AIMS Animal Science, 2 (2): 122–141.
DOI: 10.3934/aas.2026009
Received: 26 August 2026
Revised: 27 October 2026
Accepted: 06 November 2026
Published: 23 November 2026

https://www.aimspress.com/journal/aas

Review

Graph-based pangenomics in pigs: Insights into structural variation, selection, and trait architecture

Caiyun Zhang[1,2], Jian Miao[3], Xiaotian Qiu[4], Chousheng Liu[4], Yuchun Pan[3] and Peipei Ma[1,2,*]

[1] Department of Animal Science, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, 200240, PR China
[2] Shanghai Key Laboratory of Veterinary Biotechnology, Shanghai, 200240, PR China
[3] Department of Animal Breeding and Reproduction, College of Animal Science, Zhejiang University, Hangzhou, 310058, PR China
[4] National Animal Husbandry Service, Beijing, 100125, PR China

* **Correspondence:** Email: peipei.ma@sjtu.edu.cn.

Abstract: A single linear pig reference cannot represent the full diversity of non-reference sequences and structural variants (SVs) across breeds. Graph pangenomes address this limitation by encoding alternative sequence paths, but construction methods differ in input, resolution, reference dependence, and scalability. This review compares four strategies: VCF-based augmentation, iterative SV-graph construction with minigraph, reference-anchored base-level alignment with Minigraph-Cactus, and reference-free all-to-all construction with pggb. We distinguish graph construction, read mapping, SV discovery, and genotyping; summarize evaluation at assembly, graph, and application levels; and synthesize pig-specific evidence from assembly graphs, population-scale SV panels, graph genotyping, SV-GWAS, and SV-eQTL studies. Pig graph resources have revealed extensive non-reference sequence, reduced mapping bias, and linked SVs to gene expression and complex traits. Future progress requires diverse sampling, standardized benchmarking, stable coordinate systems, integrated multi-omics resources, and experimental validation of candidate functional variants.

Keywords: pig; pangenome graph; structural variation; graph genotyping; non-reference sequence

1. Introduction

Pork production is a major component of global animal agriculture, and genomic information is increasingly used to improve growth, feed efficiency, meat quality, reproductive performance, disease resistance, and environmental robustness [1–5]. The release of Sscrofa11.1 provided a substantially improved linear reference for pig genetics [6]. Subsequent long-read and chromosome-conformation technologies enabled multiple breed-specific, near-complete, and telomere-to-telomere assemblies [7–9]. These resources make it possible to examine genomic regions that were fragmented or collapsed in earlier assemblies and provide the assembly diversity required for pangenome construction.

The biological motivation for a pig pangenome is not simply that different breeds carry different single nucleotide polymorphisms (SNPs). Pig genomes contain abundant repeats, mobile elements, segmental duplications, copy-number changes, inversions, and long non-reference insertions. These features can create ambiguous alignments or remain absent when all reads are interpreted against one reference haplotype. Reference choice can consequently influence mapping rate, apparent heterozygosity, and the discovery and genotyping of structural variants [10–13]. Because domestication and breeding produced deeply differentiated Asian and European lineages and many locally adapted populations, a useful representation must capture both shared sequence and population-specific haplotypes.

Graph pangenomes should nevertheless be understood as analytical representations rather than inherently complete genomes. Their content depends on input assemblies or variant calls, filtering thresholds, the construction algorithm, sample ordering, and the coordinate system chosen for downstream analysis. A structural variants (SVs)-oriented minigraph graph, a variant call format (VCF)-augmented variation graph, and a base-level Minigraph-Cactus or pggb graph answer different questions and should not be compared as if they contained identical information. This distinction motivates the method-centered organization of this review.

Pig genomics has progressed from Sscrofa11.1 to multiple near-complete and telomere-to-telomere assemblies; however, no single haplotype represents the extensive variation created by domestication, geographic adaptation, and intensive selection [6–9]. Repetitive sequence, segmental duplication, and structurally diverse loci remain difficult to resolve with a single linear reference, while population studies centered on commercial breeds can underrepresent locally adapted populations. These limitations matter because growth, reproduction, meat quality, disease resistance, and environmental adaptation are polygenic traits influenced by variants spanning a wide range of sizes [4,5,14–17].

A pangenome integrates sequence from multiple individuals and a graph pangenome additionally represents alternative alleles as paths through connected sequence segments [10,11]. Depending on the construction method, a graph may represent known variants added to a reference, large structural differences discovered by iterative assembly alignment, or base-level variation derived directly from multiple haplotype assemblies [18–21]. Graph edges encode sequence adjacency along represented haplotypes, not gene–gene or regulatory interactions. The principal advantages are reduced reference bias, improved representation of non-reference sequence, and preservation of local haplotype context for mapping and genotyping [10,15,19,22].

Pig studies now provide direct evidence for these benefits. Graphs built from 21 assemblies revealed 105 Mb of non-reference sequence and breed-dependent differences in non-reference content [12]. A study of Chinese indigenous pigs showed that graph mapping reduced reference-related alignment bias and heterozygosity overestimation [13]. Population-scale graph genotyping has connected SVs

with expression and complex traits in hundreds to thousands of pigs [15,23,24], while SwinePan now integrates graph-derived variation with transcriptomic and phenotypic data [25]. In ref. [15], mean heritability was 0.43 using SNPs, 0.46 using SVs, and 0.56 when both were combined, indicating that SVs complement rather than replace SNPs.

This review focuses specifically on pig graph pangenomics. It compares sampling and construction designs; distinguishes graph construction from read mapping, SV discovery, and genotyping; evaluates the evidence from pig applications; and identifies remaining needs for representation, benchmarking, functional validation, and data access.

2. Designing a pig graph pangenome

2.1. Sampling and assembly quality

2.1.1. Capturing genetic and geographic diversity

Sample selection should begin with the population structure that the graph is intended to represent. Chinese indigenous breeds, European commercial breeds, wild boars, miniature pigs, and geographically isolated populations differ in demographic history and selection. Published pig pangenomes have used contrasting designs: some emphasize a small number of high-quality breed assemblies, whereas others combine more assemblies of variable continuity to maximize population coverage [12,15–17,26]. A balanced design should avoid allowing a large number of closely related commercial animals to dominate graph content while rare or geographically distinct lineages are represented by only one lower-quality assembly.

Genetic distance, pedigree, breed origin, sex, phenotype, and environmental adaptation can all guide selection. Principal-component or phylogenetic analyses based on SNPs or assembly sketches can identify redundant candidates and reveal gaps in representation. Sex chromosomes require explicit planning because male and female assemblies contribute different information, and autosomal diversity alone does not guarantee adequate representation of X- or Y-linked sequence. Where budget is limited, selecting genetically distant and phenotypically distinct individuals is generally more informative than sequencing several close relatives from the same breed [7–9,12,15–17,26,27].

