A sensitive and accurate framework for population-scale structural variant discovery and genotyping across sequence types

Guangbao Luo[1], Li Xiao[1], Zhangjun Fei[2]* & Xin Wang[1]*

[1]National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops; Hubei Hongshan Laboratory; College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, China

[2] Boyce Thompson Institute, Cornell University, Ithaca, NY, USA

*Correspondence: xinwang@mail.hzau.edu.cn; zf25@cornell.edu

Abstract

Structural variation underlies much of the phenotypic and evolutionary diversity. However, accurate discovery and genotyping of structural variants (SVs) at population scale remain challenging due to the varied characteristics of sequencing technologies and the complexity of genome architectures. Here, we introduce PSVGT, a unified framework that integrates short-read and long-read data, de novo assembled contigs, and chromosome-level assemblies to enable comprehensive SV detection and genotyping across diploid, polyploid, and allopolyploid genomes. PSVGT employs an integrated signaling module to extract precise insertion and deletion breakpoints, coupled with the ploidy-aware KLOOK clustering algorithm and local depth-adaptive filtering to resolve multi-allelic events and accommodate the uneven coverage characteristic of complex genomic regions. Benchmarking on simulated and real datasets demonstrates that PSVGT consistently outperforms state-of-the-art tools across sequence types, with advantages particularly in complex genomes and low-coverage long-read data. PSVGT fills a critical gap in scalable SV analysis by leveraging underutilized short-read data and enabling robust characterization of SVs across diverse genome architectures from diploids to polyploids–thereby facilitating population-scale analyses and pan-genome research.

Background

Genomic variation underlies phenotypic diversity and plays a crucial role in shaping evolution and gene function. Such variation spans a wide spectrum, from single-nucleotide changes to highly complex genomic rearrangements. Large structural variants (SVs), including deletions, insertions, inversions, duplications, and translocations, have profound effects on genomic architecture, altering gene function and expression[1, 2]. Consequently, population-scale studies of SVs have attracted increasing attention. Nonetheless, reliable detection and genotyping of SVs at population scale remain challenging. Advances in high-throughput sequencing have enabled the sequencing of thousands of individuals, leading to landmark initiatives such as the Human 1000 Genomes Project, the Arabidopsis 1001 Genomes Project, and the 3000 Rice Genomes Project. These efforts have produced extensive catalogs of genetic variation, greatly advancing pan-genomics, genome-wide association studies (GWAS), population genetics, and domestication research. However, most analyses have focused predominantly on single nucleotide polymorphisms (SNPs), which primarily rely on aligning short reads to a linear reference genome.

Over the past two decades, next-generation sequencing (NGS) has greatly facilitated the study of SVs. Consequently, a wide range of computational tools for SV identification have been developed. Nonetheless, these short-read-based methods often suffer from high false-positive and false-negative rates[5, 6]. Tools such as DELLY[7], LUMPY[8], and Manta[8] share similar principles by integrating paired-end, split-read, and/or read-depth signals. However, the intrinsic limitations of short reads–and the even shorter fragments derived from split reads–often lead to alignment errors and unreliable SV calls. The emergence of third-generation long-read sequencing technologies has substantially improved SV detection. Recent tools such as Sniffles2[9], cuteSV[10], VolcanoSV[11], and DeBreak[12] leverage long reads to resolve complex SVs with higher accuracy. However, the high cost of long-read sequencing remains a major barrier to population-scale studies, while vast amounts of existing short-read data remain underutilized.

At the same time, advances in sequencing and bioinformatics have enabled the generation of increasingly complete genome assemblies, some achieving telomere-to-telomere (T2T) completeness. Because assembled contigs are substantially longer and more accurate than raw reads, they enable the identification of complex rearrangements. Several tools, including AsmVar[14], Assemblytics[15],

and SyRI[16], have been developed to leverage genome assemblies for SV discovery. For haplotype-resolved or phased genome assemblies, dipcall[17] and SVIM-asm[18] enable accurate SV detection in diploid species. However, SV detection and genotyping in haplotype-resolved assemblies of polyploid species–such as rose[19], potato[20], and sweetpotato[21]–have not yet been reported. More recently, graph-based pangenomes have emerged as a powerful framework for representing genome-wide variation within a species, overcoming limitations of a single linear reference. Graph-based approaches can encode multiple haplotypes and SVs simultaneously, and effective graph-based pangenome construction depends on integrating diverse data types, including high-quality, haplotype-resolved assemblies from long-read sequencing to capture complex SVs, together with short-read datasets to refine variant breakpoints and genotypes. Despite these advances, current SV detection methods remain largely sequencing-platform specific, and no existing tool supports comprehensive SV identification and genotyping across short reads, long reads, and haplotype-resolved assemblies.

To address these systemic challenges, we developed PSVGT, an integrated framework for population-scale SV detection and genotyping across short reads, long reads, assembled contigs, and chromosome-level assemblies. PSVGT incorporates a novel KLOOK algorithm, which leverages window-based signals and local depth adaptation to efficiently resolve multiple alleles within the same genomic region. Benchmark analyses demonstrate that PSVGT outperforms existing tools in both recall and genotyping accuracy, especially for low-depth long reads and haplotype-resolved assemblies, while also delivering high-precision SV calls from short-read datasets. In addition, the genotyping modules in PSVGT support multiple sequence types, enabling rapid targeted SV genotyping at scale without reprocessing genome-wide SV signals. PSVGT also provides annotations for insertion and deletion (InDel) SVs and includes an automated primer-design module that selects optimal InDel markers for agarose gel electrophoresis analyses, thereby accelerating genetic mapping, marker-assisted selection, and population genetic studies. Collectively, PSVGT addresses the evolving demands of next-generation and long-read genomics, as well as the increasing availability of haplotype-resolved assemblies, by combining comprehensive variant detection and genotyping with marker development in a unified, high-throughput framework.

Results

Overview of PSVGT

PSVGT is designed to accommodate a wide range of sequencing data inputs, including Illumina short reads and long reads from PacBio HiFi, PacBio CLR, and ONT platforms in FASTQ format, and *de novo* assembled contigs and chromosome-level genomes in FASTA format, or pre-aligned BAM files from any of these sequence types. Short reads are first assembled into contigs, after which all input sequences are used for SV signal extraction and genotyping. PSVGT provides two modes of SV signal detection: (i) capturing large InDels signals from intra alignments, and (ii) detecting InDels from both inter- and intra-alignments, as well as duplications, inversions, and translocations from inter-alignments (**Fig. 1a**). Candidate SV signals are subsequently processed using the KLOOK clustering algorithm to resolve multi-allelic events, together with local coverage-adaptive filtering to ensure both precision and computational efficiency, particularly in complex genomic regions with uneven sequencing coverage (**Fig. 1b**).

The PSVGT workflow comprises the following steps:

Step 1. *De novo* assembly of short-read datasets using MEGAHIT[22], followed by mapping all reads and contigs from all samples to the reference genome using minimap2[23], unless pre-aligned BAM files are provided.

Step 2. Extraction of SV signatures–including CIGAR-based InDels, split reads, spanning reads, breakpoints–from BAM alignment files, followed by KLOOK clustering to resolve multiple-allelic SVs and local depth-based filtering to group and refine candidate SV signals.

Step 3. Merging of candidate SVs across all samples to generate a non-redundant SV set (for sample sizes $\geqslant 2$).

Step 4. Genotyping of candidate SVs in each sample using integrated evidence from CIGAR-based InDel signatures, breakpoints, spanning reads, and split reads.

Step 5. Optionally, refinement of genotypes for *de novo*-assembled or short-read samples using short-read mapping evidence to reduce false positives introduced by assembly errors.

Step 6. Annotation of SVs with genomic coordinates, neighboring gene context, and predicted functional impacts.

Optional: Development of PCR-based SV and CAPS markers for agarose gel electrophoresis or marker-association studies.

A schematic overview of the PSVGT pipeline is shown in **Fig. 1c**, with details provided in the "Materials and Methods" section.

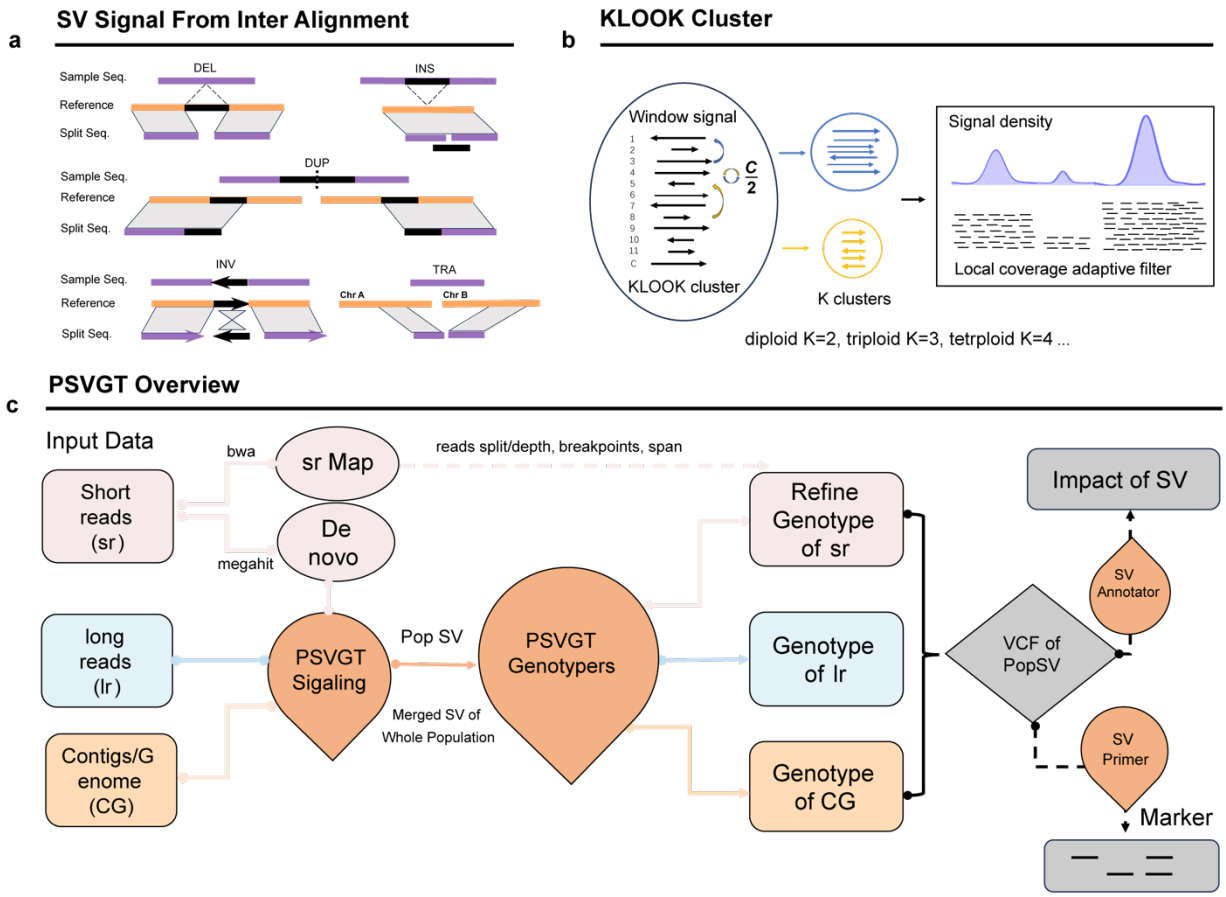


Fig. 1 | Overview of the PSVGT workflow. (a) SV inference from inter-alignments. SVs, including deletions (DEL), insertions (INS), duplications (DUP), inversions (INV), and translocations (TRA), are detected by aligning sample sequences (purple) to the reference genome (orange). Detected SVs are marked in black. **(b)** KLOOK clustering algorithm implemented in PSVGT. Signals within a window are clustered into K groups, where K is defined by ploidy, based on SV start positions, sizes, and local read depth. **(c)** Overall PSVGT workflow. PSVGT accepts diverse types of DNA sequences, including short reads (sr), long reads (lr), and assembled contigs/genomes (CG). Short reads are aligned to the reference genome using BWA and simultaneously de novo assembled into contigs. Candidate SV signals are collected from all samples and processed through KLOOK clustering and local depth-adaptative filtering within the PSVGT Signaling module. Population-level SVs (PopSVs) are then genotyped in each sample using sequence-type-specific sub-modules. Genotype calls from short-read assemblies are further refined using breakpoint information from read alignments. The final population-scale SVs in VCF format support downstream analyses, including marker development via the SV primer module and functional impact annotation via the SV annotation module.

PSVGT performance on simulated long-read datasets

To evaluate PSVGT performance on long-read data, we simulated multiple human sequencing data types, including haplotype-resolved assemblies, PacBio CLR, PacBio HiFi, and ONT reads. The simulated SV set comprised 10,000 deletions, 10,000 insertions, 1,000 inversions, and 1,000

131 duplications, following methods adapted from the DeBreak study[12], along with 365 translocations
132 based on cuteSV benchmarks[10] (**Fig. S1 & Table S2**). SV size distributions closely matched those
133 observed in the human HG002 genome. The simulated SVs were embedded into the human
134 haplotype-resolved reference genome (hs37d5.fa). Two-thirds of the simulated deletions, insertions,
135 duplications, and inversions were modeled as heterozygous. For translocations, two independent
136 simulated datasets were generated, one containing exclusively homozygous translocations and the
137 other exclusively heterozygous translocations. We benchmarked PSVGT against three state-of-the-
138 art long-read SV callers–DeBreak, cuteSV, and Sniffles2–across all sequencing data types. PSVGT
139 achieved the highest or comparable F1 scores across all SV classes (**Table 1**). Specifically, PSVGT
140 reached overall F1 scores of 99.72% (HiFi), 99.45% (CLR), and 99.67% (ONT), compared with
141 98.81%, 98.64%, and 98.66% for Sniffles2; 98.78%, 98.30%, and 96.26% for cuteSV, and 99.53%,
142 99.25%, and 99.36% for DeBreak. PSVGT also consistently delivered high recall and precision
143 across all SV classes (**Table 1**). Notably, PSVGT showed particularly strong performance in
144 resolving complex SVs, achieving recall rates of 99.9% (HiFi), 99.7% (ONT), and 99.6% (CLR) for
145 duplications, substantially exceeding those of Sniffles2 (87.0-89.2%), cuteSV (69.4-75.0%), and
146 DeBreak (95.3-97.8%).

147 In addition to SV detection accuracy, PSVGT consistently delivered superior SV genotyping
148 performance (**Table 1**). PSVGT achieved the highest genotype concordance for deletions, insertions,
149 and duplications on CLR and ONT datasets, and for duplications and translocations on HiFi datasets.
150 Specifically, PSVGT reached genotype accuracies of 98.90% (CLR), 96.79% (ONT), and 95.70%
151 (HiFi) for duplications, outperforming Sniffles2 (84.62-86.76%), cuteSV (83.50-89.48%), and
152 DeBreak (89.78-93.21%). For translocations, PSVGT achieved genotyping accuracies ranging from
153 98.90% to 99.11% across sequence types, compared with 96.67-97.57% for cuteSV, 86.10-93.83%
154 for Sniffles2, and 64.93-66.21% for DeBreak. While both PSVGT and cuteSV performed well in
155 inversion genotyping, PSVGT achieved higher recall rates and a more balanced trade-off between
156 recall and precision. Collectively, these results demonstrate that PSVGT provide robust and accurate
157 SV detection and genotyping across diverse long-read sequencing platforms, combining high recall,
158 precision, and genotype concordance.

