nature methods

ARTICLES

https://doi.org/10.1038/s41592-018-0001-7

Accurate detection of complex structural variations using single-molecule sequencing

Fritz J. Sedlazeck[1,6*], Philipp Rescheneder[2,6], Moritz Smolka[2], Han Fang[3], Maria Nattestad[3], Arndt von Haeseler[2,4] and Michael C. Schatz[3,5*]

Structural variations are the greatest source of genetic variation, but they remain poorly understood because of technological limitations. Single-molecule long-read sequencing has the potential to dramatically advance the field, although high error rates are a challenge with existing methods. Addressing this need, we introduce open-source methods for long-read alignment (NGMLR; https://github.com/philres/ngmlr) and structural variant identification (Sniffles; https://github.com/fritzsedlazeck/Sniffles) that provide unprecedented sensitivity and precision for variant detection, even in repeat-rich regions and for complex nested events that can have substantial effects on human health. In several long-read datasets, including healthy and cancerous human genomes, we discovered thousands of novel variants and categorized systematic errors in short-read approaches. NGMLR and Sniffles can automatically filter false events and operate on low-coverage data, thereby reducing the high costs that have hindered the application of long reads in clinical and research settings.

Structural variations (SVs), including insertions, deletions, duplications, inversions, and translocations at least 50 bp in size, account for the greatest number of divergent base pairs across human genomes[1]. SVs contribute to polymorphic variation; pathogenic conditions; large-scale chromosome evolution[2]; and human diseases such as cancer[3], autism[4], and Alzheimer's[5]. SVs also affect phenotypes in many other organisms[6–10]. In one of the first reports of SV prevalence, published in 2004, Sebat et al.[11] discovered in a microarray study that large-scale copy-number polymorphisms are common across healthy human genomes. Today, SV detection most often uses short paired-end reads. Copy-number variations are observed as decreases (deletions) or increases (amplifications) in aligned read coverage[12], and other types of SVs are identified by the arrangement of paired-end reads or split-read alignments[13–16]. Short-read approaches, however, have been reported to lack sensitivity (only 10%[17] to 70%[6,8] of SVs detected), exhibit very high false positive rates (up to 89%)[6,18–21], and misinterpret complex or nested SVs[6,22].

Long-read single-molecule sequencing has the potential to substantially increase the reliability and resolution of SV detection. With average read lengths of 10 kbp or higher, the reads can be more confidently aligned to repetitive sequences that often mediate the formation of SVs[22]. Long reads are also more likely to span SV breakpoints with high-confidence alignments. Despite these advantages, long reads introduce new challenges. Most important, they have a high sequencing error rate—currently 10–15% for Pacific Biosciences (PacBio) and 5–20% for Oxford Nanopore sequencing[23]—which necessitates new methods. A few aligners are available, including LAST[24], BlasR[25], BWA-MEM[26], GraphMap[27], MECAT[28], and minimap2[29]. Only one stand-alone method, PBHoney[18], is available to detect all types of SVs from long-read data, although others such as SMRT-SV[30] have been proposed for a subset of SV types.

To address these challenges, we introduce two open-source algorithms, NGMLR and Sniffles, for comprehensive long-read alignment and SV detection (Fig. 1). NGMLR is a fast and accurate aligner for long reads based on extension of our previous short-read aligner, NGM[31], with a new convex gap-cost scoring model to align long reads across SV breakpoints. Sniffles successively scans the alignments to identify all types of SVs. Its SV-scoring scheme evaluates candidate SVs on the basis of their size, position, type, coverage, and breakpoint consistency, and thus overcomes the high insertion/deletion (indel) error rates in long-read sequencing.

We applied our methods to simulated and genuine datasets for *Arabidopsis*, healthy human genomes, and a cancerous human genome to demonstrate their increased accuracy compared with that of alternate short- and long-read callers. A particularly innovative feature of Sniffles is its ability to detect nested SVs, such as inverted tandem duplications (INVDUPs) and inversions flanked by indels (INVDELs). These are poorly studied classes of SVs; although both have been previously associated with genomic disorders[32–35], they could not be routinely detected, and so their full significance is currently unknown. Finally, we show that our methods reduce the sequencing and computational costs required per sample, and thus make the application of long reads to large numbers of samples increasingly feasible.

Results

Accurate mapping and detection of SVs with long reads. Unlike most aligners, NGMLR uses a convex gap-scoring model[36] to accurately align reads spanning genuine indels in the presence of small observed indels (1–10 bp) that commonly occur as sequencing errors (Fig. 2, Methods, and Supplementary Note 1). Larger or more complex SVs are captured through split-read alignments. To achieve both high performance and accuracy, NGMLR first partitions the long reads into 256-bp subsegments and aligns them independently to the reference

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [2]Center for Integrative Bioinformatics Vienna, Max F. Perutz Laboratories, University of Vienna, Medical University of Vienna, Vienna, Austria. [3]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [4]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria. [5]Departments of Computer Science and Biology, Johns Hopkins University, Baltimore, MD, USA. [6]These authors contributed equally: Fritz J. Sedlazeck, Philipp Rescheneder. *e-mail: fritz.sedlazeck@bcm.edu; mschatz@cs.jhu.edu

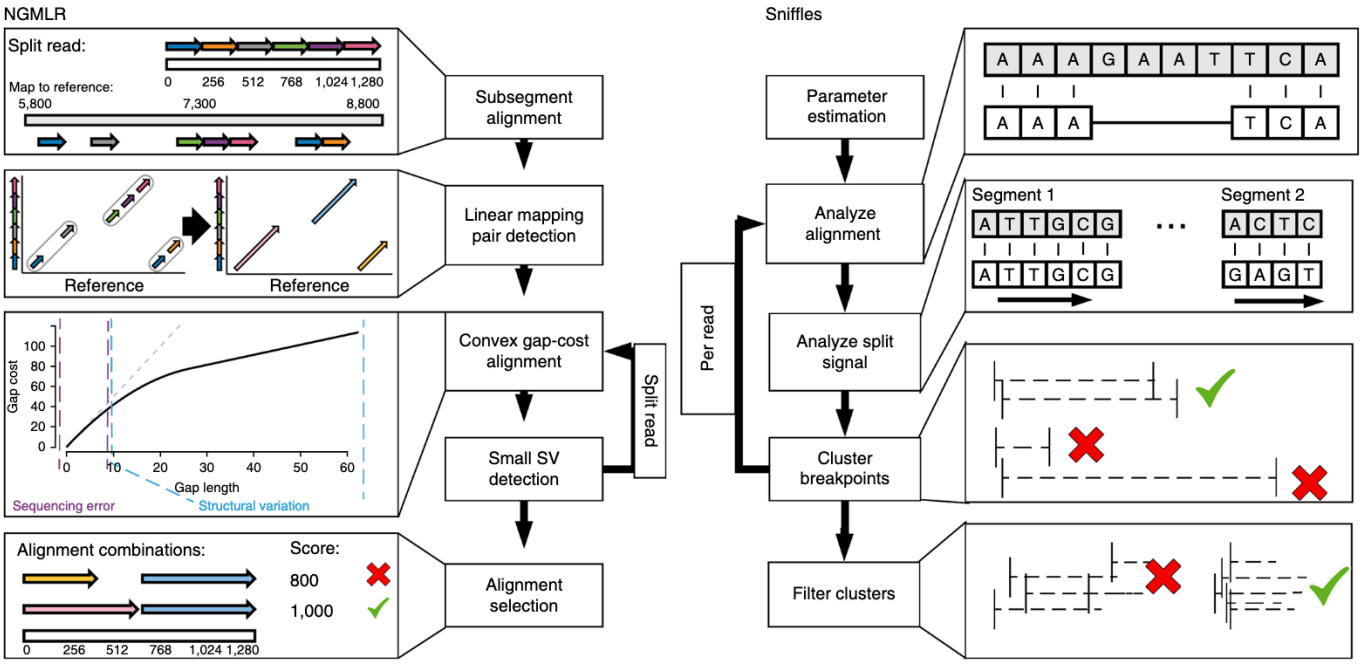

Fig. 1 | The main steps implemented in NGMLR and Sniffles. Additional details are presented in Supplementary Notes 1 and 2.

genome (Fig. 1). It then groups colinear subsegment alignments into long segments, which are aligned by dynamic programming with our convex gap-cost scoring scheme. Finally, NGMLR selects the highest-scoring nonoverlapping combination of segments per read and outputs the results in standard SAM/BAM format, thus allowing visualization in IGV[37] and other downstream processing.