2.1.2. Deciding how many assemblies are sufficient

The number of assemblies is not a quality metric by itself. Projects targeting common SVs in a well-defined breeding population may reach saturation with fewer assemblies than projects intended to represent worldwide diversity or rare breed-specific sequence. Investigators should add assemblies in a prespecified order or across repeated randomized orders and track cumulative graph length, novel non-reference sequence, new SV alleles, core and accessory sequence, and mapping improvement in held-out individuals [28]. A plateau across several of these measures is more informative than a fixed sample-number rule. Even after apparent saturation, targeted addition of an underrepresented ancestry can reveal sequence that random sampling would miss.

There is no universal minimum number of individuals for a pangenome. Published pig discovery sets already span 11 long-read genomes and 21 assemblies, the latter revealing 105 Mb of non-reference sequence [12,16]. These are observed study scales rather than minimum requirements. The

required sample size depends on population structure, breed divergence, target allele frequency, assembly quality, graph purpose, and available resources. A practical design selects unrelated representatives across major geographic and breeding groups and monitors cumulative graph length, novel sequence, new SV alleles, and held-out mapping as each assembly is added. Sampling can be considered near saturation only when these curves plateau across several addition orders; a nominal percentage alone should not be treated as a universal stopping rule [12,16,28].

2.1.3. Quality control before graph construction

Assemblies should be evaluated consistently before they enter a graph. Contig N50 is the contig length at which contigs ordered from longest to shortest first account for 50% of assembly length; NG50 instead uses an estimated genome size. BUSCO reports complete single-copy, complete duplicated, fragmented, and missing orthologs, whereas k-mer QV expresses the inferred base-error rate on a Phred scale [29,30]. The Vertebrate Genomes Project provides useful reference targets rather than pig-specific pass/fail rules: its VGP-2020 standard specifies contig NG50 above 10 Mb, phase-block NG50 above 1 Mb, QV above 50, k-mer completeness above 95%, chromosome assignment above 95%, and false duplication below 1% [29]. QV 40 corresponds to approximately one error per 10,000 bases and QV 50 to one per 100,000 bases. Recent pig assemblies illustrate the currently achieved range: the Min and Rongchang T2T assemblies reported contig N50 values of 143.37 and 143.96 Mb, BUSCO completeness of 98.2% and 98.3%, QV values of 51.17 and 47.84, and read-mapping rates above 99.73%; the near-T2T Jinhua assembly reported contig N50 of 100.5 Mb, scaffold N50 of 142.7 Mb, BUSCO completeness of 96.4%, QV 55, and HiFi-read mapping of 99.90% [7,9]. Thus, values around 100–144 Mb for contig N50, 96.4–98.3% for BUSCO completeness, QV 47.84–55, and at least 99.73% read mapping describe recent high-quality pig assemblies but should not be converted into rigid acceptance thresholds. Each project should also inspect haplotype switch errors; contamination; collapsed repeats; duplicated haplotypes; and structural concordance with Hi-C, optical maps, or independent long reads; lower-quality assemblies can create false branches and artifactual non-reference sequence [7,9,27,29,30].

2.2. Sequencing strategy

2.2.1. Short reads and population scale

Illumina short reads remain useful because they offer high per-base accuracy, mature quality-control procedures, and economical sequencing of hundreds or thousands of individuals [31]. They are well suited to SNP and small-indel calling and can genotype a panel of known SVs when the alleles are represented by informative breakpoints or k-mers. Their main limitations arise in repeats, novel insertions, segmental duplications, and complex rearrangements, where a 150-bp fragment may map to several loci or fail to align. Accordingly, short reads are most powerful in a pangenome project when used for population-scale genotyping after variant discovery from long reads and assemblies.

2.2.2. Long reads and haplotype resolved assembly

PacBio HiFi reads combine kilobase-scale read lengths with high consensus accuracy and are

widely used for phased de novo assembly and sequence-resolved SV discovery [32,33]. Oxford Nanopore sequencing can produce substantially longer reads, including reads that span extended tandem repeats, centromeric arrays, and segmental duplications [8,34]. HiFi and ultra-long ONT data are therefore complementary: HiFi reads support accurate local sequence, whereas ultra-long reads improve contiguity across difficult regions. Hi-C, parental data, or population phasing can further separate haplotypes and anchor contigs to chromosomes.

2.2.3. Project specific sequencing design

Sequencing depth should not be prescribed as a universal constant. Required depth depends on read-length distribution, heterozygosity, genome complexity, assembler, polishing strategy, and the desired confidence in haplotype separation. Pilot assemblies and k-mer-based completeness analyses can identify diminishing returns. For population projects, a practical design is to generate deep long-read data for genetically diverse discovery individuals, construct and validate an SV or pangenome panel, and genotype that panel in a much larger short-read cohort. However, this design must be updated periodically because panel-based genotyping cannot recover alleles absent from the discovery set.

Short reads remain economical for large cohorts and are effective for SNP and small-indel genotyping, but they provide limited resolution in repeats and for long insertions. PacBio HiFi reads combine long read length with high consensus accuracy, whereas Oxford Nanopore ultra-long reads can bridge centromeres, segmental duplications, and other extended repeats [8,29,33,34]. Current projects commonly combine HiFi, ultra-long ONT, and Hi-C data for haplotype-resolved assemblies; coverage should be selected from platform performance, genome complexity, phasing objectives, and empirical assembly benchmarks rather than a fixed universal threshold. Population projects can then genotype variants in larger short-read cohorts using an assembly-derived reference panel.

3. Graph construction methods and formats

3.1. VCF based reference augmentation

Reference-augmentation methods begin with a linear reference and a normalized variant set. vg construct, for example, embeds alternative alleles from VCF records into a sequence graph [18]. This approach is attractive when a well-curated cohort VCF already exists and stable reference coordinates are essential. It can incorporate SNPs, indels, and suitably represented SV alleles, and the resulting graph can support graph mapping and calling. Its completeness is limited by the discovery pipeline that produced the VCF: missing non-reference sequence, unresolved complex alleles, or inconsistent representations are not repaired automatically by graph construction.

3.2. Iterative assembly addition with minigraph

Minigraph uses a designated backbone and incrementally maps additional assemblies to the current graph [21]. It identifies long collinear chains and adds sufficiently large sequences that are not represented by an existing path. This design is fast and maintains coordinates relative to the backbone, making it practical for dozens of mammalian assemblies. It is primarily an SV graph: small variants

within aligned blocks are generally not represented at base-level resolution. The selected backbone and order of assembly addition can influence which sequences are labeled non-reference and how nested or overlapping events are represented, so construction order and parameters should be reported.

3.3. Reference anchored base level alignment with Minigraph Cactus

Minigraph-Cactus extends an initial minigraph SV graph to base-level alignment [19]. Assemblies are mapped to the SV graph; the problem is partitioned into chromosome-scale subproblems; Cactus aligns homologous sequence within those regions; and the resulting alignments are joined, clipped, and indexed. This architecture represents both large structural differences and smaller variation within aligned regions. Graph decomposition can generate a VCF for a chosen reference path, but that VCF is an output rather than the input used to build the assembly graph. Mapping with vg Giraffe relies on graph indexes, such as GBZ, minimizer, and distance indexes.