Table 1 | SV discovery and genotyping accuracy (%) on simulated long-read sequencing data

Seq Type	PSVGT				Sniffles2				cuteSV				DeBreak			
	Rec	Pre	F1	GT	Rec	Pre	F1	GT	Rec	Pre	F1	GT	Rec	Pre	F1	GT

HiFi																
DEL	**99.97**	99.96	**99.97**	99.97	99.90	99.96	99.93	**99.98**	99.57	99.98	99.77	99.94	99.92	**99.98**	99.95	97.05
INS	**99.01**	99.92	**99.46**	99.85	98.94	99.37	99.15	95.23	98.04	**99.93**	98.98	**99.86**	98.55	99.81	99.17	85.05
DUP	**99.90**	**100**	**99.95**	**95.70**	87.80	**100**	93.50	84.62	75.00	100.00	85.71	84.40	97.80	**100**	98.89	89.78
INV	**99.10**	**100**	**99.55**	98.28	78.80	99.49	87.95	90.61	97.20	99.59	98.38	**100**	98.80	**100**	99.40	85.63
TRA	98.90	**100**	99.45	**99.11**	**100**	99.86	**99.93**	86.10	**100**	98.25	99.12	96.67	99.59	99.86	99.73	64.93
Total	**99.49**	**99.95**	**99.72**	**99.64**	97.95	99.67	98.81	96.83	97.65	99.94	98.78	99.36	99.15	99.90	99.53	90.79
ONT																
DEL	**99.98**	99.91	**99.95**	**99.96**	99.90	99.58	99.74	**99.96**	95.02	**99.93**	97.41	99.95	99.94	99.92	99.93	97.06
INS	**99.07**	**99.89**	**99.48**	**99.77**	98.90	99.09	98.99	95.23	93.36	99.81	96.48	99.82	98.58	99.53	**99.05**	85.38
DUP	**99.70**	99.60	**99.65**	**96.79**	87.60	99.77	93.29	86.76	69.40	99.71	81.84	89.48	95.30	**99.79**	97.49	89.82
INV	**97.90**	99.59	**98.74**	98.47	79.50	99.62	88.43	88.68	90.90	99.78	95.13	**100**	97.40	**99.80**	**98.58**	89.12
TRA	98.90	**100**	99.45	**98.90**	**100**	99.86	**99.93**	86.93	**100**	98.25	99.12	97.37	99.59	99.86	99.73	66.21
Total	**99.46**	**99.87**	**99.67**	**99.66**	97.96	99.36	98.66	96.84	92.91	99.86	96.26	99.54	99.00	99.73	99.36	91.10
PacBio CLR																
DEL	**99.99**	99.92	**99.96**	**99.97**	99.63	**99.95**	99.79	99.94	99.25	99.94	99.59	99.91	99.78	**99.95**	99.86	98.09
INS	**98.48**	99.34	**98.91**	**99.76**	98.05	98.98	98.51	95.51	96.71	99.83	98.25	99.47	97.71	**99.89**	98.79	88.03
DUP	**99.60**	99.70	**99.65**	**98.90**	89.20	99.78	94.19	85.87	72.10	99.86	83.74	83.50	95.70	99.90	97.75	93.21
INV	**99.60**	99.40	**99.50**	99.20	85.50	99.88	92.13	84.44	97.10	**100**	98.53	**100**	98.90	99.60	**99.25**	89.08
TRA	98.63	**100**	99.31	**98.97**	**100**	96.43	98.18	93.83	**100**	98.12	99.05	97.57	**100**	99.86	**99.93**	65.85
Total	**99.27**	99.63	**99.45**	**99.79**	97.80	99.50	98.64	96.72	96.76	99.89	98.30	99.16	98.61	**99.90**	99.25	92.93

Rec, recall; Pre, precision; F1, F1 score; GT, genotype concordance. The highest recall, precision, F1 score, and genotype concordance are marked in bold.

PSVGT performance on real long-read datasets

We further benchmarked PSVGT against DeBreak, cuteSV, and Sniffles2 using real human long-read data from the HG002 reference sample, comprising 60× PacBio CLR, 50× PacBio HiFi, and 47× ONT datasets. Since high-confidence benchmark call sets for inversions and translocations in HG002 are not available, our benchmarking focused on deletions and insertions. For deletion detection, PSVGT achieved the highest F1 scores across all long-read platforms (95.10-97.52%), outperforming DeBreak (89.95-97.39%), cuteSV (95.42-97.21%), and Sniffles2 (92.94-96.58%), while maintaining high genotype concordance (96.97-98.53%). Insertion calling was generally more challenging across all tools, with lower performance than for deletions. Nevertheless, PSVGT maintained strong performance, achieving recall of 92.65-97.32%, precision of 91.99-94.90%, and F1 scores of 92.32-95.51%, comparable to–and in some cases exceeding–those of DeBreak, cuteSV, and Sniffles2 (**Table 2**). Unlike most SV callers that apply a fixed supporting-read threshold, PSVGT integrates

173 locus-specific supporting-read counts with local read-depth estimates to derive adaptive support
174 criteria. Consistent with simulation benchmarks, PSVGT and cuteSV ranked among the top
175 performers in genotype concordance, with PSVGT achieving 93.68-98.70% compared with 94.61-
176 98.56% for cuteSV (**Table 2**). Taken together, these results closely mirror the simulation-based
177 evaluations and demonstrate that PSVGT achieves a robust balance between recall, precision, and
178 genotyping accuracy, consistently outperforming or matching state-of-the-art SV callers.

Table 2 | PSVGT performance on the real data of HG002 long-reads sequencing datasets

	PSVGT				Sniffles2				cuteSV				DeBreak			
	Rec	Pre	F1	GT	Rec	Pre	F1	GT	Rec	Pre	F1	GT	Rec	Pre	F1	GT
Deletion																
CLR	**97.29**	95.96	**96.62**	96.97	93.10	**96.14**	94.60	**97.16**	94.19	**96.68**	95.42	**97.07**	**96.41**	95.93	**96.17**	93.73
HIFI	**98.67**	**96.40**	**97.52**	**98.50**	98.10	95.11	96.58	**98.69**	98.29	96.16	97.21	**98.77**	**98.62**	**96.19**	97.39	96.19
ONT	**98.50**	**91.92**	**95.10**	**98.53**	97.07	89.14	92.94	**98.60**	97.57	**93.89**	**95.69**	98.07	**98.69**	82.63	89.95	95.32
Insertion																
CLR	**92.65**	**91.99**	**92.32**	93.68	90.04	90.03	90.03	**93.88**	91.62	88.66	90.12	**94.61**	**93.17**	**92.71**	**92.94**	86.73
HIFI	**97.32**	**93.86**	**95.51**	**98.70**	96.99	91.65	94.24	96.78	96.80	91.75	94.21	**98.56**	**97.50**	**93.70**	**95.56**	90.90
ONT	**92.98**	**94.90**	**93.93**	**94.11**	90.43	87.77	89.08	92.52	91.55	**91.16**	91.36	**98.23**	**95.13**	**93.57**	**94.34**	88.93

179 Rec, recall; Pre, precision; F1, F1 score; GT, genotype concordance. The top two highest of F1 score, recall, precision
180 and genotype concordance are shown in bold.

181 PSVGT performance on downsampled long-read datasets

182 To assess the effect of sequencing depth on SV detection accuracy and sensitivity, we generated
183 downsampled datasets at 5×, 10×, 15×, 20×, 30×, and 40× coverage using simulated PacBio CLR and
184 ONT reads, as well as PacBio HiFi reads from HG002. As expected, SV detection accuracy improved
185 with increasing sequencing depth across all tools **(Fig. 2)**. Notably, PSVGT achieved ~90% recall
186 and ~95% F1 score starting at 10× coverage, demonstrating higher sensitivity than DeBreak, cuteSV,
187 and Sniffles2. At 30× coverage, most tested tools reached saturation in recall and F1 score.
188 Nevertheless, PSVGT consistently achieved the highest performance, with recall and F1 score
189 ranging from 95.75% to 99.39%, followed by DeBreak (95.90-99.20%), Sniffles2 (95.12-98.47%),
190 and cuteSV (95.13-98.19%) **(Fig. 2)**. Importantly, the genotype concordance of PSVGT was largely
191 insensitive to low sequencing depth, maintaining approximately 95% even at 10× coverage. Although
192 DeBreak showed slightly higher recall and F_1 scores than cuteSV and Sniffles2, it exhibited

substantially lower genotyping accuracy, particularly at low coverages (<30×), remaining only slightly above 90% even at 50× coverage. Overall, PSVGT consistently outperformed existing tools in both SV detection sensitivity and genotyping concordance, particularly at low sequencing depths, owing to the local depth-adaptive filtering implemented in the KLOOK clustering module, which effectively accommodates uneven read coverage.

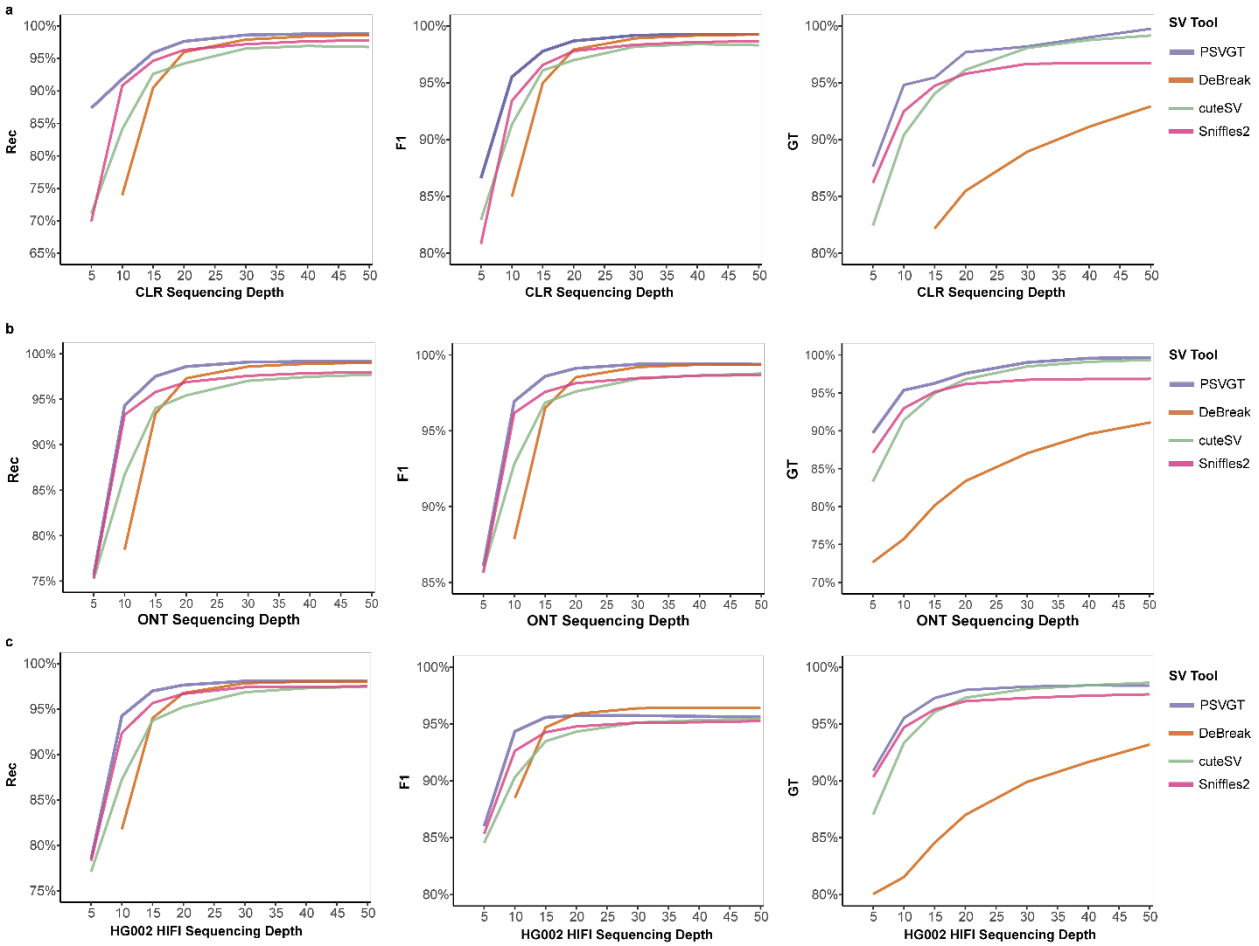


Fig. 2 | PSVGT performance on downsampled long-read datasets. **a-c**, Comparison of four SV detection tools on downsampled long-read datasets, including simulated PacBio CLR **(a)**, ONT **(b)**, and PacBio HiFi **(c)** datasets. Performance metrics–recall (Rec), F1 score (F1), and genotype concordance (GT)–are plotted against sequencing depth (x-axis).

Accuracy of SV detection using long-read datasets

SV detection sensitivity and breakpoint accuracy are often constrained by read support and signal density, prompting us to evaluate the performance across size-stratified SVs using long-read datasets. Across the full SV size spectrum, DeBreak and PSVGT consistently outperformed cuteSV and Sniffles2 in terms of F1 score (**Fig. 3a & Fig. S2a & Fig. S3a**). In particular, DeBreak excelled at

208 moderate-length insertions (100-300 bp), whereas PSVGT showed superior performance for large
209 insertions (>2 kb) (**Fig. 3a & Fig. S3a**). This advantage of PSVGT likely arises from the local depth-
210 adaptative filtering implemented in the KLOOK clustering module, which improves SV calling in
211 complex, low-coverage regions and enhances sensitivity to large SVs. For instance, PSVGT
212 successfully detected a 5,071-bp insertion (X:2238538-2238539) supported by only a single segment
213 alignment, which was missed by other state-of-the-art tools (**Fig. S4**).

214 Accurate breakpoint determination is critical for reliable SV genotyping. Using ONT data, PSVGT
215 identified 45.86% (4,159/9,069) of SVs with exact breakpoints and 58.69% (5,323/9,069) within 5-
216 bp shift, substantially outperforming Sniffles2 (38.05%), cuteSV (15.52%), and DeBreak (2.95%)
217 (**Fig. 3b**). This better performance of PSVGT was consistent across HiFi and CLR datasets, where
218 PSVGT exhibited markedly reduced breakpoint shifts compared to other tools (**Fig. S2b & Fig. S3b**).
219 Benchmarking on simulated data further corroborated these findings, demonstrating that the dynamic
220 clustering strategy implemented in PSVGT enables both accurate and efficient breakpoint resolution
221 (**Fig. S5-S7**). Together these results indicate that the dynamic KLOOK clustering algorithm in
222 PSVGT provides highly consistent breakpoint determination and effectively mitigates ambiguity in
223 SV breakpoint detection across diverse long-read sequencing platforms.

