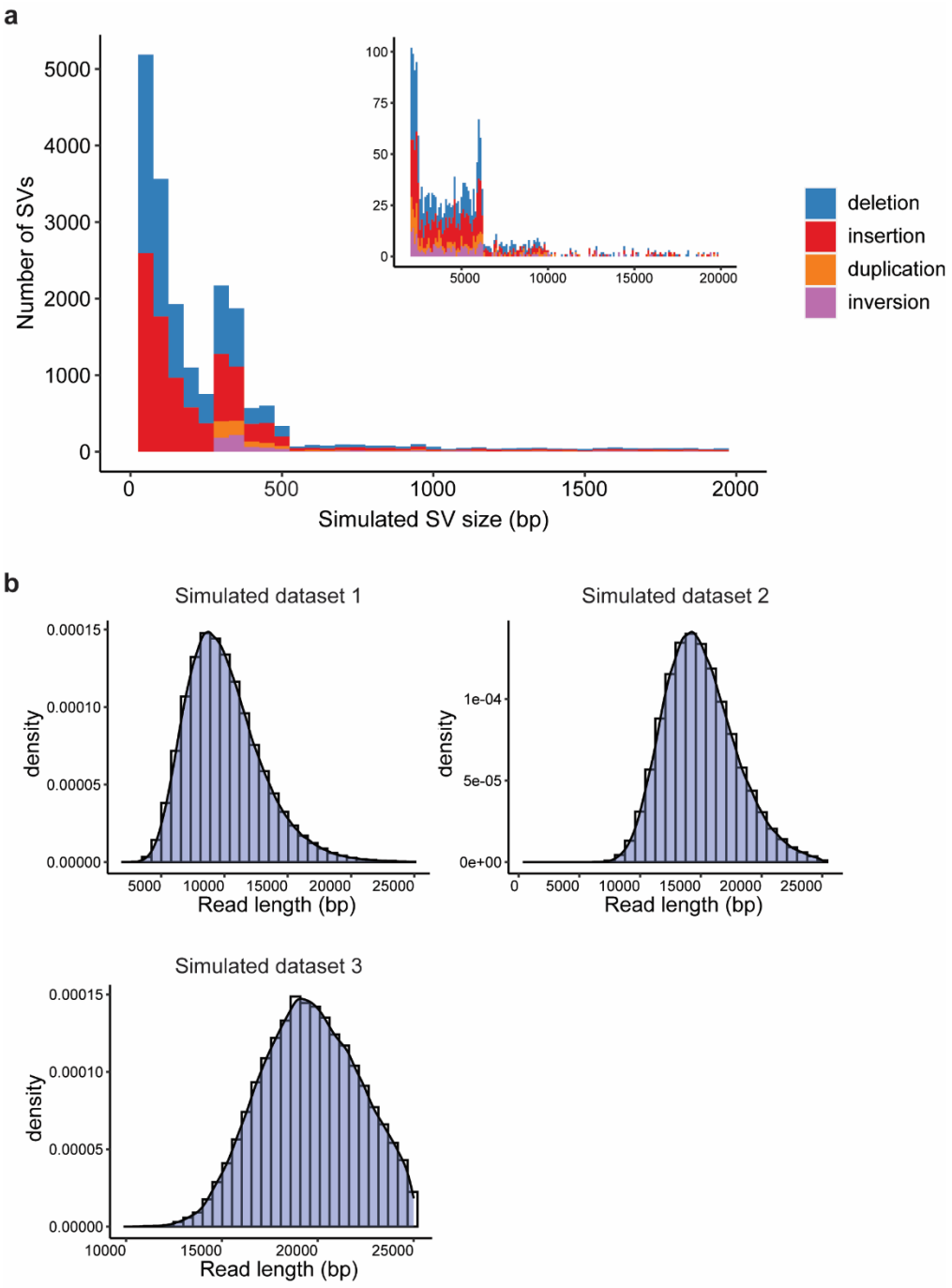

Figure S1 Characteristics of simulated datasets. a Size distributions of simulated SVs. **b** Length distributions of simulated long reads in three simulated datasets.

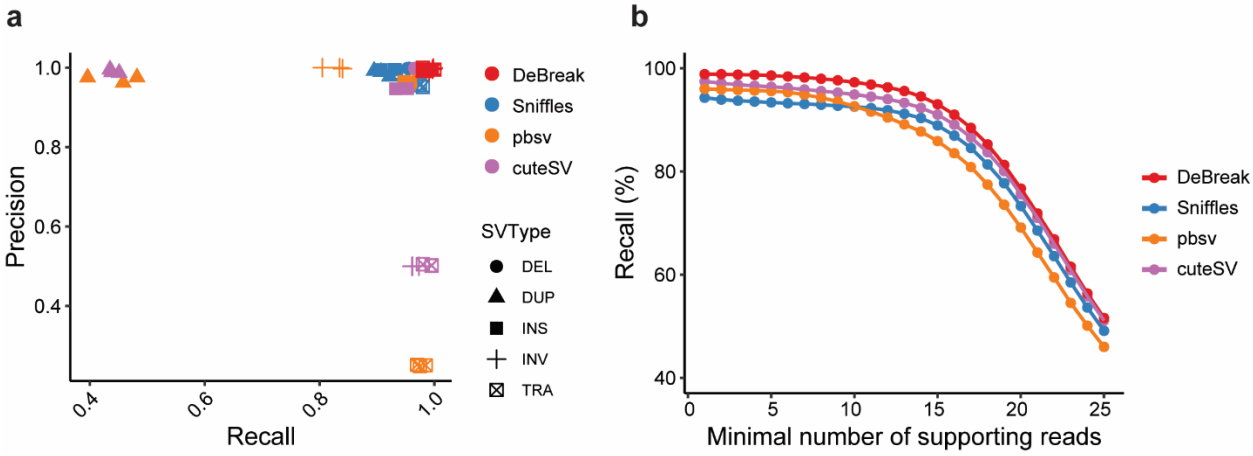

Figure S2 SV discovery accuracy in simulated datasets. a SV discovery accuracy for four tested SV callers in three simulated datasets. **b** Recall for deletion and insertion detection at different thresholds of 'minimal supporting reads' for the four SV callers in simulated datasets.

