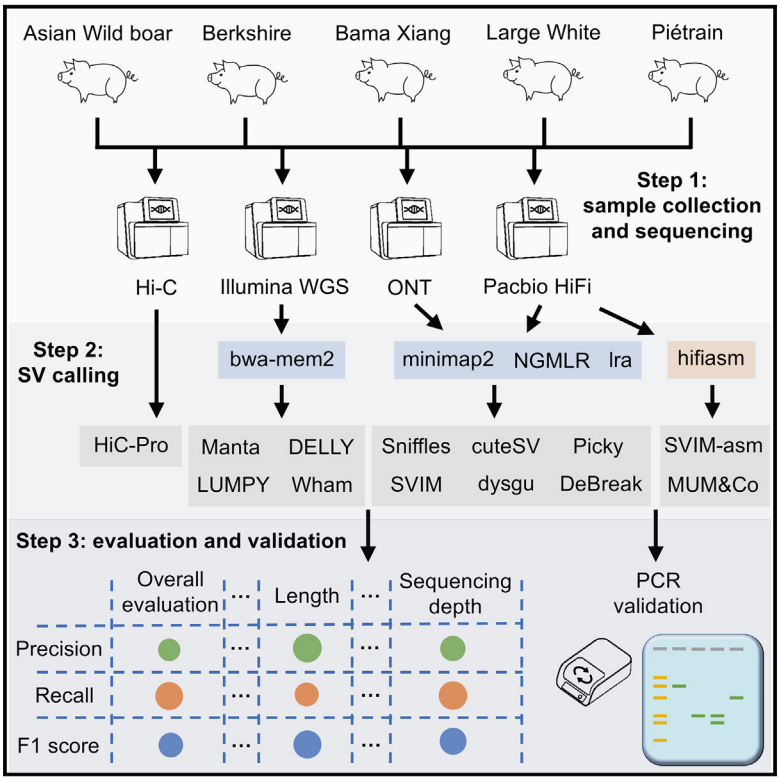
iScience

Article

Systematic benchmarking of tools for structural variation detection using short- and long-read sequencing data in pigs

Graphical abstract



Highlights

- Long reads underpin a reliable structural variation (SV) detection in pigs
- The assembly-based SVIM-asm tool is superior on both accuracy and resource use
- The alignment-based tools hold a strong detection power even at 5 × sequencing depth
- SVs in complex repeat and runs of homozygosity regions could be precisely detected

Authors

Sang He, Bangmin Song, Yueting Tang, ..., Jicai Jiang, Zhonglin Tang, Guoqiang Yi

Correspondence

tangzhonglin@caas.cn (Z.T.), yiguoqiang@caas.cn (G.Y.)

In brief

Structural biology; Bioinformatics; Omics

He et al., 2025, iScience *28*, 111983
March 21, 2025
https://doi.org/10.1016/j.isci.2025.111983

CellPress

iScience

CellPress
OPEN ACCESS

Article

Systematic benchmarking of tools for structural variation detection using short- and long-read sequencing data in pigs

Sang He,[1,2] Bangmin Song,[1] Yueting Tang,[1] Xiaolu Qu,[1] Xingzheng Li,[1] Xintong Yang,[1] Qi Bao,[1] Lingzhao Fang,[3] Jicai Jiang,[4] Zhonglin Tang,[1,2,5,*] and Guoqiang Yi[1,2,5,6,*]

[1]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Key Laboratory of Livestock and Poultry Multi-omics of MARA, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China
[2]Kunpeng Institute of Modern Agriculture at Foshan, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Foshan 528226, China
[3]Center for Quantitative Genetics and Genomics, Aarhus University, 8000 Aarhus, Denmark
[4]Department of Animal Science, North Carolina State University, Raleigh, NC 27695, USA
[5]Bama Yao Autonomous County Rural Revitalization Research Institute, Bama 547500, China
[6]Lead contact
*Correspondence: tangzhonglin@caas.cn (Z.T.), yiguoqiang@caas.cn (G.Y.)
https://doi.org/10.1016/j.isci.2025.111983

SUMMARY

Evaluating diverse structural variation (SV) detection-relevant programs leveraging different algorithms has become a pressing need in humans and farm animals. We addressed this by sequencing five genetically diverse pig individuals (breeds) with short- and long-read DNA-sequencing platforms. We created the SV benchmark set for each breed and assessed the performance of 16 SV calling-relevant tools. Results showed that long-read platforms enabled detecting many SVs missed by short-read platforms with similar precision. Benchmark SVs, mainly 200–500 bp insertions/deletions, had high validation rates. The assembly-based SV calling program SVIM-asm showed superior detection performance and resource consumption. The SVs with more supporting reads, sizes under 1 kb, outside simple repeat area, in low GC content and runs of homozygosity regions, had higher detection accuracy. Alignment-based tools performed well even at 5 × depth. Our study provides systematic guidance for an optimal SV calling pipeline in pigs and other farm animals.

INTRODUCTION

Pigs (*Sus scrofa*) have served not just as a major meat supply domestic animal but also as a critical biomedical model to human beings due to their high similarity to humans in anatomy, pathology, and physiology.[1–3] In this sense, utterly deciphering the genetic architecture of pigs is essential to meet the ever-rising need for both agricultural and medical values. Structural variations (SVs), as a critical form of genetic variants, hold great potential to account for the missing heritability that is not captured by single nucleotide polymorphisms (SNPs).[4,5] Previous studies have demonstrated their essential role in determining complex diseases in humans and shaping important agronomic traits and environmental adaption in animals and plants.[1,6–13] To date, SVs have been utilized to study domestication history and disclose genetic diversity in pigs.[13–16] Some preliminary efforts on SV-enabled genome-wide association study (GWAS) have been made to explore the relevance of SVs to phenotypes.[17,18] The renowned international Functional Annotation of Animal Genomes (FAANG)[19] and Farm animal Genotype-Tissue Expression (FarmGTEx)[20,21] projects targeting to clarify the genome function and decipher the genotype-to-phenotype link in farm animals are far-reaching, yet SVs are still scarcely considered. Bearing the vast potential of SVs in resolving genetic variance in mind, there is still immense room to exploit the merit of SVs in pigs.

Taking full advantage of SVs relies on efficient *in-silico* approaches to detect genome-wide SVs. By now, abundant SV calling programs (callers) with diverse algorithms have been developed.[22,23] These algorithms have different pros and cons, incurring a choice dilemma. The reads alignment (aligner) and genome assembly (assembler) programs also affect SV detection.[24–26] Therefore, a comprehensive evaluation of SV calling-relevant programs, including aligners, assemblers, and callers, and seeking the optimal combination is imperative,[27] and the prerequisite is to establish a high-quality SV benchmark set. In humans, the Genome In A Bottle Consortium (GIAB) has released and updated a gold standard deletion (DEL) and insertion (INS) set, greatly boosting the development, optimization, and comparison of bioinformatics software. Although many studies chose to *in silico* simulate a benchmark set,[28–31] the simulated set is too idealized to reflect the genuine blueprint of SVs in the real world due to their complex patterns.[28] Particularly, pig genomes are characterized by a higher average level

of linkage disequilibrium (LD) and homozygosity relative to humans.[32] Thus, it is envisioned that the SV calling schemes performing well in humans are not necessarily suited to handling pig genomes. In this sense, compiling a pig-specific gold standard SV benchmark set and assessing the effectiveness of various SV calling-relevant programs when dealing with pig genomes is of interest and necessity.

A reliable and confident SV disclosure should incorporate a full spectrum of SVs. The widely used next-generation sequencing (NGS) platforms, like Illumina whole genome resequencing (WGS), are advantageous on base-level resolution but short on DNA read length, which incurs the inability to accommodate both breakpoints of large and complex SVs.[33] In comparison, third-generation sequencing technologies (TGS) such as PacBio circular consensus sequencing (HiFi) and Oxford Nanopore Technologies (ONT) could produce much longer reads and endorse large and complex SV detection.[34,35] Moreover, both TGS technologies are superior in detecting SVs in repetitive genomic regions relative to NGS.[35,36] Taken together, using long reads to disclose SVs is more promising, and so is the benchmark set compilation.

Because there is still a lack of a systematic and comprehensive assessment of SV calling pipelines in pigs, in this study, we attempted to establish pig SV benchmark sets using real sequencing data and accordingly assess 16 SV calling-relevant programs in tackling pig genomes. To achieve this goal, we collected five pig individuals from five different breeds with abundant genetic diversity and deeply sequenced them using two prevailing long-read (i.e., HiFi and ONT) and two widely used short-read platforms (i.e., WGS and high-throughput chromosome conformation capture (Hi-C)). The evaluation of SV calling pipelines was not only conducted globally but also took into account SVs with varying characteristics and settled genomic regions, each being assessed individually. Our primary aim is to seek an efficient SV calling pipeline for pigs and other farm animals, facilitating breeders and geneticists to exploit the merit of SVs in practice.

RESULTS

Characterizations of SV callsets based on different platforms, aligners, and callers

To enable sufficient SV detection power, we performed high-depth sequencing for five pig individuals from five geographically diverse breeds (Table S1, Figure S1). After the quality control of call rate and linkage disequilibrium (LD) pruning, 690,855 approximately independent SNPs from WGS data were used to depict the genetic diversity. As shown in Figure S2, there were two clades: one of AW and BMA, along with the other of BKS, LW, and PTR.

To gain an insight into the difference in SV calling performance on different platforms, aligners, and callers, we inspected the number of SVs in each callset and the similarity between callsets. Concerning the number of SVs, EagleC[37] using Hi-C reads detected just eight to 26 large-size SVs, which was the lowest among all callsets (Figure 1A). This result could be attributed to the insufficient resolution for small-scale SVs and complex regions when the Hi-C technology captures chromatin interactions across the entire genome. Overall, the TGS platforms endorsed a richer detection of putative SVs relative to the NGS platforms.

Inspecting the number of overlapped SVs between different callsets for each sample (Figures 1B and S3), EagleC based on Hi-C data produced an independent callset, and so did all Picky[38] callsets using TGS data (Figure 1B). The four WGS-based callsets, especially the Manta[39] callset, presented a moderate similarity to the TGS-based callsets. Among the TGS-based callsets beside Picky ones, the HiFi-based SVIM[40] callsets formed a relatively conservative block to others probably because more SVs were detected (Figure 1A). The clustering result suggests that the primary factor driving SV detection is the caller, followed by the platform, and finally, the aligner.

Construction of the SV benchmark set in pigs

To provide robust SV benchmark sets to the scientific community in pig genetics and genomics, we integrated the SV callsets from 18 aligner-caller combinations and two assembly-based callers based on the TGS long reads of the five pig breeds (Figure 2). Over 50,000 SVs were kept in the benchmark sets of AW and BMA, which are genetically distant from the Duroc reference genome (Figure 3A). In contrast, only approximately 30,000 SVs remained in the benchmark sets of the other three European breeds. We found more overlapped SVs between two breeds with higher genetic similarity. Most SVs in the benchmark sets were less than 500 bp in size, with a peak at a size range of 200–300 bp (Figure 3B). Deletions and insertions were the two major types of SV, irrespective of the SV length considered. The other three types of SV—duplication, inversion, and translocation—were very rare in the five benchmark sets (Table S2), thus, we did not take them into consideration in the following analyses. There were more than 50,000 SVs exclusively identified in one of the five breeds. The number of SVs detected in all benchmark sets was less than 10,000 (Figure 3C).

To further describe the distribution of SVs across the genome, we took advantage of genomic annotations of different features based on the physical position of SVs and annotated regions. We found that most SVs resided outside the interspersed repeats, i.e., long (LINE) and short (SINE) interspersed nuclear elements, long terminal repeats (LTR), DNA transposons (DNA), and simple repeats (micro-satellites) regions for all breeds. LINEs, SINEs, and simple repeats were hotspot repetitive regions SVs settled in (Figure S4A). In terms of the predicted variant effect, intronic and intergenic SVs occupied the most, followed by those residing upstream and downstream of genes (Figure S4B), implying the potential regulatory role of SVs to genes.

