

Benchmarking long-read aligner-caller pipelines for structural variant detection across ten angiosperm species with PacBio high-fidelity sequencing

Zhiliang Zhang (张治梁)^{1,2} · Jun Xu (徐俊)^{1,2} · Zhizhan Wang^{1,3} · Tingxi Yu^{1,2} · Kunhui He^{1,2} · Duan Liu^{1,2} · Ying Zhao^{1,2} · Keai Wang^{1,2} · Zhaorong Xie^{1,2} · Changling Huang^{1,2} · Jiankang Wang^{1,2} · Huihui Li (李慧慧)^{1,2}

Received: 15 January 2026 / Accepted: 28 September 2026
© The Author(s) 2026

Abstract

Background Structural variants (SVs) play a central role in genome evolution and phenotypic diversity, yet the large genome size, high repeat content, and structural complexity hinder SV detection in plants. Despite extensive benchmarking on human data, the performance and cross-species generalizability of HiFi long-read SV detection pipelines for plants have been overlooked.

Results Here, we use multiple assembly-based variant calling strategies to construct SV truth sets at base-level resolution for ten representative angiosperm species, followed by systematic benchmarking of 91 HiFi-based SV calling pipelines. Cross-species evaluation reveals pronounced heterogeneity in performance across pipelines and species with deletion detection generally outperforming insertion detection. Using a stability score to quantify the across-species generalizability, we identify NGMLR/Sniffles2 and NGMLR/DeBreak as optimal cross-species pipelines for deletion detection, while Minimap2/Sniffles1, Minimap2/cuteSV2, and pbmm2/cuteSV1 exhibit robust insertion detection. We also find that detection accuracy declines sharply for large SVs and variants located within repetitive regions, particularly simple repeats. Moreover, SV calling precision shares a nonlinear relationship with sequencing depth for each individual haplotype, with F_1 scores peaking at $\sim 15\times$ coverage.

Conclusions This study establishes a benchmarking framework for plant SV detection and provides high-confidence SV truth set resources that can, together, guide the selection of robust, scalable SV calling pipelines for plant genomics.

Keywords Structural variants · Truth sets · SV calling pipelines · Benchmark · Cross-species generalizability

Zhiliang Zhang and Jun Xu contributed equally to this work.

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



Background

Structural variants (SVs), i.e., deletions, insertions, duplications, inversions, translocations, and copy number variants (CNVs), can directly alter gene dosage or disrupt gene structure, with pathological consequences in a wide range of heritable diseases [1, 2]. In humans, SVs have been implicated in intellectual disability, autism spectrum disorders, and congenital limb malformations [3, 4]. SVs exert similarly profound effects in plants [5, 6]. For example, deletion of the *sh2* gene in maize disrupts starch biosynthesis, resulting in reduced and wrinkled kernels, while the insertion of a transposable element approximately 60 kb upstream of the *tb1* locus reshaped maize plant architecture and likely constituted a hallmark event in maize domestication [7–9]. Population-scale analyses have also demonstrated that accumulation of numerous minor effect deleterious SVs can lead to genetic consequences comparable with loss-of-function mutation in protein-coding genes, underscoring the substantial evolutionary and functional impact of SVs [10, 11].

Despite their importance across diverse biological contexts, accurate SV discovery has long been constrained by their intrinsic complexity and potentially large size, which often require integration of multiple technologies and methodological approaches [12, 13]. Although early cytogenetic techniques, such as fluorescence in situ hybridization (FISH) and DNA microarrays, can be used for SV detection, these low throughput techniques are expensive and provide limited breakpoint resolution [14, 15]. While next-generation short-read sequencing dramatically increased throughput and reduced cost, the short read length precludes accurate detection of SVs in repetitive regions and for large or complex variants [12, 13]. In contrast, PacBio high-fidelity (HiFi) long-read sequencing balances read length with base-level accuracy, allowing high precision in breakpoint resolution without prohibitive cost [16]. These advantages have led to development of numerous and diverse long-read aligners and SV callers to support read-mapping-based SV detection [17]. However, numerous benchmarking studies have evaluated such tools in the context of the human genome, resulting pipelines tailored to human-specific analytical objectives, while benchmarking in plants has lagged [18–22].

Plant genomes can vary widely among species, some harboring exceptionally large, highly repetitive, and frequently polyploid genomes, thus posing bioinformatic challenges that are largely absent in human genomics [23]. Notably, repeat sequences can account for up to 85% of the wheat genome, depending on variety, compared with approximately 54% in humans [24, 25]. Moreover, intraspecific genomic diversity is dramatically higher in plants. For instance, any two maize genomes may differ by as much as 50% of their sequence content, whereas human genomes differ by only ~1.6% [26, 27]. These plant-specific characteristics often exceed the capabilities of current computational methods. Critically, nearly all widely used long-read aligners (BLEND [28], lordFAST [29], LRA [30], Minimap2 [31], NGMLR [32], pbmm2, Winnowmap2 [33]) and SV callers (cuteSV1 [34], cuteSV2 [34], DeBreak [35], DELLY [36], dysgu [37], gcSV [38], NanoVar [39], pbsv, Sniffles1 [32], Sniffles2 [40], SVDSS1 [41], SVDSS2, SVIM1 [42], SVIM2) were originally developed and optimized using human datasets. As a result, their performance, reliability, and cross-species generalizability in plants remain largely uncharacterized. Although we previously conducted a preliminary evaluation of these tools on the wheat genome, rigorous assessment of these SV detection pipelines requires systematic benchmarking across multiple plant species [43].

To address this gap in benchmarking, in the present study we collected high-quality reference genome assemblies and deep HiFi sequencing data (~39.07X) for ten angiosperm species representing five major families: Poaceae, Brassicaceae, Polygonaceae, Malvaceae, and Fabaceae. We then conducted whole-genome alignments between paired high-quality assemblies to construct species-specific SV truth sets. These truth sets were then used to evaluate the performance of 91 SV calling pipelines, comprising all compatible pairwise combinations of the above seven long-read aligners and 14 SV callers. We further investigated the influence of six factors on SV detection performance, including cross-species generalizability, aligner/caller-associated effects, SV length-associated effects, repeat-associated effects, sequencing depth-associated effects, and computational resource. These studies collectively establish a robust framework for selecting optimal SV detection pipelines and provide a foundational resource for future population-scale long-read resequencing analyses in plants.

Results

Overview of cross-species benchmarking framework for long-read SV pipelines

To ensure sufficiently wide genomic diversity for robust evaluation of pipeline accuracy and generalizability among plant species, we first selected ten representative angiosperm species, including six major crops (rice, wheat, maize, soybean, sorghum, cotton) and four model species (*Arabidopsis*, tomato, foxtail millet, broomcorn millet), from across five angiosperm families (Poaceae, Brassicaceae, Polygonaceae, Malvaceae, and Fabaceae; Fig. 1a). These genomes widely varied in their architecture and complexity, ranging from 134 Mb to 14.5 Gb in length and 24.72% to 86.51% in repeat sequence contents, reflecting the heterogeneity of analytical challenges in plant genomics (Additional file 2: Tables S1). To provide the highest possible accuracy in SV truth set construction, we obtained two high-quality reference genomes and high-depth HiFi datasets from public databases for each species (Additional file 2: Tables S1), with an average sequencing depth of 39.07.

As reliable benchmarking of SV detection pipelines fundamentally depends on robust truth sets with base-level accuracy [19], which are lacking for most plant species, we employed a three-step strategy to construct robust SV truth sets for each representative species (Fig. 1b). First, we performed whole-genome alignment between two high-quality reference assemblies, designating one as the reference and the other as the query (Additional file 2: Tables S2). Based on these alignments, we generated candidate SV call sets using five complementary assembly-based SV discovery pipelines selected to maximize sensitivity and capture diverse SVs, including Cactus-pangenome, DpCall, Minimap2 (genome alignment; parameter -ax asm5) coupled with SyRI, Minimap2 (genome alignment; parameter -ax asm5) coupled with SVIM-asm, and AnchorWave combined with SVIM-asm. Second, we merged candidate SVs based on breakpoint proximity and length similarity using Truvari collapse, thereby consolidating overlapping calls originating from multiple pipelines. Finally, an orthogonal, independent validation step was performed for all candidate SVs relying solely on high-depth HiFi sequencing data, and this workflow is entirely separated from all aligners and SV callers evaluated in our benchmark. For each candidate SV, we leveraged SVJedi-graph to construct locus-specific variation graphs for each candidate SV, mapping and genotyping all HiFi raw reads against these graphs to verify SV presence and determine precise breakpoint resolution. Only SVs supported by robust genotyping evidence were retained, resulting in high-confidence, base-level-resolved SV truth sets for evaluating the performance of long-read SV pipelines across the ten representative angiosperm species.

For the broadest possible benchmarking of current mapping-based SV discovery tools, we compiled seven long-read aligners (BLEND, lordFAST, LRA, Minimap2, NGMLR, pbmm2, Winnowmap2) and 14 SV callers (cuteSV1, cuteSV2, DeBreak, DELLY, Dysgu, gcSV, NanoVar, pbsv, Sniffles1, Sniffles2, SVDSS1, SVDSS2, SVIM1, SVIM2) from published and preprint sources (Fig. 1c). These tools spanned a variety of algorithmic paradigms, including seed-and-chain mapping, repeat-aware alignment, signal clustering, and local-assembly-based SV inference. Pairwise combinations yielded 98 possible SV calling pipelines, 91 of which successfully produced SV calls. The seven incompatible combinations, including SVDSS1/SVDSS2 with LRA, BLEND, or lordFAST, and NanoVar with pbmm2, were excluded from subsequent tests. The 91 compatible SV calling pipelines were subsequently compared in parallel SV calling analyses for each representative species with the corresponding truth sets.

Using the truth sets for each respective species, we systematically evaluated performance (precision, recall, and F_1 score) for the 91 SV calling pipelines across all species (Fig. 1d). Beyond assessing performance, we examined (i) cross-species generalizability of 91 SV calling pipelines, (ii) repeat-associated effects on SV detection performance, (iii) sequencing depth-associated effects on SV detection performance, and (iv) computational resource, including runtime and memory usage. Together, these analyses provided a comprehensive and interpretable framework for cross-species benchmarking framework for long-read SV pipelines.