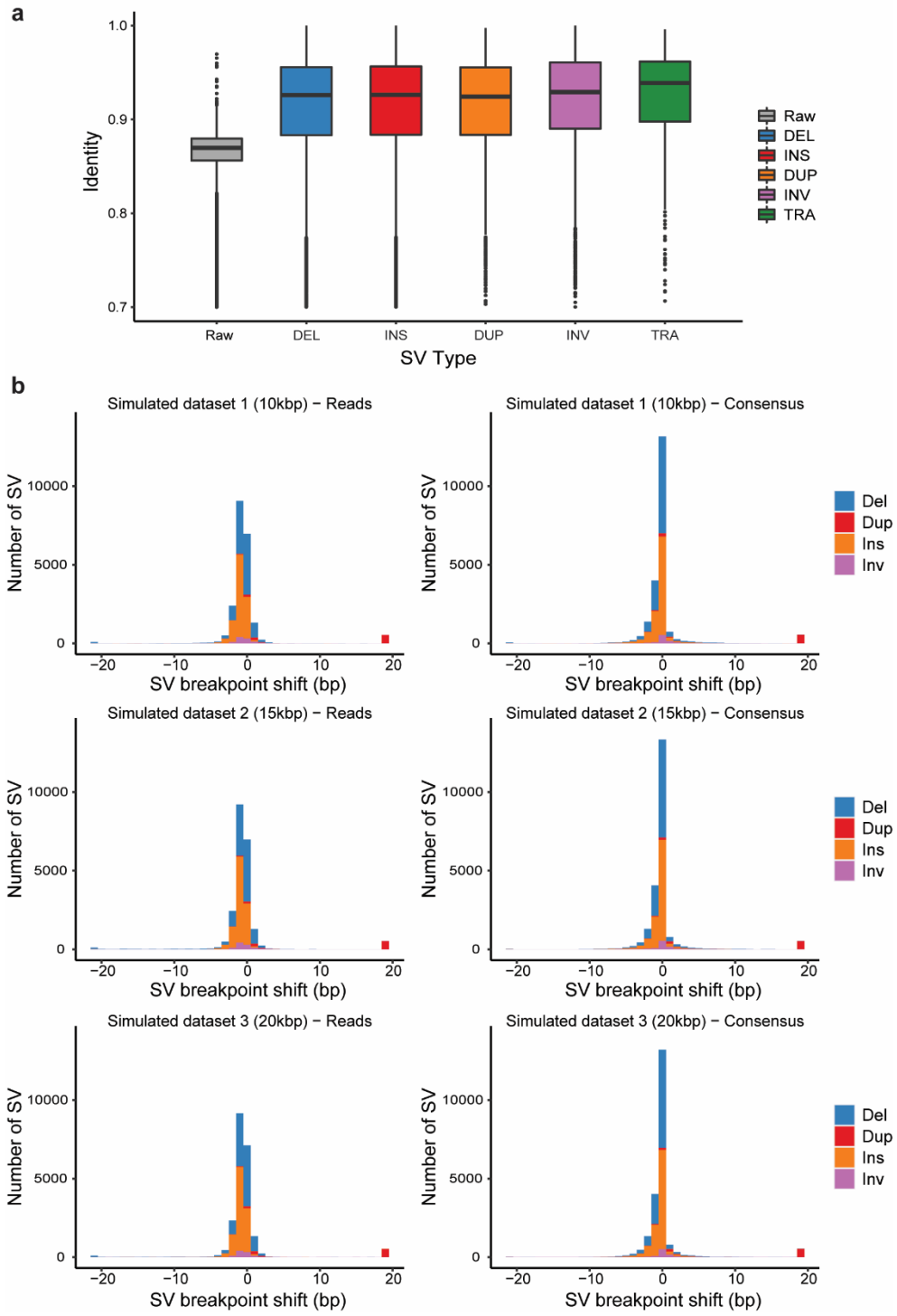

Figure S3 SV breakpoint refinement in simulated datasets. a The identity of reads/consensus sequences compared with sequences around simulated SVs for individual SV type. **b** Shift of SV breakpoints inferred from raw reads (left) and from consensus sequences (right) in three simulated datasets (top, center, and bottom). SVs with shifts more than 20bp were combined into the first and last bin.

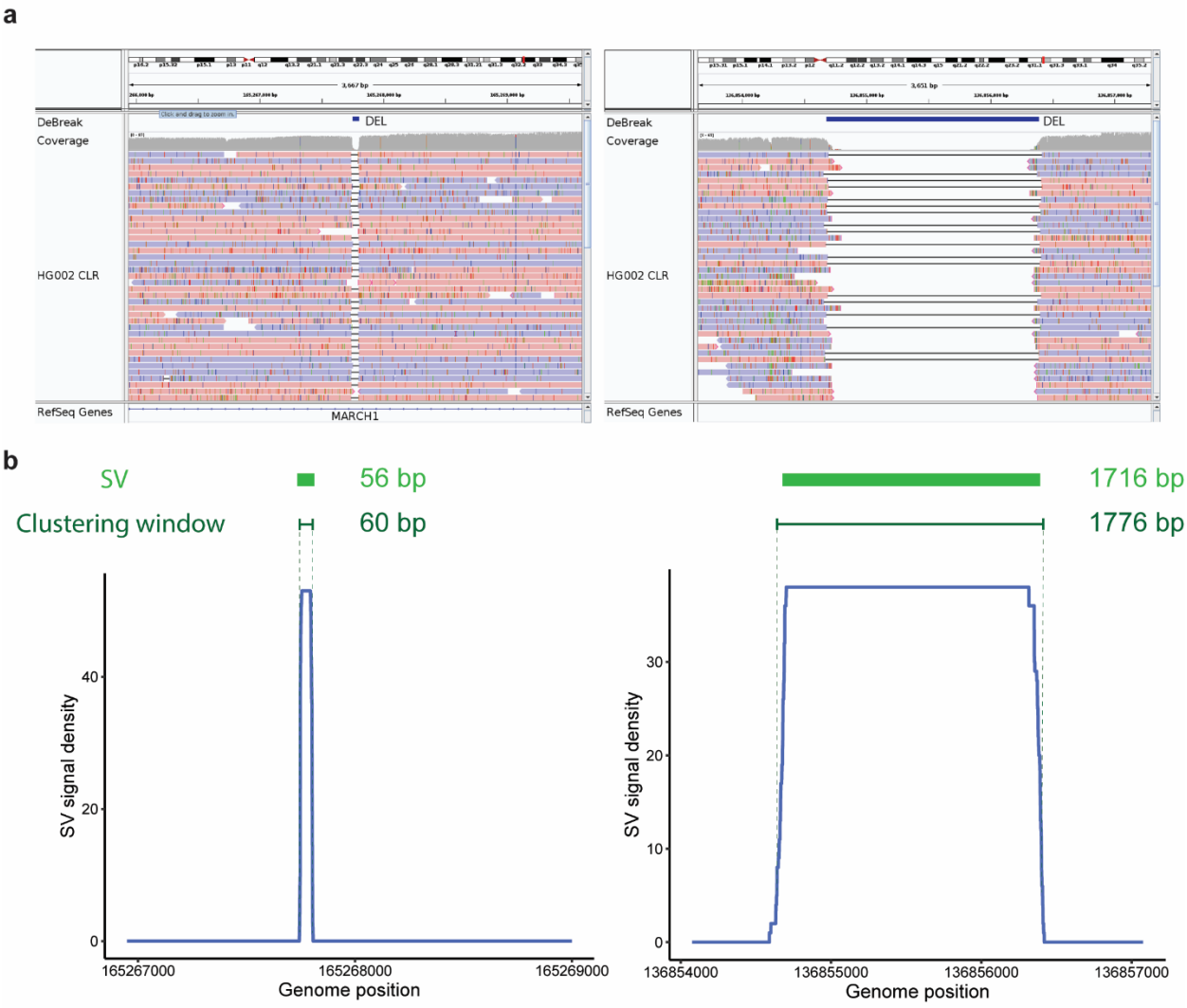

Figure S4 Adjustable clustering window size. a IGV view of read alignments flanking a small SV (left) and a large SV (right). **b** Raw SV signal density of the two SVs shown in **a**. Based on the density pattern of raw SV signals, the clustering window is smaller for shorter SVs (left) and larger for longer SVs (right).