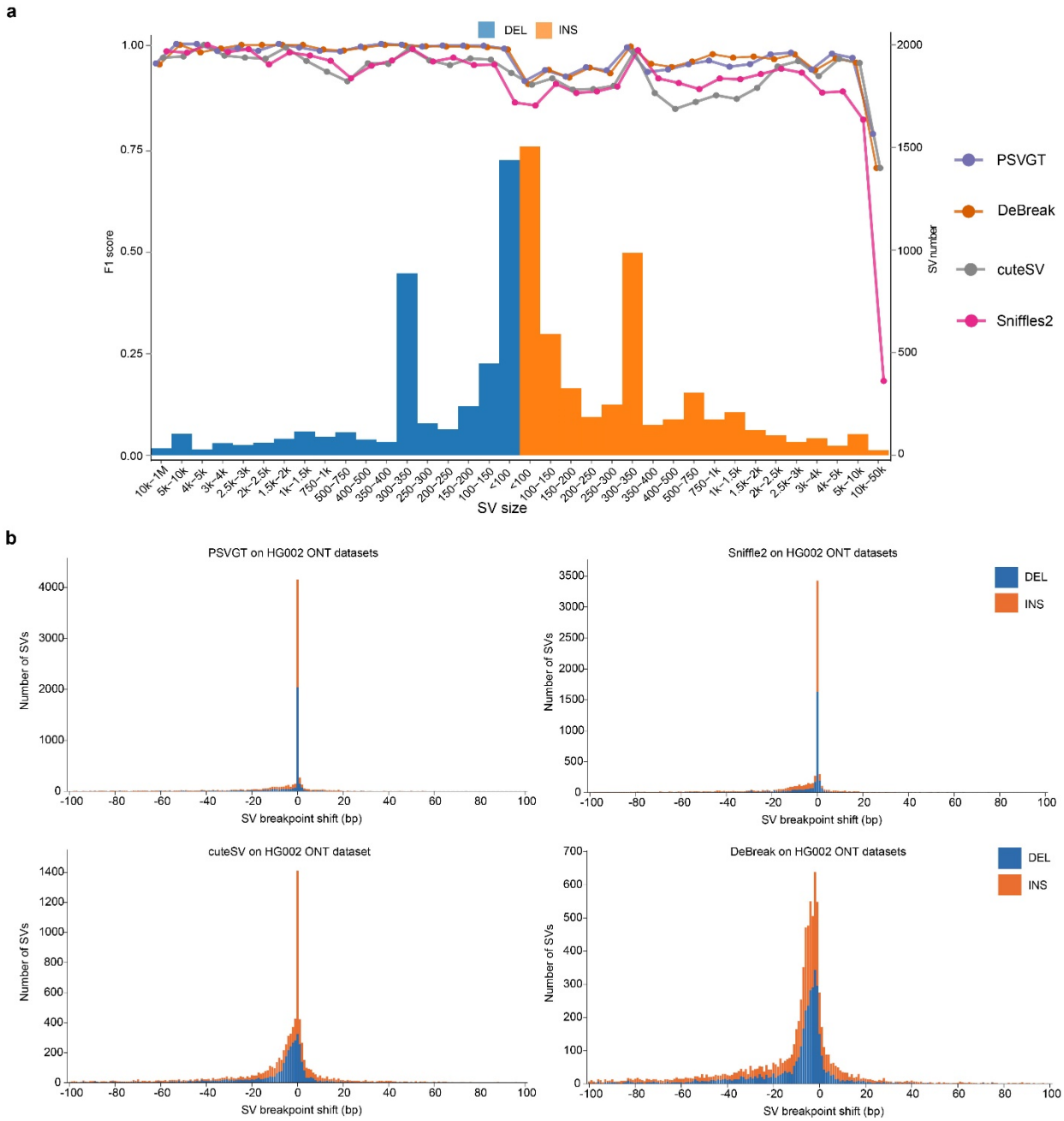


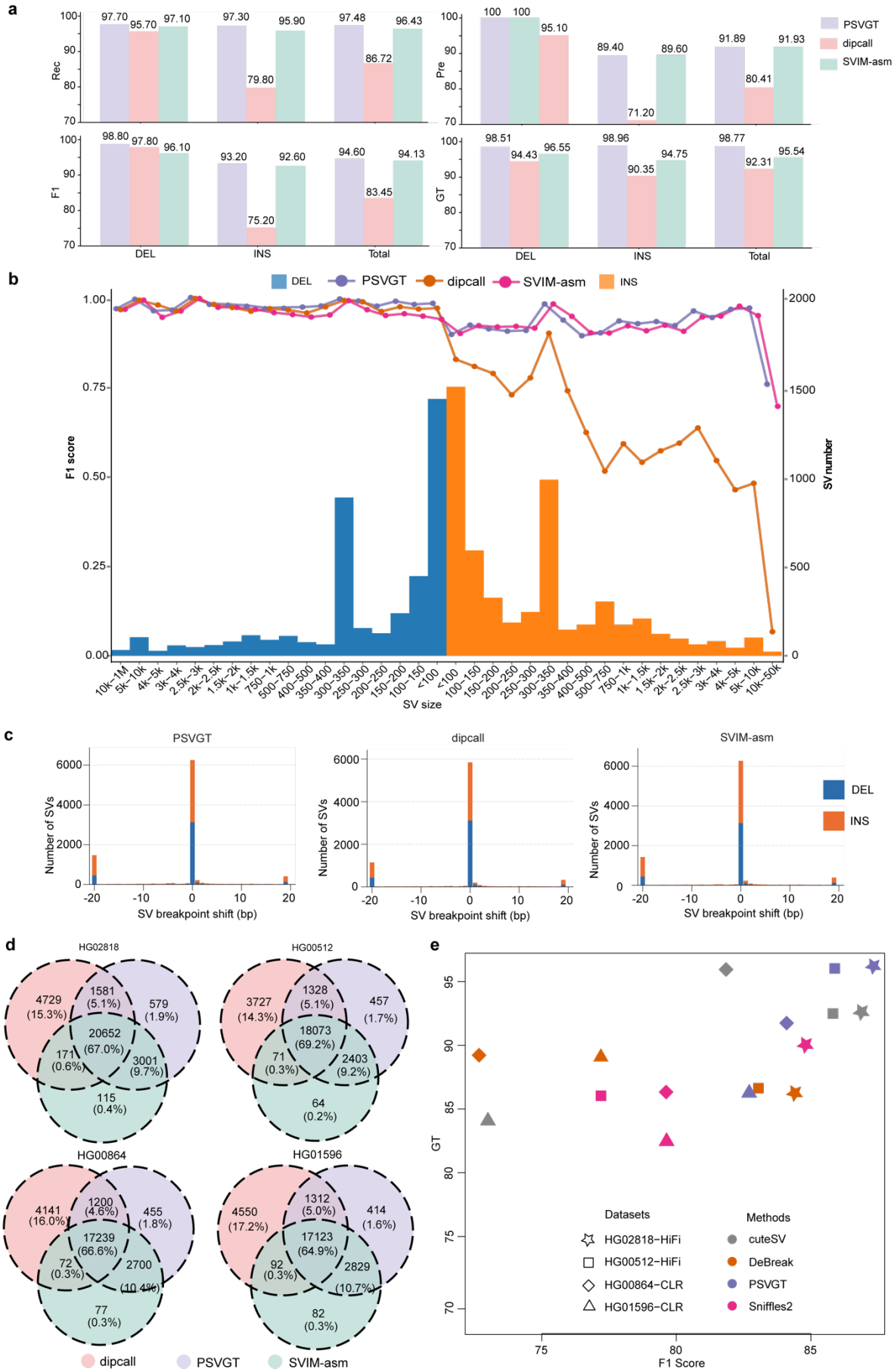
Fig. 3 | SV detection accuracy using ONT datasets. (a), F1 score distributions across SV sizes for deletions (DEL) and insertions (INS). Histograms indicate the number of SVs within each size bin, whereas lines show F1 scores. **(b)**, Breakpoint shift distributions (≤100 bp) of SVs detected from ONT datasets.

PSVGT performance on haplotype-resolved assembly

The depth-adaptive filtering implemented in PSVGT enables precise resolving of multiple sequence alignments, providing an effective strategy for discovering SVs between reference genomes and haplotype-resolved assemblies. To evaluate PSVGT performance on haplotype-resolved assemblies, we benchmarked it against two state-of-the-art tools, dipcall and SVIM-asm, using both simulated

233 data and real phased assembled genomes. All three tools achieved perfect recall, precision, and F1
234 scores on simulated benchmarks, as well as genotype concordance (exceeding 99%). In contrast,
235 performance diverged substantially on the real phased assembly (**Table S3**). The dipcall exhibited
236 the greatest decline, with recall and precision dropping to 86.72% and 80.41%, respectively, whereas
237 PSVGT and SVIM-asm maintained high recall rates of 97.48% and 96.43%, respectively. The
238 precision of both PSVGT and SVIM-asm reached ~92%. Notably, PSVGT achieved the highest
239 genotype accuracy (98.77%), outperforming SVIM-asm (95.54%) and dipcall (92.31%) (**Fig. 4a**).
240 Size-stratified analysis revealed comparable performance among the three tools for deletions, while
241 dipcall underperformed for insertions, with F1 scores falling below 75%, particularly for insertions
242 larger than 10 kb (**Fig. 4b**). This deficiency arises from an algorithmic limitation in dipcall, where
243 insertion lengths are incorrectly summed when the two haplotypes at the same locus harbor insertions
244 of different sizes (**Fig. S8**). In contrast, PSVGT showed a slight advantage in detecting large
245 insertions (>10 kb), achieving an F1 score of 75% compared with 70.59% for SVIM-asm. PSVGT
246 also demonstrated breakpoint accuracy comparable to assembly-specific callers, confirming its
247 sensitivity across both long-read and chromosome-level sequences (**Fig. 4c**).

248 Since PSVGT is designed to address SV detection and genotyping in both assembled genomes and
249 long-read datasets, we further validated its performance using haplotype-resolved genomes–
250 assembled from PacBio CLR/HiFi reads–from the HGSVC2 project[27]. Intersecting the SV call sets
251 generated by dipcall, PSVGT, and SVIM-asm revealed that 64.9-69.2% of total SVs were
252 constistently identified across samples. PSVGT showed high concordance (75.6-78.4%) with SVIM-
253 asm across all individuals, whereas dipcall produced a substantial number of likely false positives
254 due to the aforementioned insertion-length summation artifact (**Fig. 4d**, **Fig. S8**). Using the
255 intersection of PSVGT and SVIM-asm calls from phased assemblies as a provisional ground truth,
256 we next benchmarked PSVGT against long-read SV callers cuteSV, DeBreak, and Sniffles2 on the
257 corresponding PacBio CLR and HiFi datasets. Across all long-read benchmarks, PSVGT consistently
258 achieved superior performance in both F1 scores and genotype concordance. Notably, PSVGT
259 achieved F1 scores above 85% and maintained genotype concordance exceeding 96% on HiFi
260 datasets (**Fig. 4e & Table S4**). Together, these results demonstrate that PSVGT provides robust and
261 highly accurate SV detection and genotyping across both haplotype-resolved assemblies and long-
262 read sequencing data.


a
Rec
Pre
F1
GT
DEL
INS
Total
PSVGT
dipcall
SVIM-asm
b
DEL
PSVGT
dipcall
SVIM-asm
INS
F1 score
SV number
SV size
c
PSVGT
dipcall
SVIM-asm
Number of SVs
SV breakpoint shift (bp)
DEL
INS
d
HG02818
HG00512
HG00864
HG01596
dipcall
PSVGT
SVIM-asm
e
GT
F1 Score
Datasets
HG02818-HiFi
HG00512-HiFi
HG00864-CLR
HG01596-CLR
Methods
cuteSV
DeBreak
PSVGT
Sniffles2

264 **Fig. 4 | Performance of PSVGT on phased diploid assemblies and long-read datasets. (a)** Performance of PSVGT,
265 dipcall, and SVIM-asm on HG002 phased assemblies. **(b)** F1 score distributions across SV sizes for phased HG002
266 assemblies. Histograms showthe number of deletions (DEL) and insertions (INS) within each size bin, while lines indicate
267 F1 scores. **(c)** Breakpoint shift distributions for SVs detected by PSVGT, dipcall, and SVIM-asm. **(d)** Venn diagrams of
268 SV calls among assembly-based methods on phased diploid assemblies, HG02818, HG00512, HG00864, and HG01596.
269 **(e)** Performance of SV discovery tools on HGSVC2 long-read datasets. Scatter plots show F1 score versus genotype
270 concordance (GT) for cuteSV, DeBreak, PSVGT, and Sniffles2 across four datasets.

271 # PSVGT performance on real autotetraploid genome datasets

272 Given PSVGT's capacity in supporting diverse sequence types and variable ploidy levels, we further
273 benchmarked its performance using real datasets derived from tetraploid potato. We assessed the
274 performance of PSVGT using fully haplotype-resolved genome assemblies as well as coordinated
275 HiFi datasets at ~30× coverage per haplotype that were available for three distinct cultivars–Eig, Flo,
276 and BdF. These datasets provided a rigorous benchmark for SV detection because of the complex
277 genomic architecture inherent in tetraploid organisms, in which multiple homologous chromosomes
278 can harbor distinct allelic variants. In this benchmark, PSVGT dynamically adapts the number of
279 clusters according to ploidy, enabling the identification of SV signals from four haplotype assemblies
280 and precise phasing of events such as deletions and insertions that occur on different combinations
281 of homologous chromosomes. For instance, triploid SVs were exclusively detected by PSVGT from
282 HiFi read mapping, as the algorithm effectively utilized local depth adaptation to resolve SV allele
283 dosage (**Fig. 5a**). In haplotype-resolved assemblies, SV genotypes were defined in a phased format
284 such as "1|0|1|0" when the same SV was detected on haplotypes 1 and 3, providing fine-grained
285 resolution of allele dosage and inheritance patterns in polyploid genomes. Across the three potato
286 cultivars, the numbers of SVs identified at different ploidy levels revealed substantial allelic
287 complexity. Diploid SVs, present in two of the four haplotypes, ranged from 8,806 to 18,722 per
288 cultivar. Triploid SVs, in which three haplotypes carried SVs of distinct size, ranged from 447 to
289 1,761, whereas fully tetraploid SVs–distinct SVs present on all four haplotypes–were comparatively
290 rare, with 12 to 72 events per cultivar (**Fig. 5b, Fig. S9**). To further assess PSVGT sensitivity for
291 polyploid SV detection using long-read data, we treated these diploid, triploid, and tetraploid SVs
292 identified from haplotype-resolved genome assemblies as the ground truth set of multi-allelic SVs
293 (mSVs) (**Table S5, S6, S7**). Although performance was lower than that observed in diploid genomes,
294 PSVGT consistently achieved the highest recall across all three cultivars, outperforming existing
295 long-read SV callers such as cuteSV, Sniffles2, and DeBreak (**Fig. 5b**). The better performance of

PSVGT stems from its KLOOK clustering design, which dynamically adjusts cluster numbers for polyploid genomes and utilizes allele-specific local depth-adaptive filtering to recover SVs in complex, low-coverage regions. In contrast, existing SV callers such as cuteSV, Sniffles2, and DeBreak are primarily optimized for autodiploid (human) genomes and exhibit limited sensitivity when applied to polyploid data.

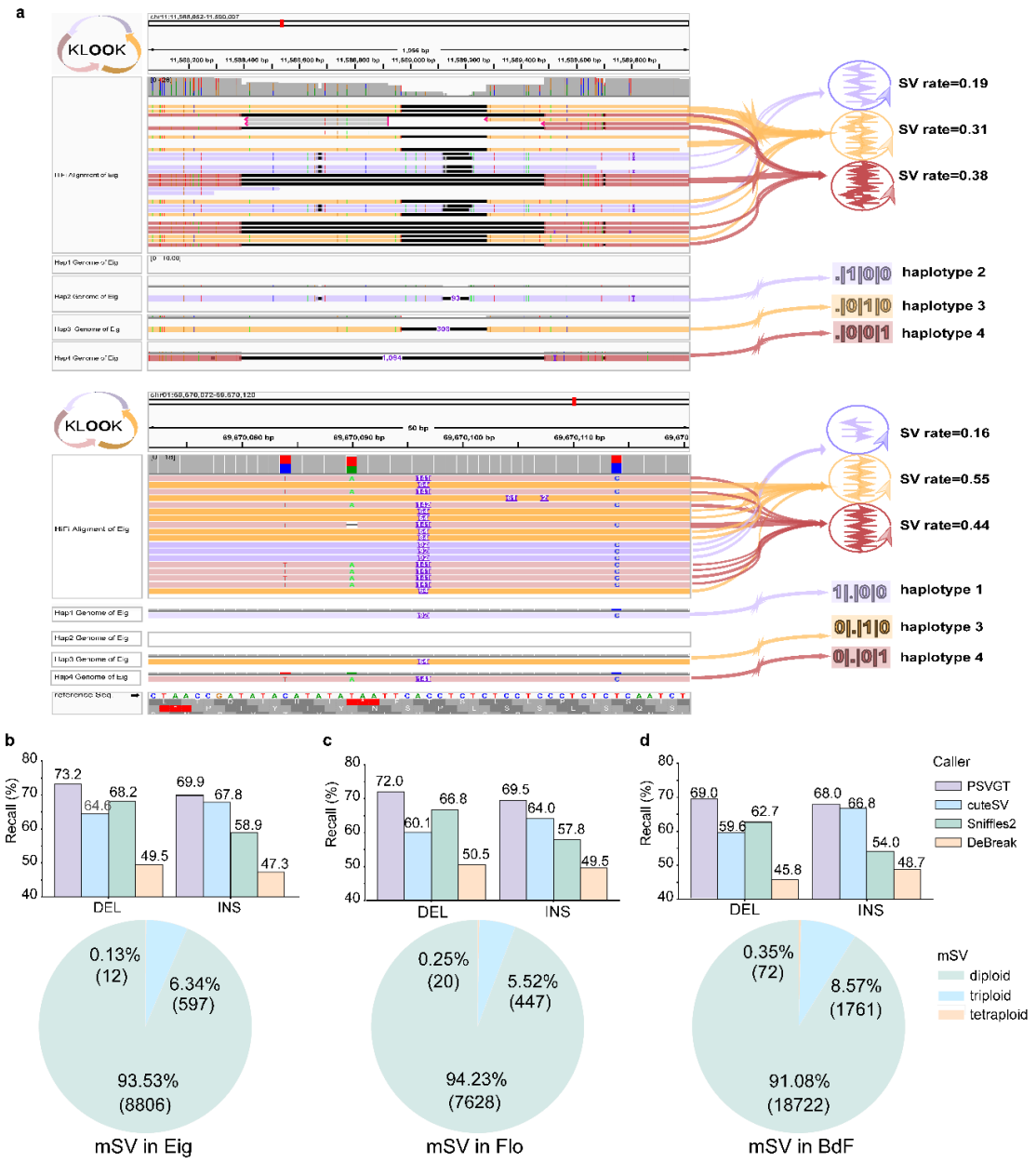


Fig. 5 | Performance of PSVGT on autotetraploid species. (a) Example of a triploid deletion and insertion event in the four-haplotype genome (top) and the corresponding HiFi read alignment (bottom). **(b-d)** Recall performance and distribution of multi-allelic SVs (mSVs) in the tetraploid potato cultivars Eig (b), Flo (c), and BdF (d). Bar plots show mSV recall rates for SV callers, while pie charts indicate the proportions of diploid, triploid, and tetraploid SVs in each cultivar.