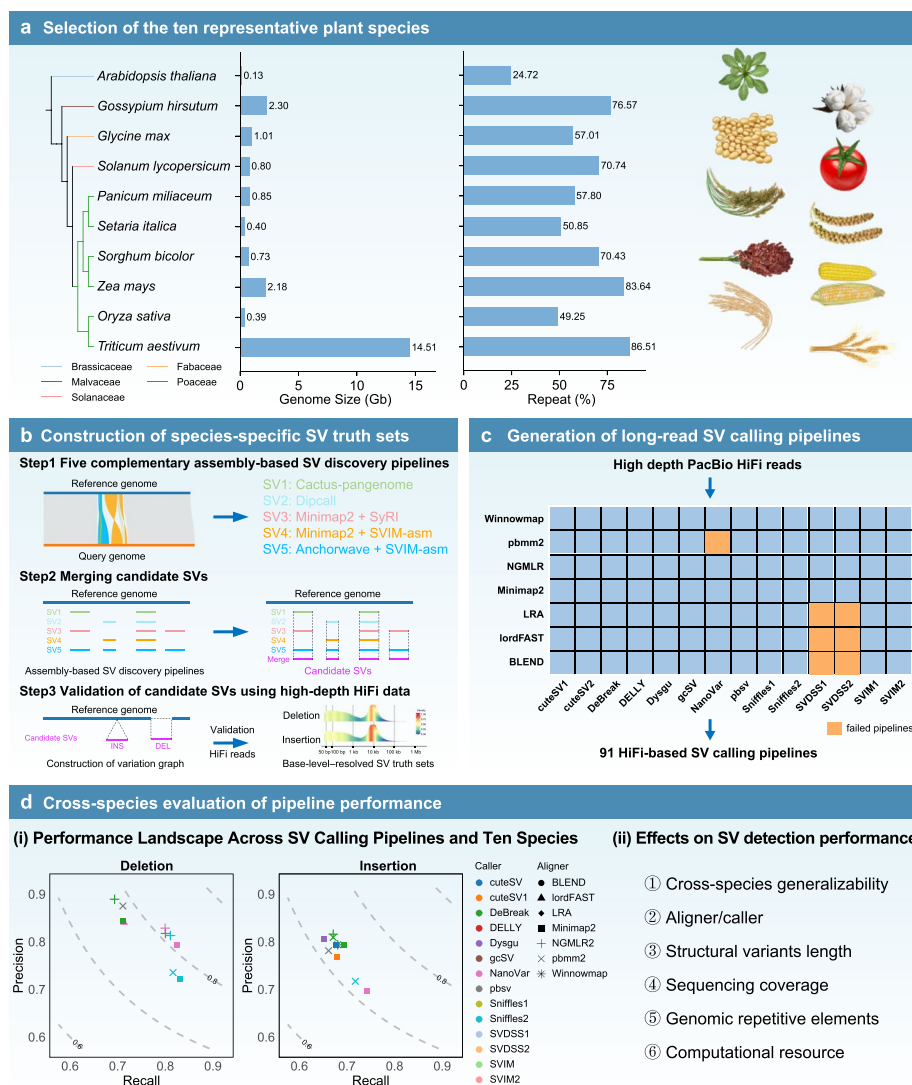


Fig. 1 Overview of cross-species benchmarking for SV detection pipelines with PacBio HiFi sequencing. **a** Phylogenetic relationships of the ten representative angiosperm species used for benchmarking (left), with genome size (middle) and repeat sequence content (right). **b** Construction of species-specific SV truth sets. For each species, (i) five independent, complementary, assembly-based SV discovery pipelines were used to call candidate SVs between reference and query genomes; (ii) candidate SVs were merged and filtered for redundancy; (iii) candidate SVs validated with high-depth HiFi data of query genome were compiled into respective truth sets. **c** Generation of long-read SV calling pipelines. Seven long-read aligners (BLEND, lordFAST, LRA, minimap2, NGMLR2, pbmm2, and Winnowmap) were paired with fourteen SV callers (cuteSV1, cuteSV2, DeBreak, Delly, Dysgu, gcSV, NanoVar, pbsv, Sniffles1, Sniffles2, SVDSS1, SVDSS2, svim1, and svim2) to construct 91 total SV calling pipelines. The pipelines labeled in orange were incompatible. **d** Cross-species evaluation of pipeline performance. (i) performance of the 91 SV calling pipelines was assessed by recall, precision, and F_1 score in each species; (ii) effects of across-species generalizability; aligner/caller; SV length; sequencing coverage; genomic repetitive elements and computational resource requirements (runtime and memory usage) on SV detection performance

Construction of high-confidence, species-specific SV truth sets with HiFi long reads

Using the above framework for constructing species-specific SV truth sets (Fig. 1b), we first generated candidate SVs for ten angiosperm species using the five complementary SV calling pipelines. We noted that total detected candidate SV numbers substantially varied among pipelines across species (Additional file 1: Fig. S1; Additional

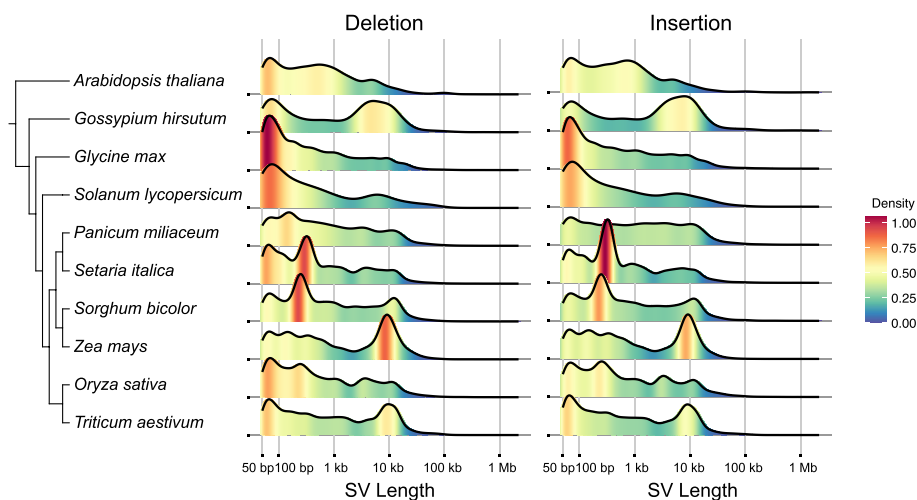
file 2: Tables S3). Overall, the AnchorWave–SVIM–asm pipeline consistently yielded the largest number of SVs, whereas minimap2–SyRI produced the fewest, with an average 3.23-fold difference in SV counts between them. Despite these quantitative differences, both deletions and insertions exhibited broadly consistent count and length distributions across pipelines and species, with each species showing a distinct SV length distribution profiles. Cotton, sorghum, maize, and wheat had bimodal distributions, characterized by a dominant class of short SVs (< 500 bp) and a second peak around ~ 10 kb. In contrast, broomcorn millet showed a relatively uniform SV length distribution, while the remaining species predominantly harbored SVs shorter than 1 kb. These patterns likely reflected species-specific genome architectures and transposable element landscapes. To obtain high-confidence SV sets, we merged the candidate SVs from all successful pipelines for each species (Additional file 2: Tables S3). Quantitative analysis of SV density (i.e., number of SVs per Mb of genome sequence) revealed an average SV density of 53.17 SVs/Mb across the ten species, with cotton (4.30 SVs/Mb) and tomato (3.31 SVs/Mb) exhibiting the lowest densities, and foxtail millet (132.56 SVs/Mb) and rice (131.51 SVs/Mb) showing the highest.

Finally, we assessed the allele-specific support of the SV captured for each species to compile the final truth sets. Genotyping of all merged candidate SVs using high-depth HiFi reads (Additional file 2: Tables S3) revealed that, on average, 73.07% of candidate SVs were supported by HiFi reads and therefore could be classified as true variants, including 73.64% of deletions and 72.55% of insertions. Importantly, the SV length distributions in the final truth sets closely mirrored those of the initial candidate SV collections generated by the five pipelines, indicating that the genotyping and filtering process did not introduce detectable length bias (Fig. 2; Additional file 1: Fig. S1). Given publicly available high-depth Oxford Nanopore Technologies (ONT) long-read datasets for maize and cotton, we further constructed three types of SV truth sets: HiFi-only, ONT-only, and integrated HiFi-ONT (Additional file 1: Fig. S2). An average of 94.9% deletions were simultaneously captured in both HiFi and ONT truth sets across the two species, while the corresponding value for insertions reached 95.0%. Overall, the SV detection performance obtained from HiFi-only, ONT-only and integrated HiFi-ONT truth sets was highly consistent without notable discrepancies, demonstrating that the HiFi-only truth set is fully reliable for downstream structural variant evaluation. Together, these results demonstrated that our three-step strategy yielded high confidence SV truth sets with base level resolution for each representative angiosperm species, and could thus provide a suitable basis for subsequent benchmarking analyses.

Generation of SV call sets for ten angiosperm species

After establishing the SV truth sets, we generated SV call sets using the 91 compatible SV calling pipelines with HiFi sequencing data from the ten angiosperm species. Although SV call sets were successfully generated for

Fig. 2 Characteristics of species-specific SV truth sets. Length distributions of deletions (left) and insertions (right) in the high-confidence SV truth sets across the ten representative angiosperm species



most species, the exceptionally large genome size of wheat (14.51 Gb) resulted in prohibitively long runtimes for the lordFAST and LRA aligners when paired with any of the 14 SV callers. Excluding these failed analyses, we obtained 67 SV call sets for wheat, and 91 sets for each of the remaining species.

Examination of the relative abundance of deletions and insertions across all SV call sets for each species (Additional file 1: Fig. S3; Additional file 2: Tables S4) revealed that all species consistently had 1.15- to 1.44-fold higher deletion counts than insertion counts, with the exception of broomcorn millet, which had slightly fewer deletions than insertions (deletion:insertion ratio = 0.93). These patterns were highly consistent across pipelines within species, suggesting that this observed imbalance between deletions and insertions reflected a common genomic characteristic rather than a computational artifact of the pipelines.

Comparison of discovery rates among the 14 SV callers across the ten species (Additional file 1: Fig. S3; Additional file 2: Tables S4) indicated that SVIM1 consistently identified the largest number of SVs, including both deletions and insertions, whereas gcSV yielded the fewest across species. On average, SVIM1 detected 4.10-fold more deletions than gcSV (range: 2.45–9.09-fold) and 2.69-fold more insertions (range: 1.68–5.68-fold) across species. These substantial differences highlighted the pronounced variability in sensitivity among SV callers, and underscored the informative value of systematic benchmarking for selecting appropriate plant genomic SV detection tools.