To verify the authenticity of the *in silico* detected SVs in the benchmark sets, we randomly selected 18 SVs (14 deletions and 4 insertions) and 8 pigs from 5 breeds (AW, BKS, BMA, LW, and Duroc) for *in vitro* PCR validation. Thus, a total of 144 PCR assays (gel lanes) were implemented. Our results revealed that 124 samples were successfully validated, achieving an overall validation rate of 88.2%, with true positive and true negative rates respectively of 90.2% and 87.1% (Table S3), which accords with some previous studies.[30,41,42] Two experimentally validated SVs, one deletion located in chromosome 7 of AW with a length of 288 bp (there was 1 bp difference at the end position between results from the caller and the comparison

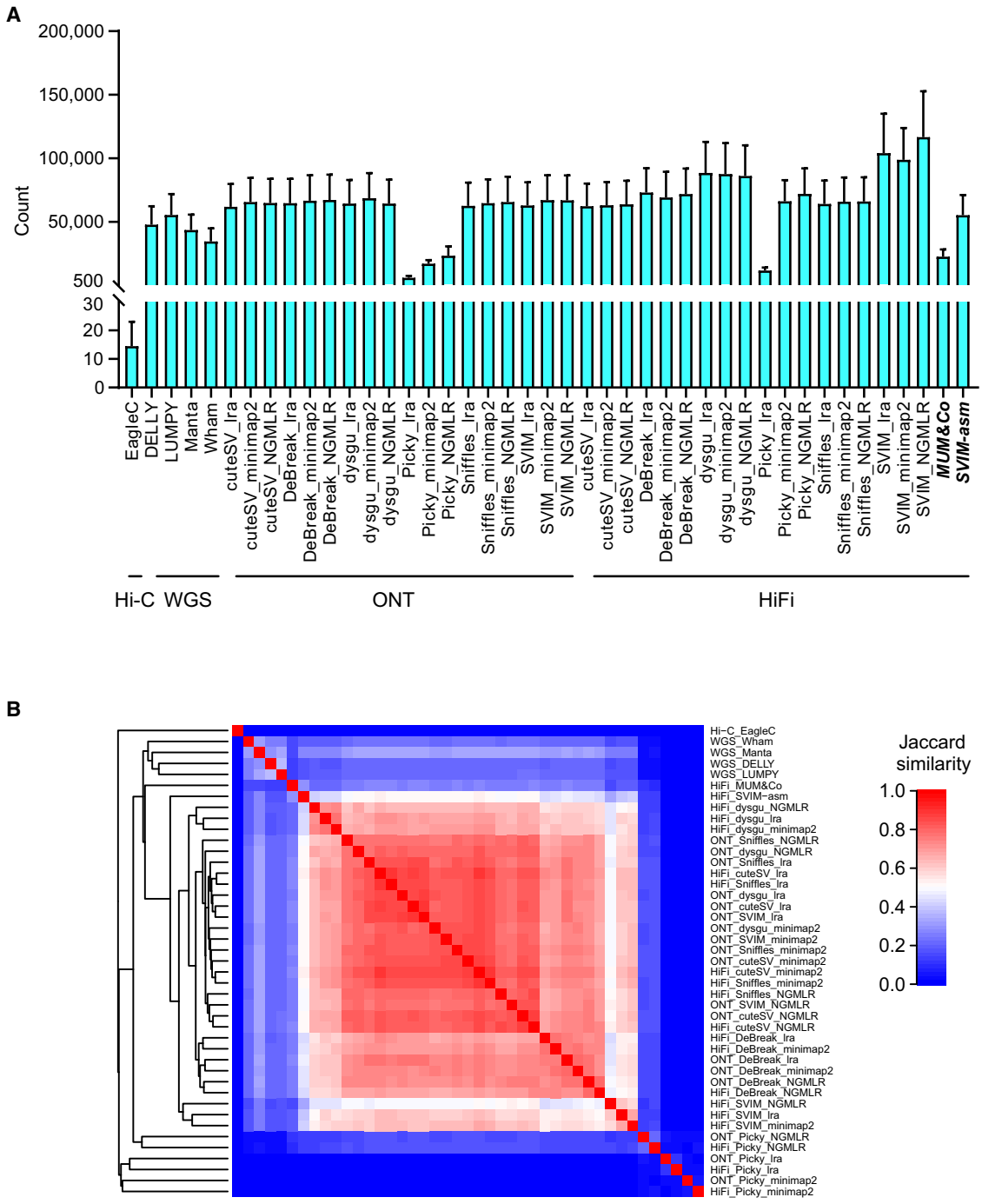


Figure 1. Characterizations of structural variation (SV) callsets using different types of sequencing platforms, reads aligners, and SV callers

(A) the number of SVs in each callset.

(B) the Jaccard similarity between different callsets in Asian wild boar. The error bars in (A) indicate the standard deviation in the five pig breeds. SVs using Hi-C data were exclusively from the EagleC callset. The two assembly-based callers in (A) are marked by bold italic labels. The callsets in (B) were named as "platform_caller_aligner". The average clustering method was used to order the callsets. Hi-C: high-throughput chromosome conformation capture; WGS: Illumina whole genome resequencing; HiFi: Pacbio circular consensus sequencing high-fidelity; ONT: Oxford Nanopore technology; DEL: deletion; INS: insertion; INV: inversion; DUP: duplication; TRA: translocation.

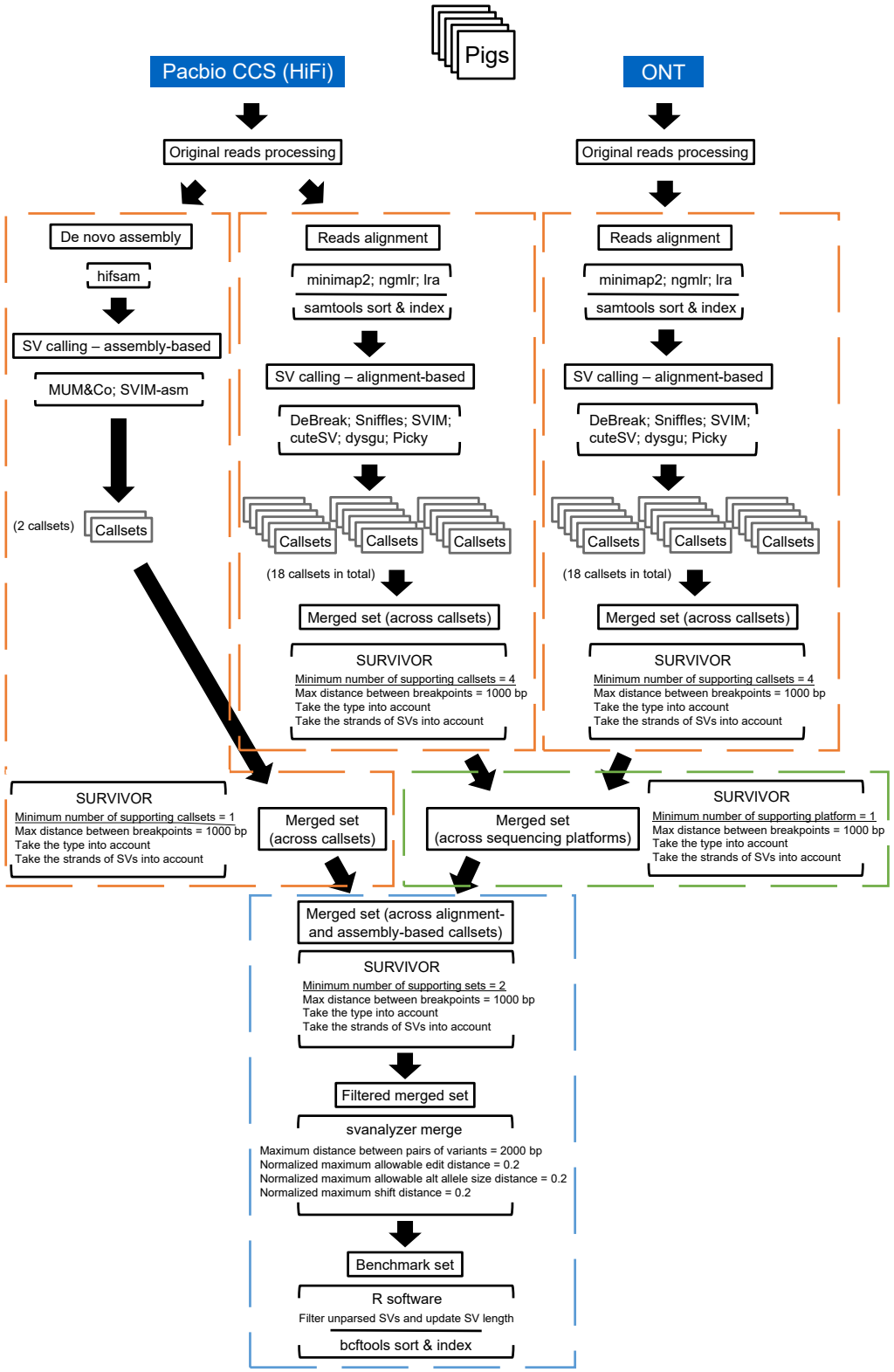


Figure 2. Workflow of structural variation benchmark set compilation applied in each pig breed

The pipeline underwent three-round merging indicated in sequence by orange, green, and blue dash boxes. HiFi: PacBio circular consensus sequencing high-fidelity; ONT: Oxford Nanopore technology.

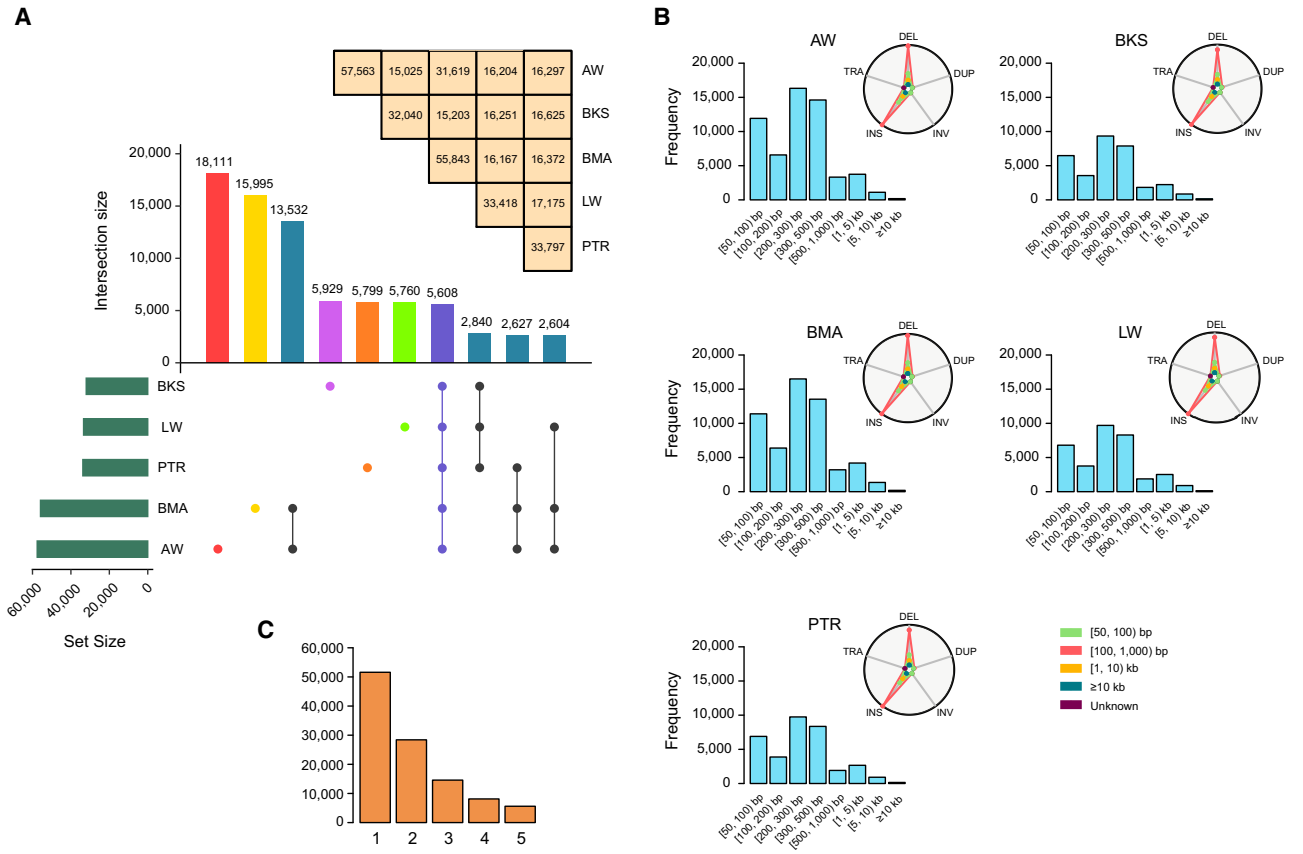


Figure 3. Constitution of structural variation (SV) benchmark sets of five pig breeds

(A) an upset plot indicating the number of specific SVs in each benchmark set and overlapped SVs exclusively present in some benchmark sets, and an upper triangle matrix showing the number of SVs in each benchmark set and overlapped SVs in benchmark set pairs.