Sniffles detects indels, duplications, inversions, translocations, and nested events and can be used with any aligner, although it performs best with NGMLR, as it produces the most accurate alignments. The principal steps consist of scanning the alignments of each read independently for potential SVs and then clustering the candidate SVs across all reads (Fig. 1, Methods, and Supplementary Note 2). Sniffles uses both within-alignment and split-read information to detect SVs, as small indels can be spanned within a single alignment, but large or complex events lead to split-read alignments. The major advance of Sniffles is its ability to filter false SV signals from the noisy reads. As with other variant detectors, minimum read support (default: 10 reads) is a critical feature, but Sniffles also considers the consistency of the breakpoint position and size. In addition, it can perform read-based phasing of SVs and report adjacent or nested events in the output VCF (Variant Call Format) file.

To establish performance, we benchmarked NGMLR and Sniffles against widely used alternative approaches, using simulated reads supplemented with SVs of different sizes and types (Supplementary Notes 3–5). We observed that NGMLR and Sniffles were among the fastest available methods and also showed the highest accuracy for alignments and SV calls overall (Fig. 3). In addition, when we used genuine sequencing reads mapped to a modified reference genome with SVs embedded at known locations, we observed similarly superior performance (Supplementary Note 5).

We next used *Arabidopsis thaliana* trio (Col-0, CVI, and Col-0×CVI F1 progeny)[38] and Ashkenazi human trio sequencing data from Genome-in-a-Bottle (GiaB)[39] to assess recall and Mendelian consistency (Table 1 and Supplementary Notes 4 and 5). Sniffles and NGMLR had the highest recall rates, that is, the percentage of homozygous parental variants found in F1 (*Arabidopsis* trio, 99.7%; GiaB trio, 97.2%). Using NGMLR or Sniffles with PacBio reads resulted in Mendelian discordance rates of 3.4% for *Arabidopsis* and 5.6% for GiaB, whereas state-of-the-art consensus calling with Illumina data gave a 21.1% discordance rate. Translocation calls were particularly erroneous for the short-read analysis and unreasonably high in number (1,550) in the son.

Comparison of PacBio and Oxford Nanopore sequencing. As a new technology, Oxford Nanopore sequencing has not yet been extensively tested for SV analysis, especially in human genomes. We investigated this in the well-studied NA12878 human genome using three datasets: 28× coverage Oxford Nanopore data (releases 3 and 4 from Jain et al.[40]) analyzed with NGMLR and Sniffles, 55× coverage PacBio data[41] analyzed with NGMLR and Sniffles, and 50× coverage Illumina data[42] analyzed by the consensus caller SURVIVOR, which incorporates Delly, Lumpy, and Manta (Table 1). We also compared our results to a published GiaB indel call set based on PacBio sequencing[41] and an Illumina-based deletion-only call set from the 1000 Genomes Project[6].

Sniffles identified a total of 15,499 SVs from the PacBio reads and 26,657 SVs from the Oxford Nanopore reads, whereas SURIVOR reported 7,275 SVs (Table 1, Supplementary Table 7). Together, the five call sets yielded a total of 40,601 SVs (Table 1, Supplementary Note 4). The majority (24,392) of the identified SVs were present in only one call set, whereas 16,209 SVs were identified in two or more call sets. Of the 15,499 PacBio calls, most (94.8%) were confirmed by Oxford Nanopore, Illumina, or the existing call sets. Oxford Nanopore had substantially worse concordance, as Sniffles reported 11,433 calls unique to Oxford Nanopore, of which 10,977 (96.0%) were deletions and the majority (92.9%) were within homopolymers or other simple repeats. In contrast, the 773 calls found only by PacBio were mainly insertions (66.5%), and only 323 (41.8%) overlapped homopolymers or repeats. This systematic bias for deletions in the Oxford Nanopore data is most likely due to base-calling errors, as also reported by Jain et al.[40]. The majority of these artifacts are small deletions; because it increased the minimum SV size to 200 bp, Sniffles reported only 38.6% of the SV calls within homopolymers and low-complexity regions. Illumina-based SV calling had relatively low concordance with alternative approaches, and 49.7% of Illumina calls were unique to the technology. The majority