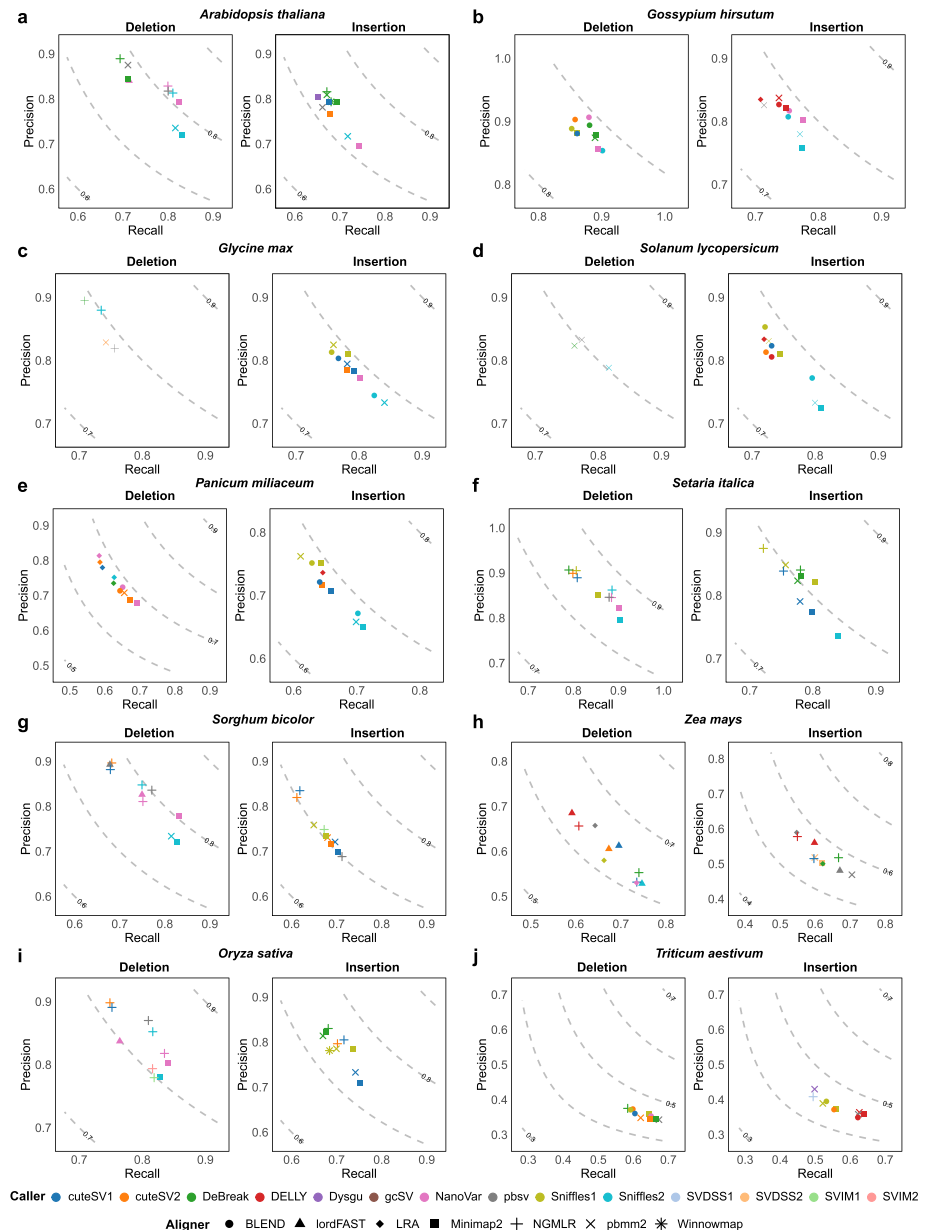
We further quantified the numbers and proportions of overlapping deletion and insertion calls among the 14 SV callers for each aligner and among the seven aligners for each SV caller (Additional file 1: Fig. S4–7). When grouped by aligner, pipelines using lordFAST exhibited the lowest overall degree of overlap among SV callers for both deletions and insertions, indicating that calls generated from lordFAST alignments had the lowest concordance across callers. Conversely, when grouped by SV caller, Dysgu-based pipelines showed the lowest degree of overlap among aligners, suggesting that Dysgu calls were particularly sensitive to the choice of upstream aligner. Together, these results demonstrate that the selection of both aligners and SV callers affects the concordance and reproducibility of SV detection across pipeline combinations.

Global performance landscape across SV calling pipelines and ten angiosperm species

After constructing species-specific SV truth sets and the corresponding SV call sets, we systematically evaluated the performance of the 91 SV calling pipelines for each species based on their precision, recall, and F_1 scores (Additional file 1: Fig. S8–11; Additional file 2: Tables S5,6). Among these metrics, F_1 score provided the most balanced measure of overall pipeline performance, as it jointly considers sensitivity and precision, with values closer to 1 indicating superior accuracy. Given the complexity of comparing 91 pipelines simultaneously, full precision-recall matrices are difficult to interpret. Therefore, we highlight the ten highest-performing pipelines for deletion and insertion detection, ranked by F_1 score, across the ten angiosperm species (Fig. 3). Interrogation of a species-by-pipeline matrix of F_1 scores revealed that wheat had consistently modest scores for both deletions (0.38 ± 0.08) and insertions (0.38 ± 0.08) across pipelines. As this relatively poor performance might reflect limited quality of the reference genome assembly quality and/or extreme genomic complexity, we excluded wheat from subsequent cross-species comparisons.

Examination of F_1 score distributions across the nine remaining species showed that average scores were higher for deletions than insertions (mean deletion F_1 score = 0.67; insertion F_1 score = 0.64; Additional file 1: Fig. S8a; Additional file 2: Tables S5,6). Among the 819 species–pipeline combinations (91 pipelines $\times$ 9 species), 101 deletion call sets achieved an F_1 score greater than 0.8. These high-performing pipelines were predominantly observed in cotton (54), foxtail millet (34), rice (8), tomato (6), Arabidopsis (4), sorghum (2), and soybean (1). In contrast, only four insertion call sets achieved F_1 scores above 0.8, all of which were confined to foxtail millet. These results demonstrated that deletions are generally easier to detect with high accuracy using HiFi data, whereas insertion detection remains more challenging these tools, particularly when balancing precision and recall.

Fig. 3 Performance landscape of the top 10 SV calling pipelines across ten representative angiosperm species. Top-ranked SV calling pipelines across *Arabidopsis thaliana* (a), *Gossypium hirsutum* (b), *Glycine max* (c), *Solanum lycopersicum* (d), *Panicum miliaceum* (e), *Setaria italica* (f), *Sorghum bicolor* (g), *Zea mays* (h), *Oryza sativa* (i), and *Triticum aestivum* (j). Colors denote SV callers, while shapes represent long-read aligners used in each pipeline



Analysis of precision further reinforced this pattern (Additional file 1: Fig. S8b; Additional file 2: Tables S5,6). Across the nine species, average precision was higher for deletions (0.72) compared to insertions (0.67). Of the 819 species–pipeline combinations, 347 deletion call sets and 207 insertion call sets achieved precision values greater than 0.8. We noted that, for each species, at least one SV calling pipeline had precision > 0.8 for specific variant types, in each species exhibited at, highlighting the potential of HiFi sequencing data for accurate SV detection when appropriate pipelines are selected. Specifically, the numbers of high-precision pipelines (deletions/insertions) per species were: *Arabidopsis* (54/30), cotton (63/56), soybean (36/27), tomato (43/25), broomcorn millet (13/6), foxtail millet (57/35), sorghum (27/9), maize (1/1), and rice (53/18) (Additional file 2: Tables S5,6). Together, these results underscored the potential for high precision in SV detection with HiFi data across diverse plant genomes.

Finally, we assessed recall across pipelines and species (Additional file 1: Fig. S8c; Additional file 2: Tables S5,6). Consistent with trends in F_1 score and precision, average recall was greater for deletions (0.68) than

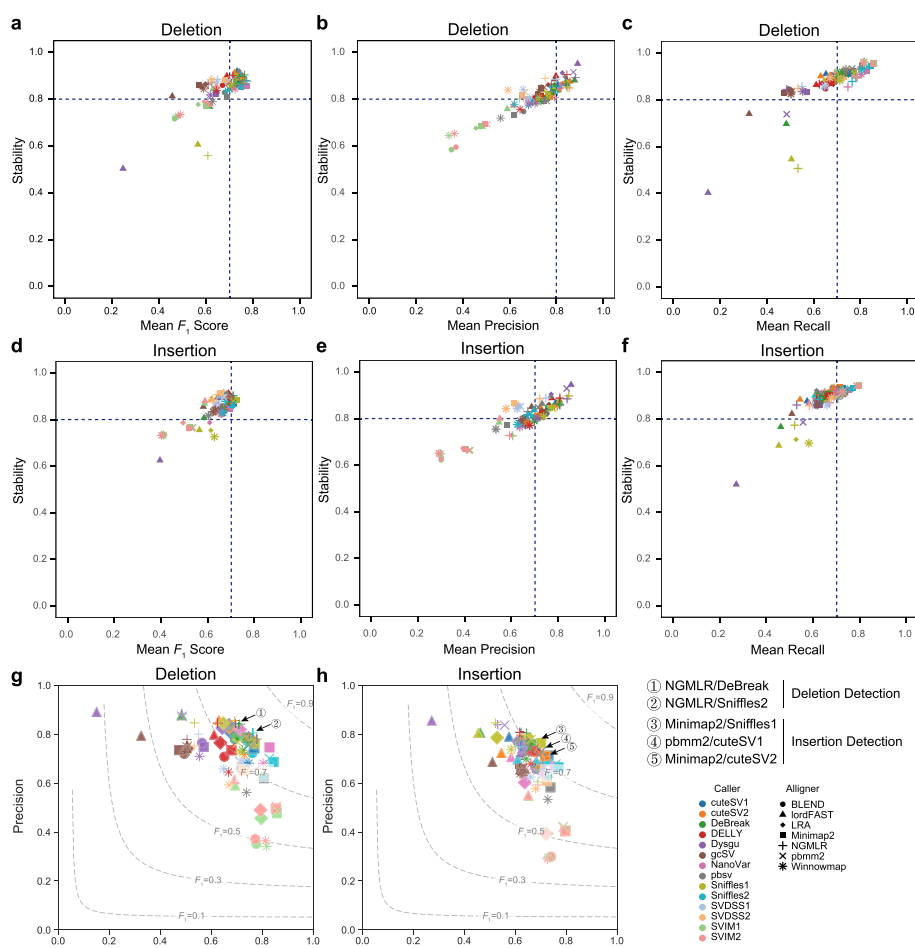
insertions (0.66). Among the 819 species–pipeline combinations, 163 deletion call sets and 40 insertion call sets achieved recall values exceeding 0.8. For deletions, all nine species had at least four pipelines with recall above 0.8, indicating robust recall across diverse plant species. In contrast, recall above 0.8 for insertions was observed in only four species—soybean (16 pipelines), tomato (9), foxtail millet (13), and maize (2)—highlighting the pronounced effects of species-specific genomic differences and overall challenge of accurate insertion detection.