(B) a histogram and pie chart showing the count distribution of SVs of different lengths and types.

(C) a histogram indicating the count distribution of SVs supported by varying number of benchmark sets. AW: Asian wild boar; BKS: Berkshire; BMA: Bama Xiang; LW: Large White; PTR: Piétrain; DEL: deletion; INS: insertion; INV: inversion; DUP: duplication; TRA: translocation.

between reference genome and mapped reads) supported by 34 reads and one insertion residing in chromosome 13 of AW, BKX, BMA, and LW with a size of 537 bp endorsed by eight reads, were displayed as representatives (Figure S5). Although we achieved a relatively high validation rate, these results from small sample size might be affected by random and sampling bias. Therefore, a larger scale of PCR experiments is preferred for a more convincing *in vitro* validation.

Global evaluation of SV detection performance based on different platforms, aligners, and callers

SVs detected based on Hi-C data were not included in the subsequent evaluation of detection performance since these SV sets were scarce and all in ultra-large sizes. For the sequencing platforms, the NGS-based callers demonstrated comparable precision to most TGS-based aligner-caller combinations, averaging around 0.6; however, they exhibited inferior performance in terms of recall. The gap in recall between NGS- and TGS-based tools was remarkable for insertion detection, which was above 0.9. The overall calling performances of aligner-caller combinations under the two TGS platforms were highly similar in terms of precision, recall, and F1 score. It is noteworthy that the Picky tool consistently performed poorly on both TGS platforms especially running with lra.[43] The precision values of dysgu[44] and SVIM[40] capitalizing on HiFi data were slightly but visibly lower than those based on ONT reads. Concerning the algorithm difference between alignment- and assembly-based callers, the assembly-based caller SVIM-asm[45] was overall superior on the calling performance. Another assembly-based caller, MUM&Co,[46] stood out with markedly higher precision than recall, unlike almost all other programs. In addition, the WGS platform proved more effective in discovering deletions compared to insertions, while the two TGS technologies showed minor detection discrepancies between the two SV types (Figure 4).

To further support our results, we benchmarked the TGS-based aligner-caller combinations using two simulated datasets with HiFi- and ONT-like synthetic reads, respectively. As shown, the calling programs performed similarly in terms of precision, recall, and F1 score as that observed in the real-data-based benchmark experiment. Notably, we found that the precision scores of various aligner-caller combinations were comprehensively higher than

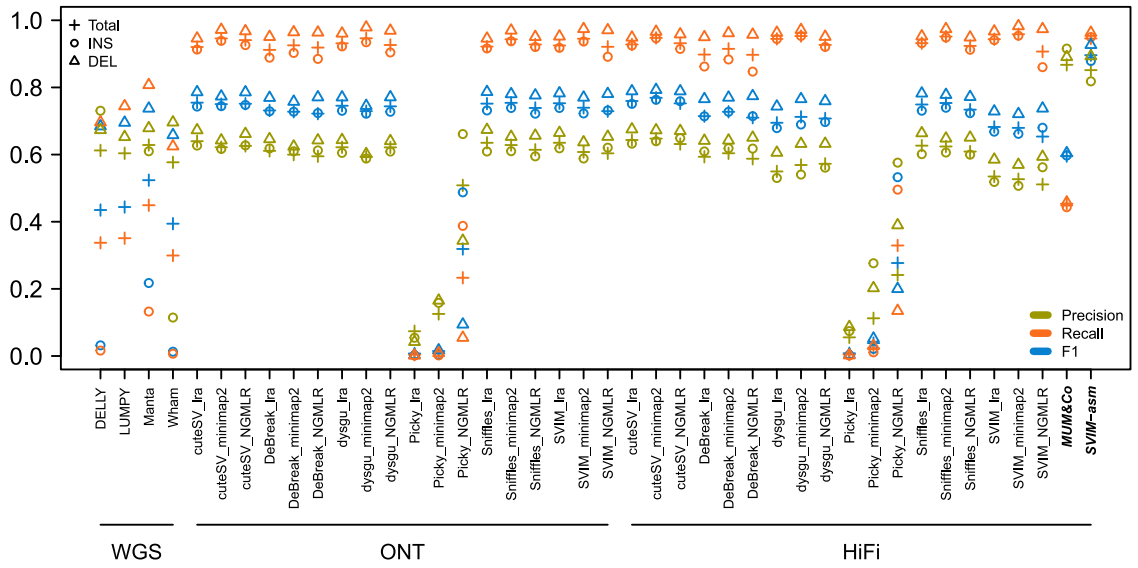


Figure 4. Precision, recall, and F1 score of structural variation (SV) detection for all SV calling programs capitalizing on three different sequencing platforms

Each point is the average value of five pig breeds. The two assembly-based callers are marked by bold italic labels. Total SVs (Total), insertions (INS), and deletions (DEL) were separately assessed. The LUMPY tool cannot detect INS. WGS: Illumina whole genome resequencing; HiFi: Pacbio circular consensus sequencing high-fidelity; ONT: Oxford Nanopore technology.

recalls in insertion detection, which is opposite to the real-data-based observation, implying the complexity of real genomes. Picky also performed inferiorly to other callers as observed in real-data experiment, but had a remarkably large improvement on precision when cooperating with minimap2[47] and NGMLR[35] aligners (Figures 4 and S6).

Impacts of number of supporting reads and length on SV detection

The number of supporting reads and SV length are key factors impacting the power and effectiveness of SV detection. For the number of supporting reads using HiFi data, the precision of total SVs detection increased overall as the number of supporting reads rose for all the aligner-caller combinations expect for Picky working with minimap2 and NGMLR (Figure 5A). Beside Picky, other callers achieved a high precision around 0.9 for SVs supported by more than 50 reads except for dysgu with a slightly lower precision of 0.8. Moreover, the dysgu tool performed weakly to other callers in detecting SVs supported by more than four reads. By inspecting the performance of aligner-caller combinations capitalizing on ONT data, surprisingly, we found that Picky cooperating with NGMLR performed well in dealing with SVs (mostly insertions, as shown in Figures S7A and S8A) supported by less than 50 reads, which could be attributed to the limited number of SVs detected with 50 or more supporting reads relative to fewer reads. The outperformance of DeBreak[30] under 5 HiFi supporting reads could also be ascribed to it as more deletions were detected with 5 supporting reads. Concerning the impact of SV length on detection power, these combinations performed overall best for SVs being 100–499 bp long, irrespective of the value or variation of calling performance metrics (Figure 5B). The combinations started to perform variedly when detecting SVs longer than 5 kb. For SVs not less than 10 kb, we found very low detection power for all aligner-caller combinations, hinting that future studies should be more cautious in explaining this kind of large genomic structural aberrations. Of note, the assembly-based tool SVIM-asm consistently exhibited robust performance for SVs less than 10 kb (Figures 5B, S7B, and S8B).

Effectiveness of SV detection in complex genomic regions

Detecting SVs in complex genomic regions is challenging yet valuable. In our study, overall, the precision and recall of detecting SVs in genomic regions of LINE and SINE was close to those outside interspersed repeats and low complexity sequences regions and notably higher (11.1% for total SVs, 9.6% for insertions, and 9.7% for deletions) than that in simple repeat regions, regardless of the platforms. Concerning the sequencing platforms, the difference between NGS-based and TGS-based programs was mainly on recall. Within the two TGS platforms, the ONT- and HiFi-based programs performed similarly expect for ONT-based dysgu and SVIM displaying superiority on precision (basically due to the superiority of insertion detection) over their HiFi counterparts. With respect to GC contents, the precisions of total SVs detection in low GC regions were conspicuously higher (9.8% for total SVs, 8.6% for insertions, and 14.7% for deletions) than those in high GC regions except for 18th chromosome. Comparatively, no distinct difference was observed between the two regions on recall. The difference on precision between ONT- and HiFi-based programs were also from dysgu and SVIM callsets. When applying our analyses to deletion and insertion separately, the gap of precision between low and high GC expanded to 20% for both insertion and deletion detections on

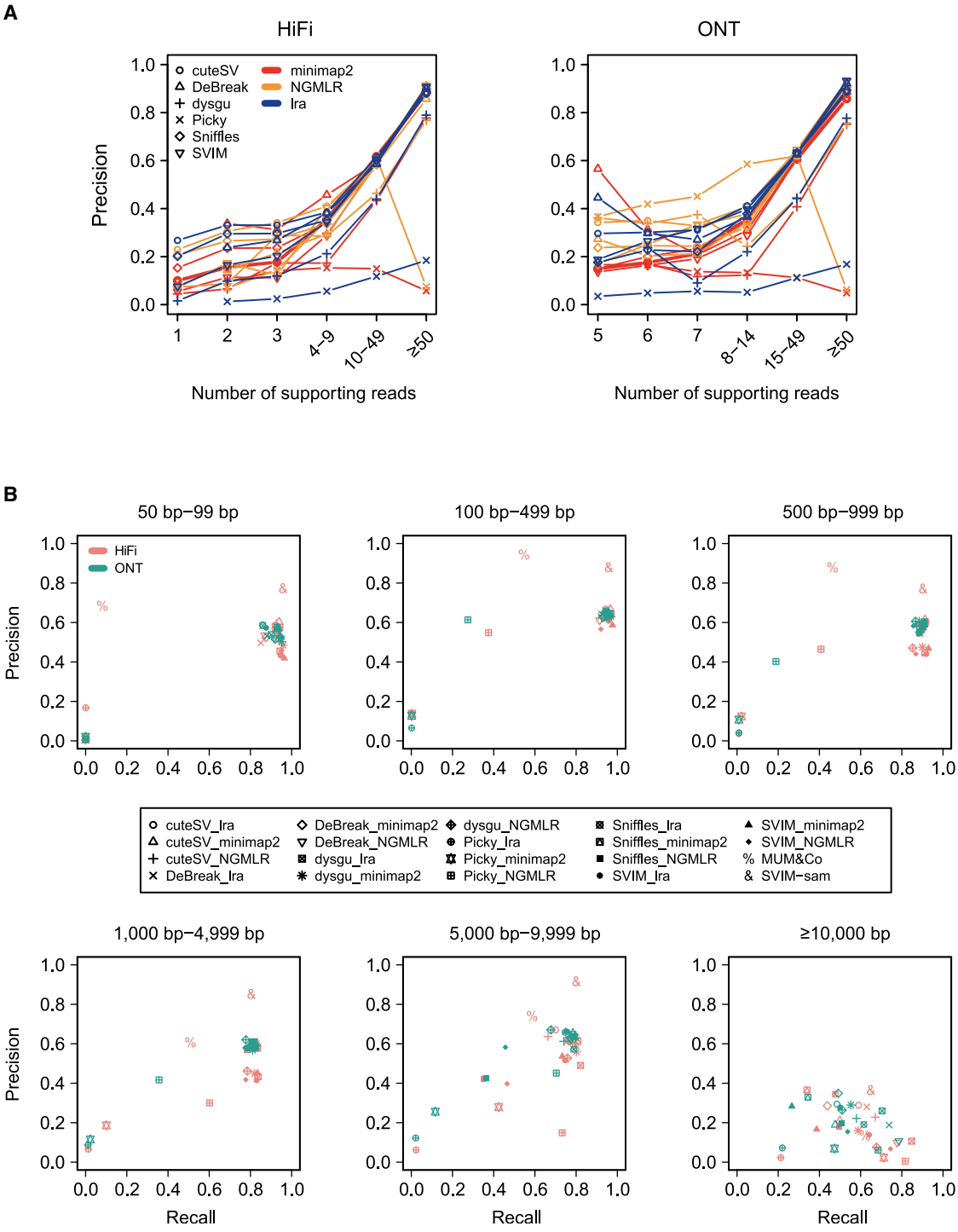


Figure 5. Impact of number of supporting reads and structural variation size on the precision and recall of total structural variations detection based on two third-generation sequencing platforms

(A) The number of supporting reads.

(B) The structural variation size.