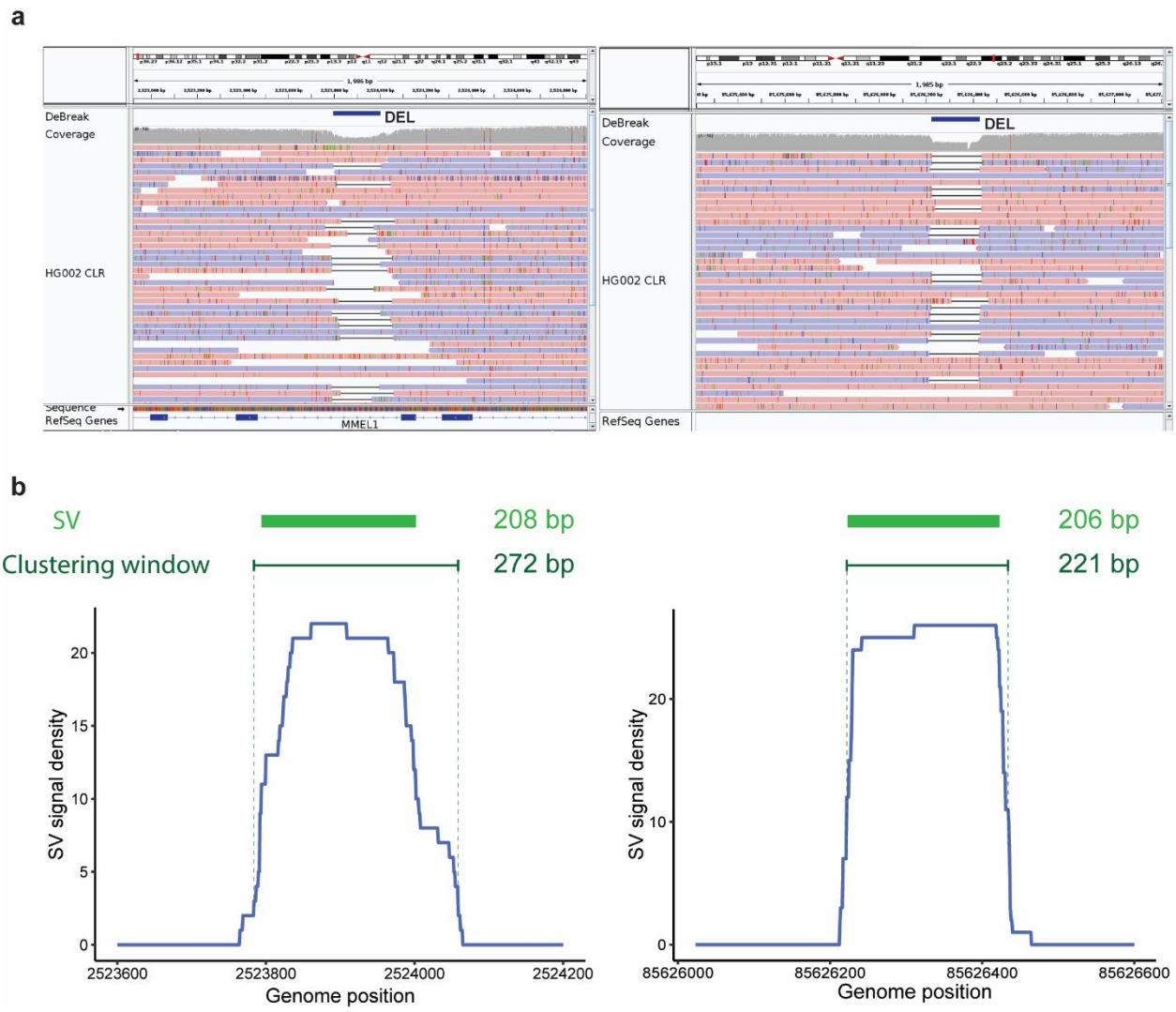

Figure S5 Adjustable clustering window for repeat regions. a IGV view of read alignments flanking SVs located within a low-complexity region (left) and a non-repeat region (right). Raw SV signals have various breakpoint positions in a low-complexity region and relatively consistent breakpoint positions in a non-repeat region. **b** Raw SV signal density of the two SVs shown in **a**. The clustering window size is larger in a low-complexity region when raw signals are diverged (left) and smaller in a non-repeat region (right) to exclude potential noise signals nearby.

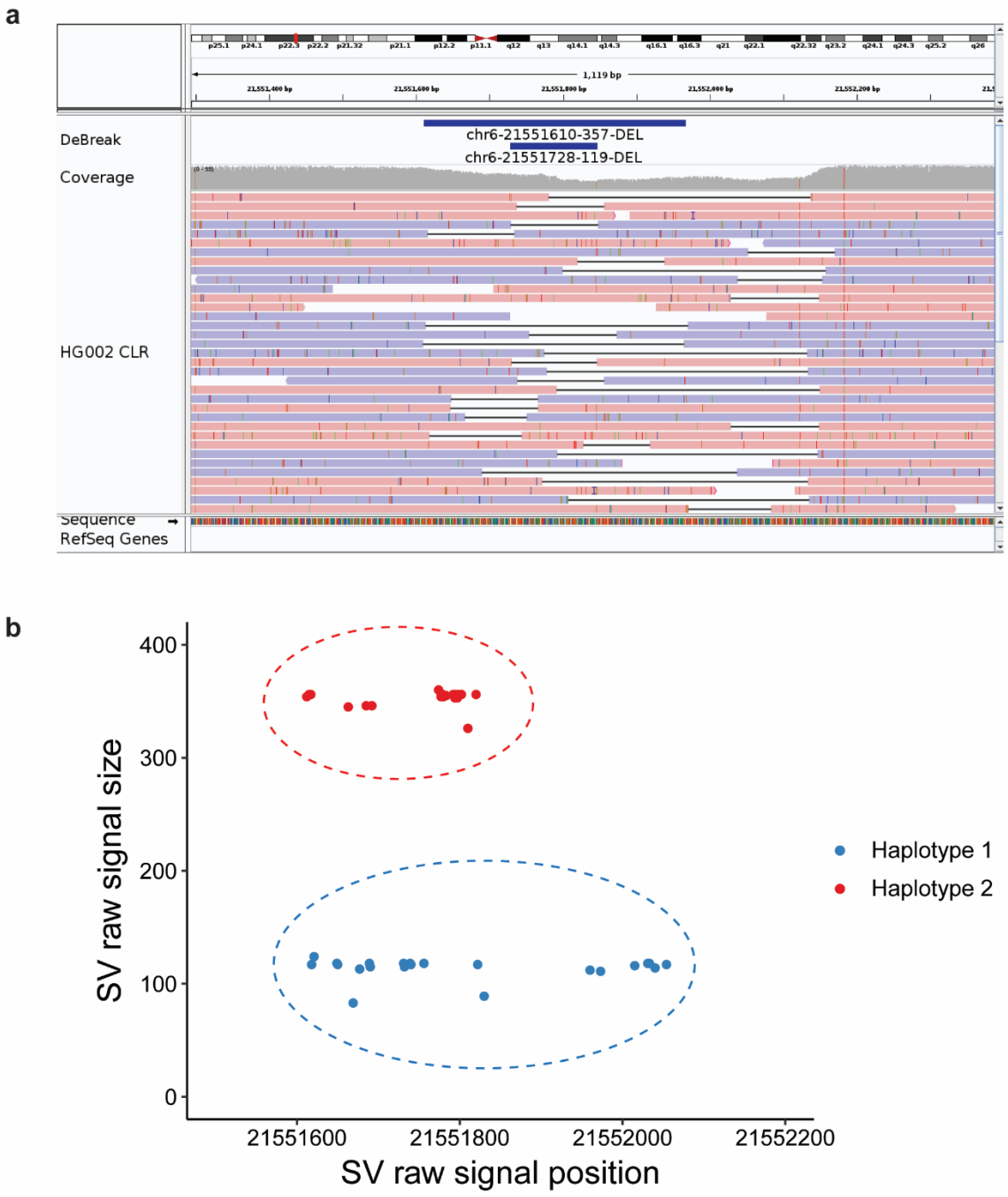

Figure S6 Example of a multi-allele SV in HG002. a IGV view of a multi-allelic SV on chromosome 6. Some reads contain ~100bp (shorter) deletion signals, and other reads contain ~300bp (longer) deletion signals. **b** k-means clustering of SV raw signals from the multi-allelic SV region shown in **a**. Based on SV size and position, SV raw signals are clustered into two groups, each representing one allele.

Table S1 SV genotyping accuracy in HG002

	PacBio CLR			PacBio HiFi			Nanopore		
	DEL	**INS**	**Total**	**DEL**	**INS**	**Total**	**DEL**	**INS**	**Total**
DeBreak	**93.22**	**84.00**	**87.98**	90.46	84.70	87.21	**92.18**	86.38	**88.92**
Sniffles	47.47	38.51	42.43	52.38	45.77	48.70	61.69	56.08	58.55
pbsv	92.35	75.09	82.14	93.26	75.19	82.34	80.22	79.81	79.99
cuteSV	92.53	78.99	84.52	**91.95**	**88.56**	**90.02**	88.05	**87.45**	87.71

The genotyping accuracy is calculated as the number of SVs with the correct genotype divided by the total number of SVs reported by each SV caller. The highest genotyping accuracy in each group is marked in bold. The unit of genotyping accuracy is %.