Identification of SV calling pipelines with cross-species generalizability

As the dramatic variability in length, repeat content and polyploidy among plant genomes make comprehensive benchmarking across all plant species infeasible, we addressed this challenge by leveraging nine representative species to identify SV calling pipelines with the strongest generalizability across diverse genomic backgrounds, thus focusing on cross-species robustness rather than optimization for individual species.

To quantify cross-species generalizability, we developed a stability score, which measures the consistency of pipeline performance across species. Stability scores range from 0 to 1, with values closer to 1 indicating greater robustness and stronger cross-species applicability. We then calculated mean stability scores for F_1 score, precision, and recall for all 91 pipelines (Additional file 2: Tables S7). For deletions, the mean stability scores across pipelines were 0.85 (F_1 score), 0.81 (precision), and 0.88 (recall) (Fig. 4a–c), while insertions had corresponding mean stability scores of 0.85, 0.81, and 0.89, respectively (Fig. 4d–f). Notably, 83.52% of the SV calling pipelines

Fig. 4 Cross-species generalizability of 91 SV calling pipelines. Relationships between stability score and F_1 score (a, d), precision (b, e), and recall (c, f) in detecting deletions or insertions. The mean F_1 score, precision, and recall across ten representative angiosperm species for deletion and insertion detection, respectively (g, h). NGMLR/DeBreak and NGMLR/Sniffles2 are optimal cross-species pipelines for deletions, while Minimap2/Sniffles1, pbmm2/cuteSV1, and Minimap2/cuteSV2 perform best for insertions. Colors indicate SV callers and shapes indicate aligners



exhibited a stability score greater than 0.8 (F_1 score) for both deletion and insertion, indicating they provided generally consistent performance across plant species.

To identify pipelines with the strongest performance in multi-species SV detection, we applied stringent performance thresholds that jointly considered accuracy and robustness. For deletions, we identified five pipelines with cross-species mean values of F_1 score ≥ 0.7 , precision ≥ 0.8 , recall ≥ 0.7 , and stability score ≥ 0.8 , including NGMLR/Sniffles2, NGMLR/DeBreak, LRA/NanoVar, lordFAST/NanoVar, and lordFAST/Sniffles2 (Additional file 2: Tables S7). For insertions, we adopted a slightly lower threshold for precision due to the consistently lower performance of all pipelines in this SV type (F_1 score ≥ 0.7 , precision ≥ 0.7 , recall ≥ 0.7 , and stability score ≥ 0.8). This analysis identified Minimap2/Sniffles1, pbmm2/cuteSV1, and Minimap2/cuteSV2 as the most suitable pipelines for detecting insertions across species (Additional file 2: Tables S7). These pipelines exhibited high stability across all species while maintaining a favorable balance among F_1 score, precision, and recall, supporting their suitability for SV discovery across diverse plant genomes.

To further validate the cross-species applicability of the selected pipelines, we tested them on the wheat genome, which was previously excluded from generalizability analyses due to its exceptional size and complexity. Among the 67 pipelines that could successfully call SVs in wheat, Minimap2/Sniffles1 achieved the highest F_1 score for detecting deletions (0.46). Among the five high-performing deletion calling pipelines we identified above, NGMLR/DeBreak achieved an F_1 score of 0.461 and NGMLR/Sniffles2 achieved 0.434, while the remaining three pipelines could not complete the analysis due to computational constraints imposed by the wheat genome size (Additional file 2: Tables S5). For insertions, the highest F_1 score was 0.462, obtained using pbmm2/Dysgu. Notably, all three of the strongest insertion calling pipelines also had F_1 scores close to this maximum: Minimap2/Sniffles1 (0.447), Minimap2/cuteSV2 (0.427), and pbmm2/cuteSV1 (0.423) (Additional file 2: Tables S6). Overall, for both deletions and insertions, the pipelines identified as optimal for multi-species SV detection achieved F_1 scores in wheat close to the theoretical upper bounds observed among all tested workflows. This validation in an independent, highly complex genome supported the robustness and generalizability of these pipelines observed in cross-species analyses.

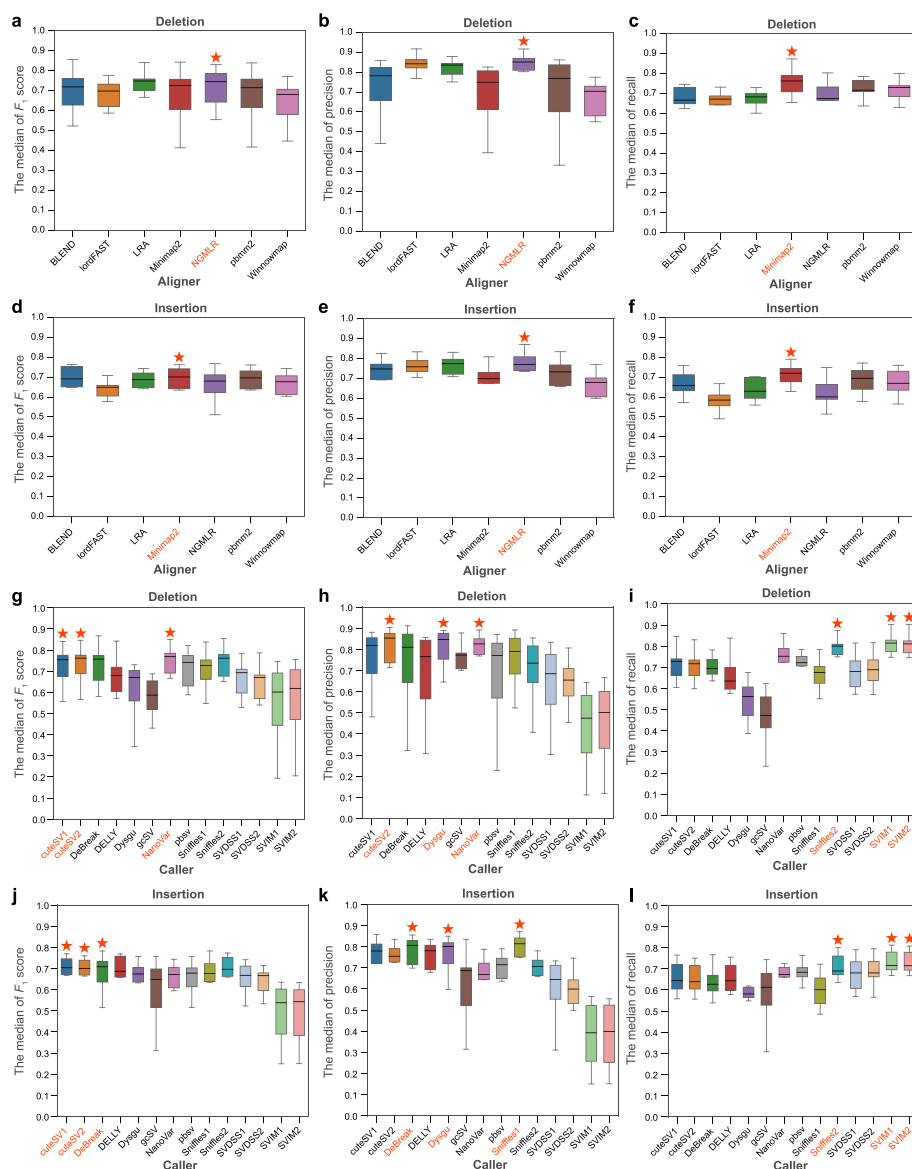
As LRA and lordFAST were either unsuitable for extremely large genomes or exhibited prohibitively slow runtimes on such datasets, these tools were excluded from the final recommendations. Consequently, we recommend NGMLR/Sniffles2 and NGMLR/DeBreak as the most reliable cross-species pipelines for deletion detection, and Minimap2/Sniffles1, Minimap2/cuteSV2, and pbmm2/cuteSV1 as optimal choices for insertion detection (Fig. 4g, h). Collectively, these pipelines demonstrated a favorable balance between accuracy, robustness, and computational feasibility, supporting their use as general-purpose workflows for SV discovery in plant genomes.

Effects of aligner and caller on SV detection performance

The accuracy of SV detection depends on the combined performance of the long-read aligner and SV caller. To determine their individual contributions to cross-species SV discovery, we grouped the 91 pipelines according to aligner or caller, calculated the median performance for each species, and compared their performance across the ten angiosperm species. Among all aligners, NGMLR produced the highest overall F_1 score for deletion detection, whereas Minimap2 performed best for insertions (Fig. 5a, d). Across both SV types, NGMLR consistently delivered the highest precision, whereas Minimap2 achieved the highest recall (Fig. 5b, c, e, f). These findings indicate that NGMLR is better suited for applications emphasizing high-confidence SV calls, whereas Minimap2 is preferable when maximizing SV discovery is the primary objective.

Because the benchmark included 14 SV callers, only the three highest-performing callers for each evaluation metric are presented. Across both deletion and insertion detection, cuteSV1 and cuteSV2 provided the best overall balance between precision and recall, resulting in the highest F_1 scores (Fig. 5g, j). By contrast, Sniffles1 achieved the greatest precision for deletion, whereas cuteSV2 ranked highest for insertion precision (Fig. 5h, k). In contrast, Sniffles2, SVIM1, and SVIM2 consistently ranked among the most sensitive callers, producing the highest recall

Fig. 5 Long-read aligners and variant callers associated effects on SV detection performance across ten angiosperm species. **a-f** Detection metrics for deletions and insertions stratified by seven aligners; **g-l** Detection metrics for deletions and insertions stratified by 14 variant callers. Boxplot shows the median F_1 score, precision and recall of all pipelines. Red asterisks indicate the best-performing tool for each metric



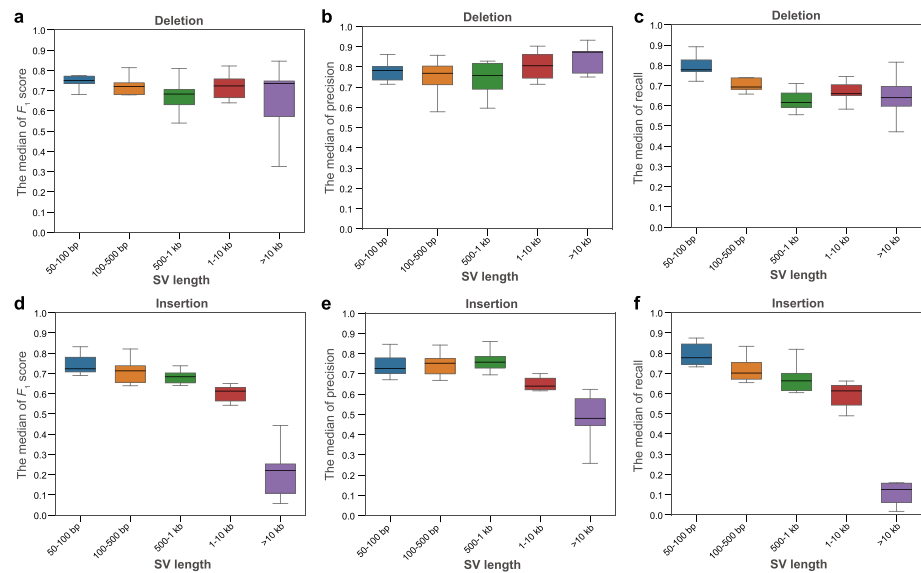
for both deletion and insertion detection (Fig. 5i, l). Together, these results demonstrate that both the aligner and caller substantially influence SV detection performance, emphasizing that optimal tool selection should be guided by whether a study prioritizes balanced performance, maximum sensitivity, or the highest possible precision.

