duplications prove to be extremely challenging to detect using short reads.

## 2.5 A catalogue of deletions and tandem duplications for the 3,202 genomes in the 1000 genomes project

We used SurVIndel2 to generate a catalogue of deletions and tandem duplications for 3,202 samples from the 1000g genomes project sequenced by the NYGC. SurVIndel2 was run on each individual sample, and the calls were left-aligned and clustered across samples (Supplementary Section 13). The final catalogue contained 211,981 deletions and 115,928

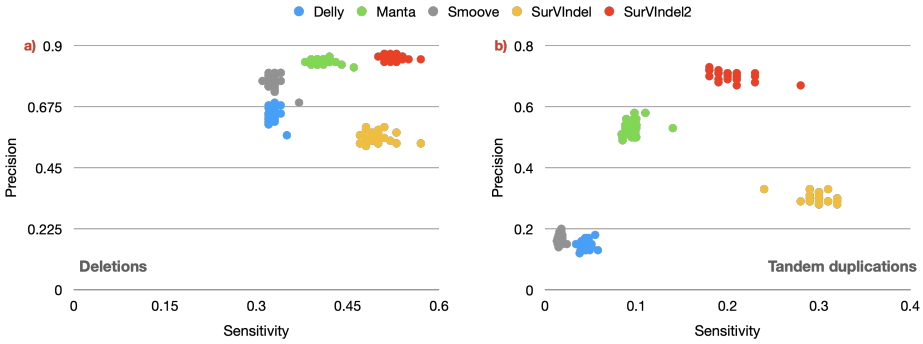


Figure 5: Sensitivity and precision for deletions (a) and tandem duplications (b) on the 34 samples in the HGSVC2 catalogue. Each sample is a dot, its $x$ coordinate is the sensitivity while its $y$ coordinate is the precision. SurVIndel2 is better than the other methods in both precision and sensitivity, with one exception: SurVIndel is more sensitive for tandem duplications. However, the low precision hinders the usefulness of the method.

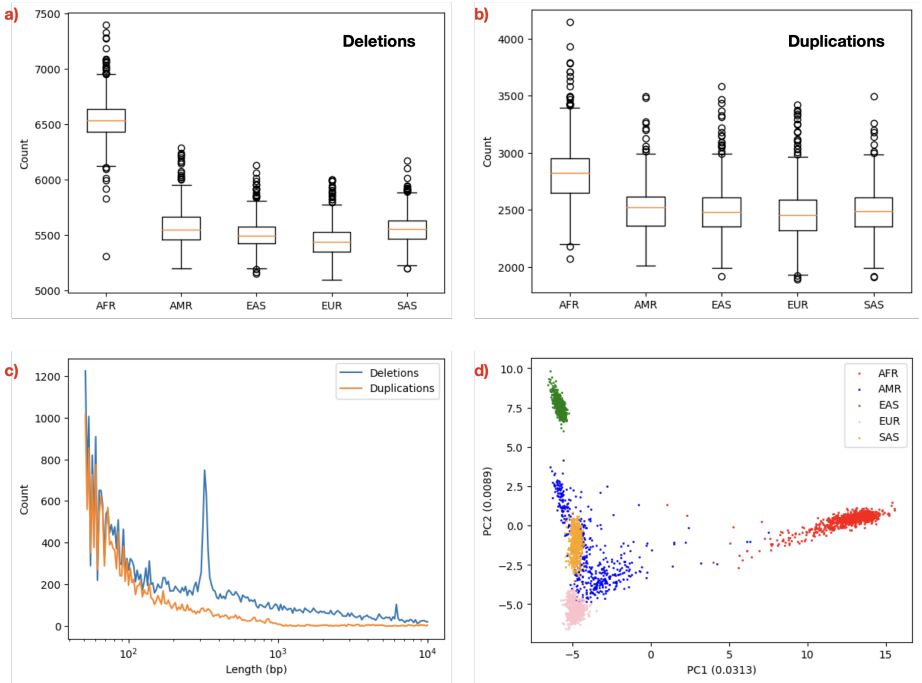


Figure 6: (a,b) Count of deletions and duplications in different superpopulations. Africans have more deletions and duplications compared to non-Africans: the median count was 6,531 deletions and 2,823 duplications for Africans vs 5,505 deletions and 2,483 duplications for non-Africans. (c) The size distribution for deletions and duplications detected in 1000g. For deletions, the expected peaks associated with ALU (∼310bp) and LINE (∼6000bp) deletions are visible. (d) PCA is able to clearly separate the superpopulations.