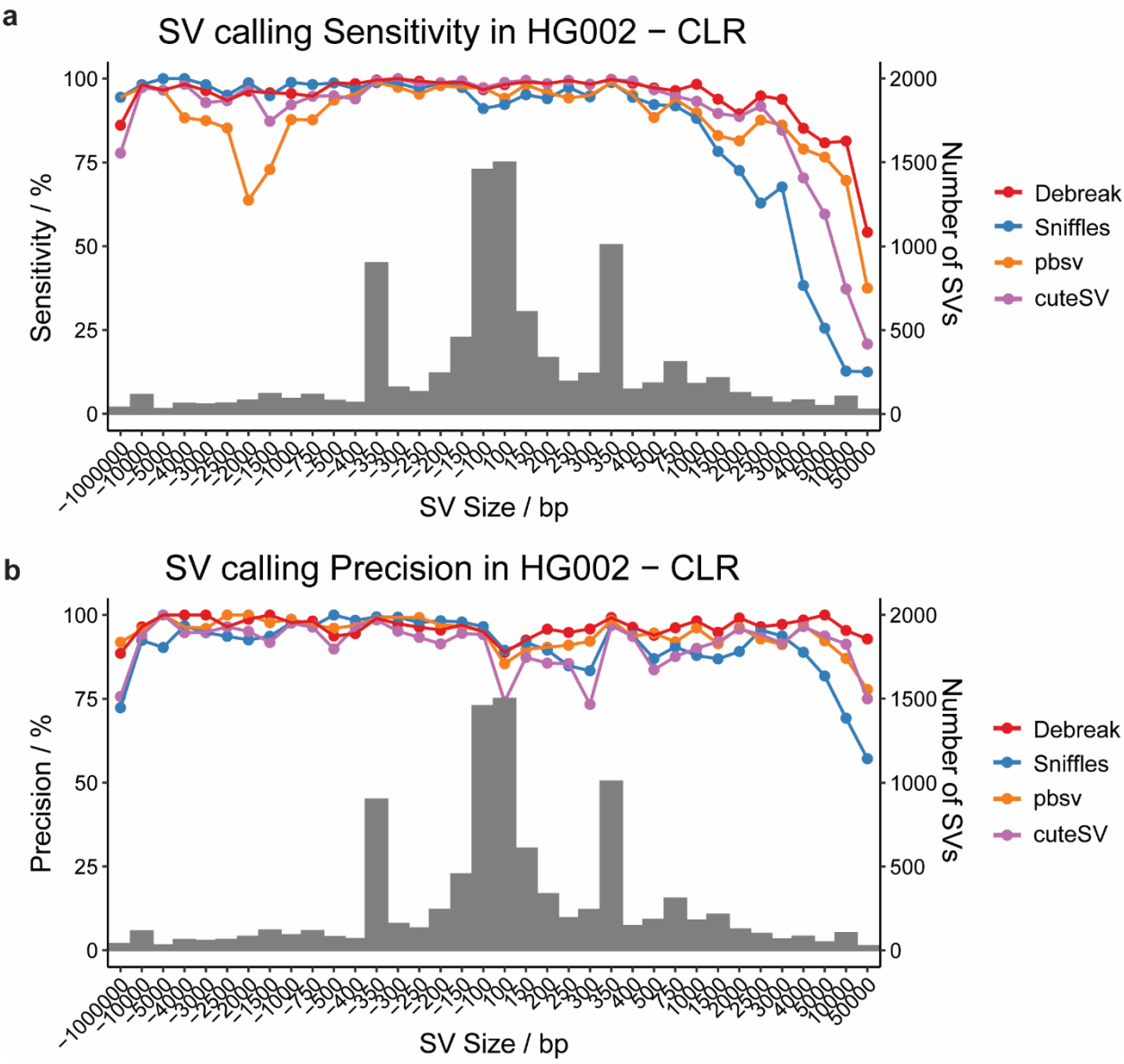

Figure S7 SV calling recall and precision for insertions (positive SV size) and deletions (negative SV size) for four tested SV callers in HG002 high-confidence regions. The bar plot indicates the number of SVs in each size range.

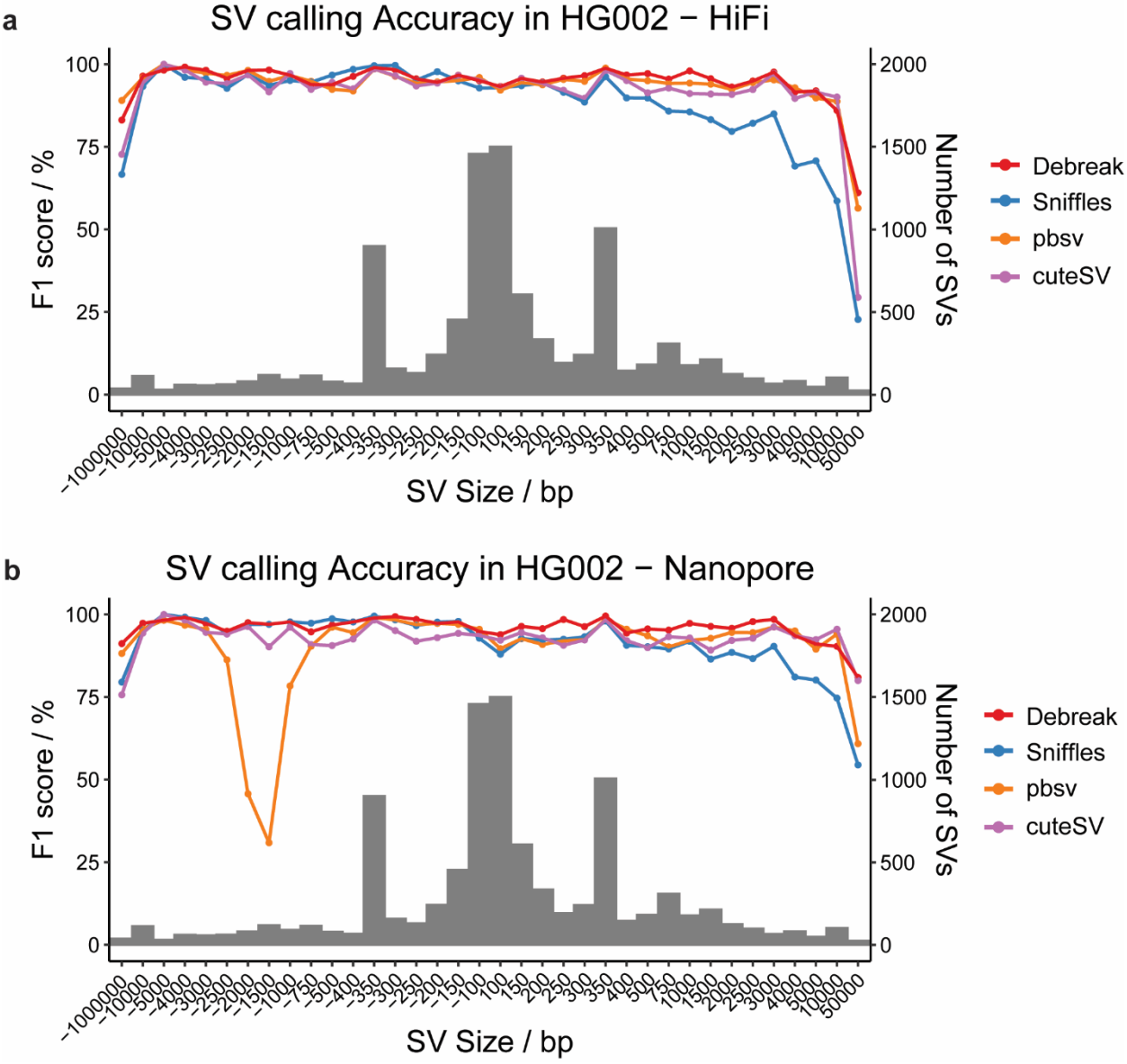

Figure S8 SV calling accuracy for insertions (positive SV size) and deletions (negative SV size) for four tested SV callers in PacBio HiFi and Nanopore datasets. The bar plot indicates the number of SVs in each size range.