Effects of SV length on SV detection performance

Although PacBio HiFi reads typically exceed 20 kb, their ability to accurately detect structural variants spanning a broad size range has not been comprehensively evaluated in plant genomes. To investigate the influence of SV length, we divided variants into five size classes according to the distribution of the species-specific truth sets: 50–100 bp, 100–500 bp, 500 bp–1 kb, 1–10 kb, and > 10 kb (Fig. 6). For each species and SV size category, the median performance across all 91 SV calling pipelines was calculated and compared.

Deletion detection became progressively less accurate as SV length increased. F_1 score, precision, and recall all declined from the 50–100 bp class through the 500 bp–1 kb interval, indicating that larger deletions are inherently more difficult to identify reliably (Fig. 6a-c). For deletions larger than 1 kb, F_1 score and recall continued

Fig. 6 SV length-associated effects on SV detection performance across ten angiosperm species. SVs were divided into five size categories: 50–100 bp, 100–500 bp, 500 bp–1 kb, 1–10 kb, and > 10 kb. **a–c** Median F_1 score, precision and recall for deletion detection across five SV length categories. **d–f** Median F_1 score, precision and recall for insertion detection across size classes. All values represent medians across all benchmarked SV calling pipelines



to decrease, whereas precision exhibited a modest increase, suggesting that larger deletions were detected more conservatively.

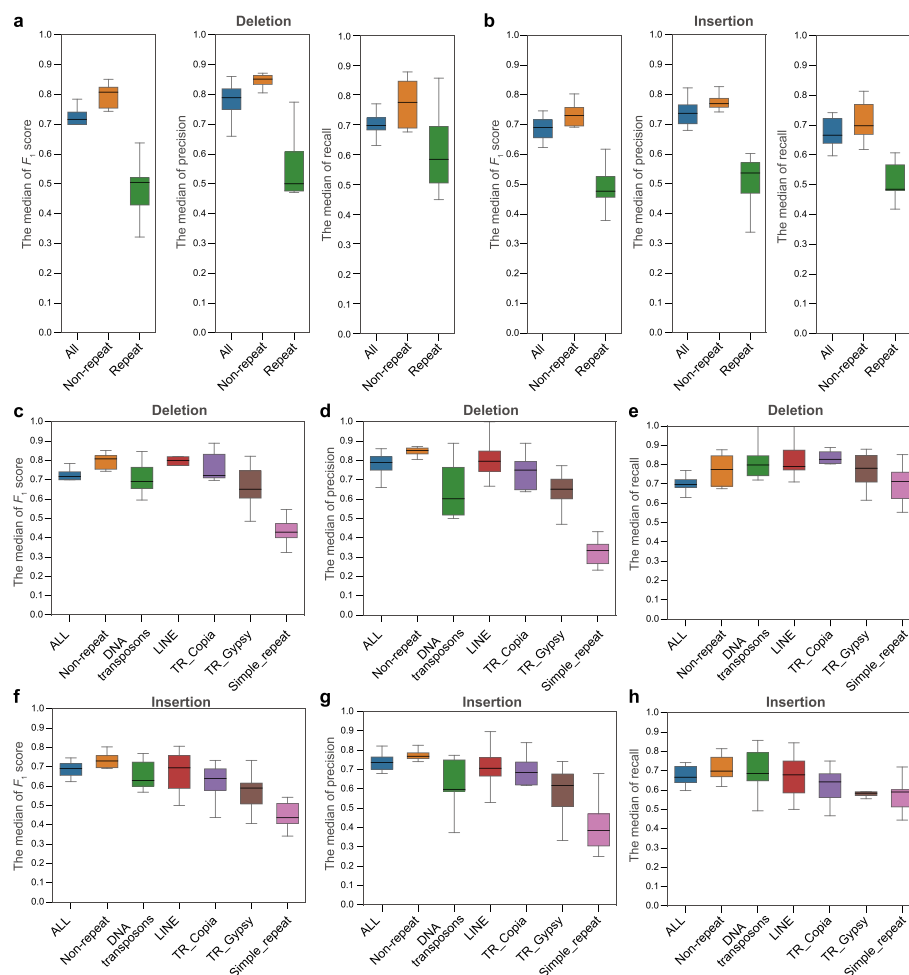
Insertion detection showed an even stronger dependence on SV size. F_1 score, precision, and recall all declined steadily with increasing insertion length (Fig. 5d-f), with the most pronounced performance reduction observed for insertions exceeding 10 kb. This likely reflects the enrichment of repetitive sequences and complex genomic architectures within large insertions, which hinder accurate read alignment and breakpoint resolution and consequently reduce calling accuracy. Overall, these findings establish SV length as a major determinant of detection performance. Although HiFi sequencing substantially improves SV discovery, the accuracy of both deletion and insertion detection decreases as variant size increases, particularly for large structural variants.

Effects of genomic repetitive elements on SV detection performance

As repetitive sequences are a major determinant of accuracy in long-read alignments, and can therefore influence SV detection, we sought to quantify this effect in our comparative analyses. To this end, we annotated genomic repetitive elements in the nine representative angiosperm species and partitioned each genome into repeat and non-repeat regions. Using species-specific SV truth sets and corresponding call sets, we systematically evaluated the performance of each SV detection pipeline in the two genomic compartments.

Across all species and pipelines, SV detection accuracy was consistently impaired in repeat regions, whereas performance was markedly improved in non-repeat regions (Fig. 7a, b; Additional file 1: Fig. S12-13; Additional file 2: Tables S8-10). This pattern was robust across both deletion and insertion variants and across F_1 score, precision, and recall evaluation metrics. For the F_1 score, deletion detection showed an average increase of $11.69 \pm 11.77\%$ in non-repeat regions but a $28.03 \pm 10.35\%$ decrease in repeat regions compared to the average in non-partitioned genomes (Fig. 7a). Insertion detection exhibited a similar trend, with F_1 score improving by $6.12 \pm 3.40\%$ in non-repeat regions and diminishing by $24.41 \pm 12.11\%$ in repeat regions (Fig. 7b). Consistent effects were observed for precision and recall, with precision increasing by $7.51 \pm 5.88\%$ and $10.61 \pm 12.68\%$, for deletions and insertions, respectively, in non-repeat regions, while recall increased by $8.92 \pm 7.95\%$ and $6.69 \pm 7.69\%$. In contrast, within repeat regions, the precision of deletion and insertion calling declined by $27.60 \pm 9.30\%$ and $14.97 \pm 13.76\%$, and recall diminished by $29.01 \pm 8.86\%$ and $23.04 \pm 7.97\%$, respectively. Overall, non-repeat regions yielded substantially higher SV detection accuracy, while repeat-rich regions were associated with pronounced reductions in performance.

Fig. 7 Repeat-associated effects on SV detection performance across ten angiosperm species. F_1 score, precision, and recall for deletion (a) and insertion (b) detection, respectively, evaluated across the whole genome, non-repeat regions, and repeat-rich regions. The median F_1 score (c, f), precision (d, g), and recall (e, h) metrics of DNA transposons, LINEs, LTR-Copia, LTR-Gypsy, and simple repeats, compared with the detection performance of the whole genome and non-repeat regions across all SV calling pipelines for each species

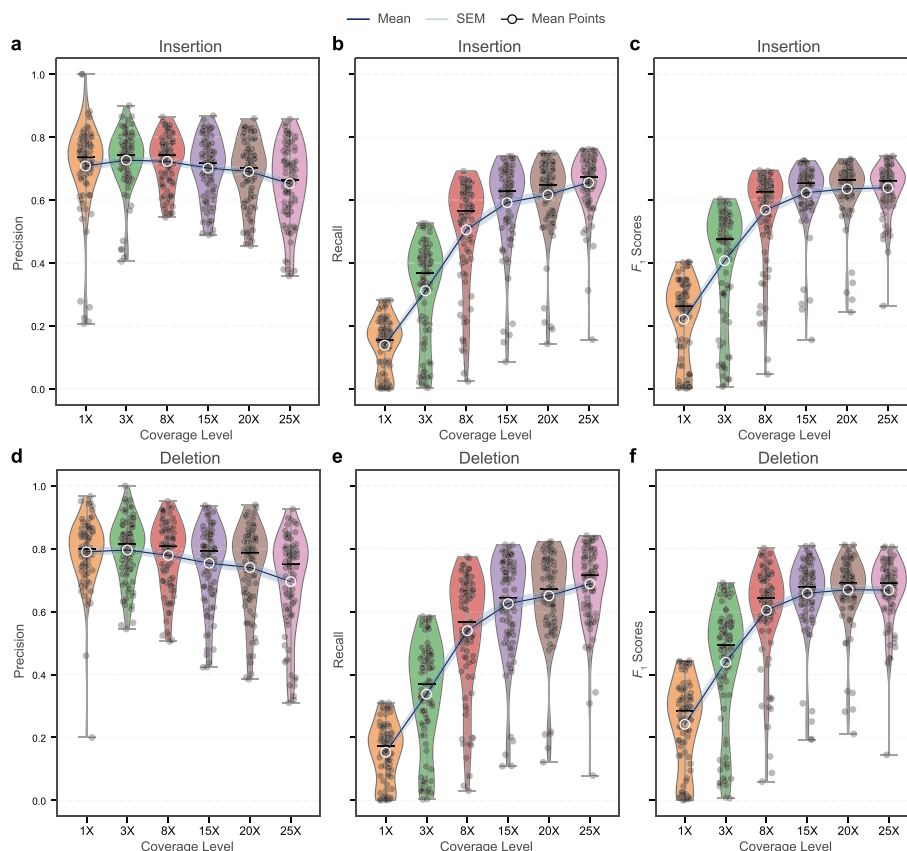


To further determine whether different repeat families exert distinct effects, repetitive elements were classified into five major categories: DNA transposons, LINEs, LTR-Copia, LTR-Gypsy, and simple repeats. Median performance across all SV calling pipelines was calculated for each species to compare SV detection accuracy among repeat classes across the ten angiosperm species. Regardless of repeat type, F_1 score, precision, and recall were consistently reduced for both deletions and insertions relative to non-repeat regions (Fig. 7c–h). Among all repeat categories, simple repeats produced the largest reduction in performance. Their highly homogeneous sequence composition and frequent tandem copy-number variation likely create substantial alignment ambiguity and reduce breakpoint resolution, thereby increasing both false-positive and false-negative SV calls. Collectively, these findings demonstrate that all major classes of repetitive DNA compromise long-read SV detection, with simple repeats representing the greatest barrier to accurate SV identification in plant genomes.