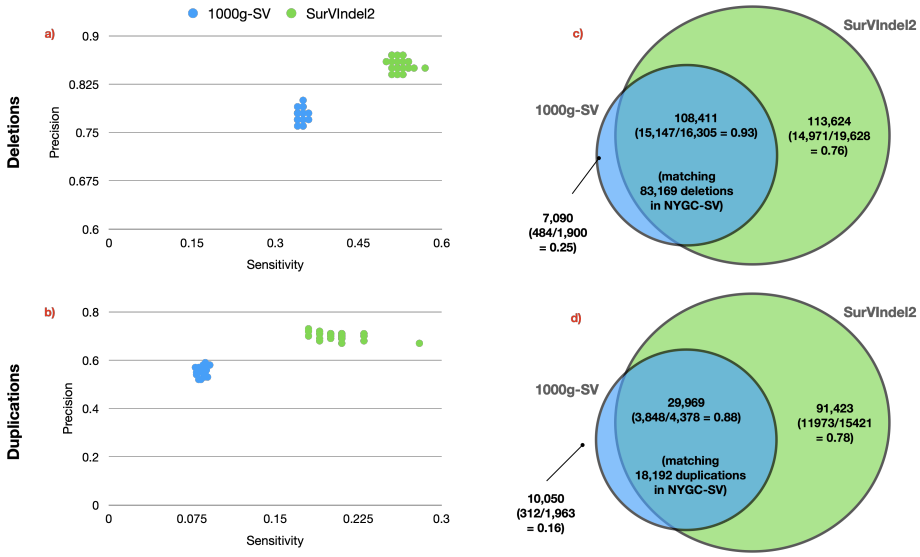


Figure 7: Comparison between the 1000g-SV and SurVIndel2 datasets. (a,b) Sensitivity and the precision for the 34 samples in HGSVC2 for deletions (a) and tandem duplications (b). The dataset produced by SurVIndel2 is both more sensitive and more precise than the 1000g-SV dataset. (c) Overlap between deletions in 1000g-SV and SurVIndel2 datasets. Between parentheses, the validation rates of calls in samples with long reads. Not only SurVIndel2 covers nearly all of the deletions in 1000g-SV, but it predicts 113,624 deletions that are not present in the latter. The validation rate for SurVIndel2-unique events is far higher than the validation rate of 1000g-SV unique events. (d) Similarly, SurVIndel2 predicts most tandem duplications in 1000g-SV, and the ones that are not predicted have very low validation rate. SurVIndel2 also predicts 91,423 duplications that are not present in 1000g-SV, with high validation rate.

tandem duplications. 175,077 deletions (82.6%) and 95,784 duplications (82.6%) were rare (detected in 1% of the population or less), and 77,127 deletions (36.4%) and 39,555 duplications (34.1%) were singletons (detected in only one individual). Africans had more deletions (Fig. 6a) and duplications (Fig. 6b) compared to non-Africans: the median count was 6,531 deletions and 2,823 duplications for Africans vs 5,505 deletions and 2,483 duplications for non-Africans. The number of events was similar in the remaining superpopulations. Fig. 6c shows the size distribution for the deletions and duplications predicted. The peaks associated with ALU (∼310 bp) and LINE (∼6 kbp) deletions are clearly visible. PCA is able to separate the superpopulations (Fig. 6d). This is also true when performing the PCA using deletions and tandem duplications separately (Fig. S5). Furthermore, PCA was able to segregate the superpopulations into subpopulation (Fig. S6).

The NYGC also generated a catalogue of SVs (including deletions and tandem duplications) [3] by integrating multiple callers, including Manta. We refer to this catalogue as 1000g-SV. Fig. 7 shows a comparison between 1000g-SV and our catalogue. Similarly to Section 2.4, we measured sensitivity and precision for 1000g-SV and for SurVIndel2 on the