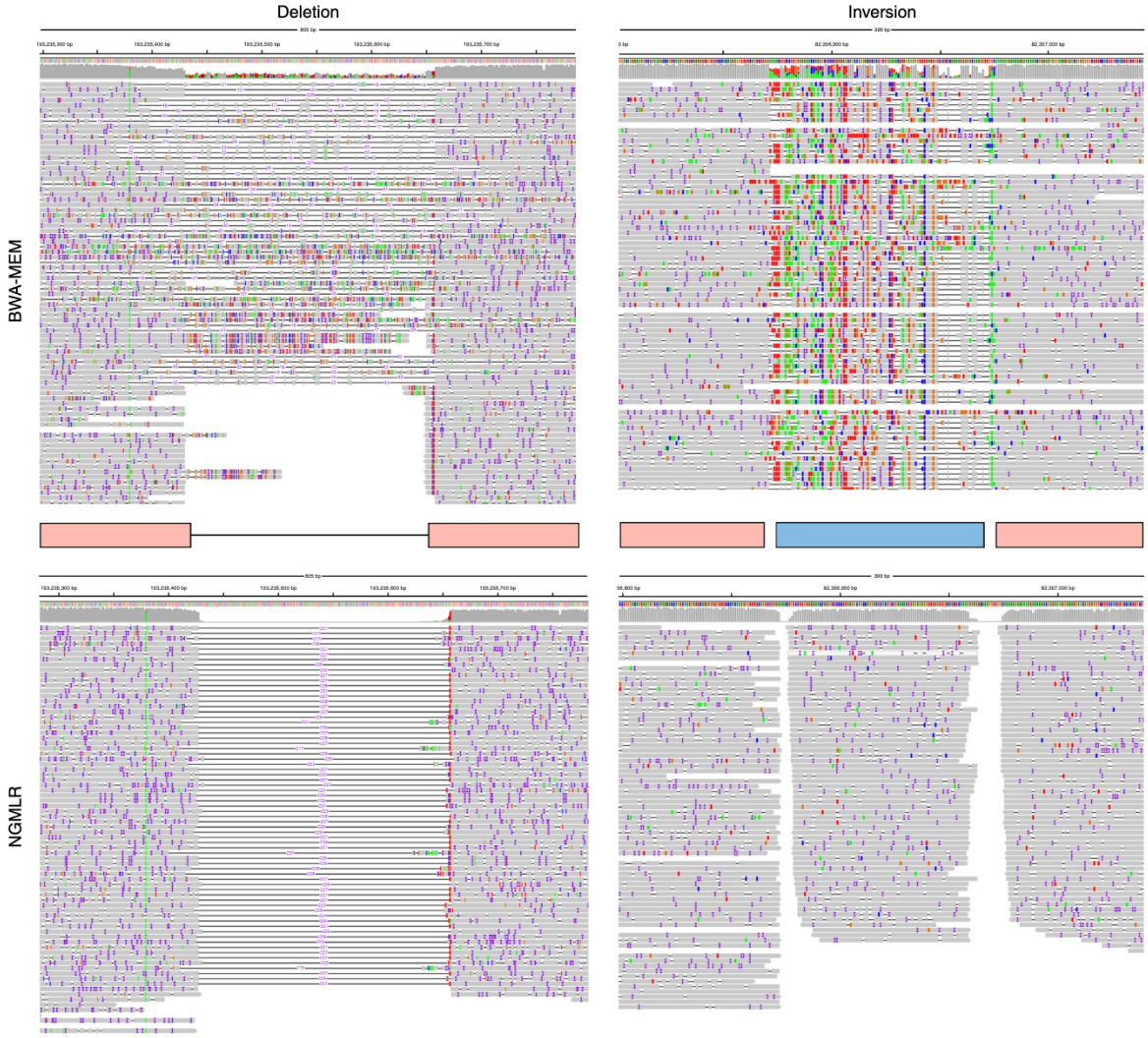

Fig. 2 | Improved alignment by NGMLR for a 228-bp deletion and a 150-bp inversion. BWA-MEM alignments (upper tracks; as shown in Integrative Genomics Viewer) indicate these events but do not localize the precise event and breakpoints. With the improved alignments from NGMLR, Sniffles can precisely pinpoint the location and type of the SV.

(54.1%) of unique calls were translocation events, and most of these appeared to be false positives (see below).

Investigation of unique short-read versus long-read events. Over all datasets, Sniffles detected far more indels than the short-read-based callers (Table 1). Conversely, short reads led to the detection of, on average, 27-fold more translocation events than detected by Sniffles in presumably healthy human datasets. We used the NA12878 genome to investigate these discrepancies.

We first investigated small insertion (50–300 bp) and deletion (50 bp–3 kbp) calls from Sniffles, using the Illumina reads as orthogonal evidence (Supplementary Note 4). We assumed that these size ranges should be captured well by the paired-end Illumina data, and we used the compression–expansion statistic[43] as an unbiased measure of the Illumina placements near predicted indels. True insertions should cause BWA-MEM to map read pairs closer than expected, and deletions farther away, with respect to the average Illumina insert size of 311 bp observed genome-wide. Using the Illumina data and a P value threshold of 0.01 (two-sided, one-sample t-test), we confirmed 3,415 (PacBio) and 3,879 (Oxford Nanopore) deletions and 2,685 (PacBio) and 1,703 (Oxford Nanopore) insertions reported by Sniffles (Supplementary Table 8). We used a SURVIVOR analysis for comparison, which confirmed 1,873 deletions and showed significant alteration in only 10% of randomly selected regions.

Next, we investigated the large number of translocations reported in the Illumina-based consensus calls (2,247) compared with those for Sniffles (PacBio, 119; Oxford Nanopore, 43) (Supplementary Note 4 and Supplementary Table 9). We noted a large overlap (48.9%) of the Illumina-based translocation sites with insertion calls from Sniffles with both long-read technologies. In one representative example, a long-read-mapped insertion fell within a low-complexity region, causing short reads to be mismapped and misreported as translocations, even when low-mapping-quality (<20) reads were excluded (Fig. 4a). In total, we were able to rule out 1,869 (83.2%) of the Illumina-based translocation calls, with most overlapping an insertion (48.9%) and others overlapping a deletion (8.9%), or other SV types (1.2%). The remaining Illumina-based translocation calls were also questionable, with 404 (18.0%) in low-complexity regions and 141 (6.3%) in regions with abnormally high coverage and without any evidence in the long-read data. Inversions showed a similar pattern: 60% of calls overlapped with a different SV type identified by long reads (Fig. 4b) or aligned to low-complexity sequences.

The majority of PacBio-based indel calls from Sniffles were thus validated by either the Oxford Nanopore or the

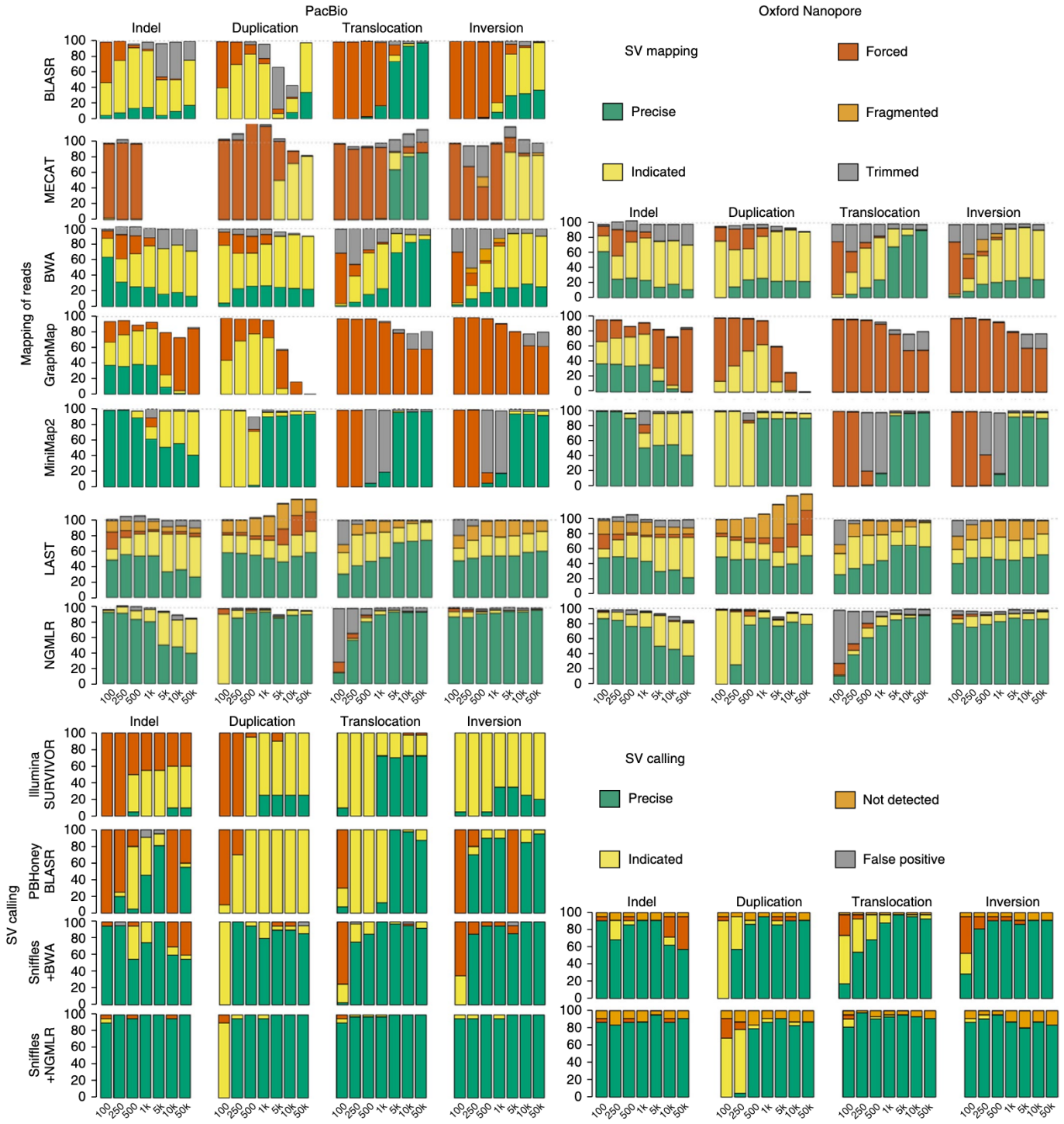

Fig. 3 | Tool evaluation with simulated data. In each plot, the *x*-axis indicates the size of the 840 simulated SVs. For read alignment (top), we simulated PacBio-like (left) and Oxford Nanopore–like reads (right) and determined whether alignments were precise, indicated, forced, unaligned, or trimmed but not aligned through the SV. For the SV analysis (bottom), we used the same alignments as before and distinguished among precise, indicated, not indicated, and false positive calls.