PSVGT performance on real datasets of allopolyploid species

Current long-read SV detection tools are predominantly designed and optimized for diploid genomes, such as humans or inbred model species. However, many important crop plants possess allopolyploid genome architecture. Allopolyploid species, arising from the hybridization of two or more divergent progenitors, present substantial challenges for accurate SV discovery and genotyping. Specifically, high sequence similarity between homologous chromosomes, uneven allele dosage, and frequent inter-subgenome rearrangements can confound alignment- or split-read-based SV detection. To assess PSVGT performance in such complex genomes, we evaluated PSVGT using haplotype-resolved genome assemblies of sour orange (*Citrus* × *aurantium* (Lour.) Engl.), an allopolyploid species composed of two divergent parental genomes, designated here as hap-M (maternal) and hap-P (paternal). Using the haplotype-resolved assemblies, PSVGT identified 13,968 deletions and 13,422 insertions in the hap-M genome, and 17,327 deletions and 16,897 insertions in the hap-P genome. These SVs corresponded to 22,977 diploid SV loci and 45,954 alternative alleles across the two subgenomes. To evaluate SV detection performance on long-read data, we treated the assembly-derived SVs as the ground truth mSV set and assessed recall across four SV callers. Sniffles2 recovered 13,599 alternative alleles, cuteSV recovered 14,265, and DeBreak recovered 12,357. In contrast, PSVGT recalled 16,254 SV alleles, outperforming the three other tools (**Table S8**). This result demonstrates that PSVGT substantially enhance sensitivity for detecting allelic variation in allopolyploid genomes. In addition to superior recall, PSVGT achieved the highest F1 score and genotyping concordance among the evaluated tools, indicating a favorable balance between precision and sensitivity (**Table S9**). Collectively, these results show that PSVGT not only captures more mSV events in allopolyploid genomes but also provides accurate genotyping of those variants. Taken together, our analyses demonstrate that PSVGT is well-suited for SV discovery and genotyping in allopolyploid plant species, underscoring its potential for population genomics and evolutionary studies in complex non-model plant systems.

PSVGT performance on population-scale short-read datasets

Population-scale resequencing predominantly relies on short-read datasets; however, accurate detection of insertions from short reads remains a major challenge. We benchmarked PSVGT against

335 two state-of-the-art short-read SV callers, Manta and DELLY, using both simulated and real HG002
336 short-read datasets. While all three callers showed comparable performance for deletion detection,
337 PSVGT demonstrated clear superiority in insertion detection. In simulated datasets, PSVGT achieved
338 an insertion recall of 94.59%, substantially higher than Manta (67.65%) and DELLY (26.34%)
339 (**Table S10**). In real HG002 short-read data, insertion detection was challenging for all tools due to
340 short read length, heterozygosity and repeats. Nevertheless, PSVGT maintained an insertion recall of
341 16.65%, approximately twice that of Manta and nearly 20-fold higher than that of DELLY (**Table**
342 **S10**). To further evaluate PSVGT across different genomic architectures, we assessed its performance
343 in plant species spanning a wide range of repeat content, including *Arabidopsis thaliana* (27%
344 repetitive)[29], rice (51%)[30], lettuce (74.2%)[31], and maize (84%)[32]. We first applied PSVGT to short-
345 read datasets from eight *A. thaliana* ecotypes, with SVs identified by both PSVGT and SVIM-asm
346 from genome assemblies and genotyping results derived from corresponding long-read alignments
347 defined as ground-truth benchmarks (**Table S11**). Across these ecotypes, PSVGT achieved the
348 highest overall recall for indels, with an average recall of 64.62% (58.57-67.86%), substantially
349 outperforming Manta (36.75%; 27.22-43.27%) and DELLY (27.28%; 21.25-32.80%) (**Fig. 6a &**
350 **Table S11**). Notably, the three tools showed pronounced performance differences in deletion and
351 insertion detection due to their distinct algorithm designs. F1 scores for insertions were consistently
352 lower than those for deletions; however, PSVGT showed only a modest decline compared with
353 DELLY and Manta. Although DELLY achieved the highest precision (91.64%) and genotype
354 concordance (99.26%), its recall for insertions was extremely low (5.66%), resulting in a substantially
355 reduced F1 score of approximately 10.62%. In contrast, PSVGT achieved a balanced performance,
356 with 56.59% insertion recall, 88.88% precision, and 94.91% genotype concordance (**Fig. 6a & Table**
357 **S11**). Importantly, PSVGT's performance on short reads (recall >62%, precision >87%, F1
358 score >73%, and genotype concordance >94%) closely mirrored that of long-read data. This
359 comparable performance likely reflects PSVGT's strategy of assembling short reads into draft contigs
360 with lengths comparable to long reads for SV detection, thereby effectively recapitulating long-read
361 signals for SV detection and genotyping. In addition, we assessed breakpoint accuracy by by refining
362 genotype (PSVGT_srGT) using short reads pre-aligned BAM and found that SV breakpoints
363 supported by short-read alignments were highly consistent with those inferred by PSVGT (**Fig. 6a)**.

Similar trends were observed in rice, maize, and lettuce short-read datasets, in which PSVGT consistently outperformed Manta and DELLY in both recall and precision, particularly for insertion detection **(Fig. S10, Table S12-S14)**. Collectively, these results demonstrate that the *de novo* assembly-based strategy implemented in PSVGT substantially improves insertion detection from short-read data, outperforming current state-of-the-art tools and exhibiting robust performance across plant genomes with different repeat content. This contrasts with the local reassembly strategy employed by Manta and the split-read/paired-end mapping approaches used by DELLY, highlighting PSVGT's versatility for population-scale SV discovery.

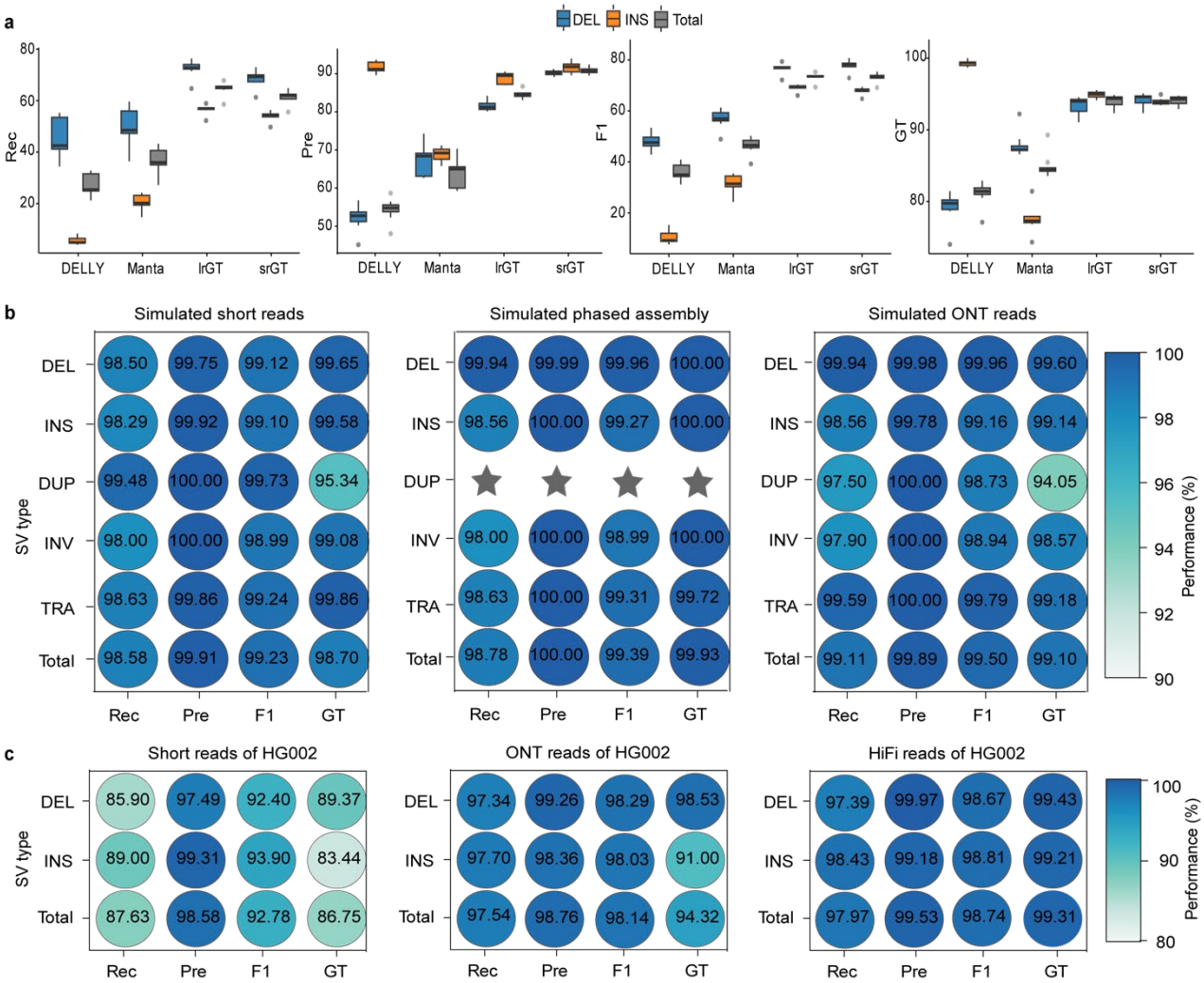


Fig. 6 | Performance of PSVGT on short-read datasets and SV genotyping. (a) Comparison of SV detection metrics across different tools. Boxplots show F1 score (F1), genotype concordance (GT), precision (Pre), and recall (Rec) for DELLY, Manta, lrGT (genotyping based on Short-read–derived contigs alignment) and srGT (genotyping based on short-read alignment), stratified by deletions (DEL), insertions (INS), and all SVs combined (Total). **(b, c)** Genotyping performance on simulated data **(b)** and real data of HG002 **(c)**. Performance metrics include recall (Rec), precision (Pre), F1 score (F1), and genotype concordance (GT) across different SV types and sequencing platforms.

Force-genotyping of PSVGT

Most of Current SV tools rely on individual signals to assign genotypes, While PSVGT is designed to enable population-scale SV detection and genotyping; its genotyping modules is specifically engineered to implement force-genotyping across large-scale population samples for predefined SVs without the computational burden of whole-genome SV discovery. To evaluate both recall and genotype concordance of PSVGT's force-genotyping module, we applied PSVGT to 21,843 candidate SVs derived from a simulated PacBio HiFi dataset and augmented this set with 11,700 synthetic false positives generated by shifting SV coordinates by 50 kb. These SVs were genotyped using 50× short reads, simulated diploid assemblies, and ONT reads. By leveraging candidate signals from the HiFi dataset, PSVGT achieved superior genotyping performance across short-read, phased assembly, and ONT datasets, maintaining recall above 97% and genotype concordance approaching or exceeding 99% across all SV types (**Fig. 6b**). For short-read datasets, genotype concordance reached 99% for most SV types, with the exception of duplications, which showed slightly reduced concordance (94.05-95.34%) due to inherent alignment limitations.

We further benchmarked SV genotyping on the real HG002 dataset using 8,906 high-confidence SVs and 6,416 synthetic false positives, using 30× short-read, 47× ONT-read, and 50× HiFi-read datasets. Deletions were genotyped more accurately than insertions, with long-read datasets achieving recall above 97% and genotype concordance exceeding 98% for most SV types. In contrast, on 30× short-read data PSVGT achieved an overall recall rate of 87.63% and genotype concordance of 86.75% (**Fig. 6c**). The reduced performance on HG002 relative to simulated datasets is primarily due to genuine genomic divergence from the reference genome, including abundant SNPs and small indels that can potentially compromise alignment accuracy.

Integrating long- and short-read sequencing is commonly used for population-scale SV discovery and genotyping, offering a balance between accuracy and cost. Using 17,895 non-redundant SVs identified from eight *Arabidopsis* ecotypes as the ground truth, we applied the force-genotyping module of PSVGT to population-scale short-read datasets. Overall, genotyping performance exceeded that observed in the HG002 benchmarks. ~~s~~. Notably, insertion genotype concordance (88.69-90.83%) was comparable to or higher than deletion concordance (84.85-87.25%), a pattern that contrasts with the HG002 results (**Table S15**). This difference likely reflects the largely

408 homozygous nature of *Arabidopsis* ecotypes, whereas human samples exhibit substantially higher
409 heterozygosity. Collectively, these results demonstrate that PSVGT enables rapid and robust SV
410 genotyping across sequencing platforms, featuring high recall and accurate genotyping in population-
411 scale short-read datasets.

412 SV annotation and marker development module in PSVGT

413 SVs can influence genome function through impacts on gene bodies, transcript architecture, and gene
414 expression. To facilitate functional interpretation of detected SVs, PSVGT incorporates an annotation
415 module that classifies putative functional impacts of SVs based on their genomic context. SVs are
416 categorized into the following functional classes: SVExp (overlaps with UTRs or promoters
417 potentially affecting gene expression), GenePAV (presence/absence variation caused by loss of
418 coding sequences), IntronExten/Shorten/Loss (intronic structural alterations), Frameshift, and
419 GeneAS (disruption of canonical splice sites leading to alternative splicing). Applying this annotation
420 framework to *Arabidopsis thaliana* ecotypes (Db-1 versus Col-0) revealed that 80.9% of detected
421 SVs were intergenic and the remaining 19.1% were predicted to have potential functional impacts,
422 with 2.5% classified as SVExp, 3.8% as Frameshift, 5.8% as intronic changes
423 (IntronExten/IntronShorten/IntronLoss), 5.4% as GeneAS, and 1.6% as GenePAV (**Fig. 7a**). These
424 annotations provide valuable insight into the functional relevance of SVs and offer a basis for
425 downstream prioritization in studies of gene regulation, trait association, and genome evolution.

426 Molecular markers are widely used in population genetics and breeding, particularly for gene
427 mapping and cloning. To further facilitate the application of SVs in genomic and genetic analyses,
428 PSVGT integrates a marker-development module-- SVInDelPrimer, enabling indel marker
429 development by selecting user-defined SV size ranges and designing primers from flanking
430 sequences. Using eight *Arabidopsis* short-read datasets, PSVGT consistently outperformed Manta
431 and DELLY in detecting InDels of experimentally optimal sizes (50-600 bp), enabling genome-wide,
432 high-density marker development (**Fig. 7b**). For deletion markers, PSVGT uniquely identified 689
433 candidates from a shared core set of 865, compared with fewer than 55 unique deletions detected by
434 either Manta or DELLY (**Fig. 7c**). For insertion markers, PSVGT reported 1,765 candidates,
435 including 967 unique events (**Fig. 7d**), whereas Manta identified 787 candidates (90% overlapping
436 with PSVGT) and DELLY identified 275 candidates (97% overlapping with PSVGT). The high

intersection rate of SV markers indicates that PSVGT captures a broader and reliable set of SV markers from short-read datasets. Primer design for candidate SV markers in PSVGT is performed using the primer3-py API, yielding optimized, high-confidence PCR primers for downstream experimental validation.