Each point is the average value of five pig breeds. The caller Picky performed poorly on both precision and recall in (B) under any scenario. HiFi: Pacbio circular consensus sequencing high-fidelity; ONT: Oxford Nanopore technology.

chromosome X. Also, we found that deletions from the NGS-used callers showed higher recall scores than insertions (Figures 6A and S9–S19).

Compared to human beings, farm animals have undergone a long history of domestication, resulting in a small effective population size and high LD extent.[32,48] In this sense, we studied the impacts of LD and runs of homozygosity (ROH) on SV detection. Detecting SVs in genomic regions with different LD patterns showed similar calling performance irrespective of assessing all SVs, insertions, or deletions (Figures 6B, S20, and S21). The difference between NGS-based and TGS-based programs primarily lay in recall, which is prominent for insertion detection but modest for deletion identification. The ONT-based dysgu and SVIM also presented a superiority over their HiFi-based counterparts in precision, especially for insertion. The aligner-caller combinations using either HiFi or ONT performed more uniformly in detecting deletions relative to insertions. ROH are contiguous blocks of homozygous genotypes and are widely used as predictors of inbreeding levels in farm animals, which are related to complex traits of economic importance.[49] Overall, most aligner-caller combinations could achieve a higher precision value on SVs in ROH rather than other regions, expect for HiFi-based dysgu and SVIM tools feeble on insertion detection. The NGS-based programs performed conspicuously superior on recall values in detecting insertions relative to deletions (Figures 6C and S22).

The impact of sequencing depths on SV calling

In general, reducing sequencing depth entails a potential drop in the accuracy of SV detection but decreases sequencing costs, which incents us to explore a trade-off between sequencing depth and detection accuracy. Inspecting the number of identified SVs under different sequencing depths, most TGS-based aligner-caller combinations held a strict increasing trend as depth deepened, disregarding the SV type (Figures 7A, S23, and S24). The exceptions were the SVIM and Picky tools based on ONT data, which maximized the detection number at 20 × depth irrespective of the aligner combined with. Picky, when working with minimap2 or lra, could only detect a markedly small number of SVs, which could account for its poor performance metrics on SV detection.

Regarding the trend in SV calling performance metrics at different sequencing depths, the F1 and recall values basically increased as the sequencing depth rose but rapidly reached the plateau, disregarding the aligners, callers, and SV types (Figures 7B and S25–S29). Among the callers, cuteSV[50] and dysgu tools capitalizing on HiFi data, and Sniffles[35] and SVIM tools based on either TGS platform plateaued at 5 × depth. As compared, DeBreak necessitated a 10 × depth data, and Picky demanded 20 × when working with NGMLR. The slightly downward tendency of F1 scores for the callers at high depths of HiFi data were attributed to the precision values with an apparent decreasing trend as the sequencing depth rose (Figuress S30–S32). Such trend could owe to the strict upgoing trend in the number of identified SVs as the sequencing depth of HiFi grew (Figure 7A), resulting in many unique SVs for each callset. These unique SVs were excluded during the benchmarking process and labeled as false positives as we adopted an intersection-based benchmarking strategy. We noticed that there was a conspicuous difference in precision between HiFi- and ONT-based detections for cuteSV, DeBreak, Sniffles, and SVIM tools when the sequencing depth was lower than 5× (Figures S30–S32). Overall, the callers could reach their maximal detection potential at a low depth of 5 × to 10 × .

Computational load of aligners and callers

The computational burden, including run time and maximum memory usage, is another critical component in evaluating the running performance of SV detection pipelines. In this sense, we further compared the computational costs of all aligner-caller combinations based on the calling performance under 10 × sequencing depth which is cost-efficient in SV detection. In addition, the two assembly-based callers were also included to show the difference from the alignment-based methods. The alignment based on either HiFi or ONT reads was overall more time- and memory-consuming than SV calling for most callers except DeBreak (Figure 8), which could account for over 99% of computation requirements at most. The DeBreak caller was exceptional in terms of both time and peak memory cost due to the reliance on a local *de novo* assembly and a partial order alignment in the algorithm.[30] The run time burden of Debreak blew up when cooperating with lra based on ONT reads. Its memory load was conspicuously heavier than other callers, disregarding the aligner combining with. Regarding the two assembly-based callers, SVIM-asm was more efficient than MUM&Co on both run time and maximal memory cost. Among the aligners, NGMLR was more time-consuming but saving on peak memory occupation.

DISCUSSION

The long history of artificial selection and domestication incurs strong imprints of human interventions in the genome of farm animals, such as long segments of ROH or high LD extent. Therefore, the SV calling pipelines tailored to the human genomes are not necessarily adapted to farm animal genomes. In this consideration, we preliminarily strived to evaluate various SV calling-relevant programs based on different sequencing platforms in dealing with pig genomes. We attempted to disclose an efficient SV calling pipeline, including the choice of sequencing platforms and depths to call SVs in pigs and other farm animals with genomic similarity. Compared to previous studies,[51–54] our work not only provides the SV benchmark set for the subsequent relevant studies but also enables a more systematic understanding of SV detection performances across various conditions. An additional merged set from the benchmark sets of the five breeds we used could be a key component of a complete pig SV reference panel, like the gnomAD-SV[55] in humans, to serve SV imputation,[56] SV-based GWAS,[18,57] and missing heritability deciphering[57] if phenotypes were available, which is consistent with the promise of the international cooperative projects like FAANG[19] and FarmGTEx.[20,21]

SVs in complex genomic regions or of complex structure are hard to decipher but of great importance.[35,58] The advantage of long reads over short reads in handling complex regions can be demonstrated in Figures 6A and S9–S11, in which the recall values of NGS-based callers when detecting SVs in repeat and high GC regions became lower than that in the global evaluation

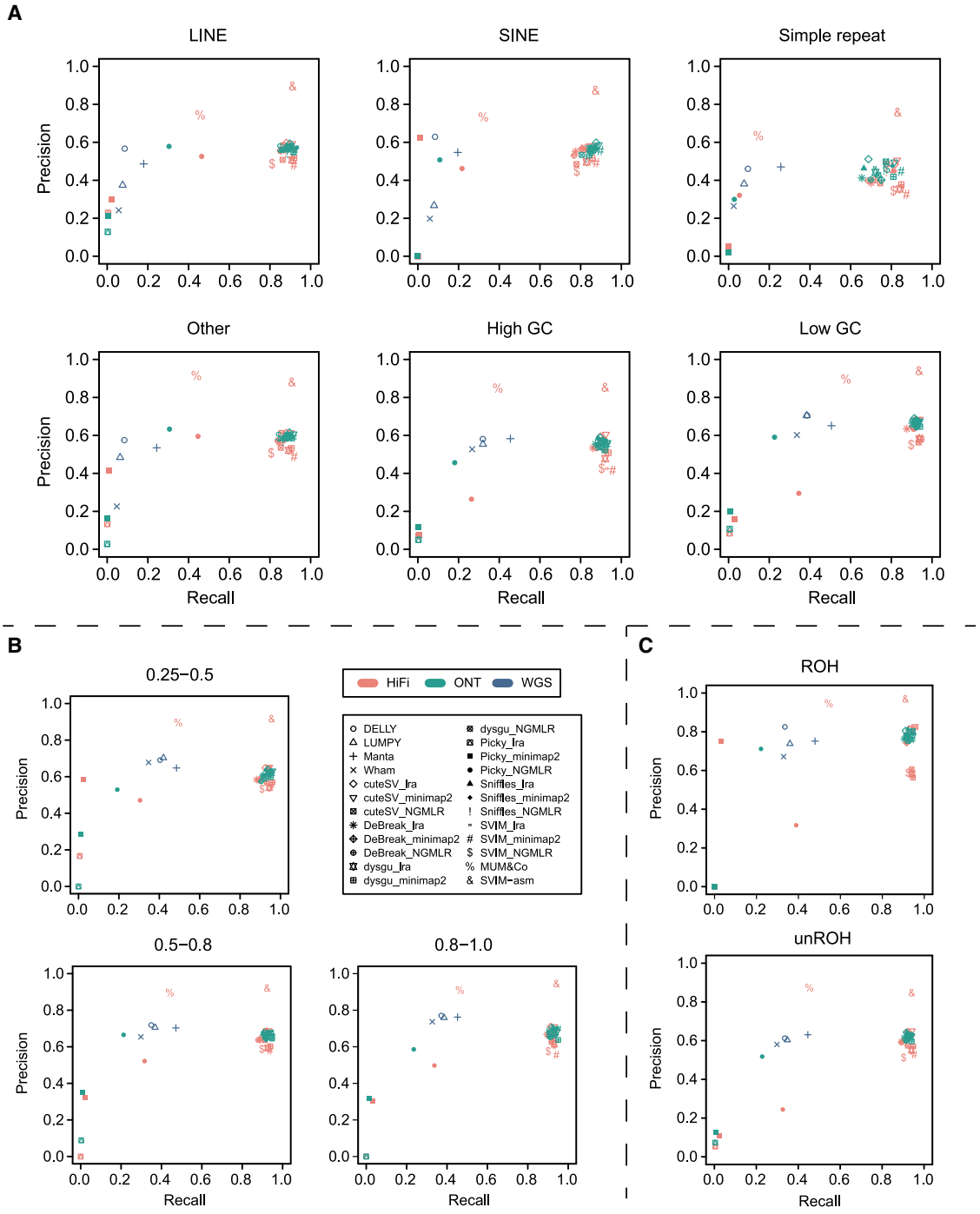


Figure 6. Precision and recall of total structural variations detection in regions with different complexities, regions with varying linkage disequilibrium (LD) extents, and regions with runs of homozygosity (ROH) segments or not (unROH)

(A) Regions with different complexities.

(B) Regions with varying LD extents.

(C) Regions with or without ROH.

Each point is the average value of five pig breeds. Complex regions include long (LINE) and short (SINE) interspersed nuclear elements, simple repeats, and high and low GC contents regions on chromosome 1. Other indicates regions outside interspersed repeats and low complexity DNA sequences areas on chromosome 1. LD extent is assessed by r^2. WGS: Illumina whole genome resequencing; HiFi: PacBio circular consensus sequencing high-fidelity; ONT: Oxford Nanopore technology.

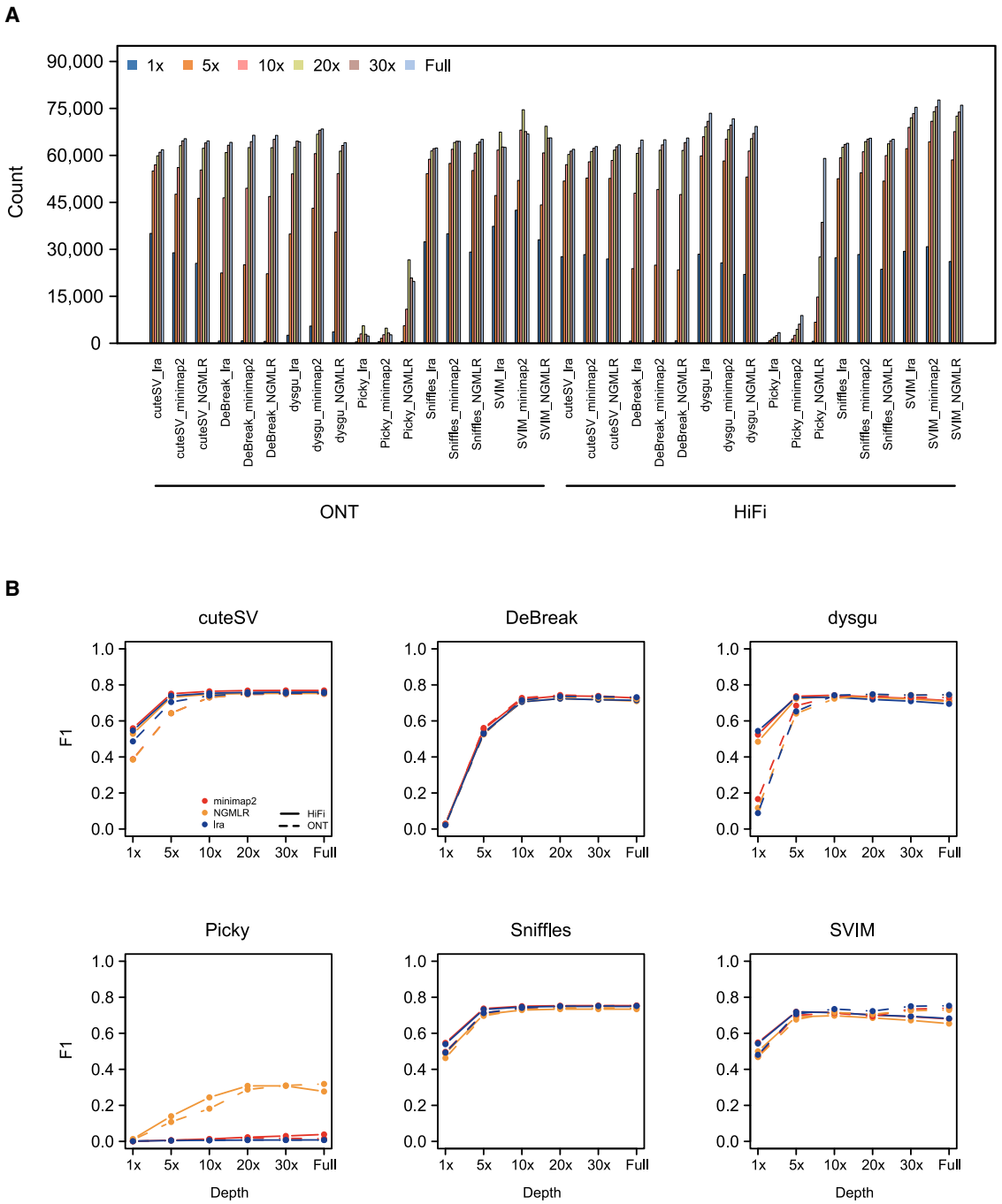


Figure 7. Impact of different sequencing depths on total structural variations (SVs) detection

The (A) number and (B) F1 score of SVs discovered based on different aligner-caller combinations under varying sequencing depths. Each point and bar are the average value of five pig breeds. HiFi: PacBio circular consensus sequencing high-fidelity; ONT: Oxford Nanopore technology.