3.4. Reference free construction with pggb

The pggb workflow avoids choosing one assembly as the privileged reference during construction [20]. It performs all-to-all alignment with wfmash, converts the alignment relation into a graph with seqwish, and applies smoothxg to normalize and order graph blocks. This design can represent variants across a broad size range and is useful when several high-quality haplotypes have comparable status. The all-to-all computation is demanding for large mammalian cohorts, and graph complexity can increase in repetitive or highly rearranged regions. Parameter selection, input assembly quality, and post-construction normalization therefore have strong effects on usability.

3.5. Graph formats coordinates and interoperability

After a construction workflow has encoded sequence variation, the resulting graph must be stored in a form that preserves both sequence and interpretable paths for downstream tools. The reviewed toolchains primarily exchange GFA1 or GFA1.1: H records store header information, S records define sequence segments, L records define links, P records describe paths, and W records introduced in GFA1.1 describe walks. Minigraph additionally uses rGFA tags to associate segments with stable coordinates on a designated reference path. GFA2 F records describe fragments of external sequences aligned to segments and should not be interpreted as containment between two graph segments. Although these formats facilitate exchange, identical biological variation can still be encoded differently across pipelines. Downstream analyses therefore require explicit graph versioning, named reference paths, documented normalization, and coordinate-liftover procedures.

Graph construction methods fall into four complementary classes (Figure 1). First, reference augmentation methods such as vg construct and GRAF add variants from a VCF to a linear reference and therefore represent variants already present in the input call set [18,35]. Second, minigraph incrementally maps assemblies to a backbone and adds sufficiently large non-reference segments, producing an SV-oriented graph while retaining backbone coordinates [21]. Third, Minigraph-Cactus begins with a minigraph SV graph; maps each assembly to that graph; partitions the input by chromosome; computes base-level Cactus alignments; and joins, clips, and indexes the result [19]. VCF is an optional output generated by graph decomposition, not an input to this assembly-based

workflow; vg Giraffe mapping uses graph and mapping indexes, such as GBZ, distance, and minimizer indexes. Fourth, pggb is reference-free: it performs all-to-all alignment with wfmash, induces a graph with seqwish, and normalizes graph structure with smoothxg [20].

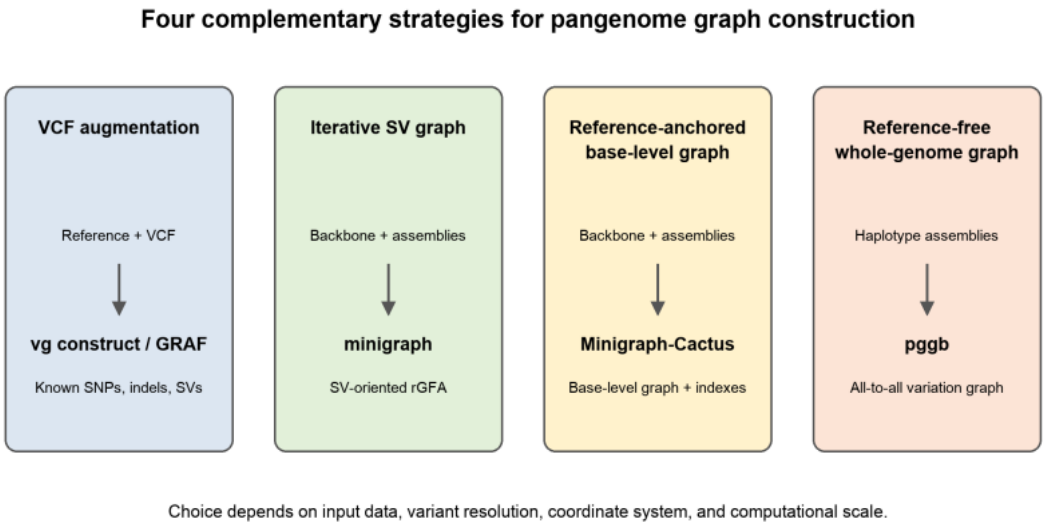

Figure 1. Overview of four complementary strategies for pangenome graph construction. This original schematic distinguishes their inputs, reference dependence, and principal outputs.

Table 1. Comparison of principal pangenome graph construction strategies [18–21,35].

Strategy	Representative tools	Primary input	Reference dependence	Typical representation	Main trade-off
VCF augmentation	vg construct; GRAF	Reference + VCF	Reference anchored	Known SNPs, indels and SV alleles	Fast and coordinate-compatible; limited to input calls
Iterative SV graph	minigraph	Backbone + assemblies	Backbone anchored	Large non-reference segments and SVs	Scalable; order and backbone can influence content
Base-level alignment	Minigraph-Cactus	Backbone + assemblies	Reference anchored	SV and small variation at base resolution	Rich representation; greater compute and indexing burden
Reference-free alignment	pggb	Haplotype assemblies	No designated reference	All-to-all variation graph	Avoids a privileged backbone; computationally intensive

4. Evaluating graph pangenomes

4.1. Topology and computational behavior

Node, edge, path, and component counts describe graph size and connectivity but do not by themselves establish biological quality. Extremely fragmented graphs may inflate node and edge counts, whereas aggressive simplification can collapse paralogous or repetitive sequence. odgi and vg utilities can summarize path length, node depth, connected components, orientation, and local complexity [18,36,37]. Runtime, peak memory, index size, and update cost should be reported because a graph that cannot be mapped or genotyped at the intended cohort scale has limited practical value. Because counts scale with represented sequence, they should also be normalized per gigabase and compared at matched sample numbers. No accepted node- or edge-count threshold separates a good graph from a poor one; the informative quantity is the change in fragmentation and complexity relative to the same input assemblies under alternative construction or normalization settings.

4.2. Representation and path fidelity

A graph should reproduce the assemblies from which it was built without losing sequence, introducing unsupported duplication, or breaking haplotype paths. K-mer comparisons can measure completeness, collapse, and duplication independently of a single linear reference. Path-based checks can quantify whether source chromosomes are represented continuously and whether rearrangements are encoded with the expected orientation. Panacus growth curves characterize core, accessory, and private sequence as samples accumulate [28], while newer tools, such as PG-SCUnK, PanSel, and graph-comparison approaches, provide complementary measures of representativeness and conservation [38–40]. Source-assembly targets provide a quantitative baseline: greater than 95% k-mer completeness and less than 1% false duplication are stringent VGP-2020 reference values [29]. For a graph, however, no universal completeness cutoff has been established. Studies should report absolute recovery and the loss relative to each input assembly, separately for unique sequence, repeats, and sex chromosomes, rather than labeling a graph complete from one aggregate percentage.