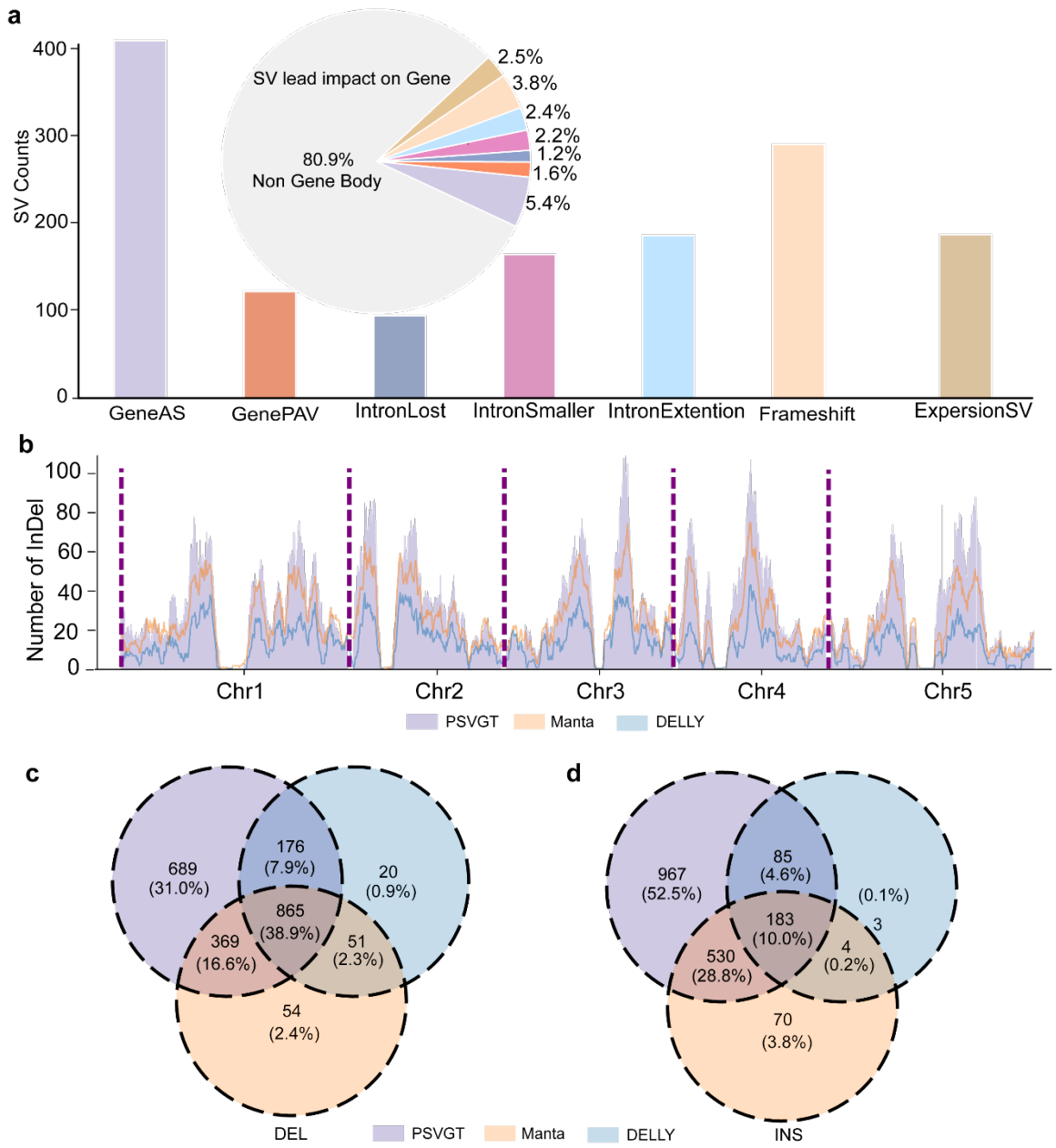


Fig. 7 | SV annotation and marker development module of PSVGT. (a) Example annotation of SVs in the PSVGT module. Annotation features are represented by distinct colors. The bar plot indicates the number of SVs annotated under each representative term, while the pie plot displays the proportion of each annotation category. **(b)** Density distribution of 50–600 bp InDels across *Arabidopsis* chromosomes. Density curves, color-coded for tools (PSVGT, purple; Manta, orange; DELLY, blue) show the density of InDel number across chromosome1–5. (**c-d)** Intersection of InDel calls among tools. Venn diagrams showing the number of deletions **(c)** and insertions **(d)** identified by PSVGT, Manta, and DELLY across all eight datasets.

Runtime and memory usage

To assess runtime and memory efficiency, we compared PSVGT with other long-read SV callers using the 50× pre-aligned PacBio HiFi BAM file from HG002. All analyses were conducted on a high-performance server equipped with an Intel Xeon Gold 6230R CPU (104 cores at 2.10 GHz) and 512 GB RAM, using 10 threads. PSVGT completed SV detection—including split-read and depth signal extraction, local depth-adaptive filtering, ploidy-aware KLOOK clustering, and genotyping—in 35 min with a peak memory usage of 3.9 GB. In comparison, cuteSV completed in 25 min using 5.1 GB RAM, DeBreak in 20 min using 10.2 GB RAM, and Sniffles2 in 15 min using 2.3 GB (**Table S16**). The modest increase in runtime for PSVGT reflects its more comprehensive filtering and clustering strategies, which improve sensitivity in regions with uneven or complex coverage.

We then evaluate runtime and memory usage using 50× human short-read dataset. *De novo* assembly using MEGAHIT required 2.5 h and a peak memory of 139.4 GB. Following assembly, PSVGT detected and genotyped SVs from the assembled contigs in 8.5 min using 3.0 GB RAM. In comparison, Manta processed the pre-aligned BAM file in 14 min with peak memory of 0.9 GB, whereas DELLY required 97 min (only supporting single thread) and 6.0 GB RAM (**Table S17**). The genotyping phase of PSVGT, which assigns allelic states to 25,561 SVs based on short reads pre-aligned BAM file, completed in under 3 min with a peak memory of 0.9 GB (**Table S16**). Overall, PSVGT completed SV discovery and genotyping in ~35 min on a standard 50× long-read human dataset and in ~8 min on assembled contigs. These results indicate that advanced filtering and clustering features based on ploidy in PSVGT do not introduce excessive computational overhead, making it well suited for high-throughput SV analysis in diploid and polyploid genomes, as well as population-scale studies.

Discussion

Structural variation is a major contributor to genomic diversity and disease etiology, exerting complex functional effects by directly altering gene structures and cis-regulatory elements and indirectly perturbing long-range regulatory interactions. Despite its importance, accurate SV discovery, genotyping, and interpretation remain challenging, particularly in species with complex genome architectures. Existing SV callers are typically optimized for specific sequencing technologies or data

477 types: DELLY and Manta are designed for short reads; cuteSV, Sniffles2, and DeBreak focus on long
478 reads; and dipcall and SVIM-asm rely on genome assemblies. In contrast to these specialized SV
479 callers, PSVGT provides a unified framework that integrates short-read, long-read, and whole-
480 genome assemblies datasets, achieving performance that matches or exceeds state-of-the-art tools
481 across all data types. Notably, PSVGT overcomes key limitations of short-read-based SV detection–
482 especially for insertions–by adopting a *de novo* assembly strategy that achieves higher precision and
483 recall than Manta and DELLY. By integrating complementary sequencing modalities, PSVGT
484 leverages the sensitivity of long reads and assemblies while remaining compatible with widely
485 available short-read data, enabling the recovery of SVs that are routinely missed by conventional
486 short-read pipelines.

487 Polyploid genomes comprise multiple homologous subgenomes that may each harbor distinct SV
488 alleles, and allopolyploids combine divergent parental haplotypes, greatly complicating sequence
489 alignment, SV discovery, and genotype inference. A major innovation of PSVGT is its local depth-
490 adaptive framework, which is specifically designed to handle complex genome regions and polyploid
491 species. In contrast to most existing SV callers, which implicitly assume diploidy and are unable to
492 estimate allele dosage beyond two haplotypes, the KLOOK clustering algorithm in PSVGT
493 dynamically infers multiple SV clusters based on user-specified ploidy. This design enables accurate
494 resolution of multi-allelic SVs. By extending SV genotype inference to arbitrary ploidy levels,
495 PSVGT provides robust SV discovery and genotyping in both autopolyploid and allopolyploid
496 species, which are common among crop plants.

497 SV analysis at the population scale requires both high sensitivity and computational efficiency. The
498 force-genotyping module implemented in PSVGT enables genotyping of predefined SV sets across
499 thousands of samples without reprocessing whole-genome SV signals, thereby facilitating
500 population-scale SV genotyping from diverse sequencing data types. In the *Arabidopsis* population
501 benchmark, force-genotyping achieved 87.81% genotype concordance with HiFi-derived genotypes,
502 highlighting its suitability for large-scale population studies. Combined with SV functional
503 annotation, PSVGT provides an integrated framework for post-GWAS analyses and functional
504 genomics studies. Moreover, the comprehensive SV sets inferred by PSVGT can be directly
505 incorporated into pangenome graph construction tools such as vg[33] and Minigraph[34]. Such integration

506 is expected to reduce reference bias and improve genotype accuracy in pangenome analyses,
507 particularly for species with extensive genomic structural diversity.

508 Despite the robustness of PSVGT across diverse sequencing data types and species, limitations
509 remain when analyzing short-read data in highly repetitive genomic regions. First, as with most SV
510 callers, PSVGT faces challenges in resolving SVs located within highly repetitive sequences,
511 primarily due to that short-read-derived contigs often fail to assemble accurately in these regions.
512 Additionally, incorporation of poorly aligned short reads in such regions can lead to missed SV calls
513 and erroneous SV genotyping. Second, although the KLOOK algorithm restricts clustering to $\frac{C}{2}$
514 comparisons (where C is the number of aligned SV signal reads) within a target window to efficiently
515 resolve multi-allelic events, regions with extreme read depth–such as duplication-rich or high-copy-
516 number loci–may still experience modest runtime overhead. Nevertheless, KLOOK remains an
517 essential component of PSVGT due to its critical role in accurately resolving alleles, particularly in
518 complex polyploid genomes where multi-allelic SVs are prevalent. Importantly, the integrated
519 processing of hybrid data types in PSVGT–including short reads, long reads, and assemblies–
520 substantially mitigates these challenges by improving assembly continuity and spanning repetitive
521 regions. Together, these capabilities advance SV discovery and genotyping in complex genomes,
522 reinforcing the utility of PSVGT for genomic research, population studies, and crop improvement.

523 Materials and Methods

524 Raw SV signal detection

525 To enable SV detection across diverse sequencing data types, PSVGT employs distinct strategies for
526 short- and long-read datasets. For short-read datasets, PSVGT implements a *de novo* assembly-based
527 workflow using MEGAHIT (v1.2.9)[22] to reconstruct contigs prior to variant calling. For long-read
528 datasets, raw FASTQ files are aligned directly to the reference genome using minimap2 (v2.24)[23].
529 PSVGT then extracts SV signatures from both intra- and inter-alignments (**Fig. 1a**). Discordant
530 alignment orientations and clipped read positions are used to detect deletions, insertions, duplications,
531 and inversions. Each raw SV signal records the start and end coordinates, SV length, supporting read
532 identifiers, mapping quality scores, and associated sequences. For contig and genome assembly
533 datasets, insertions and deletions are inferred from gaps or extensions within alignments. For long-

read datasets, deletions are inferred when reference segments are skipped between adjacent alignment blocks, while insertions are deduced from additional read sequences located between alignment blocks. Duplications are identified by overlapping alignments on the reference genome, and inversions are detected when paired segments align in opposite orientations. Translocations are identified from supplementary alignments ("SA" tags in BAM files) in which a single read maps to distinct chromosomes. All detected SV signals are subsequently clustered and refined using the KLOOK clustering algorithm.

Dynamic KLOOK clustering and local depth-adaptive filtering

To reduce background noise and handle varying signal densities across the genome, PSVGT implements a dynamic sliding-window clustering strategy coupled with local depth-adaptive filtering. Coordinates of each window are automatically adjusted according to the inferred SV length as defined follows:

SV length $\leq$ 100 bp: window = (SV_{start} - 200 - 0.2 × SV length, SV_{end} + 200 + 0.2 × SV length);

100 bp < SV length $\leq$ 500 bp: window = (SV_{start} - 400 - 0.2 × SV length, SV_{end} + 400 + 0.2 × SV length);

500 bp < SV length $\leq$ 1000 bp: window = (SV_{start} - 600 - 0.2 × SV length, SV_{end} + 600 + 0.2 × SV length);

SV length > 1000 bp: window = (SV_{start} - 1000, SV_{end} + 1000).

Within each genomic window, SV signals are first grouped by variant type and sorted by genomic coordinates. These signals are then clustered using the KLOOK algorithm, where k is set to the sample ploidy (e.g., 2 for diploid, 3 for triploid, and 4 for tetraploid). For each window, the signal will be aggregated through C/2 (C = total signals within the window) iterative comparisons, leveraging SV position, SV length, and read identifier. The top K clusters are retained as putative alleles corresponding to the ploidy level. Each allele from KLOOK is subsequently filtered based on the local coverage of the SVs ($\frac{\text{local depth}}{10} + 1 < \text{clustered SV reads}$).

Genotyping modules

PSVGT incorporates two specified genotyping modules, each tailored to unique characteristics of data types. The long-read genotyping module, applicable to long-read sequencing data, contigs, and assembled genomes, leverages two complementary sources of evidence for robust genotype inference: (1) local SV signal density and (2) breakpoint coverage information. In contrast, the short-read genotyping module, designed exclusively for short-read datasets, relies exclusively on breakpoint coverage derived from original short-read alignments (rather than assembly) as the primary evidence for genotype assignment.

Breakpoint validation

A candidate SV breakpoint is considered valid if the distance between the predicted SV boundary and clipped read ends (left/right) satisfies the following criteria:

For long-read dataset:

$$| SV_{start} - left_{clipped} | \leq 100 + (SV_{Length} \times 0.05)$$

$$| SV_{end} - right_{clipped} | \leq 100 + (SV_{Length} \times 0.05)$$

For short-read dataset:

$$| SV_{start} - left_{clipped} | \leqslant 30$$

$$| SV_{end} - right_{clipped} | \leqslant 30$$

Where SV_{start} and SV_{end} denote the predicted start and end of a candidate SV, $left_{clipped}$ and $right_{clipped}$ represent the clipped end positions of aligned reads, and SV_{length} is the length of candidate SV. This dynamic threshold account for inherent variability in long-read alignment acround breakpoints.

Breakpoint covered reads quantification

Reads are classified as breakpoint-covered reads if their alignment spans the breakpoint region with sufficient flanking sequence, as defined as below:

For long-read dataset:

$$Align_{start} + 500 \leq breakpoint_{position}$$

$$\text{Align}_{\text{end}} - 500 \geqslant \text{breakpoint}_{\text{position}}$$

For short-read dataset:

$$\text{Align}_{\text{start}} + 50 \leqslant \text{breakpoint}_{\text{position}}$$

$$\text{Align}_{\text{end}} - 50 \geqslant \text{breakpoint}_{\text{position}}$$

Where $\text{Align}_{\text{start}}$ and $\text{Align}_{\text{end}}$ are the start and end coordinates of the read alignment, and $\text{breakpoint}_{\text{position}}$ refers to either the left or right breakpoint coordinate of the SV. The 500 bp flanking requirement for long reads ensures that reads fully span the breakpoint region with high confidence.

Score calculation

Genotypes are inferred based on normalized scores, calculated as the ratio of supporting reads to local genomic depth defined as average read coverage within 1 kb window flanking the SV:

The SV allele score is calculated as:

$$\text{SV allele score} = \frac{\text{breakpoint reads} + \text{CIGAR reads}}{\text{local depth}}$$

Where breakpoint reads is the number of reads passing breakpoint validation, CIGAR reads is number of reads with CIGAR strings indicating SV-specific alignment patterns. For short-read datasets, CIGAR-based evidence is excluded due to read length limitations.