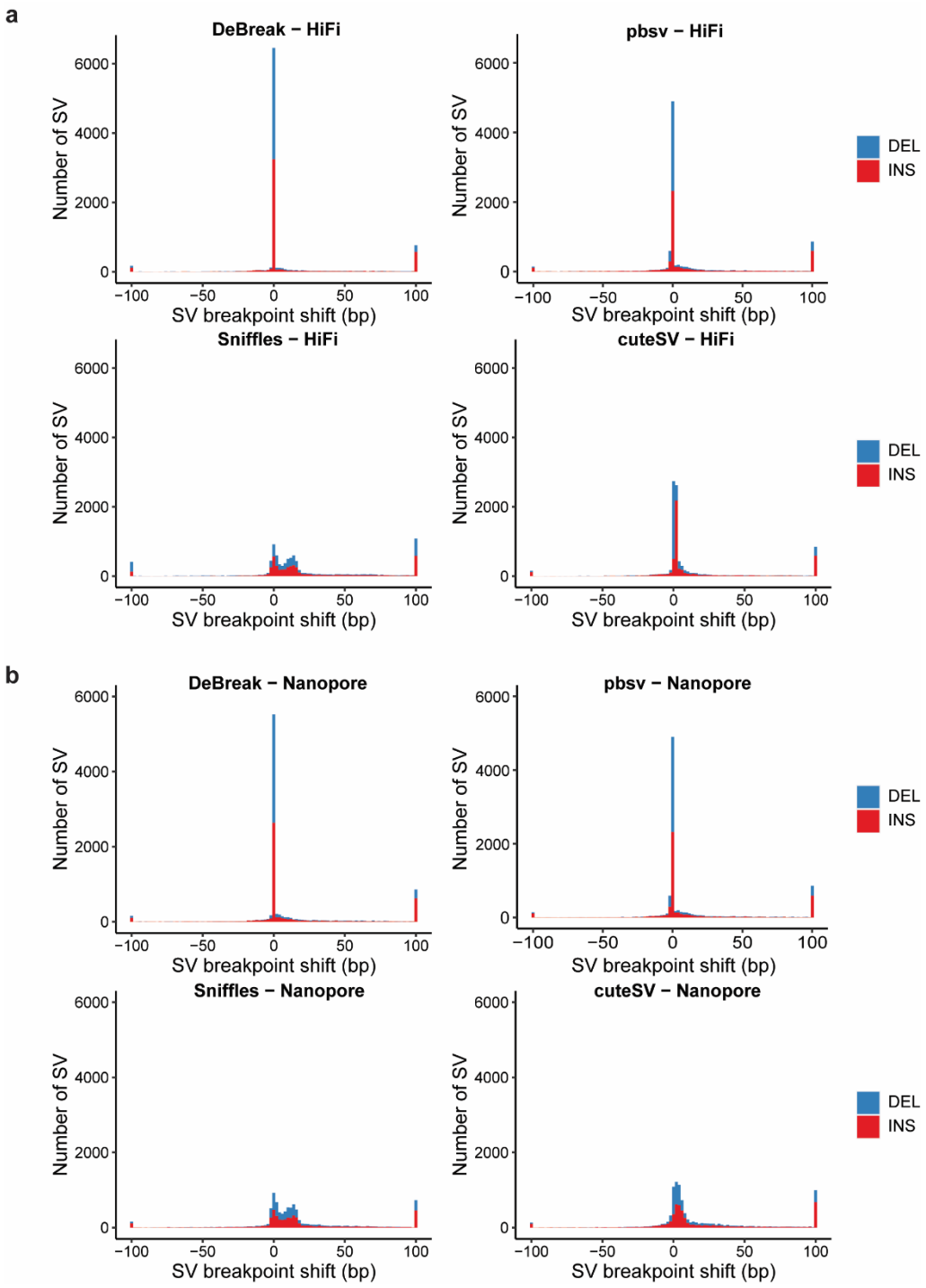

Figure S9 SV breakpoint accuracy in HG002 HiFi and Nanopore datasets. a SV breakpoint shift of four SV callers in the HiFi dataset. 64%, 57%, 5%, and 25% of SVs were identified with exact breakpoint position by DeBreak, pbsv, Sniffles, and cuteSV, respectively. **b** SV breakpoint shift of four SV callers in Nanopore dataset. 54%, 49%, 5%, and 7% of SVs were identified with exact breakpoint position by DeBreak, pbsv, Sniffles and cuteSV, respectively.

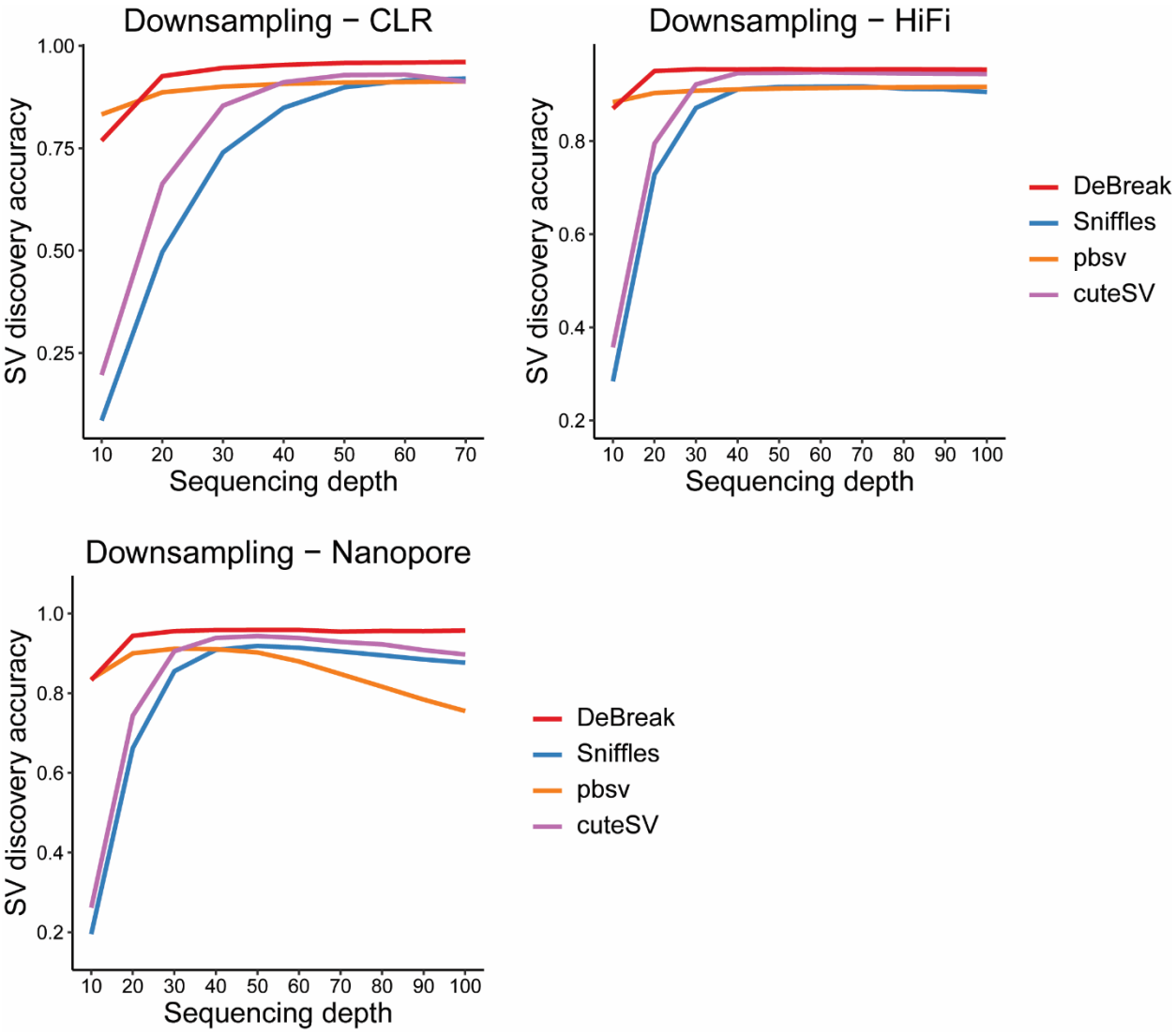

Figure S10 SV discovery accuracy in down-sampled datasets under default settings. Accuracy of Sniffles and cuteSV are worse at lower sequencing depths.

Table S2 SV discovery accuracy compared to assembly-based SV callset

	DeBreak			pbsv			cuteSV		
Sample	**DEL**	**INS**	**Total**	**DEL**	**INS**	**Total**	**DEL**	**INS**	**Total**
CLR									
HG00096	80.52	78.59	**79.34**	76.51	69.85	72.59	78.72	77.73	78.12
HG01505	80.77	78.72	**79.52**	76.96	69.51	72.57	79.18	77.51	78.16
HG01596	80.15	75.83	**77.45**	75.46	68.72	71.47	78.28	72.10	74.39
HiFi									
HG02818	81.90	80.57	**81.12**	78.43	69.05	73.19	78.45	77.08	77.65
HG03486	82.17	81.58	**81.82**	78.56	69.30	73.40	79.71	78.65	79.09
NA12878	80.78	81.66	**81.31**	77.80	69.17	72.87	77.32	76.54	76.85

SV discovery accuracy was evaluated with the assembly-based SV callset as the ground truth. The highest accuracy (F1 score) among three tested alignment-based SV callers is marked in bold in each sample. The unit of F1 score is %.

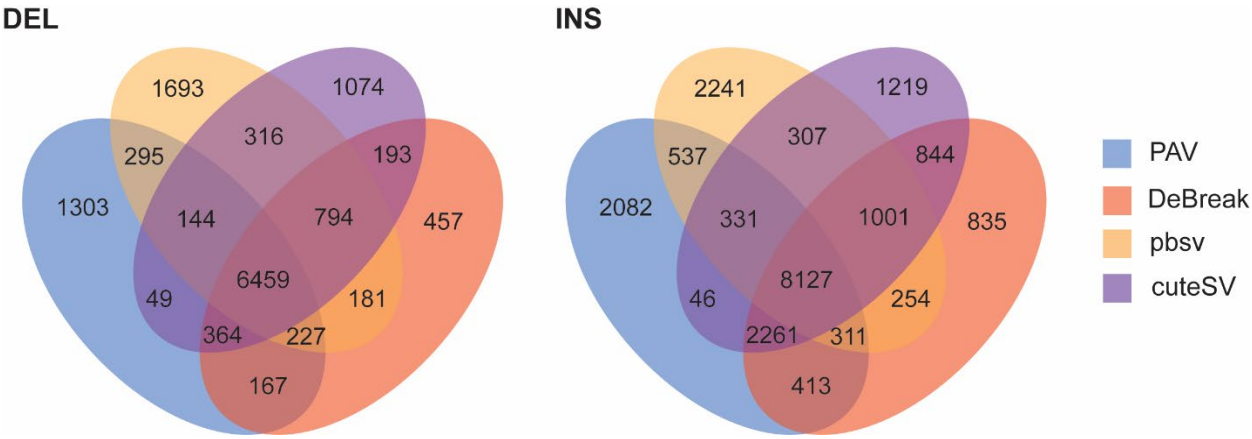

Figure S11 SV discovery consistency between alignment-based and assembly-based SV discovery approaches. Venn diagrams showing the overlap between alignment-based and assembly-based SV callsets. Numbers indicate the number of SVs in each group.