Effects of sequencing depth on SV deletion performance

Our previous work showed that low-coverage HiFi sequencing can allow high precision SV calling in wheat, although the limited available sequencing depth ($6.6\times$) prevented identification of an optimal coverage threshold [43]. To systematically assess the impact of sequencing depth on SV detection performance, we evaluated the 91 SV calling pipelines across six HiFi coverage levels ($1\times$, $3\times$, $8\times$, $15\times$, $20\times$, and $25\times$) in *Arabidopsis* (Fig. 8; Additional file 1: Fig. S14–15; Additional file 2: Tables S11). For both deletions and insertions, precision peaked at low sequencing depth, with $3\times$ HiFi data achieving the highest average precision (0.80 for deletions and 0.73

Fig. 8 Sequencing depth-associated effects on SV detection performance. Precision (**a, d**), recall (**b, e**), and F_1 score (**c, f**) for deletion and insertion detection, respectively, evaluated across six HiFi coverage levels (1×, 3×, 8×, 15×, 20×, and 25×) in *Arabidopsis thaliana*. Mean and SEM show the mean of all pipelines, and standard error of mean, respectively. Values represent the mean across all pipelines, with error bars indicating standard error of the mean (SEM)



for insertions) (Fig. 8a, d). Increasing coverage beyond this depth led to a gradual decline in precision, which decreased by 12.63% on average for deletions and 10.24% on average for insertions (Fig. 8a, d). In contrast, recall and F_1 score increased steadily with sequencing depth, reflecting improved sensitivity and breakpoint support with greater read depth (Fig. 8b, c, e and f). Notably, F_1 scores for both SV types approached saturation at approximately 15× coverage, beyond which further sequencing yielded diminishing returns (Fig. 8c, f). Given that sequencing cost share linear relationship with coverage, these results indicated that ~15× HiFi sequencing for each individual haplotype enables the most robust plant genomic SV detection while limiting costs.

Runtime and memory usage of HiFi-based aligners and callers

Computational efficiency is a key constraint for large-scale, long-read resequencing studies. Using a standardized 100-CPU environment to analyze 25× HiFi data from *Arabidopsis*, we benchmarked the runtime and memory usage of seven aligners and fourteen SV callers (Additional file 1: Fig. S16; Additional file 2: Tables S12). Among aligners, LRA (1.27 CPU·h), BLEND (1.30 CPU·h), minimap2 (1.58 CPU·h), and pbmm2 (3.67 CPU·h) were the fastest, whereas lordFAST was markedly slower (1,398.96 CPU·h), with NGMLR (49.88 CPU·h) and pbmm2 (19.37 CPU·h) showing intermediate runtimes. Memory usage was lowest for LRA (12.27 GB), while other aligners required similarly higher memory resources (24.64–77.16 GB). In contrast, SV callers exhibited generally light computational burdens. Notably, Sniffles2 had the fastest runtime (0.02 CPU·h), representing a substantial improvement over Sniffles1 (5.42 CPU·h). Additionally, most callers required relatively little memory, with SVIM1 (0.20 GB) and Sniffles2 (0.24 GB) using the least, and DeBreak showing the greatest requirement (54.46 GB). Overall, aligners dominated both runtime and memory consumption, thus constituting the primary computational bottleneck for HiFi-based SV calling pipelines.

Discussion

Structural variants are major drivers of plant genomic diversity, adaptive evolution, and agronomic trait variation [5, 7, 8]. In contrast to the human genome, plant genomes are often larger, highly repetitive, and often polyploid, thus presenting substantial challenges for achieving accuracy in SV detection [23]. Despite these influential differences, most current SV calling tools were developed and benchmarked primarily on human genomes, with little or no examination of their robustness and generalizability in plants [17]. One long-standing, critical bottleneck for plant genomics is the absence of high-quality, reusable SV truth sets, which has impeded systematic evaluation of methods across pipelines and species [19]. By constructing SV truth sets at base-level resolution across multiple representative plant species, and applying these truth sets in large-scale benchmarking, this study establishes a comprehensive cross-species framework for evaluating HiFi-based SV detection in plants (Fig. 1).

The high-confidence plant SV truth sets, developed in the present study by integrating multiple reference genome alignments, diverse SV discovery strategies, and direct validation by genotyping with HiFi reads, can greatly accelerate pipeline selection in future long-read plant resequencing studies as well as benchmarking of new alignment and SV calling tools [19]. By adopting a multi-pipeline consensus strategy, we substantially reduced algorithm-specific biases that are particularly pronounced in repeat-rich and structurally complex plant genomes (Figs. 1b, and 2). Beyond serving as a foundation for benchmarking, these truth sets can serve as a valuable resource for plant research, providing a methodological reference point for future plant SV studies. Although multiple artificial intelligence-based SV detection approaches have been developed for human genomes (e.g., SVision and SVision-pro), comparable AI-driven methods tailored to plant genomes are still largely lacking [44, 45]. Future work integrating these truth sets and AI models can further enhance the accuracy and scalability of SV detection [46].

Systematic assessment of 91 SV calling pipelines and nine angiosperm species revealed a performance landscape shaped by SV type, species genomic complexity, and detection pipelines (Fig. 3). Deletions were consistently detected with higher accuracy, precision, and recall than insertions, indicating that deletion detection from PacBio HiFi sequencing is relatively mature, whereas insertion calling remains a key bottleneck in plant genomes. Although high-performing pipelines were identified for most species, their effectiveness varied markedly across species, underscoring strong species dependence in SV calling performance. Collectively, for every species examined, at least one pipeline achieved high precision (0.8), highlighting the great potential of HiFi sequencing for accurate SV detection when paired with optimal pipelines.

Both aligner and caller contributed substantially to overall pipeline performance (Fig. 5). NGMLR consistently produced the highest precision, whereas Minimap2 maximized recall, highlighting the inherent trade-off between confidence and sensitivity. Meanwhile, several widely used SV callers achieved comparable F_1 scores by balancing these competing performance metrics. These findings emphasize that pipeline selection should be driven by the biological objectives of individual studies rather than relying on a universally optimal workflow. Variant size also emerged as a major determinant of detection performance (Fig. 6). As structural variants increased in length, both deletion and insertion detection became progressively less accurate, with the largest insertions proving particularly difficult to identify. The enrichment of repetitive sequences and structurally complex regions within large insertions likely complicates read alignment and breakpoint localization, thereby reducing calling accuracy. Although HiFi sequencing substantially improves long-read SV detection, accurately resolving very large structural variants remains a significant analytical challenge.

Through partitioning of repeat and non-repeat genomic regions and comparison across coverage levels, we found that both genomic features and sequencing strategies impose fundamental constraints on SV detection performance. Specifically, repetitive regions consistently reduce precision, recall, and F_1 score in both deletion and insertion calling, even with HiFi sequencing, indicating that repeats remain a dominant source of uncertainty (Fig. 7). Among all repeat classes, simple repeats produced the greatest loss of performance, likely because their extensive sequence homology and frequent tandem copy-number variation increase alignment ambiguity and

reduce breakpoint resolution. These findings indicate that repetitive DNA remains one of the most important obstacles to reliable long-read SV detection in plant genomes. At the same time, sequencing depth exerts a non-linear effect wherein low coverage yields high precision, consistent with prior observations in human genome benchmarks [22], but recall and F_1 score increase with depth, plateauing at approximately $15\times$ (Fig. 8c, f). This trade-off, consistent with observations in human genomes, suggests that $\sim 15\times$ HiFi coverage for each individual haplotype provides an optimal balance between accuracy and cost for population-scale plant SV studies.

In the present study, we also developed the cross-species stability score, an informative metric for evaluating the robustness of SV calling pipelines across diverse plant genomes (Fig. 4). Using this criterion, we identified NGMLR/Sniffles2 and NGMLR/DeBreak as optimal cross-species pipelines for detecting deletions, while Minimap2/Sniffles1, Minimap2/cuteSV2, and pbmm2/cuteSV1 showed high stability scores in calling insertions. Importantly, these pipelines also showed relatively stronger performance than other SV calling pipelines on wheat. As the size and complexity of the wheat genome led us to exclude it from cross-species pipeline comparisons, these results provided direct validation of the generalizability of these aligner/caller combinations. Together, these results provide a standardized benchmarking framework and practical guidance for selecting robust SV calling pipelines, which can enhance reliability in the application of HiFi sequencing data for plant genomic analyses.

An important consideration when interpreting our benchmark results is the partial overlap between the methods used to construct the truth sets and those being evaluated. Specifically, Minimap2(-ax asm5)-SVIM-asm and Minimap2(-ax asm5)-SyRI were included among the five assembly-based pipelines used to construct the truth sets, whereas Minimap2(-ax map-hifi) and SVIM were also evaluated as components of read-based pipelines. Although SVIM-asm identifies SVs from assembly-to-reference alignments and SVIM identifies SVs from long-read-to-reference alignments, the assembly- and read-based analyses share aligner/caller components and related algorithmic features. This methodological overlap could preferentially increase the concordance of read-based Minimap2-SVIM calls with the truth sets. However, analysis of *Arabidopsis thaliana* showed that only 53 deletions (1.5%) and 35 insertions (1.0%) in the truth sets were exclusively contributed by Minimap2(-ax asm5)-SVIM-asm. The remaining 98.5% of deletions and 99.0% of insertions were supported by at least one other assembly-based pipeline (Additional file 1: Fig. S17). This small pipeline-exclusive contribution suggests that direct circularity in truth-set construction was limited, although bias arising from shared algorithmic features cannot be completely excluded. In addition, the greater number of unshared read-based SVIM calls may partly reflect residual heterozygous or haplotype-specific SVs present in the long-read data but absent from the haplotype-merged assemblies, particularly in highly mutable regions such as simple repeats. Because high-quality haplotype-resolved assemblies were unavailable for the corresponding species, this effect could not be systematically quantified. Accordingly, these SV truth sets should be regarded as practical benchmarking references rather than exhaustive ground-truth representations of all haplotype-specific SVs. Accordingly, absolute performance estimates should be interpreted in light of these methodological limitations.