The reference allele score is calculated as:

$$\text{reference allele score} = \frac{\text{breakpoint covered reads}}{\text{local depth}}$$

Where breakpoint-covered reads are reads satisfying breakpoint coverage criteria and supporting the reference allele .

Benchmark metrics

To evaluate the performance of each tool, we calculated recall, precision, F1 score and genotype concordance.

607 Recall was defined as:

608 $$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

609 where TP (true positive) refers to SVs detected by the tool that are concordant with the true set, and
610 FN (false negative) refers to SVs present in the true set but not detected.

611 Precision was defined as:

612 $$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

613 where FP (false positive) refers to SVs detected by tool but not present the in true set.

614 F1 score was defined as:

615 $$\text{F1 score} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

616 Genotype concordance was defined as:

617 $$\text{genotype concordance} = \frac{\text{accurate GT}}{\text{TP}}$$

618 where accurate GT refers to true SVs whose genotypes are concordant with those in the true set.

619 Dataset simulation

620 We simulated SVs consisting of 10,000 deletions, 10,000 insertions, 1,000 duplications, and 1,000
621 inversions using scripts implemented in DeBreak, with all variants introduced into two independently
622 simulated haploid human genomes. Additionally, 365 translocation sites were collected from
623 published cuteSV datasets[10]. The size distribution of each SV type was consistent with that observed
624 in human genomes. The overall ratio of heterozygous to homozygous events was set to 2:1, with
625 alleles randomly assigned to the hs37d5 reference haplotype genome. For translocations, VISOR
626 (v1.1.2.1)[35] was used to generated both heterozygous and homozygous translocated genomes. PacBio
627 CLR datasets were simulated using PBSIM3 (v3.0.4)[24] with the parameters "--errhmm ERRHMM-
628 RSII.model --depth 25". ONT and PacBio HiFi reads were generated via Badread (v0.4.1)[36] with
629 parameters "--quantity 25× --junk_reads 0 --random_reads 0 --glitches 0,0,0" and parameters "--
630 quantity 25× --error_model pacbio2021 --qscore_model pacbio2021 --glitches 0,0,0 --junk_reads 0 -
631 -random_reads 0 --chimeras 0", respectively. For all datasets, the sequencing depth was uniformly

632 set to 25× per haploid genome. Short reads were simulated using InSilicoSeq (v2.0.1)[26] with options
633 "--model Novaseq", generating a sufficient number of reads to achieve an average sequencing depth
634 of 50×.

635 Benchmark on long-read datasets and assembled genome

636 Minimap2 (v2.24) was used to align PacBio CLR, PacBio HiFi, and ONT reads to the reference
637 genome, with parameters selected for each sequencing platform, as detailed in the minimap2 manual.
638 Sniffles2 (v2.6.0)[9], cuteSV (v2.1.1)[10], and DeBreak (v1.2)[12] were used to call SVs, with the minimum
639 number of supporting reads set from 3 to 10, and the highest F1 score of each tool was used for
640 benchmarking. Only SVs > 50 bp were retained for benchmarking. The haploid genomes were aligned
641 to the reference genome using minimap2 with parameters "-a -x asm5 --cs -t8". dipcall (v0.3)[17] and
642 SVIM-asm (v1.0.3)[18] were then applied to detect SVs from the two simulated haploid genomes with
643 default parameters. Final variant files in VCF format generated by PSVGT and other SV callers were
644 evaluated against the ground truth SV sets to compute precision, recall, F1 score, and genotype
645 concordance. For simulated data, a detected SV was considered a true positive if its breakpoint was
646 within 1 kb of the corresponding ground-truth variant and its length was within 80-120% of the true
647 SV size. Genotype concordance was recorded when the SV genotype matched the ground truth set.
648 For real datasets, SV calling performance was evaluated using Truvari (v5.1.1)[28] with parameters --
649 passonly -r 1000 -p 0.00 --dup-to-ins. High-confidence insertion and deletion variants from a publicly
650 available benchmark dataset were used as the truth set. Insertions, deletions, and duplications detected
651 by each tool were extracted and compared against the truth set using Truvari. The precision, recall,
652 F1 score, and genotype concordance were obtained from the Truvari reports.

653 For the four samples from Human Genome Structural Variation Consortium 2 (HGSVC2) resource,
654 SV detection was performed on genome assemblies using PSVGT, SVIM-asm, and dipcall, and on
655 the corresponding long-read datasets using PSVGT, cuteSV, DeBreak, and Sniffles2. Intersection SV
656 sets derived from phased assemblies were used as ground truth to assess the performance of long-
657 read SV detection. For each sample, we defined a series of minimum supporting long reads ranging
658 from 4 to 15 for SV detection and selected the optimal results from these tools for comparison with
659 the results of PSVGT. All call sets were benchmarked against the ground truth using Truvari with the

660 flags --passonly -r 1000 -p 0.00 --no-ref, generating precision, recall, F_1 scores and
661 genotype-concordance metrics for each.
662 To assess the sequencing depth effect on PSVGT, simulated PacBio CLR reads, ONT reads, and HiFi
663 reads were downsampled to depths of 5×, 10×, 15×, 20×, 30×, 40×, and 50× using samtools (v
664 1.15.1)[37]. Comparative callers (DeBreak, cuteSV, Sniffles2) were executed with a series of minimum
665 support reads based on the depth of datasets, whereas PSVGT was performed using default
666 configuration. Each downsampled call set was benchmarked against the corresponding ground truth
667 SV set using Truvari, generating precision, recall, F_1 score, and genotype concordance metrics.

668 Benchmark on tetraploid and allodiploid genome

669 For tetraploid potato, haplotype-resolved assemblies and corresponding PacBio HiFi reads were
670 aligned to DM_1-3_516_R44 (version 6.1) using minimap2. mSVs were inferred from the genome
671 alignment of the haplotype-resolved genomes when SVs were detected on different haplotypes with
672 a size fold change $\geqslant$ 0.8. A custom script was used to search mSVs from HiFi-derived VCF file to
673 identify regions harboring at least two alternative alleles, which were taken as true recall. The number
674 of clusters in PSVGT was set to four to reflect the tetraploid genome architecture of potato.

675 For allodiploid citrus, diploid genome assemblies were aligned to sweet orange reference genome
676 (SWO, version 3)[38]. The diploid SVs were inferred from the diploid genome alignments. The same
677 genomic region with SV size fold change $\geqslant$ 0.8 was defined as mSV. SV calls derived from PacBio
678 HiFi data using Sniffles2, cuteSV, DeBreak, and PSVGT were used to benchmark. The minimum
679 number of supporting reads was varied from 4 to 10, and the highest F1 score of each tool was used
680 for performance assessment.

681 Benchmark on short-read datasets.

682 Short-read–derived contigs and original short reads were aligned to reference genome using
683 minimap2(v2.24)[23] and bwa (v0.7.17)[39], respectively. To establish the ground-truth variant set for
684 benchmarking short-read performance, were first identified from haplotype-resolved assemblies
685 using SVIM-asm and the long-read module of PSVGT. Genotypes for these SVs were
686 subsequently assigned by intersecting the calls of cuteSV and PSVGT on the corresponding
687 ONT/HiFi datasets. This genotyped SV set was used as the benchmark for assessing the performance

688 of DELLY, Manta and short-read module of PSVGT. A detected SV was considered a true positive
689 if its predicted breakpoint was located within 1 kb of a ground-truth variant and its length was within
690 80–120% of the true SV size. Genotype concordance was recorded when the genotype call matched
691 the genotype in the ground-truth set.

692 Force genotyping

693 To evaluate the genotyping performance, SV calls derived from HiFi data using PSVGT were used
694 as the candidate set for genotyping. These SVs were subsequently genotyped in simulated short-read
695 datasets using the short-read genotyping module of PSVGT, and in diploid-resolved genome
696 alignments using the long-read genotyping module. For the *Arabidopsis* population, a
697 high-confidence truth set was constructed by jointly calling SVs from HiFi/ONT reads using the
698 long-read module of PSVGT (parameters: "-lr_homo_rate 0.75 -lr_ref_rate 0.15") and cuteSV. SV
699 calls from the two methods were merged, and redundant variants were removed to generate a
700 nonredundant SV set. This set was then used as the reference for forced genotyping across all
701 short-read samples using the short-read module of PSVGT (parameters: "-homo_rate 0.3 -ref_rate
702 0.15"). Genotype concordance was assessed by comparing the forced-genotyping outputs against the
703 truth set.

704 Marker module of PSVGT

705 The SVInDel_primer module enables high-throughput identification of experimentally tractable
706 InDels and produce size-distinguishable PCR amplicons for reference and variant alleles. The Primer
707 design is performed using the primer3 package implemented in Python[40, 41]. For each candidate InDel
708 locus, SVInDel_primer extracts reference sequences franking the predicted breakpoint and designs
709 forward and reverse primers from these regions. Primer design was conducted using uniform global
710 constraints, including an optimal primer length of 21 bp (rang 18-25 bp), optimal melting temperature
711 of 60 °C (range 53–65 °C), GC content 40–80%, and stringent limits on self-complementarity and
712 primer–primer interactions. No ambiguous bases were permitted in primer sequences. Primer pairs
713 are subsequently evaluated to identify breakpoint-spanning combinations that minimize amplicon
714 length while maintaining primer quality and specificity, ensuring efficient amplification of InDels
715 compatible with standard PCR assays.

716 # Data availability

717 The HiFi passed assemblies, HiFi, ONT, PacBio CLR, and Illumina sequencing datasets are available
718 at GIAB and NCBI. The high confidence benchmarks of HG002 are available at https://ftp-
719 trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/HG002
720 _SVs_Tier1_v0.6.vcf.gz. The high confidence regions is available at https://ftp-
721 trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/HG002
722 _SVs_Tier1_v0.6.bed. The phased assemblies and long read datasets of HGSVC2 are at
723 https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/assemblies/. Th
724 e raw illumina, HiFi, ONT sequencing of the 8 ecotypes of *Arabidopsis* were downloaded at EMBL-
725 ENA under the accession number PRJEB62038. And the genome assemblies of the 8 ecotypes of
726 *Arabidopsis* were downloaded at NCBI-SRA under the accession number PRJNA1033522. The rice
727 assembled genomes, long- and short-read datasets were downloaded at EMBL-ENA under accession
728 number PRJEB73710, and T2T rice genome NIP was downloaded from
729 http://www.ricesuperpir.com/web/download. The lettuce short-read datasets were downloaded at
730 China National Center for Bioinformation under accession number CRA014517. The maize short-
731 read datasets were downloaded at EMBL-ENA under accession number PRJEB31061. The haplotype
732 assemblies of the potato genome were downloaded from https://doi.org/10.5281/zenodo.14053896,
733 and the raw sequencing data were downloaded from NCBI-SRA under the accession number
734 PRJNA1074690. The diploid genome assemblies of sour orange were downloaded from
735 http://citrus.hzau.edu.cn/download.php, and raw sequencing datasets were downloaded at NCBI-
736 SRA under the accession number PRJNA1003804. All the datasets used in our study are listed in
737 **Table S1**.

738 # Code availability

739 All code is available at https://github.com/lgbTime/PSVGT under the MIT License.

740 The key custom scripts used in the manuscript are available at
741 https://github.com/lgbTime/PSVGT_Benchmarks_Code.

742 1. Wang Y, Liang N, Gao G. Quantifying the regulatory potential of genetic variants via a hybrid
743 sequence-oriented model with SVEN. *Nature Communications* **15**, 10917 (2024).

744

745 2. Collins RL*, et al.* A structural variation reference for medical and population genetics. *Nature* **581**,
746 444-451 (2020).
747
748 3. Zhang Y*, et al.* Structural variation reshapes population gene expression and trait variation in 2,105
749 Brassica napus accessions. *Nature Genetics* **56**, 2538-2550 (2024).
750
751 4. Dixon JR*, et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nature*
752 *Genetics* **50**, 1388-1398 (2018).
753
754 5. Ahsan MU, Liu Q, Perdomo JE, Fang L, Wang K. A survey of algorithms for the detection of genomic
755 structural variants from long-read sequencing data. *Nature Methods* **20**, 1143-1158 (2023).
756 6. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short
757 read general-purpose structural variant calling software. *Nature Communications* **10**, 3240 (2019).
758
759 7. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by
760 integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339 (2012).
761
762 8. Chen X*, et al.* Manta: rapid detection of structural variants and indels for germline and cancer
763 sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
764
765 9. Smolka M*, et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nature*
766 *Biotechnology* **42**, 1571-1580 (2024).
767
768 10. Jiang T*, et al.* Long-read-based human genomic structural variation detection with cuteSV. *Genome*
769 *Biology* **21**, 189 (2020).
770
771 11. Luo C, Liu YH, Zhou XM. VolcanoSV enables accurate and robust structural variant calling in diploid
772 genomes from single-molecule long read sequencing. *Nature Communications* **15**, 6956 (2024).
773
774 12. Chen Y*, et al.* Deciphering the exact breakpoints of structural variations using long sequencing reads
775 with DeBreak. *Nature Communications* **14**, 283 (2023).
776
777 13. Sedlazeck FJ*, et al.* Accurate detection of complex structural variations using single-molecule
778 sequencing. *Nature Methods* **15**, 461-468 (2018).
779
780 14. Liu S*, et al.* Discovery, genotyping and characterization of structural variation and novel sequence at
781 single nucleotide resolution from de novo genome assemblies on a population scale. *GigaScience* **4**,
782 64 (2015).
783

15. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**, 3021-3023 (2016).

16. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology* **20**, 277 (2019).

17. Li H*, et al.* A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature Methods* **15**, 595-597 (2018).

18. Heller D, Vingron M. SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* **36**, 5519-5521 (2021).

19. Zhang Z*, et al.* Haplotype-resolved genome assembly and resequencing provide insights into the origin and breeding of modern rose. *Nature Plants* **10**, 1659-1671 (2024).

20. Sun H*, et al.* The phased pan-genome of tetraploid European potato. *Nature* **642**, 389-397 (2025).

21. Wu S*, et al.* Phased chromosome-level assembly provides insight into the genome architecture of hexaploid sweetpotato. *Nature Plants*, (2025).

22. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674-1676 (2015).

23. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).

24. Ono Y, Asai K, Hamada M. PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinformatics* **29**, 119-121 (2013).

25. Wick RR. Badread: simulation of error-prone long reads. *Journal of Open Source Software* **4**, 1316 (2019).

26. Gourlé H, Karlsson-Lindsjö O, Hayer J, Bongcam-Rudloff E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics* **35**, 521-522 (2019).

27. Ebert P*, et al.* Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).

28. English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biology* **23**, 271 (2022).

29. Hou X, Wang D, Cheng Z, Wang Y, Jiao Y. A near-complete assembly of an Arabidopsis thaliana genome. *Molecular Plant* **15**, 1247-1250 (2022).

30. Shang L*, et al.* A complete assembly of the rice Nipponbare reference genome. *Molecular Plant* **16**, 1232-1236 (2023).

31. Reyes-Chin-Wo S*, et al.* Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications* **8**, 14953 (2017).

32. Wang B*, et al.* De novo genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. *Nature Genetics* **55**, 312-323 (2023).

33. Garrison E*, et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology* **36**, 875-879 (2018).

34. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biology* **21**, 265 (2020).

35. Bolognini D, Sanders A, Korbel JO, Magi A, Benes V, Rausch T. VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. *Bioinformatics* **36**, 1267-1269 (2020).

36. Wick RR. Badread: simulation of error-prone long reads. *Journal of Open Source Software*, (2019).

37. Danecek P*, et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).

38. Liu S*, et al.* Origin and de novo domestication of sweet orange. *Nature Genetics* **57**, 754-762 (2025).

39. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

40. Rozen S, Skaletsky H. Primer3 on the WWW for General Users and for Biologist Programmers. In: *Bioinformatics Methods and Protocols* (eds Misener S, Krawetz SA). Humana Press (1999).