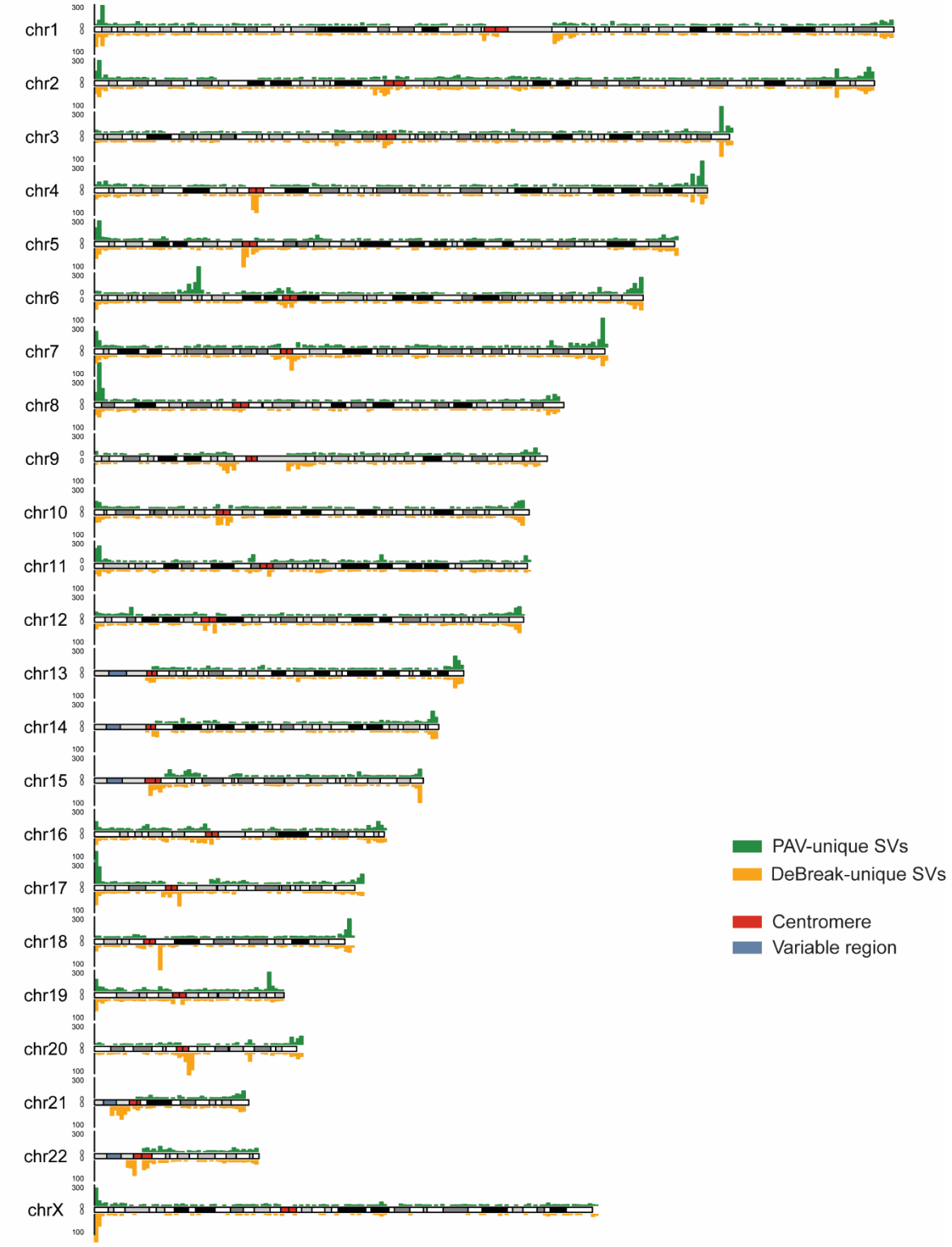

Figure S12 PAV-unique and DeBreak-unique SV distribution on the genome. The PAV-unique SVs (green) are enriched at the telomere regions. DeBreak-unique SVs (orange) are enriched at the centromere and telomere regions.

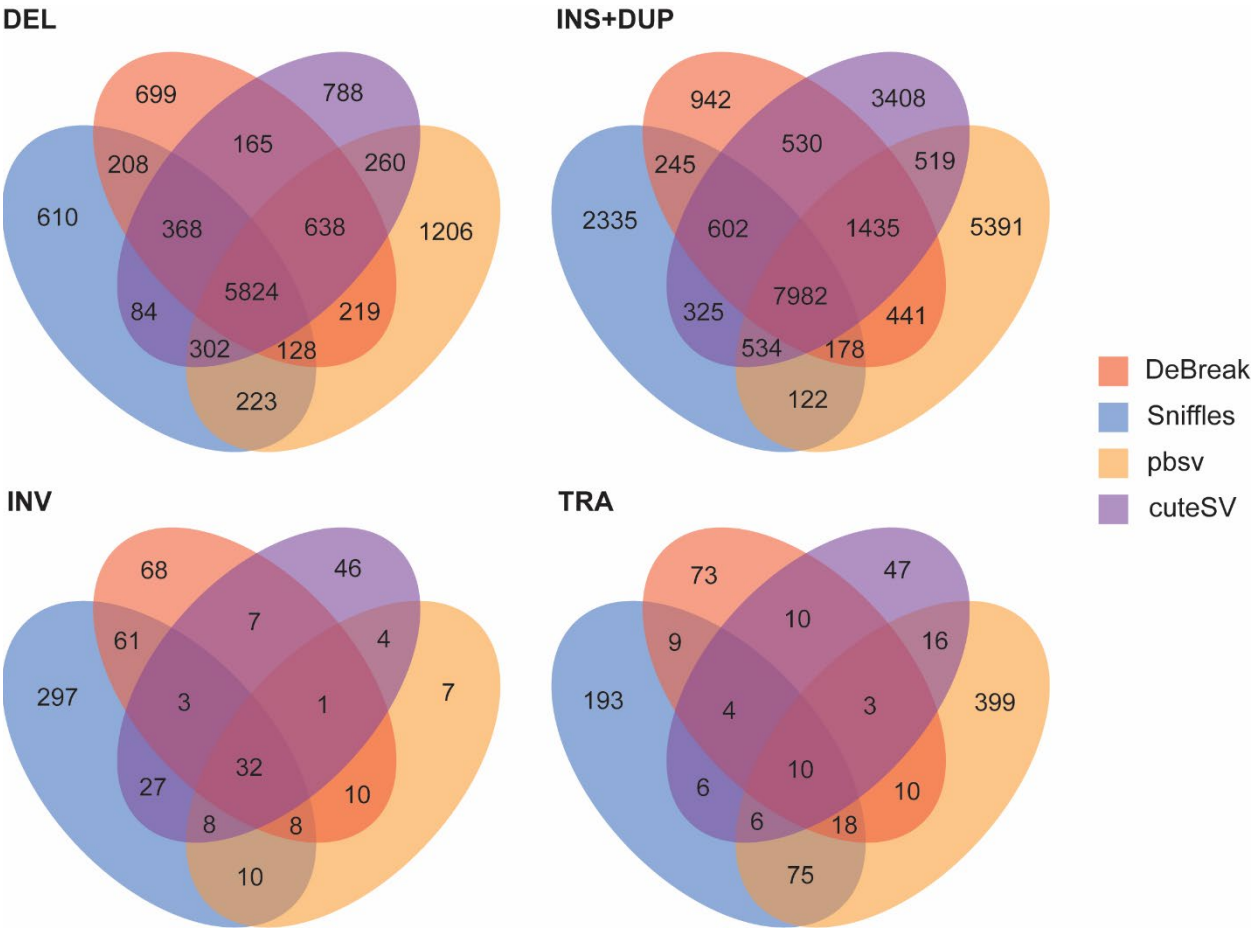

Figure S13 SV discovery in SKBR3 cell line. Venn diagrams showing the overlap between SV callsets from four SV callers. Insertions and duplications are merged together for comparison, as duplications are sometimes considered as insertions.