(Figure 4). SVs in repeat regions tend to be longer and more complex due to the presence of base repeats.[59] Long reads can enable a more accurate alignment and a more complete capture of SV alleles in contrast to short reads, necessitating an additional step of stitching short reads together to form longer segments, which may introduce extra bias.[60] This could also account for the superiority of assembly-based caller SVIM-asm when dealing with SVs in complex regions (Figures 6A and S9–S19), since the *de novo* assembly was pre-implemented to construct the longer genomic segments, i.e., contigs.[23,24,45] The superior performance of ONT-based dysgu and SVIM in detecting SVs within LINE regions, relative to their HiFi-based

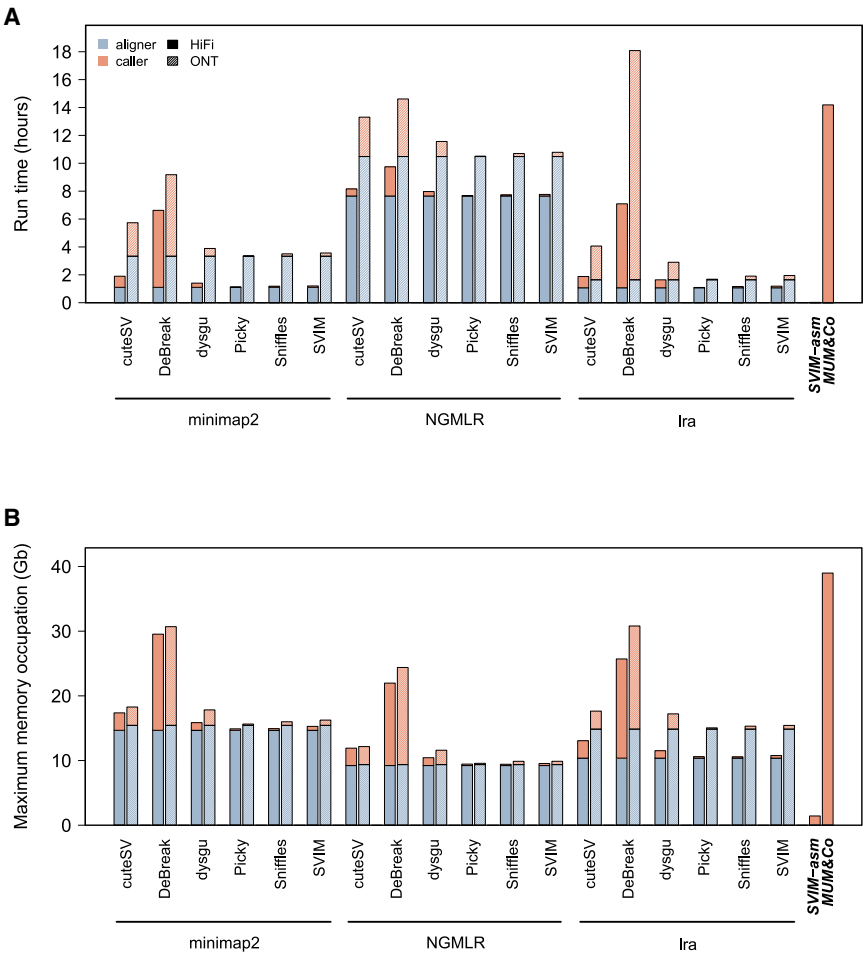


Figure 8. Computational load including run time and maximum memory occupation of each aligner-caller combination based on 10 × third-generation sequencing reads

(A) Run time.

(B) Maximum memory occupation.

Two assembly-based callers were implemented on the assembled genomes from the total HiFi reads. Each bar is the average value of five pig breeds. The two assembly-based callers are marked by bold italic labels. HiFi: PacBio circular consensus sequencing high-fidelity; ONT: Oxford Nanopore Technologies.

counterparts, can be attributed to the advantage of long reads in tackling long repeat fragments, as the overall length of ONT reads was longer than HiFi reads (Table S1). Concerning the two genomic features that farm animals differ from humans, i.e., high LD extent and long ROH, LD extents of genomic regions presented no apparent impact on SV detection performance (Figures 6B, S19, and S20). It implies the ease of choosing SV detection programs for farm animals, as different algorithms tend to result alike when dealing with SVs in high LD regions. For ROH segments, the TGS-based programs especially when capitalizing on ONT data displayed a conspicuous advantage on the precision of SV detection relative to other genomic regions, implying a more precise SV identification in these contiguous homozygous stretches, which are common regions mainly due to inbreeding in pigs and other farm animals (Figures 6C and S22).

Sequencing depth is always a pivotal concern in practice. It is essential to find an optimized trade-off between the SV detection effectiveness and sequencing cost.[61] As observed in our study, for alignment-based callers, there was an evident plateau of F1 scores as sequencing deepened. Precisions even dropped as sequencing depth increased because the number of false positives rocketed up (Figures 7 and S25–S29). Therefore, a relatively low depth of TGS is practicable and cost-effective for alignment-based SV detection. For the four callers, i.e., cuteSV, dysgu, Sniffles, and SVIM, able to perform similarly well when inspecting the F1 score at 5 × depth (Figures S30–S32), the difference in computational resource consumptions among them was not considerable and indeed tolerable. As compared for the aligners, lra and NGMLR were the most time- and memory-saving tools, respectively, but minimap2 was not stressful on both aspects. Therefore, we suggested choosing one of the four callers when using HiFi data and one of Sniffles and SVIM when ONT reads are available. Which aligner should be used depending on whether run time or maximal memory cost is a major concern. Importantly, if detection effectiveness was exclusively in pursuit without consideration of sequencing cost, the assembly-based caller SVIM-asm is absolutely recommended due to its outperformance on both calling performance and computational resource consumption. Especially for complex SVs or SVs in complex genomic regions, the pre-constructed contigs used by assembly-based callers could enhance the alignment and thereupon improve the resolution to identify large SVs.[24,45] In short, the alignment-based strategy is suggested for an at-scale application with a low third-generation sequencing depth, even down to 5 × . The assembly-based scheme is preferred for the ultimate deciphering SVs with a limited number of deeply sequenced individuals.

To sum up, we comprehensively evaluate various SV calling-relevant programs on handling SVs of varying characteristics in pigs using long and short reads from different platforms based on benchmark sets compiled from real data. The detection capacity of these programs was reinforced by a simulation with synthetic reads and SV truth set specifically adapted to pig genomes. The optimal pipeline for at-scale SV calling should be cost-effective and the alignment-based pipeline incorporating cuteSV, dysgu, Sniffles, or SVIM as the caller with any aligner from minimap2, NGMLR, and lra based on 5 × HiFi reads, or either Sniffles or SVIM tool with the availability of ONT data is

suggested. Otherwise, to achieve the best calling performance without cost concerns, the assembly-based caller SVIM-asm with in-depth sequenced genomes should be implemented.

Limitations of this study

Despite the significant findings of our work, several limitations must be noted. First, our study executed the benchmarking based on an intersection approach, which is relatively conservative. Any particular kind of SVs or genomic regions that can be well-handled by a specific caller or aligner-caller combination is possibly excluded from the benchmark set after intersecting. Using more advanced ensemble strategies, such as machine learning or deep learning, might be superior to the heuristic intersection workflow we adopted and help to enlarge the number of SVs in the benchmark set, which is worth to be experimented.[28] Second, more breeds and larger sample sizes, probably including parent-child trios, are needed to enable a more comprehensive *in silico* SV detection and a more convincing trio-binning-based *in vitro* validation. Parent-child trios can further enable the haplotype-resolved genome assembly.[62] Thereupon, the genotype of SVs detected by assembly-based callers could be phased. It is anticipated that a more robust and convincing SV benchmarking study in the future should have improvements on these two aspects.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Guoqiang Yi (yiguoqiang@caas.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The raw sequencing data processed by this study was deposited in the publicly accessible database Sequence Read Archive (SRA) with accession numbers PRJNA1091034 and PRJNA754250.
- The original codes for data processing and the benchmark sets compiled in each pig breed are available at https://github.com/Geil625/Detecting-structural-variations-in-pigs.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2021YFF1000600), the National Natural Science Foundation of China (32002150), the Agricultural Science and Technology Innovation Program of the Chinese Academy of Agricultural Sciences (CAAS-NBSCA-202301 and CAAS-ZDRW202406), the Basic and Applied Basic Research Foundation of Guangdong Province (2020B1515120053), the Shenzhen Science and Technology Innovation Commission (JCYJ20190813114401691), and Bama County Program for Talents in Science and Technology (barenke20210010). We would like to express our great gratitude to the editor and anonymous reviewers for the constructive and meaningful discussion on the conceptualization and writing of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: S.H. and G.Y.; data curation: S.H. and Q.B.; investigation: S.H.; formal analysis: S.H.; validation: B.S. and Y.T.; resources: B.S., Y.T., X.Q., X.L., and X.Y.; writing – original draft: S.H.; writing – review and editing: S.H., G.Y., L.F., J.J., and Z.T.; supervision: G.Y. and Z.T.; project administration: G.Y. and Z.T.; funding acquisition: G.Y. and Z.T.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Next- and third-generation sequencing platforms
  - SV detection based on illumina whole genome resequencing
  - SV detection based on high-throughput chromosome conformation capture technology
  - SV detection based on pacbio circular consensus sequencing and Oxford Nanopore technologies
  - Compilation of structural variation benchmark set
  - Systematic comparison on performance metrics of SV detection in different platforms
  - SV simulation
  - PCR validation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.111983.

Received: July 31, 2024
Revised: December 1, 2024
Accepted: February 5, 2025
Published: February 8, 2025

REFERENCES

1. Jiang, Y.F., Wang, S., Wang, C.L., Xu, R.H., Wang, W.W., Jiang, Y., Wang, M.S., Jiang, L., Dai, L.H., Wang, J.R., et al. (2023). Pangenome obtained by long-read sequencing of 11 genomes reveal hidden functional structural variants in pigs. iScience *26*, 106119. https://doi.org/10.1016/j.isci.2023.106119.
2. Zhang, L., Huang, Y., Wang, M., Guo, Y., Liang, J., Yang, X., Qi, W., Wu, Y., Si, J., Zhu, S., et al. (2019). Development and genome sequencing of a laboratory-inbred miniature pig facilitates study of human diabetic disease. iScience *19*, 162–176. https://doi.org/10.1016/j.isci.2019.07.025.
3. Andersson, L. (2016). Domestic animals as models for biomedical research. Upsala J. Med. Sci. *121*, 1–11. https://doi.org/10.3109/03009734.2015.1091522.
4. Génin, E. (2020). Missing heritability of complex diseases: case solved? Hum. Genet. *139*, 103–113. https://doi.org/10.1007/s00439-019-02034-4.
5. Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. Nature *606*, 527–534. https://doi.org/10.1038/s41586-022-04808-9.
6. Li, Y., Roberts, N.D., Wala, J.A., Shapira, O., Schumacher, S.E., Kumar, K., Khurana, E., Waszak, S., Korbel, J.O., Haber, J.E., et al. (2020). Patterns of somatic structural variation in human cancer genomes. Nature *578*, 112–121. https://doi.org/10.1038/s41586-019-1913-9.