Illumina paired-end reads. In contrast, the majority of calls unique to the Illumina-based methods were false, especially translocations caused by mismapped reads across insertions.

Detection of nested SVs. Next, we investigated the performance of Sniffles on complex, nested SV types such as INVDUPs and INVDELs. These variant types are poorly studied but have been associated with a number of diseases, including Pelizaeus–Merzbacher disease[33] and other diseases[32,44] for INVDUPs, and hemophilia A genetic deficiency for INVDELs by long-range PCR[35].

To start, we simulated 280 nested SVs of different sizes and types in the human genome, along with simulated PacBio-like, Oxford Nanopore–like, and Illumina-like reads (Fig. 5 and Supplementary Table 2). We evaluated each SV separately; for

Table 1 | Summary of detected SVs across 15 different datasets

Dataset	Techique	Coverage (×)	Average read length (bp)	Total SVs	DEL	DUP	INS	INV	TRA
Arabidopsis Col-0	PacBio	127	6,482	355	67	63	106	68	51
Arabidopsis CVI	PacBio	123	6,073	9,652	3,822	904	1,823	478	2,625
Arabidopsis Col-0 × CVI F1	PacBio	155	11,206	11,935	4,974	582	4,049	567	1,763
Arabidopsis Col-0 × CVI F1	Illumina	40	250	10,950	4,324	643	0	671	5,312
GiaB HG002 (son)	PacBio	69	8,540	19,131	7,957	1,084	9,656	232	202
GiaB HG002 (son)	Illumina	80	148	10,822	5,018	863	0	823	4,118
GiaB HG003 (father)	PacBio	32	6,284	11,964	5,296	408	6,048	99	113
GiaB HG003 (father)	Illumina	80	148	11,395	5,553	869	0	818	4,155
GiaB HG004 (mother)	PacBio	30	7,285	10,463	4,590	276	5,436	93	68
GiaB HG004 (mother)	Illumina	80	148	8,901	5,000	868	0	829	2,204
NA12878 (healthy female)	PacBio	55	4,334	15,499	6,734	606	7,880	160	119
NA12878 (healthy female)	Oxford Nanopore	28	6,432	26,657	19,074	761	6,376	334	112
NA12878 (healthy female)	Illumina	50	101	7,275	3,744	553	0	731	2,247
SKBR3 (breast cancer)	PacBio	69	9,872	19,165	7,268	1,019	10,391	328	159
SKBR3 (breast cancer)	Illumina	25	101	5,046	2,776	483	0	627	1,160

SVs were reported with a minimum size of 50 bp using SURVIVOR based on Delly, Lumpy, and Manta for Illumina or Sniffles for PacBio (minimum 10 reads) or Oxford Nanopore (minimum 5 reads) owing to the lower coverage (see Supplementary Table 5 for full details of all datasets used). DEL, deletion; DUP, duplication; INS, insertion; INV, inversion; TRA, translocation.

example, an inversion flanked by two deletions was evaluated as three SVs. Sniffles detected the full range of types owing to its dynamic splitting of events, and precisely called 67.9% of the nested SVs (Supplementary Note 2). This included SVs that were larger than the read length, which highlights Sniffles's ability to accurately infer complex events. With Oxford Nanopore–like reads, Sniffles was able to precisely call 67.3% of SVs on average over INVDEL and INVDUP events of different lengths. The other methods were not able to identify the full complexity of these events, and only partially called the SVs (for example, an inversion without the flanking deletions).

To highlight this capability in real data, we examined a PacBio-based dataset for the SKBR3 breast cancer cell line[45]. Sniffles and NGMLR revealed 15 gene fusions created by as many as three chained events, which were all validated by PCR. In one example (Fig. 5), short reads indicated an inversion, but the poor resolution made it impossible to detect and interpret the entire event. In contrast, Sniffles detected an INVDEL and an INVDUP, and

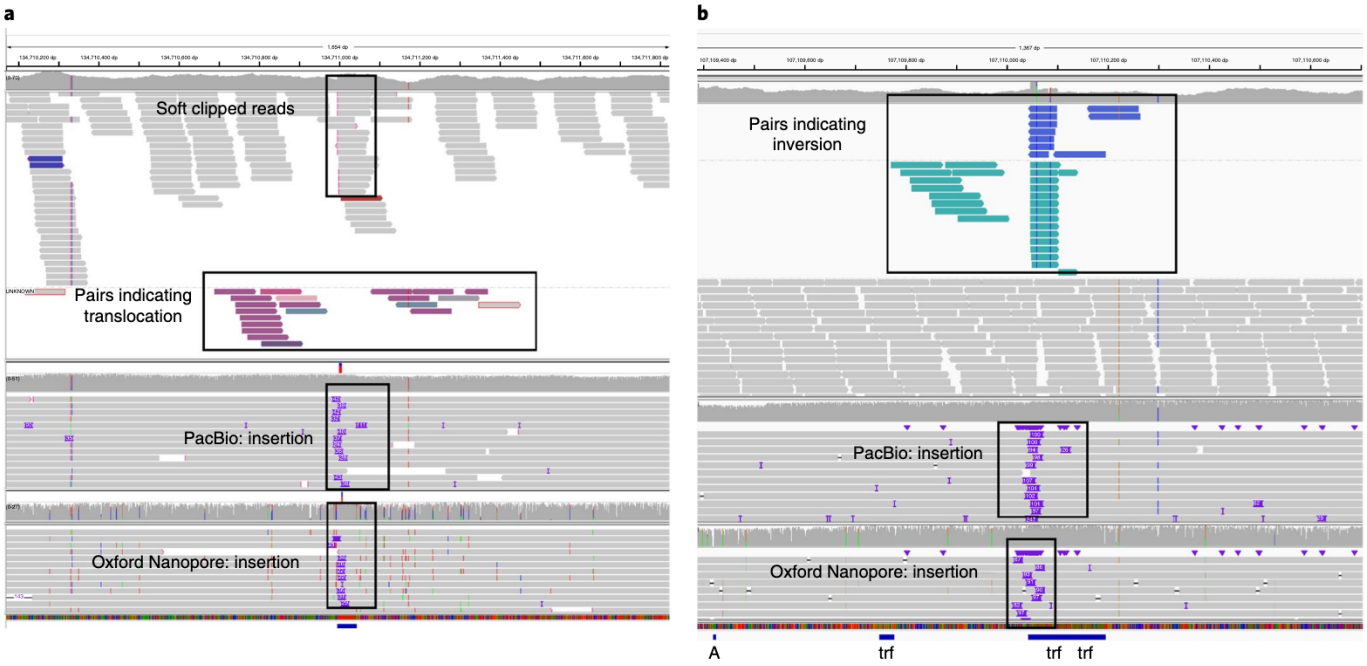

Fig. 4 | Systematic error in short-read-based SV calling. a, Example of a putative translocation identified in short-read data (top alignments) that overlaps an insertion detected by both PacBio (middle) and Oxford Nanopore sequencing (bottom). **b**, Example of a putative inversion identified in the short-read data (top) that overlaps an insertion detected in both PacBio (middle) and Oxford Nanopore reads (bottom).