4.3. Downstream benchmarking

The most informative evaluation uses samples not included in graph construction. Mapping can be compared with Sscrofa11.1 using total alignment rate, high-quality unique mappings, mismatch rate, allele balance, and performance stratified by ancestry and genomic context. Variant benchmarks should report precision, recall, genotype concordance, and errors by SV type and size. Simulated data help because the truth is known, but real held-out long-read assemblies and orthogonal validations are required to capture realistic repeats and complex alleles. Claims of reduced bias should be supported by consistent improvement across relevant populations rather than only by a higher aggregate mapping rate. Performance should be reported with denominators and uncertainty (e.g., precision and recall with 95% confidence intervals) and stratified by ancestry, SV class, size, repeat context, and allele-frequency bin. A single overall mapping rate can conceal failure in difficult regions: recent high-quality pig assemblies map at least 99.73% of validation reads [7,9], so graph evaluation should emphasize the additional high-quality mappings, reduced mismatch or reference bias, and improved variant

accuracy in held-out samples rather than merely exceeding a low global mapping threshold.

Evaluation should separate source-assembly quality from graph quality. At the topology level, vg stats and odgi stats summarize nodes, edges, paths, path lengths, and depth [18,36,37]. Panacus measures pangenome growth and core or accessory content [28]. At the representation level, k-mer completeness, duplication, collapse, path fidelity, and recovery of known variants assess whether the source haplotypes are represented without excessive redundancy [38–41]. At the application level, held-out samples and truth sets can test mapping rate, mapping bias, variant recall and precision, genotype concordance, computational cost, and stability as new assemblies are added. No single statistic defines a high-quality graph; evaluation must match the intended downstream analysis.

5. Structural variant discovery and genotyping

5.1. Discovery from short reads, long reads, and assemblies

5.1.1. Evidence used by short read callers

Short-read SV callers combine discordant paired-end orientation or distance, split-read breakpoints, local read depth, and unmapped or partially mapped sequence [42,43]. Each signal favors particular event classes. Read-depth changes detect deletions and duplications but provide imprecise breakpoints; split reads yield base-level breakpoints when unique sequence spans the junction; and paired-end evidence can identify deletions much larger than the library insert size but cannot reconstruct long novel insertions. Integrating signals improves sensitivity, although repetitive sequence and reference-absent alleles remain major limitations.

5.1.2. Long read alignment callers

Long reads often span an entire SV and its unique flanks, enabling sequence-resolved insertions and improved analysis of repeats [44]. Sniffles2, cuteSV, and SVIM begin with long-read alignments and cluster signatures across supporting reads [45–47]. Their algorithms differ in signature extraction, clustering, genotyping, and handling of nested or population-scale events. SKSV uses a skeleton-based strategy optimized for PacBio CCS reads [48]. Benchmarking shows that the aligner, sequencing platform, depth, and variant class can change relative caller performance [49]; consequently, studies should report full software versions and parameters and consider orthogonal confirmation or consensus call sets.

5.1.3. Assembly based discovery

Assembly-based discovery compares phased haplotypes with a reference, another assembly, or a pangenome alignment. It can reconstruct long insertions, multi-breakpoint events, and sequence in regions where individual reads map ambiguously. However, assembly errors can be misinterpreted as biological SVs, and different whole-genome aligners or decomposition algorithms can produce different representations of the same rearrangement. Confidence increases when an event is supported by raw long reads, consistent haplotype coverage, and concordant sequence or breakpoint evidence.

SVs are commonly defined as variants of at least 50 bp and include deletions, insertions,

duplications, inversions, translocations, and complex rearrangements [42]. Mechanisms include non-allelic homologous recombination, non-homologous end joining, fork stalling and template switching, and mobile-element insertion. Deletions, insertions, and duplications are unbalanced SVs, whereas inversions and reciprocal translocations are balanced. Copy-number variants are a dosage-changing subset rather than a synonym for all unbalanced SVs.

Short-read callers combine discordant paired-end mapping, split reads, and read depth [42,43]. Paired-end signals can detect deletions larger than the library insert size, whereas insertions longer than the insert size are difficult to resolve. Long-read callers analyze mappings to a linear reference: Sniffles2 integrates long-read alignment signatures and supports population-scale calling; cuteSV clusters signatures, such as split mappings and indel signals; SVIM clusters five classes of signatures from mapped long reads; and SKSV uses a skeleton-based alignment strategy optimized for PacBio CCS reads [45–48]. These tools do not use short-read pairs, graph traversal, or alignment-free k-mer classification. Assembly-based discovery compares haplotype assemblies with a reference or with one another and can better resolve long insertions and complex alleles, although results depend on assembly and alignment quality.

Table 2. Structural variant discovery methods [45–49].

Tool or approach	Task	Input	Core method	Important limitation
Sniffles2	SV discovery	Mapped long reads	Clustering and genotyping long-read SV signatures	Dependent on long-read alignment and coverage
cuteSV	SV discovery	Mapped long reads	Heuristic collection and clustering of SV signatures	Performance varies by aligner and SV class
SVIM	SV discovery	Mapped long reads	Clustering of linear-alignment signatures	Not a graph-traversal method
SKSV	SV discovery	PacBio CCS reads	Skeleton-based alignment and SV extraction	Designed for CCS rather than mixed short/long reads
Assembly comparison	SV discovery	Haplotype assemblies	Whole-genome or graph-based assembly alignment	Requires high-quality phased assemblies

5.2. Genotyping known variants

Paragraph builds local graphs for candidate variants and aligns short reads to alternative alleles, allowing focused genotyping without constructing one whole-genome graph [43]. GraphTyper2 similarly uses local variation graphs and population information for scalable genotyping [50]. These alignment-based approaches can reduce linear-reference bias near known SVs, but results depend on breakpoint accuracy and on whether the candidate allele is represented correctly.

BayesTyper and PanGenie illustrate a different strategy. BayesTyper uses k-mer evidence in a Bayesian framework and can genotype SNVs, indels, and SVs [51]. PanGenie relates short-read k-mer

counts to haplotypes in an assembly panel and infers genotypes for variants represented by those haplotypes [52]. K-mer approaches avoid conventional read alignment and can perform well in population cohorts, but repetitive k-mers and absent panel alleles limit identifiability. Neither method should be described as discovering arbitrary new SVs from the sample.

Whole-genome pangenome mapping provides another route. vg Giraffe maps reads rapidly to an indexed graph, after which vg call or another compatible caller produces genotypes [18,53]. Keeping mapping and calling conceptually separate is important when comparing methods: a mapper can improve placement of reads without by itself defining alleles or estimating genotypes. Evaluation should therefore distinguish mapping accuracy, allele representation, calling accuracy, and computational cost.

Genotyping should be distinguished from discovery. Paragraph aligns short reads to local graphs built from candidate SVs supplied in a VCF [43]. GraphTyper2 maps short reads to local variation graphs and performs population-scale genotyping [50]. BayesTyper is an alignment-free, k-mer-based Bayesian genotyper for SNVs, indels, and SVs [51]. PanGenie also avoids read alignment: it uses short-read k-mer counts and a panel of haplotype-resolved assemblies to genotype variants represented in that panel [52]. vg Giraffe is a mapper; genotypes require a downstream caller, such as vg call [53]. General pangenome graphs are not necessarily directed acyclic graphs and may encode cycles or inversions.