7. Shale, C., Cameron, D.L., Baber, J., Wong, M., Cowley, M.J., Papenfuss, A.T., Cuppen, E., and Priestley, P. (2022). Unscrambling cancer genomes via integrated analysis of structural variation and copy number. Cell Genom. *2*, 100112. https://doi.org/10.1016/j.xgen.2022.100112.

8. Rees, E., and Kirov, G. (2021). Copy number variation and neuropsychiatric illness. Curr. Opin. Genet. Dev. *68*, 57–63. https://doi.org/10.1016/j.gde.2021.02.014.

9. Glessner, J.T., Li, J., Wang, D., March, M., Lima, L., Desai, A., Hadley, D., Kao, C., Gur, R.E., Cohen, N., et al. (2017). Copy number variation meta-analysis reveals a novel duplication at 9p24 associated with multiple neurodevelopmental disorders. Genome Med. *9*, 106. https://doi.org/10.1186/s13073-017-0494-1.

10. Gill, R.A., Scossa, F., King, G.J., Golicz, A.A., Tong, C., Snowdon, R.J., Fernie, A.R., and Liu, S. (2021). On the role of transposable elements in the regulation of gene expression and subgenomic interactions in crop genomes. CRC Crit. Rev. Plant Sci. *40*, 157–189. https://doi.org/10.1080/07352689.2021.1920731.

11. Gabur, I., Chawla, H.S., Snowdon, R.J., and Parkin, I.A.P. (2019). Connecting genome structural variation with complex traits in crop plants. Theor. Appl. Genet. *132*, 733–750. https://doi.org/10.1007/s00122-018-3233-0.

12. Xia, X., Zhang, F., Li, S., Luo, X., Peng, L., Dong, Z., Pausch, H., Leonard, A.S., Crysnanto, D., Wang, S., et al. (2023). Structural variation and introgression from wild populations in East Asian cattle genomes confer adaptation to local environment. Genome Biol. *24*, 211. https://doi.org/10.1186/s13059-023-03052-2.

13. Yang, L., Yin, H., Bai, L., Yao, W., Tao, T., Zhao, Q., Gao, Y., Teng, J., Xu, Z., Lin, Q., et al. (2024). Mapping and functional characterization of structural variation in 1060 pig genomes. Genome Biol. *25*, 116. https://doi.org/10.1186/s13059-024-03253-3.

14. Yang, R., Fang, S., Wang, J., Zhang, C., Zhang, R., Liu, D., Zhao, Y., Hu, X., and Li, N. (2017). Genome-wide analysis of structural variants reveals genetic differences in Chinese pigs. PLoS One *12*, e0186721. https://doi.org/10.1371/journal.pone.0186721.

15. Li, Z., Liu, X., Wang, C., Li, Z., Jiang, B., Zhang, R., Tong, L., Qu, Y., He, S., Chen, H., et al. (2023). The pig pangenome provides insights into the roles of coding structural variations in genetic diversity and adaptation. Genome Res. *33*, 1833–1847. https://doi.org/10.1101/gr.277638.122.

16. Du, H., Zheng, X., Zhao, Q., Hu, Z., Wang, H., Zhou, L., and Liu, J.-F. (2021). Analysis of Structural Variants Reveal Novel Selective Regions in the Genome of Meishan Pigs by Whole Genome Sequencing. Front. Genet. *12*, 550676. https://doi.org/10.3389/fgene.2021.550676.

17. Blaj, I., Tetens, J., Bennewitz, J., Thaller, G., and Falker-Gieske, C. (2022). Structural variants and tandem repeats in the founder individuals of four F2 pig crosses and implications to F2 GWAS results. BMC Genom. *23*, 631. https://doi.org/10.1186/s12864-022-08716-0.

18. Zong, W., Wang, J., Zhao, R., Niu, N., Su, Y., Hu, Z., Liu, X., Hou, X., Wang, L., Wang, L., and Zhang, L. (2023). Associations of genome-wide structural variations with phenotypic differences in cross-bred Eurasian pigs. J. Anim. Sci. Biotechnol. *14*, 136. https://doi.org/10.1186/s40104-023-00929-x.

19. Andersson, L., Archibald, A.L., Bottema, C.D., Brauning, R., Burgess, S.C., Burt, D.W., Casas, E., Cheng, H.H., Clarke, L., Couldrey, C., et al. (2015). Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. Genome Biol. *16*, 57. https://doi.org/10.1186/s13059-015-0622-4.

20. Liu, S., Gao, Y., Canela-Xandri, O., Wang, S., Yu, Y., Cai, W., Li, B., Xiang, R., Chamberlain, A.J., Pairo-Castineira, E., et al. (2022). A multi-tissue atlas of regulatory variants in cattle. Nat. Genet. *54*, 1438–1447. https://doi.org/10.1038/s41588-022-01153-5.

21. Teng, J., Gao, Y., Yin, H., Bai, Z., Liu, S., Zeng, H., PigGTEx Consortium, Bai, L., Cai, Z., Zhao, B., et al. (2024). A compendium of genetic regulatory effects across pig tissues. Nat. Genet. *56*, 112–123. https://doi.org/10.1038/s41588-023-01585-7.

22. Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. *20*, 117. https://doi.org/10.1186/s13059-019-1720-5.

23. Liu, Y.H., Luo, C., Golding, S.G., Ioffe, J.B., and Zhou, X.M. (2024). Trade-offs in alignment and assembly-based methods for structural variant detection with long-read sequencing data. Nat. Commun. *15*, 2447. https://doi.org/10.1038/s41467-024-46614-z.

24. Lin, J., Jia, P., Wang, S., Kosters, W., and Ye, K. (2023). Comparison and benchmark of structural variants detected from long read and long-read assembly. Brief. Bioinform. *24*, bbad188. https://doi.org/10.1093/bib/bbad188.

25. Yildiz, G., Zanini, S.F., Afsharyan, N.P., Obermeier, C., Snowdon, R.J., and Golicz, A.A. (2023). Benchmarking Oxford Nanopore read alignment-based insertion and deletion detection in crop plant genomes. Plant Genome *16*, e20314. https://doi.org/10.1002/tpg2.20314.

26. Bolognini, D., and Magi, A. (2021). Evaluation of Germline Structural Variant Calling Methods for Nanopore Sequencing Data. Front. Genet. *12*, 761791. https://doi.org/10.3389/fgene.2021.761791.

27. Liu, Z., Xie, Z., and Li, M. (2024). Comprehensive and deep evaluation of structural variation detection pipelines with third-generation sequencing data. Genome Biol. *25*, 188. https://doi.org/10.1186/s13059-024-03324-5.

28. Cameron, D.L., Di Stefano, L., and Papenfuss, A.T. (2019). Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat. Commun. *10*, 3240. https://doi.org/10.1038/s41467-019-11146-4.

29. Bolognini, D., Sanders, A., Korbel, J.O., Magi, A., Benes, V., and Rausch, T. (2020). VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. Bioinformatics *36*, 1267–1269. https://doi.org/10.1093/bioinformatics/btz719.

30. Chen, Y., Wang, A.Y., Barkley, C.A., Zhang, Y., Zhao, X., Gao, M., Edmonds, M.D., and Chong, Z. (2023). Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak. Nat. Commun. *14*, 283. https://doi.org/10.1038/s41467-023-35996-1.

31. Dierckxsens, N., Li, T., Vermeesch, J.R., and Xie, Z. (2021). A benchmark of structural variation detection by long reads through a realistic simulated model. Genome Biol. *22*, 342. https://doi.org/10.1186/s13059-021-02551-4.

32. Qanbari, S. (2019). On the Extent of Linkage Disequilibrium in the Genome of Farm Animals. Front. Genet. *10*, 1304. https://doi.org/10.3389/fgene.2019.01304.

33. Olson, N.D., Wagner, J., Dwarshuis, N., Miga, K.H., Sedlazeck, F.J., Salit, M., and Zook, J.M. (2023). Variant calling and benchmarking in an era of complete human genome sequences. Nat. Rev. Genet. *24*, 464–483. https://doi.org/10.1038/s41576-023-00590-0.

34. De Roeck, A., De Coster, W., Bossaerts, L., Cacace, R., De Pooter, T., Van Dongen, J., D'Hert, S., De Rijk, P., Strazisar, M., Van Broeckhoven, C., and Sleegers, K. (2019). NanoSatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on PromethION. Genome Biol. *20*, 239. https://doi.org/10.1186/s13059-019-1856-3.

35. Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., and Schatz, M.C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. Nat. Methods *15*, 461–468. https://doi.org/10.1038/s41592-018-0001-7.

36. Hon, T., Mars, K., Young, G., Tsai, Y.-C., Karalius, J.W., Landolin, J.M., Maurer, N., Kudrna, D., Hardigan, M.A., Steiner, C.C., et al. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. Sci. Data *7*, 399. https://doi.org/10.1038/s41597-020-00743-4.

37. Wang, X., Luan, Y., and Yue, F. (2022). EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell

contact maps. Sci. Adv. *8*, eabn9215. https://doi.org/10.1126/sciadv.abn9215.

38. Gong, L., Wong, C.-H., Cheng, W.-C., Tjong, H., Menghi, F., Ngan, C.Y., Liu, E.T., and Wei, C.-L. (2018). Picky comprehensively detects high-resolution structural variants in nanopore long reads. Nat. Methods *15*, 455–460. https://doi.org/10.1038/s41592-018-0002-6.

39. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

40. Heller, D., and Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. Bioinformatics *35*, 2907–2915. https://doi.org/10.1093/bioinformatics/btz041.

41. Fujimoto, A., Wong, J.H., Yoshii, Y., Akiyama, S., Tanaka, A., Yagi, H., Shigemizu, D., Nakagawa, H., Mizokami, M., and Shimada, M. (2021). Whole-genome sequencing with long reads reveals complex structure and origin of structural variation in human genetic variations and somatic mutations in cancer. Genome Med. *13*, 65. https://doi.org/10.1186/s13073-021-00883-1.

42. Becker, T., Lee, W.-P., Leone, J., Zhu, Q., Zhang, C., Liu, S., Sargent, J., Shanker, K., Mil-homens, A., Cerveira, E., et al. (2018). FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. Genome Biol. *19*, 38. https://doi.org/10.1186/s13059-018-1404-6.

43. Ren, J., and Chaisson, M.J.P. (2021). lra: A long read aligner for sequences and contigs. PLoS Comput. Biol. *17*, e1009078. https://doi.org/10.1371/journal.pcbi.1009078.

44. Cleal, K., and Baird, D.M. (2022). Dysgu: efficient structural variant calling using short or long reads. Nucleic Acids Res. *50*, e53. https://doi.org/10.1093/nar/gkac039.

45. Heller, D., and Vingron, M. (2021). SVIM-asm: structural variant detection from haploid and diploid genome assemblies. Bioinformatics *36*, 5519–5521. https://doi.org/10.1093/bioinformatics/btaa1034.

46. O'donnell, S., and Fischer, G. (2020). MUM&Co: accurate detection of all SV types through whole-genome alignment. Bioinformatics *36*, 3242–3243. https://doi.org/10.1093/bioinformatics/btaa115.

47. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

48. Hayes, B.J., Visscher, P.M., McPartlan, H.C., and Goddard, M.E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Res. *13*, 635–643. https://doi.org/10.1101/gr.387103.

49. Sumreddee, P., Toghiani, S., Hay, E.H., Roberts, A., Aggrey, S.E., and Rekaya, R. (2020). Runs of homozygosity and analysis of inbreeding depression. J. Anim. Sci. *98*, skaa361. https://doi.org/10.1093/jas/skaa361.

50. Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., and Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. Genome Biol. *21*, 189. https://doi.org/10.1186/s13059-020-02107-y.

51. Gao, Y., Ma, L., and Liu, G.E. (2022). Initial Analysis of Structural Variation Detections in Cattle Using Long-Read Sequencing Methods. Genes *13*, 828. https://doi.org/10.3390/genes13050828.