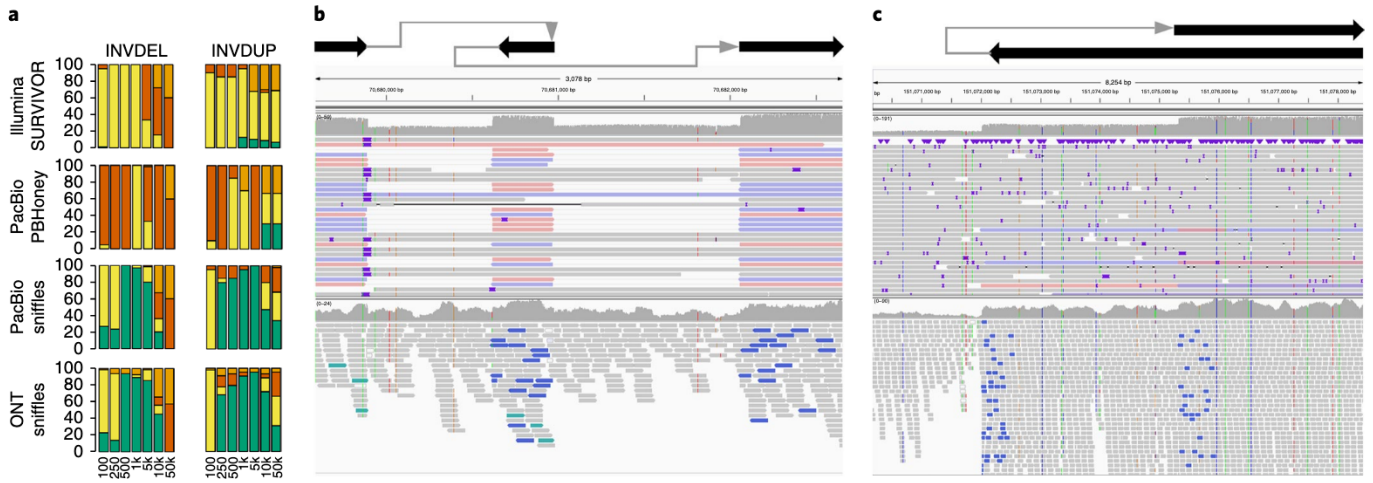

Fig. 5 | Nested SVs in the SKBR3 cancer cell line. a, Evaluation of Sniffles and NGMLR with simulated data to identify nested SVs. **b**, A 3-kb region including two deletions flanking an inverted sequence is clearly visible and was detected by Sniffles with NGMLR (top), but was not detected by the Illumina methods (bottom). **c**, The start of an inverted duplication. Breakpoints were reported by Sniffles as the start of an inverted duplication (top) but were not correctly detected by short-read methods (bottom).

the read phasing allowed the complex regions to be fully resolved (Fig. 5 and Supplementary Note 2). Although these were the only two nested types in this sample, Sniffles is capable of detecting and reporting any combination of SVs based on the IDs assigned in the reported VCF file.

How much coverage is required?. Finally, we assessed how much coverage is required to detect SVs. This is an important consideration because long-read technologies are more expensive than short-read methods for the same coverage[23]. From a purely statistical standpoint, about 10× coverage should be sufficient to infer all SV breakpoints using 10-kbp reads, whereas about 25× coverage is needed for 2 × 100-bp short reads (Fig. 6a and Supplementary Note 4). However, this represents an idealized case (e.g., a lack of repeats or coverage biases) that underestimates the required coverage.

To investigate this, we subsampled reads from the NA12878 PacBio and Oxford Nanopore datasets and the more complex SKBR3 PacBio sample to 5×, 10×, 15×, 20×, and 30× coverage. We analyzed these subsets with NGMLR and Sniffles with different parameters (-s 1 to -s 10) to vary the minimum number of reads, and measured precision and recall with respect to the full-coverage dataset (Fig. 6b–d). As expected, using a minimum support of only one or two reads led to many false positives.

When we focused on settings with a precision rate of 80% or higher, we found that 15× PacBio read coverage led to recall of 69.6% and 67.2% for NA12878 and SKBR3 for homozygous and heterozygous SVs of any type, respectively (Fig. 6b,d). The difference in recall was largely due to the complexity of the SKBR3 cancer sample, which has some extreme copy amplifications (>20-fold). When we increased the coverage to 30×, Sniffles showed 80.0% and 76.6% recall with a precision of ~85% for NA12878 and SKBR3, respectively.

For the Oxford Nanopore NA12878 dataset, the highest recall rate (84.2%) had a precision of 82.2% for 20× coverage (Fig. 6c). This higher apparent accuracy is largely attributable to the fact that the original dataset has only 28× coverage, so this constitutes a less dramatic downsampling. We observed a greater loss in precision than with the PacBio data, due to the stringent minimum number of supporting reads (-s 10) used throughout the study. Overall, our analysis shows that NGMLR and Sniffles can detect the vast majority of heterozygous and homozygous SVs with only a fraction of the original coverage.

Discussion

NGMLR and Sniffles enable an unprecedented view of SVs with long-read sequencing, by outperforming existing tools in terms of both sensitivity and specificity with simulated and real data. In particular, we demonstrated that they can overcome the sensitivity issues reported for short-read callers, which miss 30%[6,8] to 90%[17] of SVs. This allowed us to detect many thousands of additional variants beyond what has been reported for large-scale short-read sequencing projects such as the 1000 Genomes Project. Indeed, prototype versions of our methods were recently used to identify the causal, pathogenic SV in a person with multiple neoplasia and cardiac myxomata[46]. We also used the long-read data to identify systematic errors in short-read SV analysis, for which the vast majority (>85%) of identified translocations were false positives due to mismapped reads.

The identification of SVs from long reads is challenging chiefly because of the high underlying error rates. In addition to numerous small indels, we discovered that PacBio introduces larger false insertions at a low but noticeable rate (Supplementary Note 2). We control for this artifact by requiring that the size and composition of candidate SVs be consistent across the spanning reads. Within the Oxford Nanopore dataset, we highlighted systematic artifacts in base-calling that generate deletions in low-complexity repeats. Although we fully expect accuracy to improve with improved base-calling, it is currently necessary to exclude most small SV calls when using Oxford Nanopore sequencing. Beyond sequencing errors, alignment artifacts can lead to miscalled SVs. For example, some long-read mappers falsely align reads through an SV without indication of the underlying event. Although Sniffles recognizes the increase in mismatches, NGMLR alignments correct these issues more directly. Finally, we showed a deficiency in the detection of nested variations such as INVDUPs and INVDELs for all methods except Sniffles. Several diseases are already known to be associated with these SV types, and we expect that their importance will increase as more are detected.

A key remaining barrier to routine analysis of SVs across large numbers of samples is cost. Long-read technologies are becoming less expensive every year, but they remain more expensive than short-read sequencing[23]. We addressed this by showing that high accuracy is possible with as little as 15× to 30× coverage for healthy or cancerous human genomes. These requirements will be further reduced as read lengths increase and error rates drop. We expect