34 samples in HGSVC2. SurVIndel2 had consistently higher sensitivity and precision. For deletions (Fig. 7a), SurVIndel2 was on average 1.5× more sensitive (average sensitivity was 0.345 for 1000g-SV and 0.526 for SurVIndel2), while being more precise (average precision was 0.777 for 1000g-SV and 0.856 for SurVIndel2). For tandem duplications (Fig. 7b), SurVIndel2 was 2.4× more sensitive (average sensitivity was 0.084 for 1000g-SV and 0.204 for SurVIndel2) and significantly more precise (0.554 for 1000g-SV, 0.702 for SurVIndel2). When comparing the full catalogues (Fig. 7c, 7d), SurVIndel2 captures the vast majority of the deletions in 1000g-SV, and deletions in 1000g-SV but not predicted by SurVIndel2 have low validation rates according to HGSVC2 (0.25). Furthermore, SurVIndel2 predicts 113,624 deletions not present in 1000g-SV, with high validation rates (0.76). The difference in detected tandem duplications is even more striking. Duplications predicted by 1000g-SV but not by SurVIndel2 have very low validation rate (0.16). The two datasets share 29,969 duplications, and SurVIndel2 predicts an additional 91,423, with high validation rate (0.78). Overall, it appears that SurVIndel2 is able to predict a more accurate and significantly more complete catalogue of both deletions and tandem duplications.

# 3 Discussion

We introduced SurVIndel2, a local CNV caller that uses split reads and discordant pairs as well as a new type of evidence, hidden split reads, to produce a comprehensive set of candidate CNVs. SurVIndel2 borrows the idea of statistical testing from its predecessor, SurVIndel, and uses it together with depth-based filters to effectively filter the set of candidate CNVs.

We used publicly available benchmarks to show that SurVIndel2 outperforms the state of the art in SV calling. On the HG002 GIAB benchmark [16], SurVIndel2 achieved the highest sensitivity and precision for both deletions and tandem duplications. When tested on the HGSVC2 benchmark, SurVIndel2 had the highest sensitivity and precision on deletions. It detected less true positive tandem duplications than SurVIndel, but its precision was twice as high. Finally, we used SurVIndel2 to generate a catalogue of deletions and tandem duplications for the 1000 Genome Project, and compared it to a recently published catalogue generated using state-of-the-art pipelines. Our catalog revealed numerous events that were not present in the existing one, and the long read-based validation rates for those events were high. We expect that SurVIndel2 will be beneficial to the community, and will help uncover a large number of events that would be missed by using existing tools, while maintaining the number of false positives to a minimum.

Detecting tandem duplications is extremely challenging, and the sensitivity of SurVIndel2 is still very low. In fact, it is lower than what SurVIndel was able to achieve. We believe that it is crucial to close this gap by either showing why tandem duplications are so challenging to predict, or by providing a significantly improved algorithm.

# References

[1] Haley J Abel, David E Larson, Allison A Regier, Colby Chiang, Indraniel Das, Krishna L Kanchi, Ryan M Layer, Benjamin M Neale,

William J Salerno, Catherine Reeves, Steven Buyske, NHGRI Centers for Common Disease Genomics, Tara C Matise, Donna M Muzny, Michael C Zody, Eric S Lander, Susan K Dutcher, Nathan O Stitziel, and Ira M Hall. Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, 583(7814):83–89, Jul 2020.

[2] G Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic Acids Res*, 27(2):573–80, Jan 1999.

[3] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, Susan Fairley, Alexi Runnels, Lara Winterkorn, Ernesto Lowy, Human Genome Structural Variation Consortium, Paul Flicek, Soren Germer, Harrison Brand, Ira M Hall, Michael E Talkowski, Giuseppe Narzisi, and Michael C Zody. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440.e19, Sep 2022.

[4] Daniel L Cameron, Leon Di Stefano, and Anthony T Papenfuss. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun*, 10(1):3240, Jul 2019.

[5] 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*, 32(8):1220–2, 04 2016.

[6] Ryan L Collins, Harrison Brand, Konrad J Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C Francioli, Amit V Khera, Chelsea Lowther, Laura D Gauthier, Harold Wang, Nicholas A Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Mark Walker, Christopher W Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M Laricchia, Valentin Ruano-Rubio, Christine Stevens, Namrata Gupta, Caroline Cusick, Lauren Margolin, Genome Aggregation Database Production Team, Genome Aggregation Database Consortium, Kent D Taylor, Henry J Lin, Stephen S Rich, Wendy S Post, Yii-Der Ida Chen, Jerome I Rotter, Chad Nusbaum, Anthony Philippakis, Eric Lander, Stacey Gabriel, Benjamin M Neale, Sekar Kathiresan, Mark J Daly, Eric Banks, Daniel G MacArthur, and Michael E Talkowski. A structural variation reference for medical and population genetics. *Nature*, 581(7809):444–451, May 2020.