52. Zhang, J., Nie, C., Li, X., Zhao, X., Jia, Y., Han, J., Chen, Y., Wang, L., Lv, X., Yang, W., et al. (2022). Comprehensive analysis of structural variants in chickens using PacBio sequencing. Front. Genet. *13*, 971588. https://doi.org/10.3389/fgene.2022.971588.

53. Liu, Y., Zhang, M., Sun, J., Chang, W., Sun, M., Zhang, S., and Wu, J. (2020). Comparison of multiple algorithms to reliably detect structural variants in pears. BMC Genom. *21*, 61. https://doi.org/10.1186/s12864-020-6455-x.

54. Luan, M.-W., Zhang, X.-M., Zhu, Z.-B., Chen, Y., and Xie, S.-Q. (2020). Evaluating structural variation detection tools for long-read sequencing datasets in saccharomyces cerevisiae. Front. Genet. *11*, 159. https://doi.org/10.3389/fgene.2020.00159.

55. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature *581*, 444–451. https://doi.org/10.1038/s41586-020-2287-8.

56. Beyter, D., Ingimundardottir, H., Oddsson, A., Eggertsson, H.P., Bjornsson, E., Jonsson, H., Atlason, B.A., Kristmundsdottir, S., Mehringer, S., Hardarson, M.T., et al. (2021). Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nat. Genet. *53*, 779–786. https://doi.org/10.1038/s41588-021-00865-4.

57. Chen, L., Pryce, J.E., Hayes, B.J., and Daetwyler, H.D. (2021). Investigating the effect of imputed structural variants from whole-genome sequence on genome-wide association and genomic prediction in dairy cattle. Animals *11*, 541. https://doi.org/10.3390/ani11020541.

58. Zook, J.M., Hansen, N.F., Olson, N.D., Chapman, L., Mullikin, J.C., Xiao, C., Sherry, S., Koren, S., Phillippy, A.M., Boutros, P.C., et al. (2020). A robust benchmark for detection of germline large deletions and insertions. Nat. Biotechnol. *38*, 1347–1355. https://doi.org/10.1038/s41587-020-0538-8.

59. Wagner, J., Olson, N.D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.-C., Gupta, R., Wenger, A.M., Rowell, W.J., et al. (2022). Curated variation benchmarks for challenging medically relevant autosomal genes. Nat. Biotechnol. *40*, 672–680. https://doi.org/10.1038/s41587-021-01158-1.

60. Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., and Sedlazeck, F.J. (2019). Structural variant calling: the long and the short of it. Genome Biol. *20*, 1–14. https://doi.org/10.1186/s13059-019-1828-7.

61. Jiang, T., Liu, S., Cao, S., Liu, Y., Cui, Z., Wang, Y., and Guo, H. (2021). Long-read sequencing settings for efficient structural variation detection based on comprehensive evaluation. BMC Bioinf. *22*, 552. https://doi.org/10.1186/s12859-021-04422-y.

62. Cheng, H., Concepcion, G.T., Feng, X., Zhang, H., and Li, H. (2021). Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. Nat. Methods *18*, 170–175. https://doi.org/10.1038/s41592-020-01056-5.

63. Liu, L., Yi, G., Yao, Y., Liu, Y., Li, J., Yang, Y., Liu, M., Fang, L., Mo, D., Zhang, L., et al. (2024). Multiomics analysis reveals signatures of selection and loci associated with complex traits in pigs. iMeta *3*, e250. https://doi.org/10.1002/imt2.250.

64. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. j. *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.

65. Vasimuddin, M., Misra, S., Li, H., and Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems (IEEE), pp. 314–324.

66. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

67. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

68. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

69. Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339. https://doi.org/10.1093/bioinformatics/bts378.

70. Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84. https://doi.org/10.1186/gb-2014-15-6-r84.

71. Kronenberg, Z.N., Osborne, E.J., Cone, K.R., Kennedy, B.J., Domyan, E.T., Shapiro, M.D., Elde, N.C., and Yandell, M. (2015). Wham: identifying structural variants of biological consequence. PLoS Comput. Biol. *11*, e1004572. https://doi.org/10.1371/journal.pcbi.1004572.

72. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259. https://doi.org/10.1186/s13059-015-0831-x.

73. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

74. Wolff, J., Bhardwaj, V., Nothjunge, S., Richard, G., Renschler, G., Gilsbach, R., Manke, T., Backofen, R., Ramírez, F., and Grüning, B.A. (2018). Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. Nucleic Acids Res. *46*, W11–W16. https://doi.org/10.1093/nar/gky504.

75. Abdennur, N., and Mirny, L.A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics *36*, 311–316. https://doi.org/10.1093/bioinformatics/btz540.

76. Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F.J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat. Commun. *8*, 14061. https://doi.org/10.1038/ncomms14061.

77. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

78. English, A.C., Menon, V.K., Gibbs, R.A., Metcalf, G.A., and Sedlazeck, F.J. (2022). Truvari: refined structural variant comparison preserves allelic diversity. Genome Biol. *23*, 271. https://doi.org/10.1186/s13059-022-02840-6.

79. Smit, A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0. http://www.repeatmasker.org.

80. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

81. Hall, M. (2022). Rasusa: Randomly subsample sequencing reads to a specified coverage. J. Open Source Softw. *7*, 3941. https://doi.org/10.21105/joss.03941.

82. Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinf. *13*, 134. https://doi.org/10.1186/1471-2105-13-134.

83. Dombroski, B.A., Feng, Q., Mathias, S.L., Sassaman, D.M., Scott, A.F., Kazazian, H.H., Jr., and Boeke, J.D. (1994). An *in vivo* assay for the reverse transcriptase of human retrotransposon L1 in Saccharomyces cerevisiae. Mol. Cell Biol. *14*, 4485–4492. https://doi.org/10.1128/mcb.14.7.4485-4492.1994.

84. Elbarbary, R.A., Lucas, B.A., and Maquat, L.E. (2016). Retrotransposons as regulators of gene expression. Science *351*, aac7247. https://doi.org/10.1126/science.aac7247.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw WGS and ONT data	This paper	SRA: PRJNA1091034
Raw Hi-C and HiFi data	Liu et al.[63]	SRA: PRJNA754250
Software and algorithms		
Cutadapt (v4.1)	Martin[64]	https://github.com/marcelm/cutadapt
bwa-mem2 (v2.2.1)	Vasimuddin et al.[65]	https://github.com/bwa-mem2/bwa-mem2
Samtools (v1.6)	Danecek et al.[66]	https://github.com/samtools/samtools
Picard (v2.27.4)	Picard toolkit	http://broadinstitute.github.io/picard/
GATK (v4.2.6.1)	McKenna et al.[67]	https://gatk.broadinstitute.org/hc/en-us
Plink (v1.90)	Purcell et al.[68]	https://www.cog-genomics.org/plink/
Manta (v1.6.0)	Chen et al.[39]	https://github.com/Illumina/manta
DELLY (v1.1.6)	Rausch et al.[69]	https://github.com/dellytools/delly
LUMPY (v0.2.12)	Layer et al.[70]	https://github.com/arq5x/lumpy-sv
Wham (v1.7.0)	Kronenberg et al.[71]	https://github.com/zeeev/wham
HiC-Pro	Servant et al.[72]	https://github.com/nservant/HiC-Pro
Bowtie2	Langmead and Salzberg[73]	https://github.com/BenLangmead/bowtie2
hicConvertFormat	Wolff et al.[74]	https://github.com/deeptools/HiCExplorer/blob/master/hicexplorer/hicConvertFormat.py
cooler	Abdennur and Mirny[75]	https://github.com/open2c/cooler
EagleC	Wang et al.[37]	https://github.com/XiaoTaoWang/EagleC
minimap2 (v2.24-r1122)	Li[47]	https://github.com/lh3/minimap2
NGMLR (v0.2.7)	Sedlazeck et al.[35]	https://github.com/philres/ngmlr
lra (v1.3.4)	Ren and Chaisson[43]	https://github.com/ChaissonLab/LRA
Sniffles (v2.0.7)	Sedlazeck et al.[35]	https://github.com/fritzsedlazeck/Sniffles
cuteSV (v1.0.13)	Jiang et al.[50]	https://github.com/tjiangHIT/cuteSV
Picky (v1.0)	Gong et al.[38]	https://github.com/TheJacksonLaboratory/Picky
SVIM (v1.4.2)	Heller and Vingron[40]	https://github.com/eldariont/svim
dysgu (v1.3.16)	Cleal and Baird[44]	https://github.com/kcleal/dysgu
DeBreak (v1.0.2)	Chen et al.[30]	https://github.com/Maggi-Chen/DeBreak
Hifiasm (v0.19.4)	Cheng et al.[62]	https://github.com/chhylp123/hifiasm
MUM&Co (v3.8)	O'donnell and Fischer[46]	https://github.com/SAMtoBAM/MUMandCo
SVIM-asm (v1.0.3)	Heller and Vingron[45]	https://github.com/eldariont/svim-asm
SURVIVOR (v1.0.7)	Jeffares et al.[76]	https://github.com/fritzsedlazeck/SURVIVOR
SVanalyzer	Hansen	https://svanalyzer.readthedocs.io
bcftools (v1.8)	Danecek et al.[66]	https://github.com/samtools/bcftools
VEP (v112.0)	McLaren et al.[77]	https://github.com/Ensembl/ensembl-vep
Truvari (v4.0.0)	English et al.[78]	https://github.com/ACEnglish/truvari
RepeatMasker	Smit et al.[79]	http://www.repeatmasker.org
bedtools (v2.30.0)	Quinlan and Hall[80]	https://github.com/ryanlayer/bedtools
rasusa (v0.7.1)	Hall[81]	https://github.com/mbhall88/rasusa
Sim-it (v1.3.6)	Dierckxsens et al.[31]	https://github.com/ndierckx/Sim-it
SnapGene (v4.3.6)	SnapGene	https://www.snapgene.com/
Primer-BLAST	Ye et al.[82]	https://www.ncbi.nlm.nih.gov/tools/primer-blast/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We collected blood samples from five animals, representing one Asian wild boar (AW), one Chinese native breed (BMA: Bama Xiang), and three breeds from Europe (LW: Large white, BKS: Berkshire, and PTR: Piétrain). These blood samples were used for DNA extraction and sequencing with two TGS (HiFi and ONT) and two NGS sequencing platforms (WGS and Hi-C). Detailed information regarding the raw DNA reads from the four platforms is shown in Table S1. The genetic diversity of the five breeds was disclosed using SNPs called by short-read data. We confirm that the committee approved the experiments and all experiments conformed to the relevant regulatory standards. The pig individuals used in this study were all females. All the results derived in this study were genome-based and would not be impacted by participants' status such as gender, age/developmental stage, and sampling time.

METHOD DETAILS

Next- and third-generation sequencing platforms

Two next-generation sequencing platforms, Illumina whole genome resequencing and high-throughput chromosome conformation capture (Hi-C),[63] and two third-generation sequencing platforms, Pacbio circular consensus sequencing[63] and Oxford Nanopore technology, were deployed on the five collected pig breeds. The overall sequencing depth for Hi-C was approximately 100 × , and for the other three platforms, it was around 50 × , taking the pig genome size and total dosage of raw reads into account. For ONT data, the reads with a quality score > seven were used. Detailed information regarding the raw DNA reads from the four platforms is shown in Table S1.

SV detection based on illumina whole genome resequencing

Firstly, the adapters of raw paired-end reads from the WGS platform were removed using cutadapt (v4.1).[64] Then the adapter-removed reads were aligned to the pig reference genome Sscrofa11.1 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000003025.6) using bwa-mem2 (v2.2.1).[65] With respect to SV detection, only the reads with a minimum mapping quality of 20 were used. In contrast, no mapping quality control was exerted on reads when calling SNPs as there would be no SNP identified in some chromosomes if the minimum mapping quality of reads was set to 20. The samtools suite (v1.6)[66] was used to sort and index the bam files containing mapped reads. Picard (v2.27.4) (http://broadinstitute.github.io/picard/) module *MarkDuplicates* was used to mark the duplicate reads. GATK (v4.2.6.1)[67] modules *HaplotypeCaller*, *GenomicsDBImport*, and *GenotypeGVCFs* were in sequence used to call SNPs. Only autosomal, fully genotyped, and independent SNPs were used to dissect the genetic relatedness between different pig breeds. Plink (v1.90)[68] with flags *–geno 0* and *–indep-pairwise 50 5 0.1* was applied to identify the completely fingerprinted and independent SNPs. For SV calling, four alignment-based programs, i.e., Manta (v1.6.0),[39] DELLY (v1.1.6),[69] LUMPY (v0.2.12),[70] and Wham (v1.7.0),[71] were implemented. The discordant and split mapped reads required as inputs to LUMPY were obtained respectively using samtools flag *-F 1294* and LUMPY built-in Python script *extractSplitReads_BwaMem*. The minimal number of reads supporting an SV detection was set to 5.