Despite the breadth of this benchmarking effort, several limitations should be acknowledged. First, all benchmark species belong to the angiosperms. Other major plant lineages, including gymnosperms, monilophytes, lycophytes, bryophytes, and green algae, could not be evaluated because matched high-quality reference genomes and PacBio HiFi datasets remain scarce. Expanding taxonomic sampling as additional genomic resources become available will be essential for assessing the broader applicability of our conclusions. Second, to ensure consistent comparisons across all species, the SV truth sets were constructed exclusively from PacBio HiFi datasets shared by every species. Consequently, ONT data, which were available for only a subset of taxa, were not incorporated. Future truth sets integrating HiFi, ultra-long ONT sequencing, short-read sequencing, and orthogonal experimental validation should provide even greater accuracy and completeness. Third, our analyses focused exclusively on deletions and insertions, the two most abundant structural variant classes in plant genomes. Other important variant types, including inversions and translocations, remain to be systematically benchmarked. In addition, the benchmark species consisted primarily of well-characterized cultivated accessions with relatively high homozygosity. Highly heterozygous wild populations and autopolyploid species were not included because

suitable genomic resources are currently unavailable. Finally, because comprehensive ONT datasets were lacking for most benchmark species, this study focused exclusively on PacBio HiFi-based workflows. As high-quality ONT datasets continue to accumulate, future benchmarking efforts should incorporate ONT-specific aligners and callers to enable comprehensive cross-platform comparisons. Together, these future developments will further strengthen benchmarking frameworks for structural variant detection across the full diversity of plant genomes.

Conclusions

This study established a benchmarking framework for SV detection in plant genomes using PacBio HiFi long-read sequencing. By constructing SV truth sets across ten representative angiosperm species and evaluating 91 SV calling pipelines, we showed that detection performance varied substantially across methods, SV types, and genomic contexts. Our analyses highlighted major limitations in large SVs and repeat-rich regions and uncovered nonlinear effects of sequencing depth on detection precision. Importantly, we identified SV calling pipelines with robust cross-species performance, offering practical guidance for plant genomics applications. Together, the high-confidence SV truth sets and standardized evaluation framework provide valuable community resources and lay a foundation for reproducible, scalable, and comparative SV analyses in plants.

Methods

Plant species selection and genome characterization

Ten representative angiosperm species, including model organisms and staple crops, were selected for this study, including *Arabidopsis thaliana*, *Gossypium hirsutum*, *Glycine max*, *Solanum lycopersicum*, *Panicum miliaceum*, *Setaria italica*, *Sorghum bicolor*, *Zea mays*, *Oryza sativa*, and *Triticum aestivum* (Additional file 2: Tables S1). For each species, two high-quality reference genome assemblies and high-coverage PacBio HiFi sequencing data were available for truth set construction and pipeline evaluation (Additional file 2: Tables S2).

Phylogenetic reconstruction of the selected species based on predicted protein sequences was performed using OrthoFinder (version 2.5.5) [47]. A species tree was inferred with 1,000 bootstrap replicates, and the resulting phylogeny was visualized using iTOL (<https://itol.embl.de/>). To characterize genome repetitiveness, repeat libraries were generated for each species using RepeatModeler (version 2.0.7) [48], followed by repeat annotation with RepeatMasker (version 4.2.3) using default parameters [49]. The proportion of repetitive sequences in each genome was calculated based on RepeatMasker annotations.

Construction of species-specific SV truth sets

For each species, five assembly-based SV calling pipelines were performed between two high-quality reference assemblies, designating one as the reference and the other as the query (Additional file 2: Tables S2), resulting five SV candidate datasets.

1. SV1: Cactus-pangenome (version 3.0.0) [50] was applied directly to the reference and query genome to generate SV calls using default parameters.
2. SV2: Dipcall (version 0.3) [51] was used to identify SVs directly from the the reference and query genome with default settings.
3. SV3: After aligning query genome to the reference genome with Minimap2 (version 2.30) [31] using the parameters -ax asm5 -cs -r2k, SAM files were converted to BAM files using SAMtools (version 1.21) [52]. SVs were then detected using SVIM-asm (version 1.0.3).

4. SV4: Query genome was aligned to the reference genome using Minimap2 with parameters -ax asm5 -eqx, and SVs were identified using SyRI (version 1.7.1) [53] with default parameters.
5. SV5: Gene annotations were processed using AnchorWave (version 1.2.6) [54] (gff2seq) and aligned with minimap2. MAF files were generated from genome alignments using AnchorWave genoAli, converted to SAM format using BioConvert (version 1.1.1) [55], then transformed into BAM format with SAMtools. SVs were ultimately detected using SVIM-asm.

SV candidates shorter than 50 bp were filtered out with BCFtools (version 1.21) [52] and the remaining SV call sets from each respective pipeline were combined using the merge function in BCFtools. Duplicate SVs were then eliminated with Truvari [56] collapse to generate integrated SV candidate sets for each species. High-depth HiFi sequencing data were then used to genotype each candidate SV using SVJedi-graph (version 1.2.1) [57]. SVs genotyped as heterozygous (0/1) or homozygous alternate (1/1) were retained as high-confidence SV truth sets. These truth sets were further separated into deletion and insertion datasets for downstream benchmarking analyses.

Construction of SV calling pipelines

Seven long-read aligners (BLEND, lordFAST, LRA, minimap2, NGMLR2, pbmm2, and Winnowmap) and fourteen SV callers (cuteSV1, cuteSV2, DeBreak, DELLY, Dysgu, gcSV, NanoVar, pbsv, Sniffles1, Sniffles2, SVDSS1, SVDSS2, SVIM1, and SVIM2) were collected from published or preprint studies and detailed software versions and parameters are provided (Additional file 2: Tables S13). All possible aligner–caller combinations were generated, resulting in 98 SV calling pipelines. Seven incompatible combinations (SVDSS1/SVDSS2 with LRA, BLEND, or lordFAST; NanoVar with pbmm2) that yielded no results in the above preliminary tests were excluded from subsequent analyses. The remaining 91 pipelines were applied to HiFi sequencing data for SV detection. SVs shorter than 50 bp were filtered out using BCFtools, and the resulting SV call sets were used for performance benchmarking.

Benchmarking analysis

Benchmarking was performed separately for deletions and insertions. For each species, the SV call sets generated by the 91 pipelines were compared against the corresponding SV truth set using Truvari [56] bench with the parameter –sizemax 1,000,000. Precision, recall and F_1 score performance metrics were respectively calculated as follows:

$$Precision = \frac{TPs}{TPs + FPs}$$

$$Recall = \frac{TPs}{TPs + FNs}$$

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

where TPs , FPs , and FNs represent the numbers of true positives, false positives, and false negatives, respectively.

To assess the impact of repetitive sequences on SV detection, repeat regions longer than 1,000 bp were identified with RepeatMasker and subsequently extracted. Benchmarking was independently conducted on the partitioned repeat and non-repeat genomic regions. To evaluate the effect of sequencing depth on SV detection performance, HiFi reads from *Arabidopsis thaliana* were randomly downsampled to six coverage levels (1×, 3×,

8×, 15×, 20×, and 25×). SV calling, filtering, and benchmarking were conducted for each downsampled dataset using the same procedures described above.

Assessment of cross-species generalizability among SV calling pipelines

To quantify the robustness of each SV calling pipeline across species, we estimated stability scores for each performance metric in all pipelines using the following formula:

$$\text{Stability score} = 1 - \frac{\sigma}{\mu}$$

where σ represents the standard deviation and μ represents the mean of a given performance metric across species. The stability score ranges from 0 to 1, with values closer to 1 indicating higher cross-species stability.

To identify optimal pipelines for multi-species SV detection, pipelines with F_1 score ≥ 0.7 , precision ≥ 0.8 , recall ≥ 0.7 , and stability score ≥ 0.8 were considered optimal for deletion detection. For insertion detection, thresholds of F_1 score ≥ 0.7 , precision ≥ 0.7 , recall ≥ 0.7 , and stability score ≥ 0.8 were applied. Pipelines meeting these criteria and exhibiting high stability scores (0.8) were designated as optimal cross-species SV calling pipelines.

Runtime and memory usage assessment

Runtime and memory consumption of aligners and callers were evaluated using 25× HiFi public sequencing data from *Arabidopsis thaliana*. All tests were conducted on a server equipped with 2*AMD EPYC 9754 128-core processors, 3 TB of memory, and 131 TB of storage. Each aligner and caller was executed independently using 100 threads and repeated ten times. Runtime (elapsed wall-clock time) and memory usage (maximum resident set size) were recorded using the Linux command “/usr/bin/time”. The average runtime and memory consumption across replicates were used for downstream comparisons.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1186/s13059-026-04300-x>.

Acknowledgements Not applicable.

Peer review information Fritz Sedlazeck and Veronique van den Berghe were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions HL provided overall supervision of the work. HL and ZZ developed the analytical workflow. ZZ, JX and HL performed data analysis. ZW, TY, KH, DL, YZ, KW, ZX, CH and JW helped with data analysis. ZZ, HL and JX wrote and revised the manuscript. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the Biological Breeding-National Science and Technology Major Project (2023ZD04073), the National Natural Science Foundation of China (32361143514), Project of Sanya Yazhou Bay Science and Technology City (SKIC-JYRC-2025-44), the Research Program of Zhongyuan Research Center, the Chinese Academy of Agricultural Sciences (CAAS-ZRC-ZYZX202501) and the Agricultural Science and Technology Innovation Program (CAAS-CSIAF-202303).