41. Untergasser A*, et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Research* **40**, e115-e115 (2012).

Acknowledgements

This research was supported by the National Key Research and Development Program of China (2023YFF1000100), the scientific research start-up funding from Hubei Hongshan Laboratory award no. 11020102, the "Young Scientist Fostering Funds for the National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops". The computations in this paper were run on the bioinformatics computing platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University.

Author information

Authors and Affiliations

National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops, Huazhong Agricultural University, Wuhan, P. R. China

Guangbao Luo, Li Xiao, Xin Wang

Boyce Thompson Institute, Ithaca, NY, 14853, USA

Zhangjun Fei

Contributions

X.W. and Z.F. conceived and led this work. G.L and X.W. designed the framework. G.L. implemented the framework. G.L. and X.L. performed data analysis. G.L., X.W., and Z.F. wrote the manuscript.

Corresponding author

Correspondence Xin Wang or Zhangjun Fei.

Ethics declarations

Competing interests

The authors declare no competing interests.

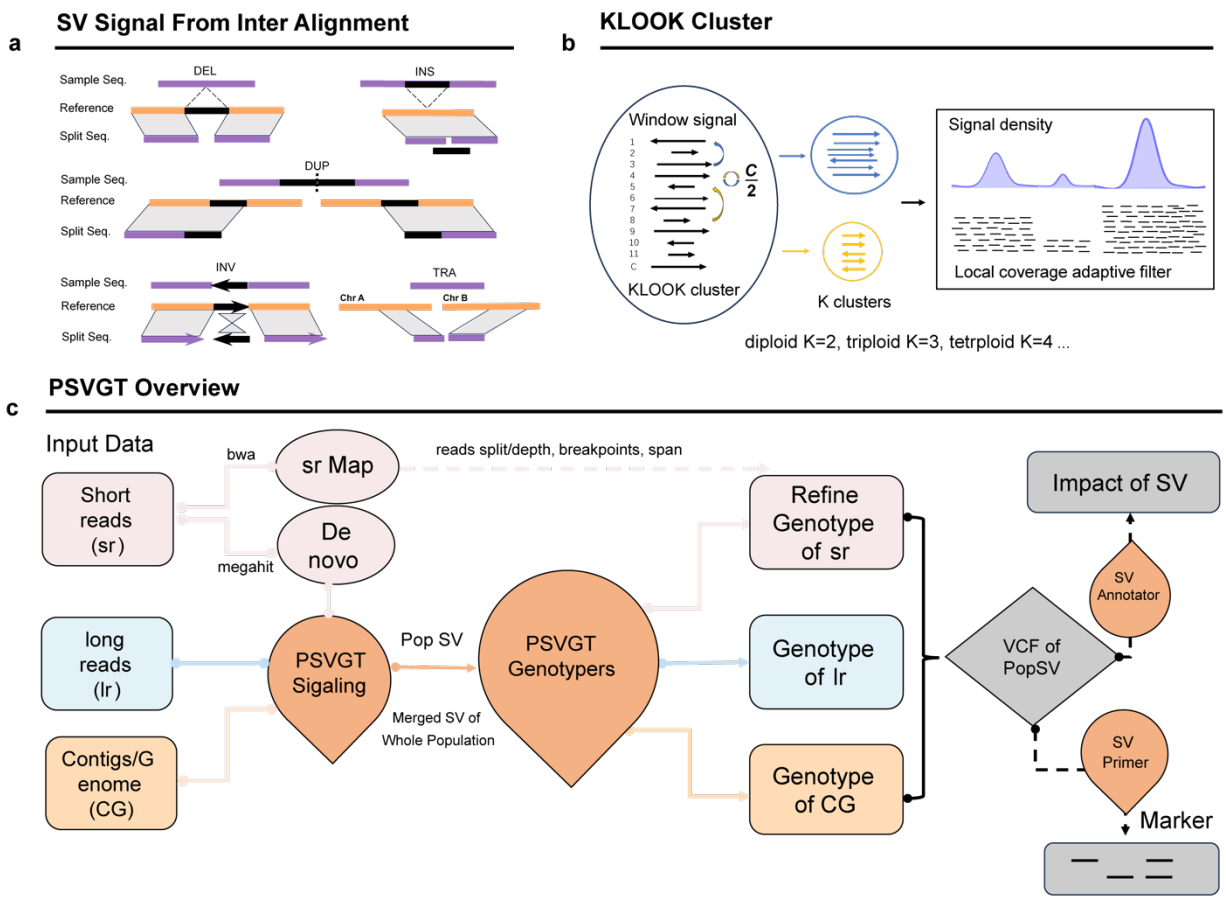


Fig. 1 | Overview of the PSVGT workflow. (a) SV inference from inter-alignments. SVs, including deletions (DEL), insertions (INS), duplications (DUP), inversions (INV), and translocations (TRA), are detected by aligning sample sequences (purple) to the reference genome (orange). Detected SVs are marked in black. **(b)** KLOOK clustering algorithm implemented in PSVGT. Signals within a window are clustered into K groups, where K is defined by ploidy, based on SV start positions, sizes, and local read depth. **(c)** Overall PSVGT workflow. PSVGT accepts diverse types of DNA sequences, including short reads (sr), long reads (lr), and assembled contigs/genomes (CG). Short reads are aligned to the reference genome using BWA and simultaneously de novo assembled into contigs. Candidate SV signals are collected from all samples and processed through KLOOK clustering and local depth-adaptative filtering within the PSVGT Signaling module. Population-level SVs (PopSVs) are then genotyped in each sample using sequence-type-specific sub-modules. Genotype calls from short-read assemblies are further refined using breakpoint information from read alignments. The final population-scale SVs in VCF format support downstream analyses, including marker development via the SV primer module and functional impact annotation via the SV annotation module.

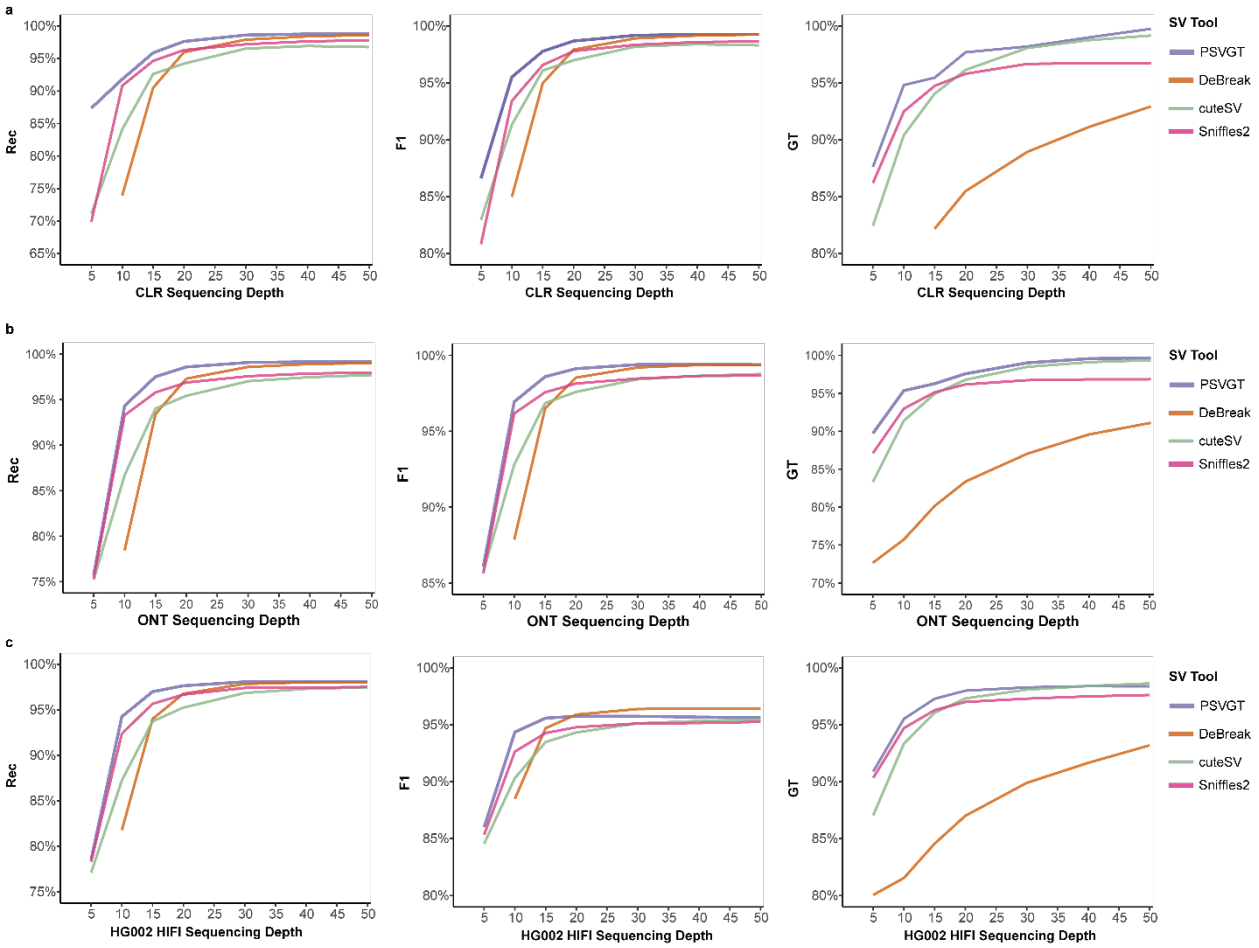


Fig. 2 | PSVGT performance on downsampled long-read datasets. **a-c**, Comparison of four SV detection tools on downsampled long-read datasets, including simulated PacBio CLR **(a)**, ONT **(b)**, and PacBio HiFi **(c)** datasets. Performance metrics–recall (Rec), F1 score (F1), and genotype concordance (GT)–are plotted against sequencing depth (x-axis).

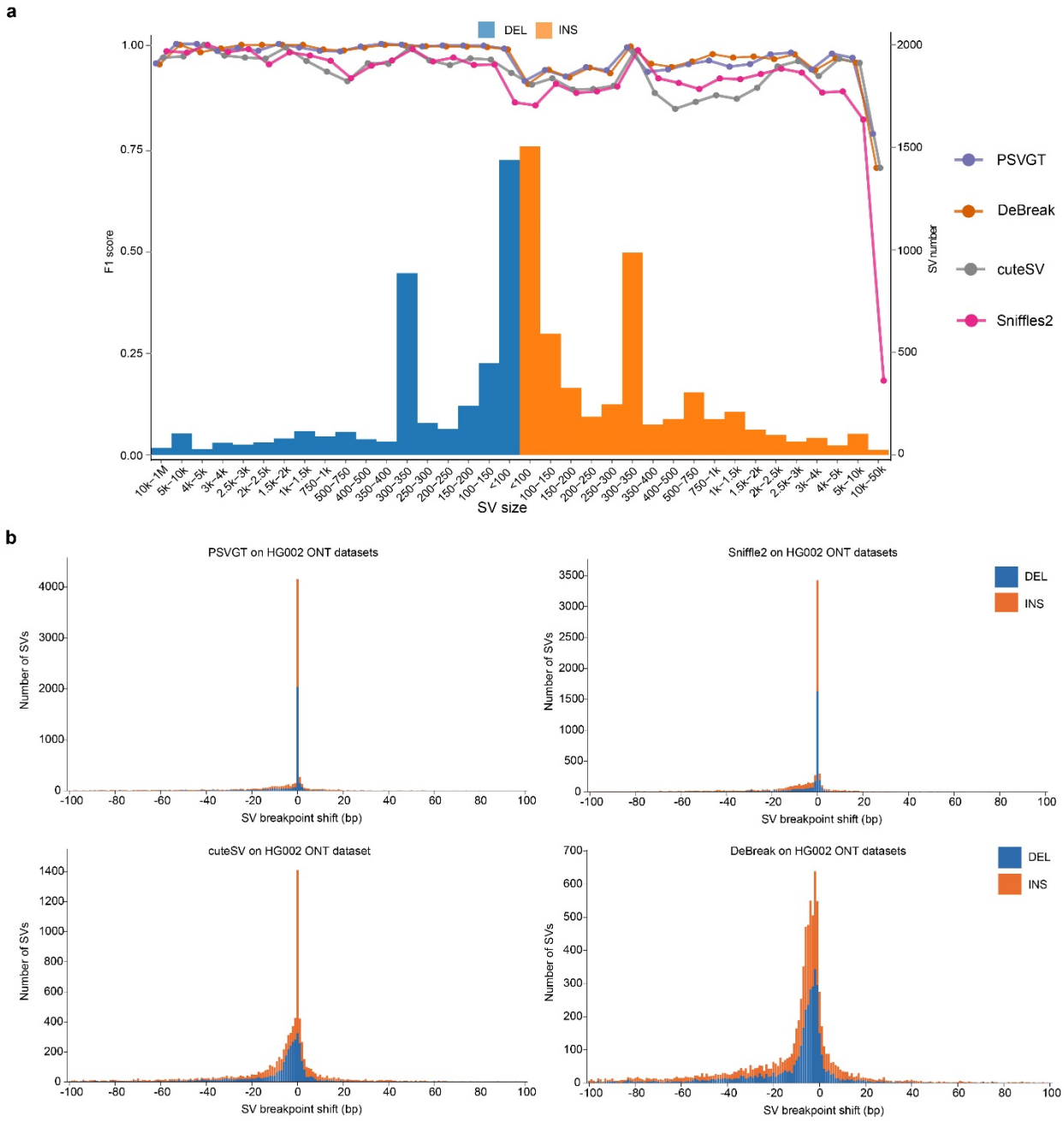


Fig. 3 | SV detection accuracy using ONT datasets. (a), F1 score distributions across SV sizes for deletions (DEL) and insertions (INS). Histograms indicate the number of SVs within each size bin, whereas lines show F1 scores. **(b)**, Breakpoint shift distributions (≤100 bp) of SVs detected from ONT datasets.