[7] Peter Ebert, Peter A Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, Jana Ebler, Weichen Zhou, Rebecca Serra Mari, Feyza Yilmaz, Xuefang Zhao, PingHsun Hsieh, Joyce Lee, Sushant Kumar, Jiadong Lin, Tobias Rausch, Yu Chen, Jingwen Ren, Martin Santamarina, Wolfram Höps, Hufsah Ashraf, Nelson T Chuang, Xiaofei Yang, Katherine M Munson, Alexandra P Lewis, Susan Fairley, Luke J Tallon, Wayne E Clarke, Anna O Basile, Marta Byrska-Bishop, André Corvelo, Uday S Evani, Tsung-Yu Lu, Mark J P Chaisson, Junjie Chen, Chong Li, Harrison Brand, Aaron M Wenger, Maryam Ghareghani, William T Harvey, Benjamin Raeder, Patrick Hasenfeld, Allison A Regier, Haley J Abel, Ira M Hall, Paul Flicek, Oliver Stegle, Mark B Gerstein,

Jose M C Tubio, Zepeng Mu, Yang I Li, Xinghua Shi, Alex R Hastie, Kai Ye, Zechen Chong, Ashley D Sanders, Michael C Zody, Michael E Talkowski, Ryan E Mills, Scott E Devine, Charles Lee, Jan O Korbel, Tobias Marschall, and Evan E Eichler. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 372(6537), 04 2021.

[8] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. Lumpy: a probabilistic framework for structural variant discovery. *Genome Biol*, 15(6):R84, Jun 2014.

[9] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem, 2013.

[10] Dong-Xu Liu, Ramesh Rajaby, Lu-Lu Wei, Lei Zhang, Zhi-Quan Yang, Qing-Yong Yang, and Wing-Kin Sung. Calling large indels in 1047 arabidopsis with indelensembler. *Nucleic Acids Res*, 49(19):10879–10894, Nov 2021.

[11] Ramesh Rajaby and Wing-Kin Sung. Survindel: improving cnv calling from high-throughput sequencing data through statistical testing. *Bioinformatics*, Apr 2019.

[12] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, Sep 2012.

[13] Varuni Sarwal, Sebastian Niehus, Ram Ayyala, Minyoung Kim, Aditya Sarkar, Sei Chang, Angela Lu, Neha Rajkumar, Nicholas Darfci-Maher, Russell Littman, Karishma Chhugani, Arda Soylev, Zoia Comarova, Emily Wesel, Jacqueline Castellanos, Rahul Chikka, Margaret G Distler, Eleazar Eskin, Jonathan Flint, and Serghei Mangul. A comprehensive benchmarking of wgs-based deletion structural variant callers. *Brief Bioinform*, 23(4), Jul 2022.

[14] Brett Trost, Susan Walker, Zhuozhi Wang, Bhooma Thiruvahindrapuram, Jeffrey R MacDonald, Wilson W L Sung, Sergio L Pereira, Joe Whitney, Ada J S Chan, Giovanna Pellecchia, Miriam S Reuter, Si Lok, Ryan K C Yuen, Christian R Marshall, Daniele Merico, and Stephen W Scherer. A comprehensive workflow for read depth-based identification of copy-number variation from whole-genome sequence data. *Am J Hum Genet*, 102(1):142–155, Jan 2018.

[15] Joachim Weischenfeldt, Orsolya Symmons, François Spitz, and Jan O Korbel. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet*, 14(2):125–38, Feb 2013.

[16] Justin M Zook, Nancy F Hansen, Nathan D Olson, Lesley Chapman, James C Mullikin, Chunlin Xiao, Stephen Sherry, Sergey Koren, Adam M Phillippy, Paul C Boutros, Sayed Mohammad E Sahraeian, Vincent Huang, Alexandre Rouette, Noah Alexander, Christopher E Mason, Iman Hajirasouliha, Camir Ricketts, Joyce Lee, Rick Tearle, Ian T Fiddes, Alvaro Martinez Barrio, Jeremiah Wala, Andrew Carroll, Noushin Ghaffari, Oscar L Rodriguez, Ali Bashir, Shaun Jackman, John J Farrell, Aaron M Wenger, Can Alkan, Arda Soylev, Michael C Schatz, Shilpa Garg, George Church, Tobias Marschall, Ken Chen, Xian Fan, Adam C English, Jeffrey A Rosenfeld, Weichen Zhou, Ryan E Mills, Jay M Sage, Jennifer R Davis, Michael D Kaiser, John S Oliver, Anthony P Catalano, Mark J P Chaisson,

Noah Spies, Fritz J Sedlazeck, and Marc Salit. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol*, 38(11):1347–1355, 11 2020.