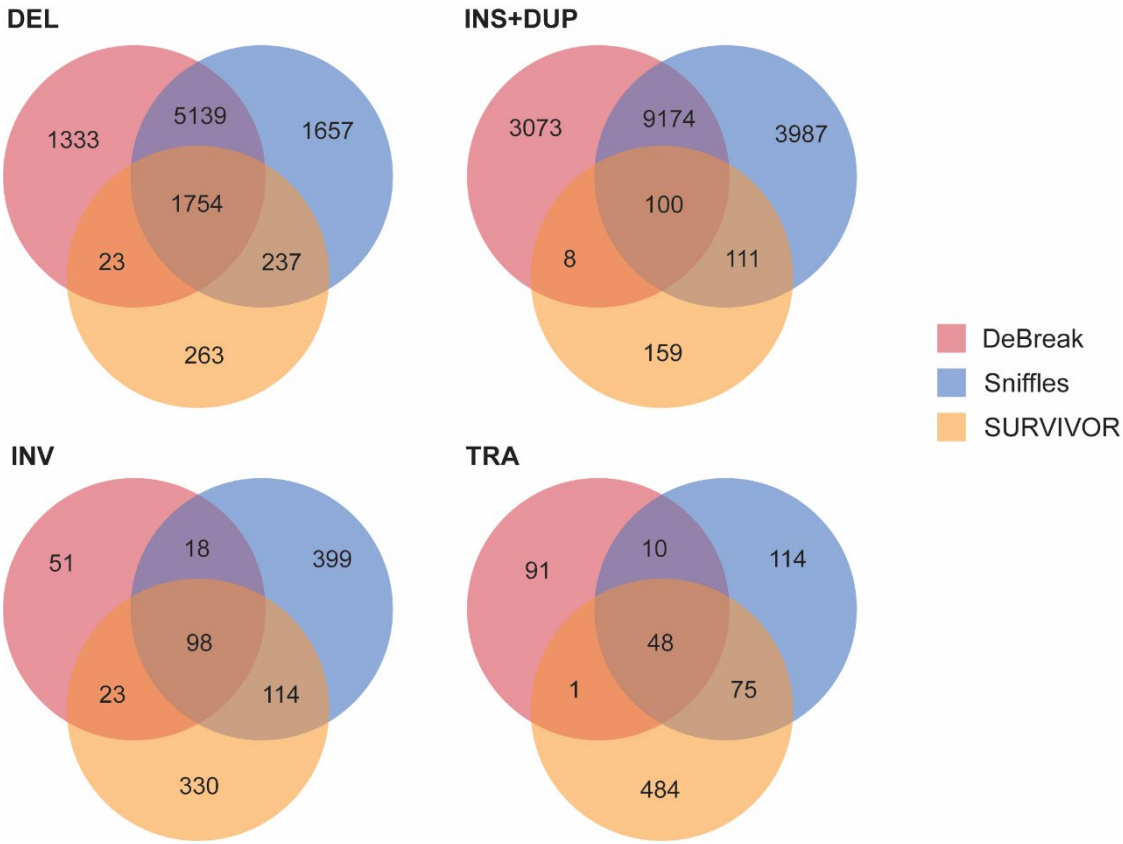

Figure S14 SV discovery in SKBR3 cell line. Venn diagrams showing the overlap between DeBreak SV callset and SV calls previously reported from short-read (SURVIVOR) and long-read data (Sniffles). Numbers indicate the number of SVs in each category.

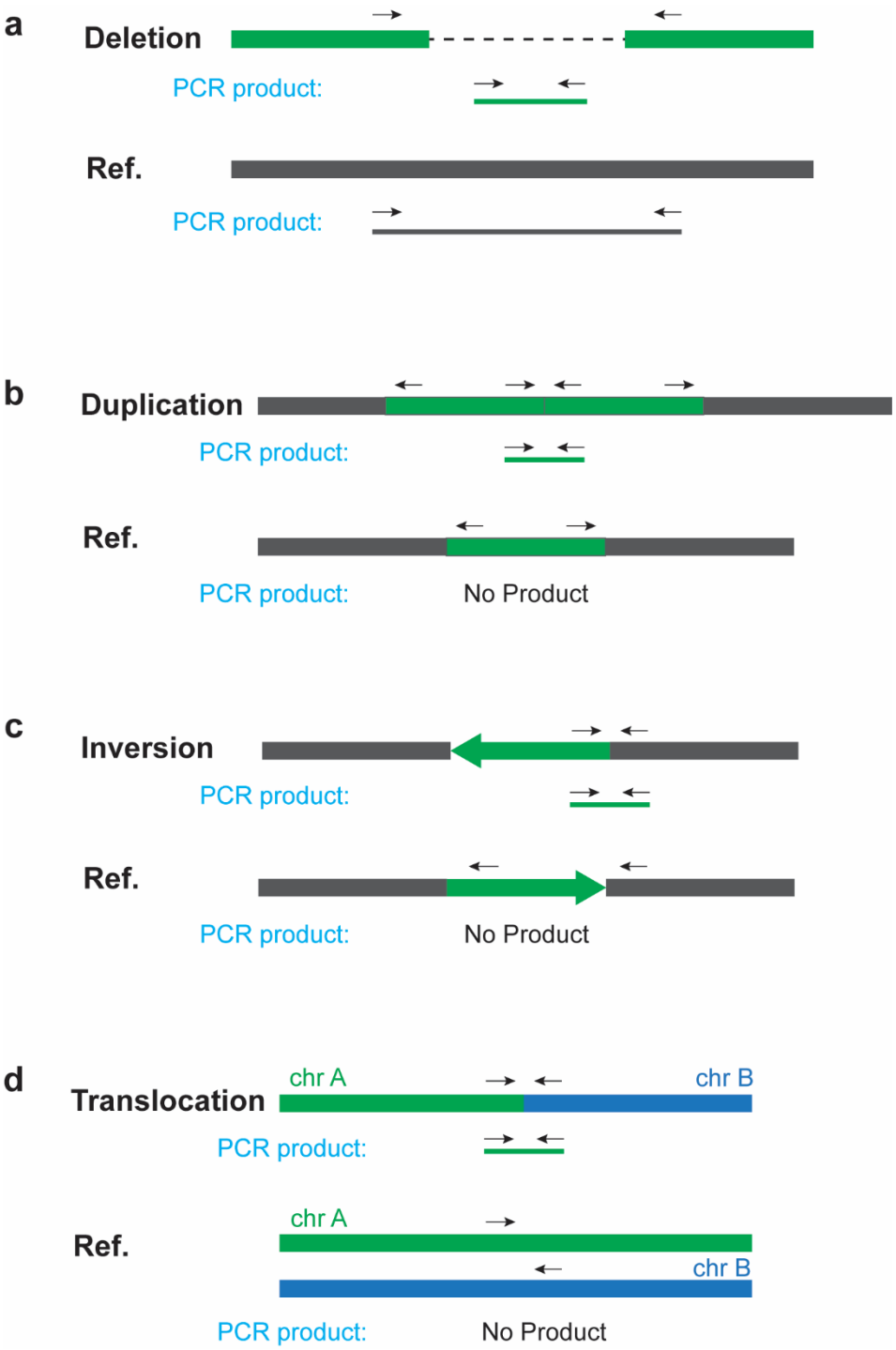

Figure S15 PCR primer for validation. PCR primer design for deletions (**a**), duplications (**b**), inversions (**c**), and translocations (**d**). For deletions, the PCR product size is much smaller when the SV is a true event than for false positive (reference allele). For duplications, inversions, and translocations, the PCR product is expected only when SV is a true event.