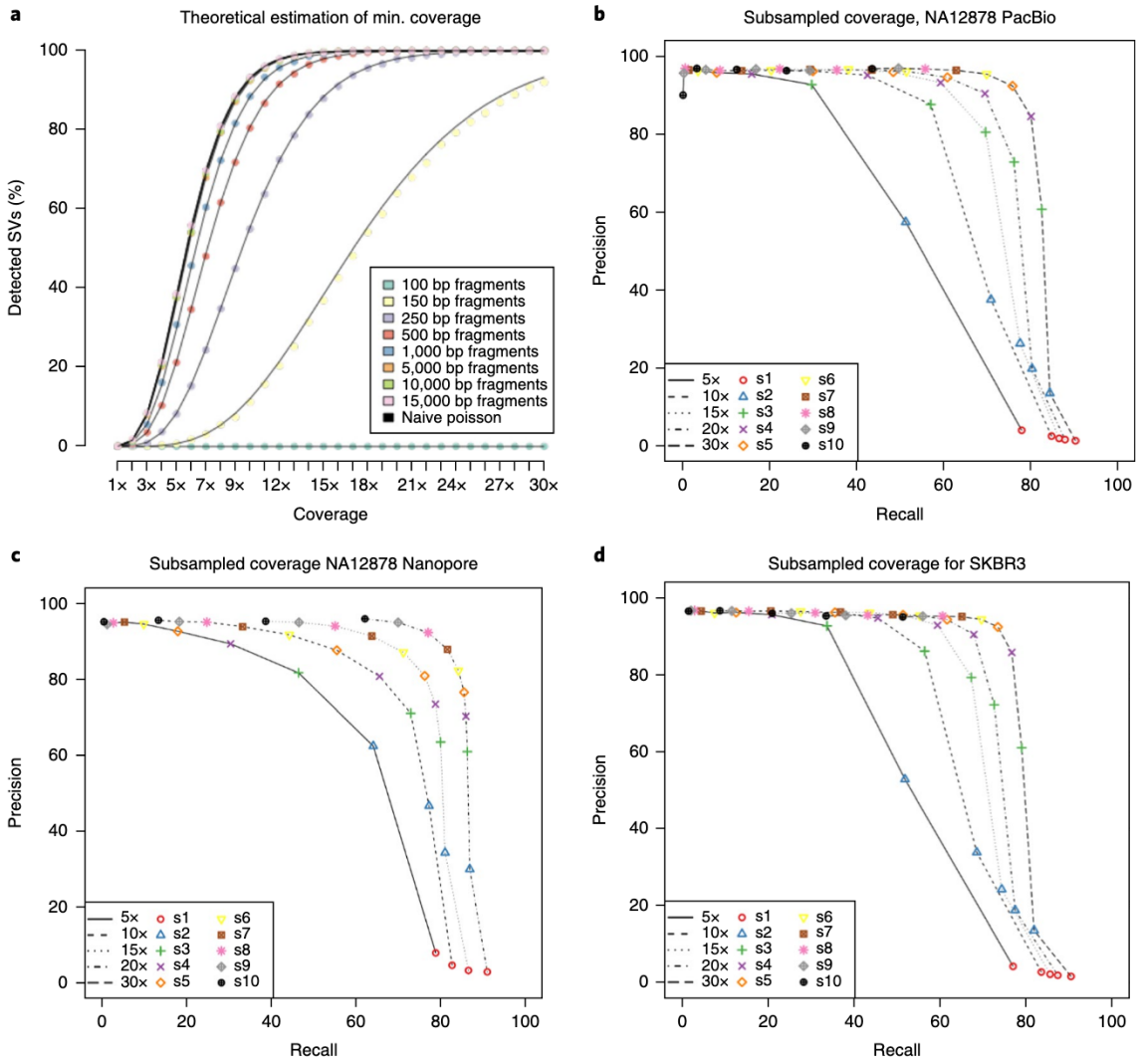

Fig. 6 | Dependence of SV detection accuracy on the level of coverage. a, Theoretical assessment of recall versus coverage for different read lengths requiring 50-bp overlap of each breakpoint for SV events. Min., minimum. **b**, Subsampling results for the 55× PacBio NA12878 data. **c**, Subsampling results for 28× Oxford Nanopore NA12878 data. **d**, Subsampling results for the 70× PacBio SKBR3 breast cancer cell line dataset. For **b**–**d**, Sniffles and NGMLR were run on subsampled data (rates are indicated by lines) and using different thresholds for Sniffles (s: 1–10, indicated by symbols and color-coding). With every dataset we were able to demonstrate success for Sniffles using NGMLR with only 10–30× coverage, which recovered around 80% of the calls with precision of ~80% or higher.

that these improvements, aided by NGMLR and Sniffles, will usher in a new era of high-quality genome sequences for a broad range of research and clinical applications.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41592-018-0001-7.

Received: 11 August 2017; Accepted: 16 March 2018;
Published online: 30 April 2018

References

1. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013).
2. Lupski, J. R. Structural variation mutagenesis of the human genome: impact on disease and evolution. *Environ. Mol. Mutagen.* **56**, 419–436 (2015).
3. Macintyre, G., Ylstra, B. & Brenton, J. D. Sequencing structural variants in cancer for precision therapeutics. *Trends Genet.* **32**, 530–542 (2016).
4. Hedges, D. J. et al. Evidence of novel fine-scale structural variation at autism spectrum disorder candidate loci. *Mol. Autism* **3**, 2 (2012).
5. Rovelet-Lecrux, A. et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat. Genet.* **38**, 24–26 (2006).
6. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
7. Dennenmoser, S. et al. Copy number increases of transposable elements and protein-coding genes in an invasive fish of hybrid origin. *Mol. Ecol.* **26**, 4712–4724 (2017).
8. Jeffares, D. C. et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
9. Zichner, T. et al. Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res.* **23**, 568–579 (2013).
10. Imprialou, M. et al. Genomic rearrangements in *Arabidopsis* considered as quantitative traits. *Genetics* **205**, 1425–1441 (2017).
11. Sebat, J. et al. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).

12. Kadalayil, L. et al. Exome sequence read depth methods for identifying copy number changes. *Brief. Bioinform.* **16**, 380–392 (2015).
13. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
14. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
15. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
16. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
17. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
18. English, A. C., Salerno, W. J. & Reid, J. G. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* **15**, 180 (2014).
19. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
20. Tattini, L., D'Aurizio, R. & Magi, A. Detection of genomic structural variants from next-generation sequencing data. *Front. Bioeng. Biotechnol* **3**, 92 (2015).
21. Teo, S. M., Pawitan, Y., Ku, C. S., Chia, K. S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28**, 2711–2718 (2012).
22. Lucas Lledó, J. I. & Cáceres, M. On the power and the systematic biases of the detection of chromosomal inversions by paired-end genome sequencing. *PLoS One* **8**, e61292 (2013).
23. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
24. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
25. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
26. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* Preprint at https://arxiv.org/abs/1303.3997 (2013).
27. Sović, I. et al. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat. Commun.* **7**, 11307 (2016).
28. Xiao, C. L. et al. MECAT: fast mapping, error correction, and *de novo* assembly for single-molecule sequencing reads. *Nat. Methods* **14**, 1072–1074 (2017).
29. Li, H. Minimap2: fast pairwise alignment for long nucleotide sequences. *arXiv* Preprint at https://arxiv.org/abs/1708.01492 (2017).
30. Chaisson, M. J. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
31. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29**, 2790–2791 (2013).
32. Carvalho, C. M. et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* **43**, 1074–1081 (2011).
33. Shimojima, K. et al. Pelizaeus-Merzbacher disease caused by a duplication-inverted triplication-duplication in chromosomal segments including the PLP1 region. *Eur. J. Med. Genet.* **55**, 400–403 (2012).
34. Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
35. Mühle, C., Zenker, M., Chuzhanova, N. & Schneider, H. Recurrent inversion with concomitant deletion and insertion events in the coagulation factor VIII gene suggests a new mechanism for X-chromosomal rearrangements causing hemophilia A. *Hum. Mutat.* **28**, 1045 (2007).
36. Gusfield, D. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge Univ. Press, Cambridge, UK, 1997).
37. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
38. Chin, C. S. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* **13**, 1050–1054 (2016).
39. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
40. Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
41. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
42. Eberle, M. A. et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.* **27**, 157–164 (2017).
43. Zimin, A. V., Smith, D. R., Sutton, G. & Yorke, J. A. Assembly reconciliation. *Bioinformatics* **24**, 42–45 (2008).
44. Beri, S., Bonaglia, M. C. & Giorda, R. Low-copy repeats at the human VIPR2 gene predispose to recurrent and nonrecurrent rearrangements. *Eur. J. Hum. Genet.* **21**, 757–761 (2013).
45. Nattestad, M. et al. Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *bioRxiv* Preprint at https://www.biorxiv.org/content/early/2017/08/10/174938 (2017).
46. Merker, J. D. et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet. Med.* **20**, 159–163 (2017).