Table 3. Methods for genotyping known variants [18,43,50–53].

Method	Primary input	Alignment based	Variant scope	Output constraint
Paragraph	Short reads + candidate VCF	Yes, local graphs	Candidate SVs	Genotypes supplied alleles
GraphTyper2	Short reads + candidate variants	Yes, local graphs	SNVs, indels, SVs	Genotypes supplied alleles
BayesTyper	Short-read k-mers + variant panel	No	SNVs, indels, SVs	Limited by panel and informative k-mers
PanGenie	Short-read k-mers + phased assembly panel	No	Panel variants including SVs	Cannot discover variants absent from panel
vg Giraffe + vg call	Reads + indexed pangenome graph	Yes, graph mapping	Variants represented in graph	Mapper and caller are separate steps

6. Pig graph-pangenome and population-scale SV studies

6.1. Assembly graphs and non reference sequence

Early pig pangenome studies used de novo assemblies or population sequencing to identify sequence absent from the linear reference [15–17,26]. Miao et al. constructed a minigraph graph from

21 assemblies and identified 23,831 non-reference sequences totaling 105 Mb [12]. Non-reference content increased with genetic divergence from the Duroc-derived reference, and insertions showed enrichment in immune-related functions. This study demonstrates how graph construction can organize assembly-specific sequence, while also illustrating that graph content depends on which assemblies and quality levels are included.

Other pig pangenome resources have emphasized gene content, breed adaptation, or a larger collection of genomes [15–17,26]. Comparisons across these projects require caution because their definitions of non-reference sequence, minimum event size, assembly inclusion criteria, and backbone differ. Counts of novel genes or SVs should not be compared directly without harmonized filtering. Nevertheless, the collective studies show that a substantial fraction of pig sequence and structural diversity is poorly represented by a single Duroc reference.

6.2. Mapping bias and population genotyping

The benefit of a graph reference can be tested directly by mapping the same reads to linear and graph representations. Miao et al. showed that a pig graph reduced reference-related mapping bias and heterozygosity overestimation in Chinese indigenous pigs [13]. This result is especially relevant because inflated heterozygosity can arise when divergent reads are forced onto an incomplete reference or when paralogous sequence is mismapped. Future studies should extend such comparisons across breeds, coverage levels, and difficult genomic regions and report whether improvements persist after stringent mapping-quality filters.

Li et al. genotyped assembly-derived SVs in 402 pig genomes and connected several enhancer deletions with expression of FSD2, LMOD2, and WASL [23]. Yang et al. developed a large SV catalog from 1,060 pig genomes [54], and PSVRP provides a reference panel intended to support SV imputation and complex-trait studies [55]. These resources represent related but distinct stages: discovery defines candidate alleles, graph or panel methods genotype them in populations, and association or functional analyses test their consequences.

6.3. Multiomics resources and recent scale up

Wu et al. integrated 60 long-read samples and 2,445 high-quality short-read samples from 120 breeds or populations, using two graph-based genotyping approaches to create a broad SV map [24]. Multi-tissue eQTL analysis associated 49.01% of expressed genes with cis-region SVs, and integrated analyses identified differentiated regulatory candidates related to adaptation and growth. These results extend the field from cataloging variation toward testing effects on expression and complex traits, although the authors also note limitations caused by heterogeneous public data and incomplete discovery of complex alleles.

SwinePan integrates a graph-derived variome covering 2,598 individuals from 35 breeds with transcriptomic data; phenotypic records; and tools for GWAS, eQTL mapping, and colocalization [25]. Its emergence changes the perspective from a need to create any pig pangenome database to a need for interoperable, versioned, and independently benchmarked resources. Key priorities are transparent update policies, coordinate translation, downloadable graph and index files, provenance for each allele, and interfaces connecting variation to tissue-specific regulation and breeding phenotypes.

Pig graph research has moved from non-reference sequence discovery to population-scale

genotyping and functional interpretation. The studies below are not methodologically equivalent: some construct assembly graphs, some develop SV panels, and others genotype graph-represented variants in short-read cohorts. Linear-reference SV studies remain informative but should not be presented as graph-derived evidence.

Table 4. Selected pig graph-pangenome and population-scale structural-variation studies. A dash indicates that population genotyping was not the principal design.

Study	Assemblies or discovery set	Genotyped samples	Approach	Principal contribution
Miao et al. 2024 [12]	21 assemblies	—	minigraph	105 Mb in 23,831 non-reference sequences; immune and breed-related patterns
Miao et al. 2025 [13]	Pig graph panel	Chinese indigenous pigs	Graph read mapping	Reduced mapping bias and heterozygosity overestimation
Li X et al. 2024 [23]	Assembly-derived SV panel	402 genomes	Graph-based SV genotyping	Enhancer deletions associated with FSD2, LMOD2, and WASL expression
Yang et al. 2024 [54]	Long- and short-read discovery	1,060 genomes	Population SV mapping	Large SV catalog and functional characterization
Li X et al. 2025 [55]	PSVRP panel	Population-scale resource	SV reference panel	Supports SV imputation and complex-trait analysis
Wu et al. 2026 [24]	60 long-read samples	2,445 short-read samples	Two graph genotyping approaches	SV-eQTL, adaptation, and growth-related associations
Lin et al. 2026 [25]	Graph-derived variome	2,598 individuals, 35 breeds	SwinePan database	Integrates SVs, SNPs, indels, transcriptomes, and phenotypes

7. Selection, regulation, and complex traits

7.1. Population differentiation and selection

Graph-derived genotypes can be analyzed with nucleotide diversity, FST, haplotype-based statistics, and other selection scans, but the graph itself does not establish selection. Population structure, demographic history, ascertainment, and uneven genotype quality can create apparent differentiation. A robust analysis should compare graph and linear results, control relatedness and ancestry, report allele frequencies by population, and distinguish a differentiated locus from a validated adaptive mechanism. Structural alleles are particularly challenging because linked SNPs may tag the same haplotype and breakpoint uncertainty can affect genotype accuracy.

7.2. Regulatory effects and trait association

SVs can alter coding sequence, dosage, promoters, enhancers, chromatin accessibility, and the distance or contact between regulatory elements and genes [56]. A statistical association between an SV and expression is therefore biologically plausible but not automatically causal. Strong evidence combines accurate genotypes, tissue-matched eQTL data, overlap with regulatory annotation, colocalization or fine-mapping, and ideally experimental perturbation. The enhancer deletions reported in the 402-genome graph-genotyping study and the multi-tissue SV-eQTL results from 2,445 pigs illustrate this progression from allele representation to functional hypothesis [23,24].