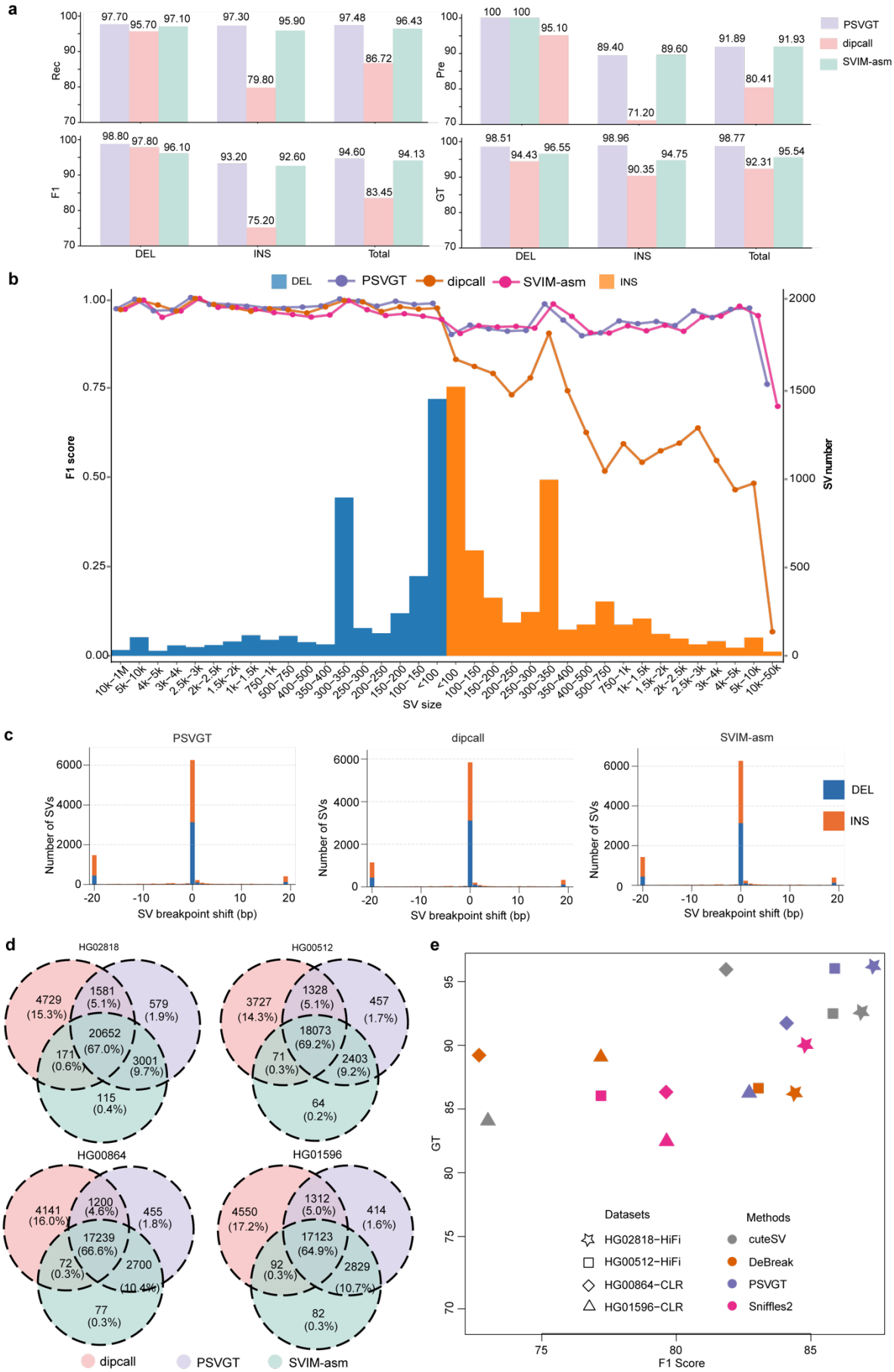

a
Rec
Pre
F1
GT
PSVGT
dipcall
SVIM-asm
DEL
INS
Total
b
DEL
PSVGT
dipcall
SVIM-asm
INS
F1 score
SV number
SV size
c
PSVGT
dipcall
SVIM-asm
Number of SVs
SV breakpoint shift (bp)
DEL
INS
d
HG02818
HG00512
HG00864
HG01596
dipcall
PSVGT
SVIM-asm
e
GT
F1 Score
Datasets
HG02818-HiFi
HG00512-HiFi
HG00864-CLR
HG01596-CLR
Methods
cuteSV
DeBreak
PSVGT
Sniffles2

Fig. 4 | Performance of PSVGT on phased diploid assemblies and long-read datasets. (a) Performance of PSVGT, dipcall, and SVIM-asm on HG002 phased assemblies. **(b)** F1 score distributions across SV sizes for phased HG002 assemblies. Histograms showthe number of deletions (DEL) and insertions (INS) within each size bin, while lines indicate F1 scores. **(c)** Breakpoint shift distributions for SVs detected by PSVGT, dipcall, and SVIM-asm. **(d)** Venn diagrams of SV calls among assembly-based methods on phased diploid assemblies, HG02818, HG00512, HG00864, and HG01596. **(e)** Performance of SV discovery tools on HGSVC2 long-read datasets. Scatter plots show F1 score versus genotype concordance (GT) for cuteSV, DeBreak, PSVGT, and Sniffles2 across four datasets.

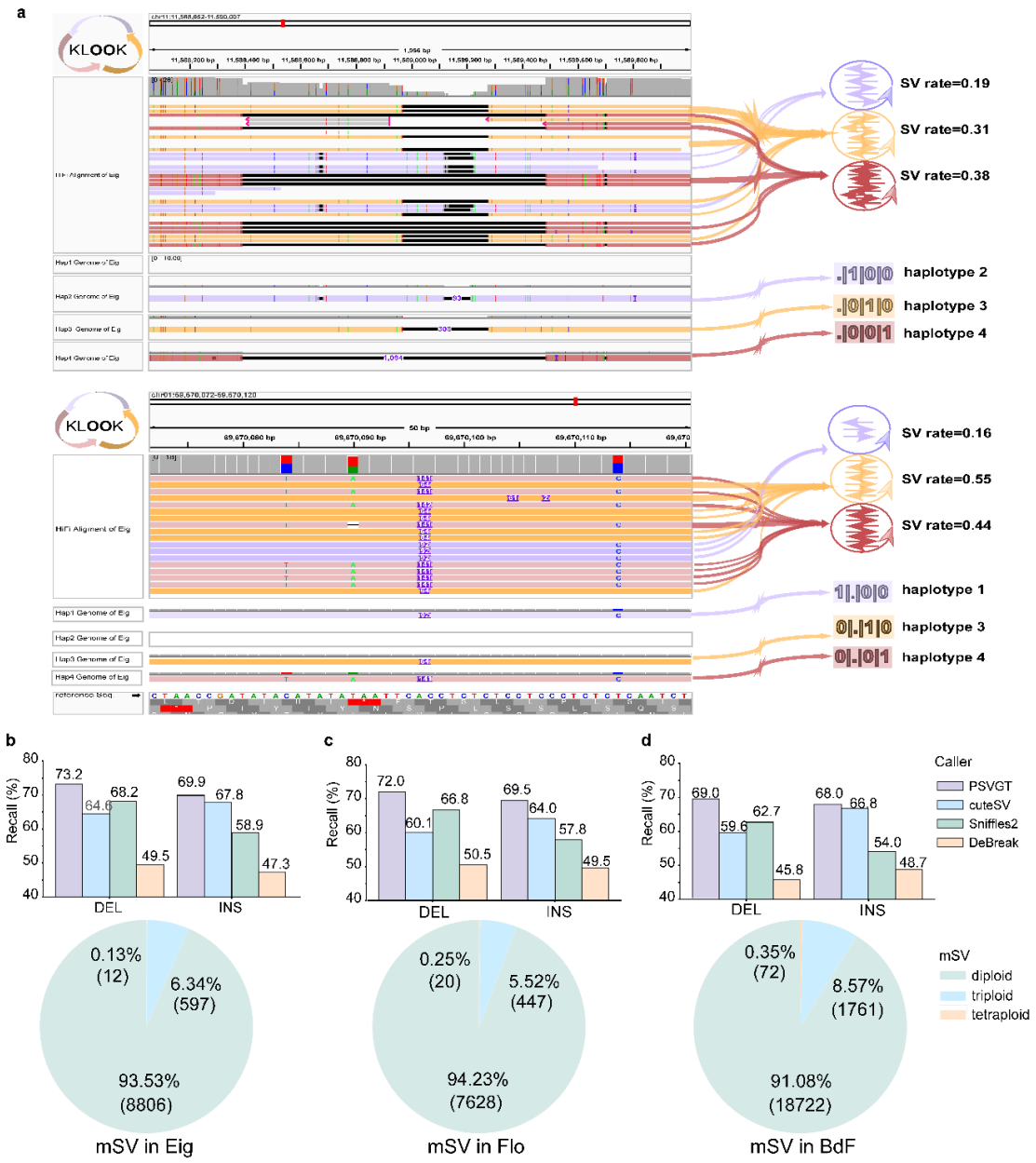


Fig. 5 | Performance of PSVGT on autotetraploid species. (a) Example of a triploid deletion and insertion event in the four-haplotype genome (top) and the corresponding HiFi read alignment (bottom). **(b-d)** Recall performance and distribution of multi-allelic SVs (mSVs) in the tetraploid potato cultivars Eig (b), Flo (c), and BdF (d). Bar plots show mSV recall rates for SV callers, while pie charts indicate the proportions of diploid, triploid, and tetraploid SVs in each cultivar.

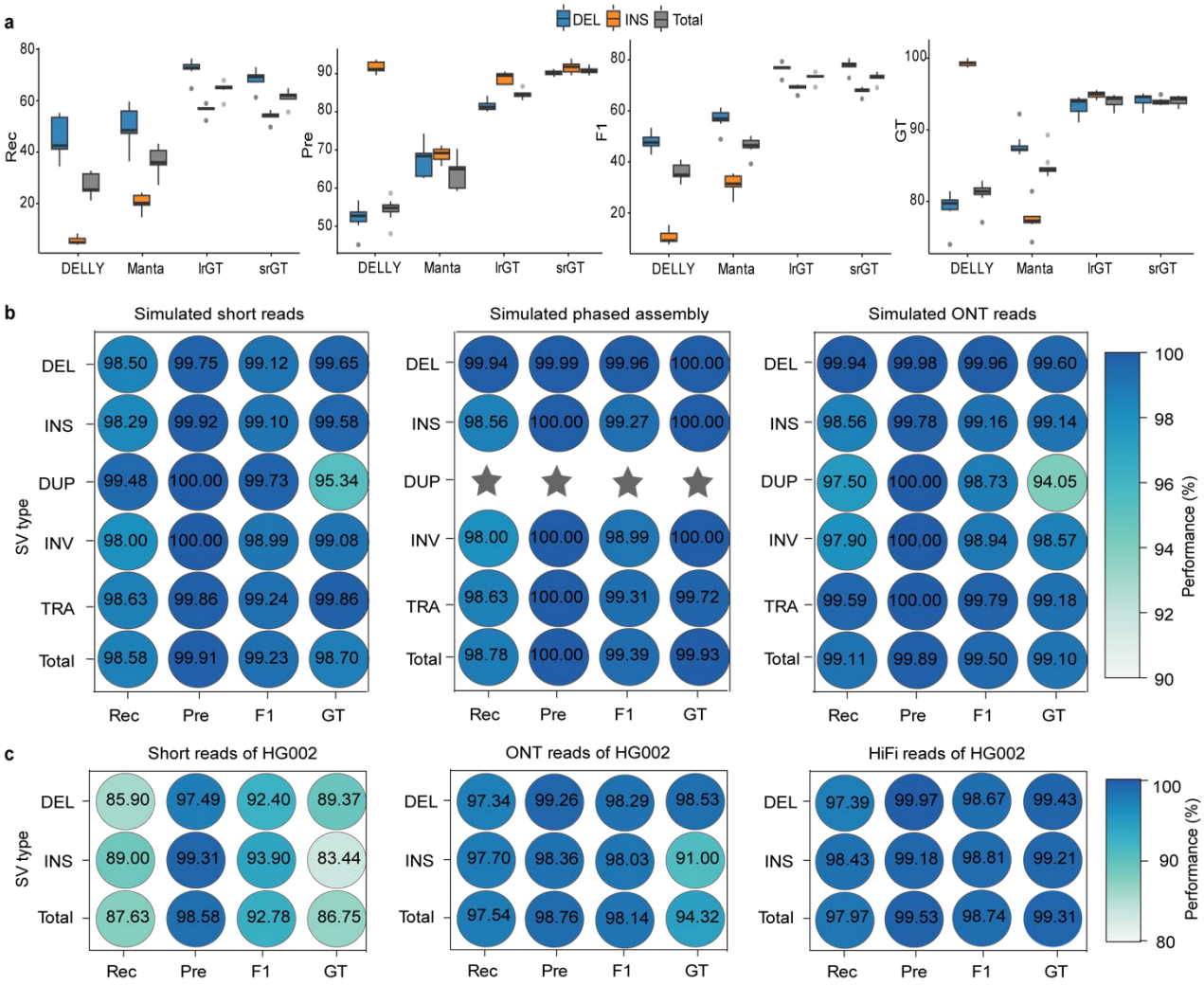


Fig. 6 | Performance of PSVGT on short-read datasets and SV genotyping. (a) Comparison of SV detection metrics across different tools. Boxplots show F1 score (F1), genotype concordance (GT), precision (Pre), and recall (Rec) for DELLY, Manta, lrGT (genotyping based on Short-read–derived contigs alignment) and srGT (genotyping based on short-read alignment), stratified by deletions (DEL), insertions (INS), and all SVs combined (Total). **(b, c)** Genotyping performance on simulated data **(b)** and real data of HG002 **(c)**. Performance metrics include recall (Rec), precision (Pre), F1 score (F1), and genotype concordance (GT) across different SV types and sequencing platforms.

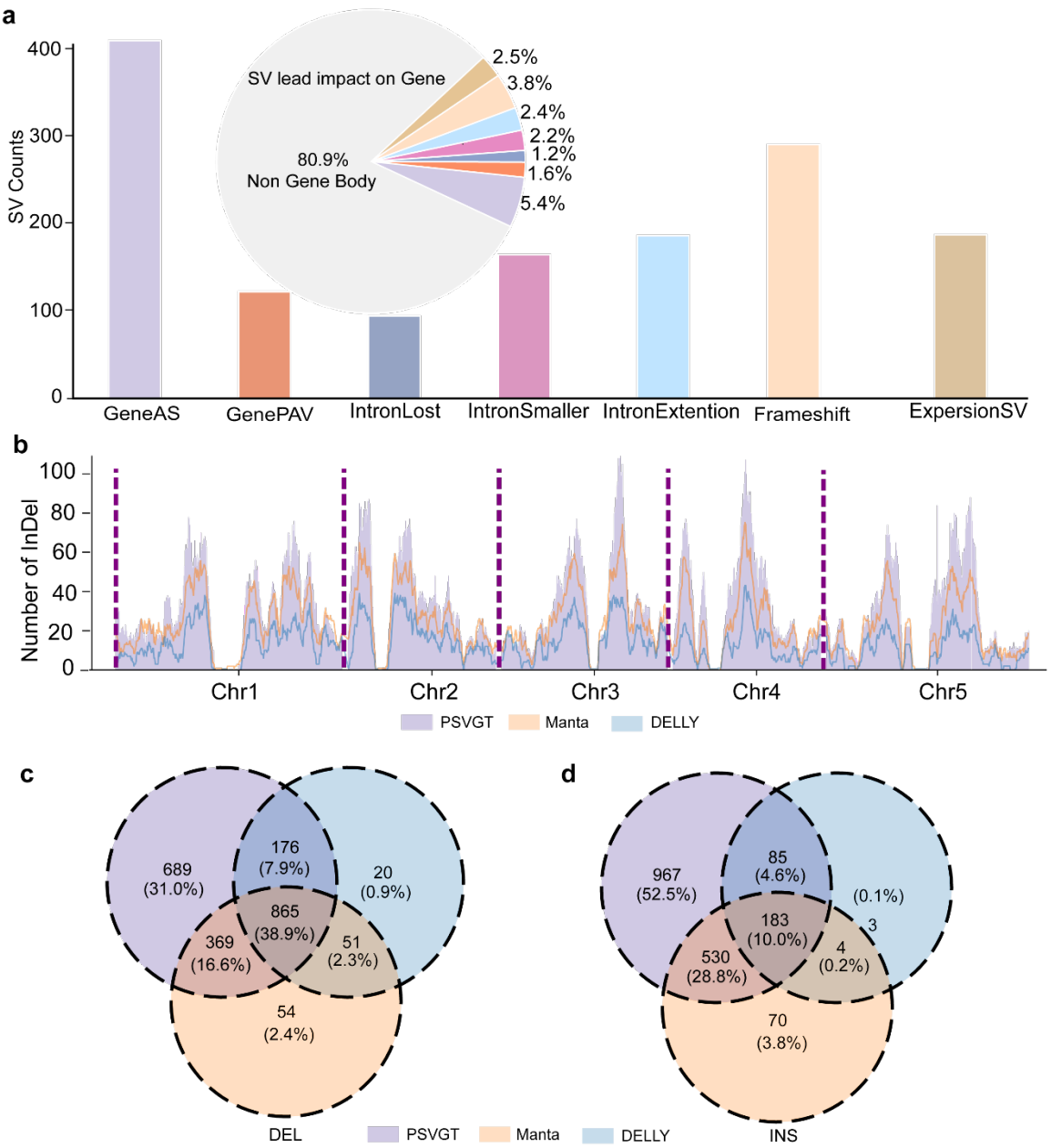


Fig. 7 | SV annotation and marker development module of PSVGT. (a) Example annotation of SVs in the PSVGT module. Annotation features are represented by distinct colors. The bar plot indicates the number of SVs annotated under each representative term, while the pie plot displays the proportion of each annotation category. **(b)** Density distribution of 50–600 bp InDels across *Arabidopsis* chromosomes. Density curves, color-coded for tools (PSVGT, purple; Manta, orange; DELLY, blue) show the density of InDel number across chromosome1–5. **(c-d)** Intersection of InDel calls among tools. Venn diagrams showing the number of deletions **(c)** and insertions **(d)** identified by PSVGT, Manta, and DELLY across all eight datasets.