Acknowledgements

We thank W.R. McCombie, S. Wheelan, S. Goodwin, H. Li, and B.Q. Minh for helpful discussions. This work was supported by the National Science Foundation (DBI-1350041, IOS-1732253, and IOS-1445025 to M.C.S.) and the US National Institutes of Health (R01-HG006677 and UM1 HG008898 to M.C.S. and F.J.S.). P.R. acknowledges support from DK RNA Biology (W1207-B09). A.v.H. and M.S. acknowledge financial support from the University of Vienna and the Medical University of Vienna.

Author contributions

F.J.S., P.R., and M.S. developed the software. F.J.S., P.R., M.S., H.F., and M.N. performed analysis. F.J.S., P.R., M.C.S., and A.v.H. wrote the manuscript. M.C.S. and A.v.H. directed the project. All authors read and approved the final manuscript.

Competing interests

M.C.S. and F.J.S. have participated in PacBio-sponsored meetings over the past few years and have received travel reimbursement and honoraria for presenting at these events. Since the initial submission of this paper, P.R. has become an employee of Oxford Nanopore. PacBio and Oxford Nanopore had no role in decisions related to the study/work, data collection, or analysis of data described in this paper.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-018-0001-7.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to F.J.S. or M.C.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

NGMLR: fast, accurate mapping of long single-molecule reads. NGMLR is designed to accurately map long single-molecule sequencing reads from either PacBio or Oxford Nanopore sequencing to a reference genome, with the goal of enabling precise SV calls. We use terminology as in the SAM specification[47], whereby a read mapping consists of either one linear alignment covering the full read length or multiple linear alignments covering nonoverlapping segments of the read (i.e., split reads).

The main challenge when mapping high-error long reads is to evaluate whether a read should be mapped to the reference genome with one linear alignment, or must be split. For example, the correct mapping for a read that spans an inversion can be found only when the read is split into three segments. Conversely, reads that do not span an SV should always be mapped with a single linear alignment. However, error rates are high and are not always uniform, with some regions having an error rate of over 30%. These segments can cause read mappers to falsely split a read. Furthermore, the high indel sequencing error of long-read technologies causes current read aligners to falsely split large SVs into several smaller ones, and makes it difficult to detect exact break points.

To address these challenges, NGMLR implements the following workflow (Fig. 1):

1. NGMLR identifies subsegments of the read and of the reference genome that show high similarity and can be aligned with a single linear alignment. These segments can contain small indels but must not span a larger SV breakpoint. In reference to BLAST's high-scoring segment pairs, we call these segment linear mapping pairs.
2. For each linear mapping pair, NGMLR extracts the read sequence and the reference sequence and uses the Smith–Waterman algorithm to compute a pairwise sequence alignment using a convex gap-cost model that accounts for sequencing error and SVs at the same time.
3. NGMLR scans the sequence alignments for regions of low sequence identity to identify small SVs that were missed in steps 1 and 3.
4. Finally, NGMLR selects the set of linear alignments with the highest joint score, computes a mapping quality (MQ) for each alignment, and reports these values as the final read mapping in a SAM/BAM file.

Convex scoring model. When aligning high-error long reads, it is crucial to choose an appropriate gap model, because there are two sources of indels. Sequencing error predominantly causes very short, randomly distributed indels (1–5 bp), whereas longer indels (20+ bp) are caused by genomic SVs. A linear gap model appropriately models indels that originate from sequencing error but cannot account for longer indels from genomic variation, as these large blocks occur as a single unit, rather than as combinations of multiple single-base indels. With affine gap models, the gap-open penalty falsely causes short indels from sequencing error to cluster together for noisy long reads, and has only minimal effect on longer indels, especially in repetitive regions of the genome. With the convex scoring model of NGMLR, extension of an indel is penalized proportionally less the longer the indel is. Therefore, the convex scoring model encourages large alignment gaps, such as those that occur from an SV, to be grouped together into contiguous stretches (extension of a large indel has a relatively low cost), whereas small indels, which commonly occur as sequencing errors, remain separate (costs are almost the same for extension of a 1-bp gap and opening of a new gap).

Using a convex gap model to compute optimal alignments increases the computation time substantially, as each cell in the alignment matrix depends not only on three other cells but also on the full row and column it is located in[36]. This would make it infeasible to use convex gap costs to align large long-read datasets, so we adapted a heuristic implementation of the convex gap-cost algorithm found in the swalign library (https://github.com/mbreese/swalign). Instead of scanning the full cell and row while filling the alignment matrix, we use two additional matrixes to store indel length estimations for each cell. Furthermore, we use the initial subsegment alignments to identify the part of the alignment matrix that is most likely to contain the correct alignment and skip all other cells of the matrix during alignment computation (Supplementary Note 1).

Sniffles: robust detection of structural variations from long-read alignments. Sniffles operates within and between the long-read alignments to infer SVs. It applies five major steps (Fig. 1):

1. Sniffles first estimates the parameters to adapt itself to the underlying dataset, such as the distribution in alignment scores and distances between indels and mismatches on the read, as well as the ratios of the best and second-best alignment scores.
2. Sniffles then scans the read alignments and segments to determine whether they potentially represent SVs.
3. Putative SVs are clustered and scored according to the number of supporting reads, the type and length of the SV, the consistency of the SV composition, and other features.
4. Sniffles optionally genotypes the variant calls to identify homozygous and heterozygous SVs.
5. Sniffles optionally provides a clustering of SVs based on the overlap with the same reads, especially to detect nested variants.

Details on each step are included in Supplementary Note 2. In the following subsections, we focus on the methods that are unique to Sniffles, which are the detection and analysis of alignment artifacts to reduce falsely called variants and the clustering of variants.

Putative variant scoring. The high error rate of long reads induces many alignments that falsely appear as SVs. Sniffles addresses these by scoring each putative variant on the basis of several characteristics that we have determined to be the most relevant for detecting SVs. The two main user thresholds are the number of high-quality reads supporting the variant (set using the -s parameter) and the s.d. of the coordinates in the start and stop breakpoints across all supporting reads. The minimum variant size reported defaults to 50 bp but can be adjusted using the -l parameter. To account for additional noise in the data and imprecision of the breakpoints, we use quantile filtering to ignore outliers given a coverage of more than eight reads. The computed s.d. values for both breakpoints are compared to the s.d. of a uniform distribution representing spurious SV breakpoints reported in the region. SVs are reported only if both breakpoints are below this threshold. If the s.d. for both breakpoints is <5 bp, the coordinates are marked as 'precise' in the VCF file. Details are presented in Supplementary Note 2.

Variant scoring and genotyping. At the start of the program, the user may specify that the VCF output should be genotyped. In this case, Sniffles stores summary information (coordinates and orientation) about all high-quality reads that do not include SVs in a binary file. This includes those reads that support the reference sequence that pass the thresholds for MQ and alignment score ratio. After the detection of SVs, the VCF file is read in, and Sniffles constructs a self-balancing tree of the variants. With this information, Sniffles then computes the fraction of reads that support each variant versus those that support the reference. Variants below the minimum allele frequency (default: below 30%) are considered unreliable, variants with high allele frequency (default: above 80%) are considered homozygous, and variants with an intermediate allele frequency are considered heterozygous. Note that Sniffles does not currently consider higher ploidy, although this will be the focus of future work. Details are presented in Supplementary Note 2.