7.3. Contribution to genomic prediction

SVs can complement SNPs when they tag functional variation not well captured by SNP arrays or linear-reference calls. In ref. [15], mean heritability was 0.43 using SNPs, 0.46 using SVs, and 0.56 when both were combined. The combined estimate supports complementarity but does not imply that SV-only models universally outperform SNP models. Prediction gains will depend on allele frequency, linkage disequilibrium, training-population size, genotype accuracy, trait architecture, and transfer across breeds. Prospective breeding evaluations are needed before graph-derived SVs can be recommended routinely for selection.

Graph-derived genotypes can be analyzed with familiar population-genetic and quantitative-genetic frameworks, including nucleotide diversity, FST, selection scans, GWAS, eQTL mapping, and heritability partitioning. Interpretation must distinguish statistical association, selection evidence, regulatory support, and demonstrated causality. The NR6A1 example previously used in this review did not satisfy this requirement: Rubin et al. [57] used pooled resequencing against a linear reference, and the established NR6A1 candidate is a missense variant rather than a graph-discovered SV. That example is therefore removed.

Pig-specific graph evidence is now available. Graph genotyping in 402 genomes linked enhancer deletions to expression of FSD2, LMOD2, and WASL [23]. The study in ref. [15] showed that SVs and SNPs contributed complementary heritability. Wu et al. [24] combined graph-based SV genotyping with multi-tissue eQTL and trait analyses in 2,445 pigs; 49.01% of expressed genes were associated with cis-region SVs, and integrated analyses identified regulatory candidates, including a deletion associated with PLAG1 expression and body weight. These findings support graph pangenomes as infrastructure for testing regulatory hypotheses, but experimental validation remains necessary before an association is called causal.

8. Challenges and future priorities

8.1. Representation and standards

Pig graph resources currently differ in assemblies, backbones, filters, graph formats, allele decomposition, and coordinate systems. Reproducible comparison requires reporting source assembly accessions, haplotype status, software versions, parameters, graph hashes or releases, and the exact indexes used for mapping. Community benchmarks should include common and rare variants, repeats, segmental duplications, insertions absent from Sscrofa11.1, and several divergent pig populations. The

evaluation should also test whether graph expansion improves underrepresented groups without reducing mapping specificity in already well-represented breeds.

8.2. Computational and analytical limitations

Graph size and local branching affect indexing, visualization, mapping, and calling. Base-level graphs can encode richer variation than SV-only graphs but may require more storage and computation and can accumulate redundant paths. Complex rearrangements are difficult to decompose into consistent alleles, while repeated sequence can generate ambiguous paths or k-mers. No single construction or genotyping method is optimal for every task. A practical ecosystem will probably retain several representations: a stable reference-coordinate graph for routine population analysis, richer assembly graphs for discovery, and local graphs for difficult trait loci.

8.3. Functional validation and translation

The number of graph-genotyped associations is likely to grow faster than the capacity for validation in live pigs. CRISPR editing, cloning, embryo transfer, and animal phenotyping require substantial resources and raise welfare and background-genotype considerations. Prioritization should therefore integrate allele frequency, predicted molecular consequence, tissue-specific expression, chromatin data, colocalization, and replication across populations. Cell models, reporter assays, organoids, and naturally segregating families can provide intermediate evidence before generating edited animals. Functional claims should remain proportional to the validation level.

8.4. Data access and multiomics integration

PSVRP, SwinePan, PigBiobank, and PigGTEx provide complementary genotype, expression, phenotype, and annotation resources [25,55,58,59]. The next challenge is not simply to create more databases but to connect them through stable identifiers, shared metadata, graph-aware browsers, and programmatic access. Single-cell transcriptomic and epigenomic data can help identify the cell types in which an SV acts, while proteomic, metabolomic, microbiome, and phenomic measurements can trace downstream effects. These integrations require careful control of tissue, age, environment, population structure, and multiple testing.

The first challenge is representation. Increasing the number of assemblies is useful only when sampling adds underrepresented ancestry, geography, sex chromosomes, or phenotype and when assembly quality is sufficient. Pangenome growth curves, held-out mapping, and ancestry-stratified evaluation should guide expansion. Second, construction and indexing costs increase with graph size and local complexity; repeat-rich loci and long or nested rearrangements remain difficult, and tool performance varies by platform, aligner, coverage, and SV class [19,40,41,49]. Third, population genotyping is usually limited to variants represented in the discovery panel, so deep long-read discovery and broad short-read genotyping should be iterated rather than treated as a one-time workflow.

Functional validation is a separate bottleneck. CRISPR-edited pigs are expensive and time-consuming, and editing efficiency, mosaicism, welfare requirements, and background effects complicate interpretation. Prioritization should therefore combine graph genotypes with eQTL,

chromatin accessibility, epigenomic annotation, and trait association before animal experiments. Current resources have improved substantially: PSVRP supports SV-panel analyses [55], and SwinePan provides graph-derived variation, multi-tissue expression, phenotypes, and online association tools [25]. Remaining needs include interoperable coordinates, stable graph versions, richer rare-breed representation, standardized benchmarks, and reproducible interfaces linking variant paths to functional annotations.

9. Conclusions

Graph pangenomes provide a practical framework for representing pig genomic diversity beyond a single linear reference. Their value depends on matching the construction method to the scientific question; validating source assemblies and graph behavior separately; and distinguishing discovery, mapping, and genotyping tools. Pig-specific studies now show reduced mapping bias, improved access to non-reference sequence, population-scale SV genotyping, and integration with expression and complex traits [12,13,15,23–25,54,55,60]. The next phase should emphasize diverse sampling, transparent benchmarking, versioned public resources, and functional follow-up that distinguishes association from causation.

Use of AI tools declaration

The authors used an AI-assisted language and document-editing tool during revision to help reorganize text and improve clarity. All scientific statements, citations, interpretations, and final wording were reviewed and approved by the authors, who take full responsibility for the content.

Acknowledgment

This work was funded by the National Agricultural Science and Technology Major Project (2022).

Conflict of interest

All authors declare no conflicts of interest.