SV detection based on high-throughput chromosome conformation capture technology

HiC-Pro[72] was exclusively used to directly procure the final intra- and inter-chromosomal contact matrices of varying resolutions of 5 kb, 10 kb, and 50 kb using the raw Hi-C reads. Bowtie2[73] was embedded in HiC-Pro as the reads aligner. The reads with a minimum mapping quality of 20 were also exclusively used. The format of contact matrices was transformed from *hicpro* to *cool* using *hicConvertFormat*.[74] The contact matrices of the cool format were further iteratively corrected, and eigenvector decomposition (ICE) was normalized by cooler[75] module *balance*. For SV discovery, an up-to-date program EagleC,[37] capable of identifying a wide spectrum of SV types, was implemented based on the 5 kb-, 10 kb-, and 50 kb-resolution contact matrices derived from HiC-Pro.

SV detection based on pacbio circular consensus sequencing and Oxford Nanopore technologies

Three read alignment programs, i.e., minimap2 (v2.24-r1122),[47] NGMLR (v0.2.7),[35] and lra (v1.3.4),[43] were used to align the raw reads from HiFi and ONT to the Sscrofa11.1 pig reference genome. In minimap2, the flag *-x* was set to *map-hifi* and *map-ont*, respectively, for HiFi and ONT reads to build the index and then map. In NGMLR, the similar flag *-x* was set to *pacbio* and *ont*. In lra, the corresponding flags are *-CCS* and *-ONT* for HiFi and ONT reads for both index building and alignment. The mapped reads were sorted and indexed by samtools (v1.6).[66] For SV calling, six alignment-based callers, namely Sniffles (v2.0.7),[35] cuteSV (v1.0.13),[50] Picky (v1.0),[38] SVIM (v1.4.2),[40] dysgu (v1.3.16),[44] and DeBreak (v1.0.2)[30] were implemented. The flags in each caller indicating minimal mapping quality and SV size were set to 20 and 50 bp, respectively, disregarding HiFi or ONT reads. The minimal number of supporting reads for a detected SV was set to 1 and 5, respectively, for HiFi and ONT reads, considering the different qualities between HiFi and ONT platforms. The genomes of different pig breeds were *de novo* assembled using hifiasm (v0.19.4)[62] based on HiFi reads. The assembled genomes were compared to the Sscrofa11.1 reference genome to detect SVs using two assembly-based programs MUM&Co (v3.8)[46] and SVIM-asm (v1.0.3),[45] which are capable of identifying a full spectrum of SV types. SVIM-asm requires mapped reads as the input, and the HiFi reads aligned by minimap2 were used as suggested by the program developers.

Compilation of structural variation benchmark set

We used the callsets from two TGS platforms to compile the pig SV benchmark set, with the advantage of long reads over short reads on SV detection in mind. Only SVs with the FILTER "PASS" in the output *vcf* file were adopted. The compilation involved three rounds of merging: in each pig individual (breed), respectively in the HiFi and ONT repositories, the 18 callsets from the combinations of six callers and three aligners were merged using SURVIVOR (v1.0.7)[76] by setting the flags indicating the max distance between breakpoints, the minimum number of supporting callsets, and the minimum size of SVs to be taken into account respectively to 1 kb, 4, and 50 bp. The type and strands of SVs were both taken into account in SURVIVOR. The two preliminary merged SV sets from various aligner-caller combinations, respectively, in the HiFi and ONT repositories were then merged using SURVIVOR, assigning the minimum number of supporting platforms to one, i.e., SVs detected by either preliminary merged set were included.

Meanwhile, the callsets from two assembly-based callers based on HiFi data, i.e., MUM&Co and SVIM-asm, were merged by SURVIVOR, maintaining SVs detected by either of the assembly-based programs. Then, the alignment-based merged set between HiFi and ONT platforms and the assembly-based merged set between the two assembly-based callers were further merged by SURVIVOR. The final merged callset was additionally filtered by SVanalyzer (https://svanalyzer.readthedocs.io) to cluster SVs and remove duplicate SVs following the default settings to produce the final benchmark set. The length of SV indicated by "SVLEN" in the INFO column of *vcf* file was recalculated according to the start and end positions. SVs that could not be parsed by SURVIVOR, i.e., "SVTYPE=NA", were excluded from the final benchmark set. The bcftools package (v1.8)[66] was used to sort and index the benchmark sets. The variant effects of SVs in the benchmark sets were predicted using VEP (v112.0).[77]

Systematic comparison on performance metrics of SV detection in different platforms

In each pig breed, SV callsets based on Hi-C, WGS, HiFi, and ONT data were compared to the benchmark set to evaluate SV detection performance using three metrics: precision, recall, and F1 score. Precision is computed as the number of true positives (TPs) divided by the total number of positives, i.e., true positives + false positives (FPs). The true or false positives indicate SVs in or not in the benchmark set. Recall is computed as the ratio between the number of TP and the sum of TPs and false negatives (FN). False negatives imply the true SVs in the benchmark set that are not detected by the callers. F1 score is a parameter integrating precision and recall, formulated as $\frac{2\times(\text{precision}\times\text{recall})}{\text{precision}+\text{recall}}$. Truvari (v4.0.0)[78] module *bench* was used to measure the three metrics by setting flags *-p 0 -r 1000* without consideration of the reference genome. As insertion and deletion are two major SV types, we individually computed the precision and recall for insertions and deletions.

Besides the global evaluation encompassing all kinds of SVs, we also investigated the performance metrics when detecting specific SVs of different sizes and numbers of supporting reads. Six ranges of SV lengths, including 50-99 bp, 100-499 bp, 500-999 bp, 1,000-4,999 bp, 5,000-9,999 bp, and $\geq$ 10,000 bp, were considered. In regard to the number of supporting reads, SVs supported by 1, 2, 3, 4-9, 10-49, and $\geq$ 50 reads for HiFi data and 5, 6, 7, 8-14, 15-49, and $\geq$ 50 reads for ONT data were individually assessed. When studying the number of supporting reads, only precision could be assessed as there was no definite number of supporting reads for an SV in the benchmark set, which was derived from different platforms and aligners most likely with varying numbers of supporting reads. In this context, the false negatives belonging to a specific number of supporting reads were unknown.

The impact of the genomic regions accommodating SVs was also studied. Complex genomic regions with LINEs, SINEs, and simple repeats (micro-satellites) characterized by different repeat extents, high GC content (high GC) and low GC content (low GC) characterized by varying GC volumes, runs of homozygosity (ROH), and varying linkage disequilibrium (LD) extents were respectively considered. LINEs and SINEs are non-LTR retrotransposons, a kind of interspersed repeats across genome taking effect by converting their transcribed RNA back to DNA, i.e., reverse transcription.[83] LINEs are normally of thousands base-pair length and SINEs are hundreds of base-pair long.[84] Simple repeats are micro-satellites, a kind of tandem repeats, with a simple repeat pattern. High and low GC refer to a track of genomic regions with high and low Guanine (G) and Guanine (C) base-pair volumes. The coordinates of repeat regions were obtained from RepeatMasker[79] program embedded in https://genome.ucsc.edu/cgi-bin/hgTables using the pig Sscrofa11.1 assembly. To distinguish genomic regions with different GC volumes, the whole genome was firstly divided into 50 kb segments using bedtools (v2.30.0)[80] with flag *–makewindows*. The GC volume in each segment was calculated using bedtools flag *–nuc*. The high or low GC contents implied the GC volume was higher than 90% or lower than 10% quantiles of all segments GC volumes, respectively. To facilitate the calling efficiency in these specific regions, we focused on three representative autosomes of varying sizes, including the 1st, 4th, and 18th chromosomes, and the X chromosome. The regions of ROH and with different LD extents were disclosed based on SNP data from a population of 175 pig genomes consisting of the five breeds we used.[53] The WGS platform was applied to this population and is used to call the SNPs. The ROH regions were captured using Plink flag *–homozyg*. For regions with different LD extents, we first identified the LD blocks using Plink with flag *–block*. The flags indicating block maximal size (*–blocks-max-kb*) and confidence interval tops of r^2 (*–blocks-recomb-highci*) were set to 1,000 bp and 0.9, respectively. The pairwise r^2 of SNPs in specific blocks was calculated by Plink flag *–r^2*. The median value of r^2 was used to represent the overall LD extent of the block. Three LD extent ranges with lower and upper limits of r^2 being (0.25, 0.5), (0.5, 0.8), and (0.8, 1) were assigned. To facilitate the calling efficiency, the 1st, 4th, and 18th autosomes and the X chromosome were also exclusively considered.

For the investigation of the impact of sequencing depth on SV detection, we downsampled the total HiFi and ONT reads in each breed to depths respectively of 1 ×, 5 ×, 10 ×, 20 ×, and 30 × using rasusa (v0.7.1).[81] All the aligners and callers used in processing the full HiFi and ONT data were also used for the downsampled reads. The minimum number of supporting reads for alignment-based callers at each depth of HiFi data was set 1. For ONT reads, this value was set to 1 for 1 ×, 5 ×, and 10 ×, 2 for 20 × data, and 3 for 30 × data.

Considering the sensitivity of SVIM tool that a large number of SVs with only one supporting read would be reported, the minimum number of supporting reads for 5 × and 10 × was set to 2. The computational load, including time consumption and maximum memory usage for each aligner and caller, was recorded at a sequencing depth of 10 × except for the two assembly-based callers, i.e., SVIM-asm and MUM&Co, which processed the assembled genome based on total HiFi reads.

SV simulation

The SV detection capability of various aligners and callers for pig genomes was also investigated via a simulation trial. Sim-it (v1.3.6)[31] was used to simulate HiFi and ONT long reads respectively and SV benchmark sets based on the Sscrofa11.1 reference genome. The SV heterozygosity was set to 30% and 50% respectively for HiFi- and ONT-based detection as observed in the real-data experiment. The accuracy of ONT reads remained to 88% as default. A higher value of 95% was applied to HiFi reads considering the technical advantage of HiFi platform on sequencing accuracy. The median length and length range of reads were designated to 15,000 bp and 500-70,000 bp for HiFi reads, and 20,000 bp and 500-300,000 bp for ONT data according to the real sequencing data. The default error profiles for HiFi and ONT reads were used. Four types of SV including deletions, insertions, tandem duplications, and inversions were simulated with the number of 20,000, 20,000, 500, and 150 referring to the number in the benchmark sets of five pig breeds.

PCR validation

A total of 14 deletions and four insertions were randomly selected from the benchmark sets of AW, BKS, BMA, and LW (PTR samples were not used in the PCR validation due to the absence of DNA samples). The PCR primers of the 18 selected SVs were designed using SnapGene (v4.3.6) software and the Primer-BLAST program[82] on the National Center of Biotechnology Information (NCBI) website. The specificity of the PCR products was validated in NCBI and University of California Santa Cruz (UCSC) databases. The DNA samples of eight pig individuals from five breeds, including AW, BKS, BMA, LW, and Duroc (reference genome), were collected to *in vitro* validate the 18 SVs via gel electrophoresis. In total, $18 \times 8 = 144$ validations (gel lanes) were processed. A successful validation is defined as the SV *in silico* identified in a specific breed can be *in vitro* verified in the gel lane of that breed, i.e., true positive (TP), or the SV not *in silico* discovered in a specific breed is absent in the gel lane of that breed, i.e., true negative (TN). The total number of positives (P) or negatives (N) indicates the number of breeds with or without *in silico* identified SVs. The TP and TN rates are calculated as $\frac{\text{No. TP}}{\text{No. P}}$ and $\frac{\text{No. TN}}{\text{No. N}}$, respectively, in which $\text{No. P} + \text{No. N} = 144$, and the overall validation rate is $\frac{\text{No. TP} + \text{No. TN}}{144}$.

QUANTIFICATION AND STATISTICAL ANALYSIS

The quantification in regard to SV detection power, i.e., precision, recall, and F1 score, and PCR validation rate have been specified in the "method details" section. No statistical analysis was implemented in this study.