Clustering and nested SVs. To enable the study of closely positioned or nested SVs, Sniffles optionally clusters SVs that are supported by the same set of reads. Note that Sniffles does not fully phase the haplotypes, as it does not consider single-nucleotide polymorphisms or small indels, but rather identifies SVs that occur together. If this option is enabled, Sniffles stores the name of each read that supports an SV in a hash table keyed by the read name, with the list of SVs associated with that read name as the value. The hash table is used to find reads that span more than one event, and later to cluster reads that span one or more of the same variants. In this way Sniffles can cluster two or more events, even if the distance between the events is larger than the read length. Future work will include a full phasing of hapolotypes including SVs, single-nucleotide polymorphisms, and other small variants. Details are presented in Supplementary Note 2.

Mapping and SV evaluation. *Simulation of SVs and reads.* SVs were randomly simulated on chromosomes 21 and 22 of the human genome (GRCh37). Datasets were generated with 20 variants for each type of SV (tandem duplication, indel, inversion, translocation, and nested) and different sizes of these events (100 bp, 250 bp, 500 bp, 1 kbp, 5 kbp, 10 kbp, and 50 kbp). Illumina reads were simulated as 100-bp paired-end reads using the default parameters of dwgsim. For PacBio and Oxford Nanopore sequences, we scanned the alignments of HG002 (GiaB) and NA12878, respectively, and measured their error profiles using SURIVOR (option 2). The measured error profiles and read lengths were then used to simulate the reads for each simulated SV dataset (Supplementary Note 3).

Modified reference analysis. To allow for a more realistic scenario, we also modified the human reference (GRCh37) and analyzed real reads to assess the introduced SVs. Here we were able to simulate only a subset of SV types as insertions, deletions, inversions, and translocations. We simulated 140 variants of each type on the human genome (GRCh37) using SURVIVOR (option 1) (Supplementary Note 5).

Evaluation of long-read mappings. All simulated reads were mapped to the human reference genome (GRCh37) using BWA-MEM[26], BLASR[25], GraphMap[27], MECAT[28], LAST[24], Minimap2[29], and NGMLR. Reads that overlapped or mapped in close proximity to a simulated SV were extracted from the BAM files and used for evaluation. For the genuine datasets, we first mapped the reads to the unmodified reference genome (without SVs) using BWA-MEM and extracted all reads that would span our simulated SV by at least 500 bp. Only these reads were then mapped to the modified reference genome by the four read mappers and used for evaluation.

Both simulated and genuine reads were then divided into six categories (Supplementary Note 3):

1. Read mappings are considered 'precise' if they fully identify the SV they cover. To be placed in this category, read mappings have to cover all parts

of the SV that are required for identification—for example, a read mapping to an inversion has to cover the inverted part of the genome as well as the noninverted sequences flanking the inversion. Furthermore, correct mappings have to be split at the simulated breakpoints (±10 bp) of the SV.
2. Read mappings are considered 'indicated' if they show the presence of the correct SV but identify it as the wrong type (for example, a duplication that is represented as an insertion), or show the correct SV but not the exact borders (>10 bp away).
3. Read mappings are considered 'forced' if they indicate the wrong number of SVs (for example, several small insertions instead of a single long insertion) or contain a significant portion of mapping artifacts (for example, not simulated mismatches) over >10% of the SV length. These include mappings such as a read that is aligned through a deletion or inversion (Fig. 2, top).
4. Read mappings are considered 'trimmed' if they have been soft-clipped or otherwise trimmed so that they cannot indicate the SV and do not contain randomly aligned base pairs (i.e., noisy regions).
5. Read mappings that are split into more parts than required to cover the underlying SV are classified as 'fragmented'.
6. Read mappings that are supposed to map across the SV but are not mapped are considered 'unaligned'.

For all simulated SV types and sizes and all mappers, we count the number of reads that fall into the above categories, normalize by the number of simulated reads, and visualize the resulting data in bar plots.

Evaluation of SV callers. After the SV calling, each VCF file was evaluated with SURVIVOR[48] with appropriate parameter sets to compare the variants to the truth set. An SV is considered precise if its start–stop coordinate is within 10 bp of the simulated start–stop coordinate and if the caller predicted the correct type. An SV is considered indicated if the start–stop coordinate of the SV is within ±1 kb of the simulated event regardless of the inferred type of SV. A simulated SV is considered not detected if there is no call that fulfills the two previous criteria. An SV is considered to be a false positive if the event was not simulated.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The source code, documentation, and test datasets are available at https://github.com/philres/ngmlr and https://github.com/fritzsedlazeck/Sniffles for the mapping and SV-calling method, respectively.

Software versions and parameter settings. BWA-MEM (version 0.7.12-r1039)[26] was used with the -M parameter to map the short reads and with "-X pacbio -M" to follow the recommended parameter settings for PacBio reads. The parameter -M is used to mark only one alignment as primary and the subsequent alignments as secondary. BlasR (version 1.3.1)[25] was run using the parameters "-sam -bestn 1 -nproc 15" to obtain only the best alignment in SAM format using 15 threads. Furthermore, Blasr was run with the parameters suggested by PBHoney[18], "-nproc 15 -bestn 1 -sam -clipping subread -affineAlign -noSplitSubreads -nCandidates 20 -minPctIdentity 75 -sdpTupleSize 6." SAMTools (version 0.1.19–44428 cd)[47] was used to convert the SAM alignment files to BAM and to sort the aligned reads.

We used Delly (version v0.7.3)[15], Lumpy (version 0.2.13)[14], and Manta (version 1.0.3)[16] to call SVs over the high-mapping-quality aligned Illumina reads (MQ: 20+), followed by SURVIVOR (version 0.0.1)[48] to combine the calls and report the consensus variants. To allow for the uncertainty with short-read variant positioning, SVs were considered the same if their start–stop coordinates fell within 1 kb of each other and were of the same type. PBHoney (version PBSuite_15.8.24)[18] with default parameters was used to infer SVs on the basis of the specified BlasR alignments. The output was converted into a VCF file with SURVIVOR (option 10).

Data availability. The raw sequencing data used in this study are available from the respective publications listed in Supplementary Table 5. The alignments and SV calls produced in this study for NGMLR and Sniffles are available at https://github.com/fritzsedlazeck/Sniffles.

References

47. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
48. Jeffares, D. C. et al. Transient structural variations alter gene expression and quantitative traits in *Schizosaccharomyces pombe*. *Nat. Commun.* **8**, 14061 (2017).

nature research

Corresponding author(s): Michael C Schatz, Fritz J Sedlazeck

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

   Describe how sample size was determined. — Not applicable to the study

2. Data exclusions

   Describe any data exclusions. — No data was excluded.

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings. — We tested the proposed methods on 85 different simulated data sets based on a published simulation method for SV. Furthermore, we evaluated the reproducibility by using trios for Arabdidopsis and Human data and benchmarked the effect of downsizing the initial data sets for various technologies as well as healthy and diseased humans. All attempts at replication were successful,

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups. — Not relevant as no selection of samples were preformed no cross validation is appropriate.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. — Data were not partitioned into groups

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	We used BWA-MEM, NGMLR (proposed method in this manuscript), Graphmap and BlasR to align reads. Lumpy, Delly, Manta, Sniffles (proposed method), PBHoney to infer structural Variations. SURVIVOR (v2) to simulate, evaluate and compare methods. All methods (apart from Sniffles (v1.0.6) and NGMLR (v0.2.6) that are the topic of this paper) are peer reviewed and published. Sniffles and NGMLR are available on Github and links are included in the manuscript.

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	no eukaryotic cell lines were used

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	no eukaryotic cell lines were used
b. Describe the method of cell line authentication used.	no eukaryotic cell lines were used
c. Report whether the cell lines were tested for mycoplasma contamination.	no eukaryotic cell lines were used
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	no eukaryotic cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	No animals were used.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	No new data were generated for this paper; human sequence data are from publicly available sources