References

1. OECD/FAO (2024) OECD-FAO Agricultural Outlook 2024-2033. Paris/Rome: OECD Publishing/ FAO. https://doi.org/10.1787/4c5d2cfb-en
2. FAO (2024) World Food and Agriculture—Statistical Yearbook 2024. Rome: FAO. https://doi.org/10.4060/cd2971en
3. National Bureau of Statistics of China (2024) China Statistical Yearbook 2024. Beijing: China Statistics Press.
4. Tan X, He Z, Fahey AG, et al. (2023) Research progress and applications of genome-wide association study in farm animals. *Anim Res One Health* 1: 56–77. https://doi.org/10.1002/aro2.14
5. Georges M, Charlier C, Hayes B (2019) Harnessing genomic information for livestock improvement. *Nat Rev Genet* 20: 135–156. https://doi.org/10.1038/s41576-018-0082-2

6. Warr A, Affara N, Aken B, et al. (2020) An improved pig reference genome sequence to enable pig genetics and genomics research. *GigaScience* 9: giaa051. https://doi.org/10.1093/gigascience/giaa051
7. Zong W, Chen L, Zhang D, et al. (2025) Two telomere-to-telomere pig genome assemblies and pan-genome analyses provide insights into genomic structural landscape and genetic adaptations. *iMeta* 4: e70013. https://doi.org/10.1002/imt2.70013
8. Luo Y, Huang N, Zha C, et al. (2026) Telomere-to-telomere genome assembly of a male pig provides insight into population structure and selection for body stature. *Nat Genet* 58: 195–205. https://doi.org/10.1038/s41588-025-02433-6
9. Cao C, Miao J, Xie Q, et al. (2025) A near telomere-to-telomere genome assembly of the Jinhua pig: enabling more accurate genetic research. *GigaScience* 14: giaf048. https://doi.org/10.1093/gigascience/giaf048
10. Paten B, Novak AM, Eizenga JM, et al. (2017) Genome graphs and the evolution of genome inference. *Genome Res* 27: 665–676. https://doi.org/10.1101/gr.214155.116
11. Ebert P, Audano PA, Zhu Q, et al. (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372: eabf7117. https://doi.org/10.1126/science.abf7117
12. Miao J, Wei X, Cao C, et al. (2024) Pig pangenome graph reveals functional features of non-reference sequences. *J Anim Sci Biotechnol* 15: 32. https://doi.org/10.1186/s40104-023-00984-4
13. Miao J, Wang Q, Zhang Z, et al. (2025) Pangenome graph mitigates heterozygosity overestimation from mapping bias: a case study in Chinese indigenous pigs. *BMC Biol* 23: 89. https://doi.org/10.1186/s12915-025-02194-y
14. Liu Y, Du H, Li P, et al. (2020) Pan-genome of wild and cultivated soybeans. *Cell* 182: 162–176.e13. https://doi.org/10.1016/j.cell.2020.05.023
15. Du H, Zhuo Y, Lu S, et al. (2024) Pangenome reveals gene content variations and structural variants contributing to pig characteristics. *Genomics Proteomics Bioinformatics* 22: qzae081. https://doi.org/10.1093/gpbjnl/qzae081
16. Jiang YF, Wang S, Wang CL, 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
17. Li D, Wang Y, Yuan T, et al. (2024) Pangenome and genome variation analyses of pigs unveil genomic facets for their adaptation and agronomic characteristics. *iMeta* 3: e257. https://doi.org/10.1002/imt2.257
18. Garrison E, Sirén J, Novak AM, et al. (2018) Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat Biotechnol* 36: 875–879. https://doi.org/10.1038/nbt.4227.
19. Hickey G, Monlong J, Ebler J, et al. (2024) Pangenome graph construction from genome alignments with Minigraph-Cactus. *Nat Biotechnol* 42: 663–673. https://doi.org/10.1038/s41587-023-01793-w
20. Garrison E, Guarracino A, Heumos S, et al. (2024) Building pangenome graphs. *Nat Methods* 21: 2008–2012. https://doi.org/10.1038/s41592-024-02430-3
21. Li H, Feng X, Chu C (2020) The design and construction of reference pangenome graphs with minigraph. *Genome Biol* 21: 265. https://doi.org/10.1186/s13059-020-02168-z
22. Wang S, Qian YQ, Zhao RP, et al. (2023) Graph-based pan-genomes: Increased opportunities in

plant genomics. *J Exp Bot* 74: 24–39. https://doi.org/10.1093/jxb/erac412
23. Li X, Liu Q, Fu C, et al. (2024) Characterizing structural variants based on graph-genotyping provides insights into pig domestication and local adaption. *J Genet Genomics* 51: 394–406. https://doi.org/10.1016/j.jgg.2023.11.005
24. Wu Z, Gui L, Hu J, et al. (2026) The influence of structural variants from 2445 pigs on gene expression and complex traits. *Nat Commun* 17: 9092. https://doi.org/10.1038/s41467-026-75525-4
25. Lin M, Liu L, Cai G, et al. (2026) SwinePan for pig graph-based pangenome and multiomics data mining. *Genome Res* 36: gr.281750.125. https://doi.org/10.1101/gr.281750.125
26. Li Z, Liu X, Wang C, 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
27. Liao WW, Asri M, Ebler J, et al. (2023) A draft human pangenome reference. *Nature* 617: 312–324. https://doi.org/10.1038/s41586-023-05896-x
28. Parmigiani L, Garrison E, Stoye J, et al. (2024) Panacus: Fast and exact pangenome growth and core size estimation. *Bioinformatics* 40: btae720. https://doi.org/10.1093/bioinformatics/btae720
29. Rhie A, McCarthy SA, Fedrigo O, et al. (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592: 737–746. https://doi.org/10.1038/s41586-021-03451-0
30. Simão FA, Waterhouse RM, Ioannidis P, et al. (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351
31. Bentley DR, Balasubramanian S, Swerdlow HP, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53–59. https://doi.org/10.1038/nature07517
32. Eid J, Fehr A, Gray J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323: 133–138. https://doi.org/10.1126/science.1162986
33. Wenger AM, Peluso P, Rowell WJ, et al. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 37: 1155–1162. https://doi.org/10.1038/s41587-019-0217-9
34. Jain M, Olsen HE, Paten B, et al. (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 36: 338–345. https://doi.org/10.1038/nbt.4060
35. Rakocevic G, Semenyuk V, Lee WP, et al. (2019) Fast and accurate genomic analyses using genome graphs. *Nat Genet* 51: 354–362. https://doi.org/10.1038/s41588-018-0316-4
36. Grytten I, Rand KD, Nederbragt AJ, et al. (2020) Assessing graph-based read mappers against a baseline approach highlights strengths and weaknesses of current methods. *BMC Genomics* 21: 282. https://doi.org/10.1186/s12864-020-6685-y
37. Guarracino A, Heumos S, Nahnsen S, et al. (2022) ODGI: understanding pangenome graphs. *Bioinformatics* 38: 3319–3326. https://doi.org/10.1093/bioinformatics/btac308
38. Cumer T, Milia S, Leonard AS, et al. (2026) PG-SCUnK: Measuring pangenome graph representativeness using single-copy and universal K-mers. *BMC Bioinf* 27: 29. https://doi.org/10.1186/s12859-025-06355-2
39. Zytnicki M (2025) Assessing genome conservation on pangenome graphs with PanSel. *Bioinform Adv* 5: vbaf018. https://doi.org/10.1093/bioadv/vbaf018