Data availability Details of the high-quality reference and query genome assemblies and high-coverage PacBio HiFi sequencing datasets used in this study are provided in Additional file 2: Table S2. The analysis scripts and corresponding parameters are provided in Additional file 2: Table S13.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare that they have no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet.* 2020;21:171–89. <https://doi.org/10.1038/s41576-019-0180-9>.
2. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. *Nature.* Nat Publ Group. 2015;526:75–81. <https://doi.org/10.1038/nature15394>.
3. McDonald-McGinn DM, Sullivan KE, Marino B, Philip N, Swillen A, Vorstman JAS, et al. 22q11.2 deletion syndrome. *Nat Rev Dis Primers.* 2015;1:15071. <https://doi.org/10.1038/nrdp.2015.71>.
4. Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. *Nature.* 2020;581:444–51. <https://doi.org/10.1038/s41586-020-2287-8>.
5. Liang Y, Liu H-J, Yan J, Tian F. Natural variation in crops: realized understanding continuing promise. *Annu Rev Plant Biol.* 2021;72:357–85. <https://doi.org/10.1146/annurev-arplant-080720-090632>.
6. Wang K, Abid MA, Rasheed A, Crossa J, Hearne S, Li H. DNNP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol Plant.* 2023;16:279–93. <https://doi.org/10.1016/j.molp.2022.11.004>.
7. Hu Y, Colantonio V, Müller BSF, Leach KA, Nanni A, Finegan C, et al. Genome assembly and population genomic analysis provide insights into the evolution of modern sweet corn. *Nat Commun.* 2021. <https://doi.org/10.1038/s41467-021-21380-4>.
8. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat Genet.* 2011;43:1160–3. <https://doi.org/10.1038/ng.942>.
9. Dong Z, Xiao Y, Govindarajulu R, Feil R, Siddoway ML, Nielsen T, et al. The regulatory landscape of a core maize domestication module controlling bud dormancy and growth repression. *Nat Commun.* 2019;10:3810. <https://doi.org/10.1038/s41467-019-11774-w>.
10. Jay P, Chouteau M, Whibley A, Bastide H, Parrinello H, Llaurens V, et al. Mutation load at a mimicry supergene sheds new light on the evolution of inversion polymorphisms. *Nat Genet.* 2021;53:288–93. <https://doi.org/10.1038/s41588-020-00771-1>.
11. Cicconardi F, Milanetti E, de Pinheiro Castro EC, Mazo-Vargas A, Van Belleghem SM, Ruggieri AA, et al. Evolutionary dynamics of genome size and content during the adaptive radiation of Heliconiini butterflies. *Nat Commun.* 2023;14:5620. <https://doi.org/10.1038/s41467-023-41412-5>.
12. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. *Genome Biol.* 2019;20:246. <https://doi.org/10.1186/s13059-019-1828-7>.
13. Porubsky D, Eichler EE. A 25-year odyssey of genomic technology advances and structural variant discovery. *Cell.* 2024;187:1024–37. <https://doi.org/10.1016/j.cell.2024.01.002>.
14. Reid-Nicholson MD, Ramalingam P, Adeagbo B, Cheng N, Peiper SC, Terris MK. The use of Urovysion™ fluorescence in situ hybridization in the diagnosis and surveillance of non-urothelial carcinoma of the bladder. *Mod Pathol.* 2009;22:119–27. <https://doi.org/10.1038/modpathol.2008.179>.
15. Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. *Nat Genet.* 1999;21:48–50. <https://doi.org/10.1038/4475>.

16. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* 2019;37:1155–62. <https://doi.org/10.1038/s41587-019-0217-9>.
17. Ahsan MU, Liu Q, Perdomo JE, Fang L, Wang K. A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat Methods.* 2023;20:1143–58. <https://doi.org/10.1038/s41592-023-01932-w>.
18. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 2019;20:117. <https://doi.org/10.1186/s13059-019-1720-5>.
19. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol.* 2020;38:1347–55. <https://doi.org/10.1038/s41587-020-0538-8>.
20. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun.* 2019;10:3240. <https://doi.org/10.1038/s41467-019-11146-4>.
21. Liu YH, Luo C, Golding SG, Ioffe JB, Zhou XM. Tradeoffs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. *Nat Commun.* 2024;15:1–59. <https://doi.org/10.1038/s41467-024-46614-z>.
22. Liu Z, Xie Z, Li M. Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data. *Genome Biol.* 2024;25:188. <https://doi.org/10.1186/s13059-024-03324-5>.
23. He B, Liu W, Li J, Xiong S, Jia J, Lin Q, et al. Evolution of plant genome size and composition. *Genomics Proteomics Bioinform.* 2024;22:qzae078. <https://doi.org/10.1093/gpbjnl/qzae078>.
24. (IWGSC) TIWGSC, Appels R, Eversole K, Stein N, Feuillet C, Keller B, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science.* 1979;361:eaar7191. <https://doi.org/10.1126/science.aar7191>.
25. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science.* 1979;376:44–53. <https://doi.org/10.1126/science.abj6987>.
26. Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell.* 2005;17:343–60. <https://doi.org/10.1105/tpc.104.025627>.
27. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 2010;11:R52. <https://doi.org/10.1186/gb-2010-11-5-r52>.
28. Firtina C, Park J, Alser M, Kim JS, Cali DS, Shahroodi T, et al. BLEND: a fast, memory-efficient and accurate mechanism to find fuzzy seed matches in genome analysis. *NAR Genom Bioinform.* 2023;5:Iqad004. <https://doi.org/10.1093/nargab/lqad004>.
29. Haghshenas E, Sahinalp SC, Hach F. LordFAST: sensitive and fast alignment search tool for long noisy read sequencing data. *Bioinformatics.* 2019;35:20–7. <https://doi.org/10.1093/bioinformatics/bty544>.
30. Ren J, Chaisson MJP. Lra: a long read aligner for sequences and contigs. *PLoS Comput Biol.* 2021;17:e1009078. <https://doi.org/10.1371/journal.pcbi.1009078>. (Public Library of Science).
31. Li H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics.* 2021;1–3. <https://doi.org/10.1093/bioinformatics/btab705>.
32. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods.* 2018;15:461–8. <https://doi.org/10.1038/s41592-018-0001-7>.
33. Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods.* 2022. <https://doi.org/10.1038/s41592-022-01457-8>.
34. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, et al. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.* 2020;21:189. <https://doi.org/10.1186/s13059-020-02107-y>.
35. Chen Y, Wang AY, Barkley CA, Zhang Y, Zhao X, Gao M, et al. Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. *Nat Commun.* 2023;14:283. <https://doi.org/10.1038/s41467-023-35996-1>.
36. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics.* 2012;28:i333–9. <https://doi.org/10.1093/bioinformatics/bts378>.
37. Cleal K, Baird DM. Dysgu: efficient structural variant calling using short or long reads. *Nucleic Acids Res.* 2022;50:e53–e53. <https://doi.org/10.1093/nar/gkac039>.
38. Li G, Liu Y, Liu B, Qian L, Wang Y. gcSV: a unified framework for comprehensive structural variant detection. *bioRxiv.* 2025;2025.02.10.637589. <https://doi.org/10.1101/2025.02.10.637589>.
39. Tham CY, Tirado-Magallanes R, Goh Y, Fullwood MJ, Koh BTH, Wang W, et al. Nanovar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. *Genome Biol.* 2020;21:56. <https://doi.org/10.1186/s13059-020-01968-7>.
40. Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol.* 2024;42:1571–80. <https://doi.org/10.1038/s41587-023-02024-y>.
41. Denti L, Khorsand P, Bonizzoni P, Hormozdiari F, Chikhi R. SVDSS: structural variation discovery in hard-to-call genomic regions using sample-specific strings from accurate long reads. *Nat Methods.* 2023;20:550–8. <https://doi.org/10.1038/s41592-022-01674-1>.

42. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. *Bioinformatics*. 2019;35:2907–15. <https://doi.org/10.1093/BIOINFORMATICS/BTZ041>. (Oxford Academic).
43. Zhang Z, Zhang J, Kang L, Qiu X, Xu S, Xu J, et al. Structural variation discovery in wheat using PacBio high-fidelity sequencing. *Plant J*. 2024;120:687–98. <https://doi.org/10.1111/tpj.17011>.
44. Lin J, Wang S, Audano PA, Meng D, Flores JJ, Kusters W, et al. SVision: a deep learning approach to resolve complex structural variants. *Nat Methods*. 2022;19:1230–3. <https://doi.org/10.1038/s41592-022-01609-w>.
45. Wang S, Lin J, Jia P, Xu T, Li X, Liu Y, et al. De novo and somatic structural variant discovery with SVision-pro. *Nat Biotechnol*. 2025;43:181–5. <https://doi.org/10.1038/s41587-024-02190-7>.
46. Wang S, Ye K. Deep-learning based representation and recognition for genome variants—from SNVs to structural variants. *Natl Sci Rev*. 2024;11:nwae335. <https://doi.org/10.1093/nsr/nwae335>.
47. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20:238. <https://doi.org/10.1186/s13059-019-1832-y>.
48. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 2020;117:9451–7. <https://doi.org/10.1073/pnas.1921046117>.
49. Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 2004;5:4.10.1–4.10.14. <https://doi.org/10.1002/0471250953.bi0410s05>.
50. Hickey G, Monlong J, Ebler J, Novak AM, Eizenga JM, Gao Y, et al. Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol*. 2024;42:663–73. <https://doi.org/10.1038/s41587-023-01793-w>.
51. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, et al. A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods*. 2018;15:595–7. <https://doi.org/10.1038/s41592-018-0054-7>.
52. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10:giab008. <https://doi.org/10.1093/gigascience/giab008>.
53. Goel M, Sun H, Jiao W-B, Schneeberger K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol*. 2019;20:277. <https://doi.org/10.1186/s13059-019-1911-0>.
54. Song B, Marco-Sola S, Moreto M, Johnson L, Buckler ES, Stitzer MC. AnchorWave: sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. *Proc Natl Acad Sci U S A*. 2022;119:e2113075119. <https://doi.org/10.1073/pnas.2113075119>.
55. Caro H, Dollin S, Biton A, Brancotte B, Desvillechabrol D, Dufresne Y, et al. BioConvert: a comprehensive format converter for life sciences. *NAR Genom Bioinform*. 2023;5:Iqad074. <https://doi.org/10.1093/nargab/lqad074>.
56. English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol*. 2022;23:271. <https://doi.org/10.1186/s13059-022-02840-6>.
57. Romain S, Lemaitre C. SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph. *Bioinformatics*. 2023;39:i270–8. <https://doi.org/10.1093/bioinformatics/btad237>.

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

Authors and Affiliations

Zhiliang Zhang (张治梁)^{1,2} · Jun Xu (徐俊)^{1,2} · Zhizhan Wang^{1,3} · Tingxi Yu^{1,2} · Kunhui He^{1,2} · Duan Liu^{1,2} · Ying Zhao^{1,2} · Keai Wang^{1,2} · Zhaorong Xie^{1,2} · Changling Huang^{1,2} · Jiankang Wang^{1,2} · Huihui Li (李慧慧)^{1,2}

✉ Huihui Li (李慧慧)
lihuihui@caas.cn

Zhiliang Zhang (张治梁)
zhangzhiliang@caas.cn

Jun Xu (徐俊)
xujune29@gmail.com

Zhizhan Wang
wangzhizhan227@163.com

Tingxi Yu
yutingxi@caas.cn

Kunhui He
hekunhui@caas.cn

Duan Liu
821012550110@caas.cn

Ying Zhao
zy17303914292@163.com

Keai Wang
15729223329@163.com

Zhaorong Xie
xiezhazhaorong0520@163.com

Changling Huang
huangchangling@caas.cn

Jiankang Wang
wangjiankang@caas.cn

- ¹ State Key Laboratory of Crop Gene Resources and Breeding, National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China
- ² Present Address: Chinese Academy of Agricultural Sciences, Nanfan Research Institute, Sanya 572024, Hainan, China
- ³ Zhongyuan Research Center, Chinese Academy of Agricultural Sciences, Xinxiang 453500, China