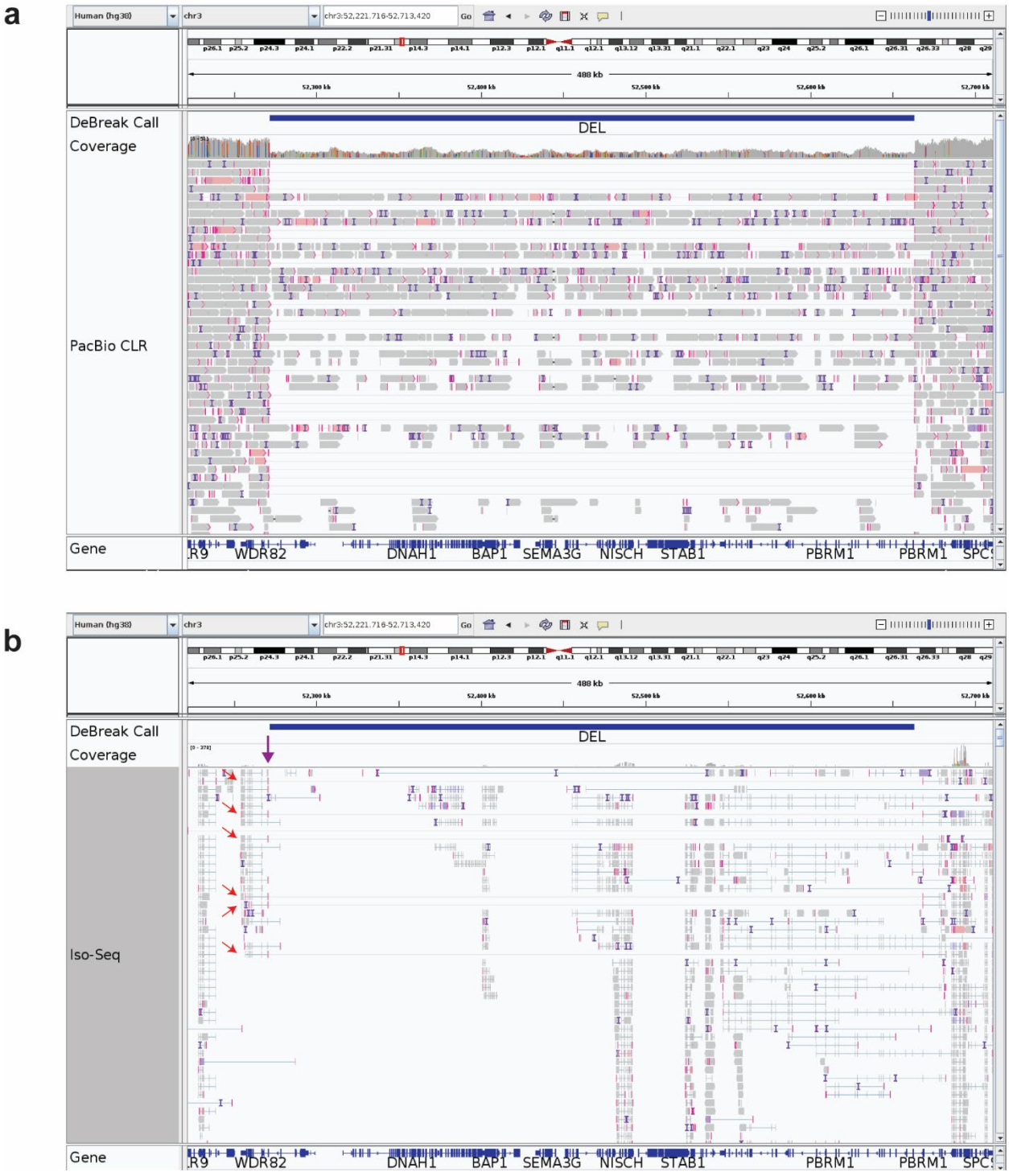

Figure S16 Gene fusion of WDR82 and PBRM1. IGV view of PacBio CLR (**a**) and Iso-Seq data (**b**) at the gene fusion junction. The 'DeBreak Call' panel shows SVs identified from PacBio data with DeBreak. Red arrows indicate IsoSeq reads that contain sequences from both WDR82 and PBRM1. The purple arrow indicates the gene fusion junction position inferred from IsoSeq reads.

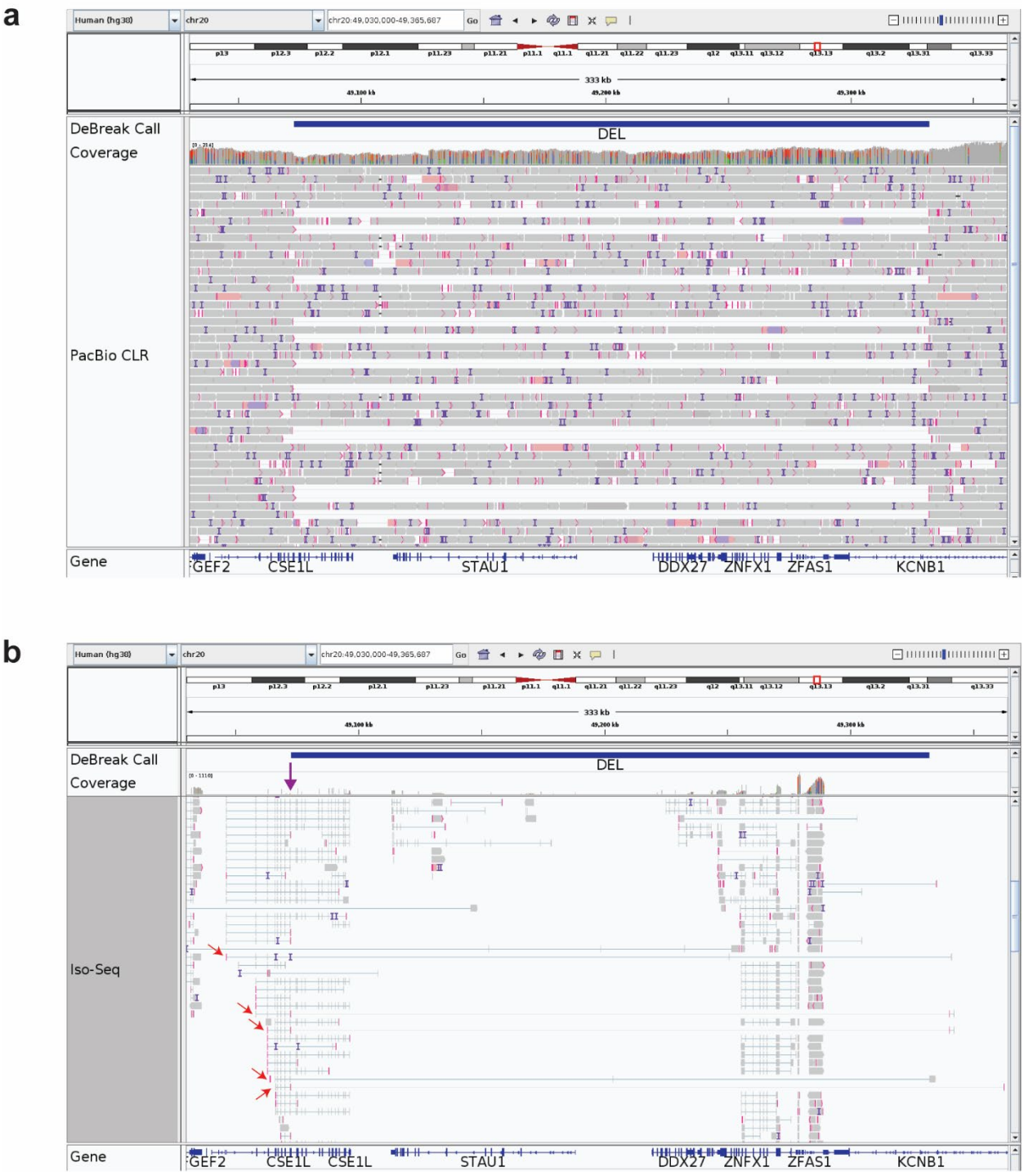

Figure S17 Gene fusion of CSE1L and KCNB1. IGV view of PacBio CLR (**a**) and Iso-Seq data (**b**) at the gene fusion junction. The 'DeBreak Call' panel shows SVs identified from PacBio data with DeBreak. Red arrows indicate the IsoSeq reads that contain sequences from both CSE1L and KCNB1. The purple arrow indicates the gene fusion junction position inferred from IsoSeq reads.

Table S3 Runtime and memory usage of SV callers

	DeBreak	Sniffles	pbsv	cuteSV
CPU x node	12 x 1	12 x 1	12 x 1	12 x 1
Wall-clock time	12:22:48	03:02:37	1-21:06:15	01:29:37
CPU time	1-19:11:22	1-08:28:26	1-22:42:03	08:10:08
Peak Memory (GB)	63.00	12.62	71.84	3.39