40. Dubois S, Zytnicki M, Lemaitre C, et al. (2025) Pairwise graph edit distance characterizes the impact of the construction method on pangenome graphs. *Bioinformatics* 41: btaf291. https://doi.org/10.1093/bioinformatics/btaf291
41. Kopalli V, Arslan K, Morales-Díaz N, et al. (2025) Toward a standardized framework for pangenome graph evaluation: assessing crop plant pangenome variation graph construction from multiple assemblies. *GigaScience* 14: giaf121. https://doi.org/10.1093/gigascience/giaf121
42. Quan C, Lu H, Lu Y, et al. (2022) Population-scale genotyping of structural variation in the era of long-read sequencing. *Comput Struct Biotechnol J* 20: 2639–2647. https://doi.org/10.1016/j.csbj.2022.05.047
43. Chen S, Krusche P, Dolzhenko E, et al. (2019) Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol* 20: 291. https://doi.org/10.1186/s13059-019-1909-7
44. Ahsan MU, Liu Q, Perdomo JE, et al. (2023) A survey of algorithms for the detection of genomic structural variants from long-read sequencing data. *Nat Methods* 20: 1143–1158. https://doi.org/10.1038/s41592-023-01932-w
45. Smolka M, Paulin LF, Grochowski CM, et al. (2024) Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* 42: 1571–1580. https://doi.org/10.1038/s41587-023-02024-y
46. Heller D, Vingron M (2019) SVIM: structural variant identification using mapped long reads. *Bioinformatics* 35: 2907–2915. https://doi.org/10.1093/bioinformatics/btz041
47. Jiang T, Liu Y, Jiang Y, et al. (2020) Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* 21: 189. https://doi.org/10.1186/s13059-020-02107-y
48. Liu Y, Jiang T, Su J, et al. (2021) SKSV: ultrafast structural variation detection from circular consensus sequencing reads. *Bioinformatics* 37: 3647–3649. https://doi.org/10.1093/bioinformatics/btab341.
49. Helal AA, Saad BT, Saad MT, et al. (2024) Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. *Sci Rep* 14: 6160. https://doi.org/10.1038/s41598-024-56604-2
50. Eggertsson HP, Kristmundsdottir S, Beyter D, et al. (2019) GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun* 10: 5402. https://doi.org/10.1038/s41467-019-13341-9
51. Sibbesen JA, Maretty L, Danish Pan-Genome Consortium, et al. (2018) Accurate genotyping across variant classes and lengths using variant graphs. *Nat Genet* 50: 1054–1059. https://doi.org/10.1038/s41588-018-0145-5
52. Ebler J, Ebert P, Clarke WE, et al. (2022) Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nat Genet* 54: 518–525. https://doi.org/10.1038/s41588-022-01043-w
53. Chang X, Novak AM, Eizenga JM, et al. (2025) Rapid, accurate long- and short-read mapping to large pangenome graphs with vg Giraffe. *bioRxiv*. https://doi.org/10.1101/2025.09.29.678807
54. Yang L, Yin H, Bai L, 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
55. Li X, Quan Y, Li D, et al. (2025) PSVRP: a pig structural variant reference panel for complex trait genomics and precision breeding. *BMC Genomics* 26: 1073. https://doi.org/10.1186/s12864-025-12273-7

56. Wang YJ, Wang XP, Yu L (2024) Research progress on dissecting complex animal traits based on genomic structural variations. (in Chinese) *Sci Sin Vitae* 54: 283–299. https://doi.org/10.1360/ssv-2023-0216
57. Rubin CJ, Megens HJ, Martinez Barrio A, et al. (2012) Strong signatures of selection in the domestic pig genome. *Proc Natl Acad Sci USA* 109: 19529–19536. https://doi.org/10.1073/pnas.1217149109
58. Zeng H, Zhang W, Lin Q, et al. (2024) PigBiobank: a valuable resource for understanding genetic and biological mechanisms of diverse complex traits in pigs. *Nucleic Acids* Res 52: D980–D989. https://doi.org/10.1093/nar/gkad1080
59. Teng J, Gao Y, Yin H, 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
60. Gong Y, Li Y, Liu X, et al. (2023) A review of the pangenome: how it affects our understanding of genomic variation, selection and breeding in domestic animals? *J Anim Sci Biotechnol* 14: 73. https://doi.org/10.1186/s40104-023-00860-1
61. Bayer PE, Golicz AA, Scheben A, et al. (2020) Plant pan-genomes are the new reference. *Nat Plants* 6: 914–920. https://doi.org/10.1038/s41477-020-0733-0
62. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74: 5463–5467. https://doi.org/10.1073/pnas.74.12.5463
63. Logsdon GA, Vollger MR, Eichler EE (2020) Long-read human genome sequencing and its applications. *Nat Rev Genet* 21: 597–614. https://doi.org/10.1038/s41576-020-0236-x
64. Li H, Durbin R (2024) Genome assembly in the telomere-to-telomere era. *Nat Rev Genet* 25: 658–670. https://doi.org/10.1038/s41576-024-00718-w
65. Shi L, Guo Y, Dong C, et al. (2016) Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* 7: 12065. https://doi.org/10.1038/ncomms12065
66. Skiena SS (2008) The Algorithm Design Manual. 2 Eds., London: Springer. https://doi.org/10.1007/978-1-84800-070-4
67. Compeau PEC, Pevzner PA, Tesler G (2011) Why are de Bruijn graphs useful for genome assembly? *Nat Biotechnol* 29: 987–991. https://doi.org/10.1038/nbt.2023
68. Marco-Sola S, Eizenga JM, Guarracino A, et al. (2023) Optimal gap-affine alignment in O(s) space. *Bioinformatics* 39: btad074. https://doi.org/10.1093/bioinformatics/btad074
69. Hickey G, Heller D, Monlong J, et al. (2020) Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* 21: 35. https://doi.org/10.1186/s13059-020-1941-7
70. Li H (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34: 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
71. Sedlazeck FJ, Rescheneder P, Smolka M, et al. (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
72. Sohn JI, Nam JW (2018) The present and future of de novo whole-genome assembly. *Brief Bioinform* 19: 23–40. https://doi.org/10.1093/bib/bbw096
73. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* 12: 363–376. https://doi.org/10.1038/nrg2958
74. Ameur A (2019) Goodbye reference, hello genome graphs. *Nat Biotechnol* 37: 866–868. https://doi.org/10.1038/s41587-019-0199-7
75. Ma H, Jiang J, He J, et al. (2022) Long-read assembly of the Chinese indigenous Ningxiang pig

genome and identification of genetic variations in fat metabolism among different breeds. *Mol Ecol Resour* 22: 1508–1520. https://doi.org/10.1111/1755-0998.13550

76. Du H, Zheng X, Zhao Q, et al. (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
77. Yan H, Sun M, Zhang Z, et al. (2023) Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. *Nat Genet* 55: 507–518. https://doi.org/10.1038/s41588-023-01302-4
78. Yang R, Fang S, Wang J, et al. (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
79. Gong H, Liu W, Wu Z, et al. (2022) Evolutionary insights into porcine genomic structural variations based on a novel-constructed dataset from 24 worldwide diverse populations. *Evol Appl* 15: 1264–1280. https://doi.org/10.1111/eva.13455
80. Liu C, Ran X, Wang J, et al. (2018) Detection of genomic structural variations in Guizhou indigenous pigs and the comparison with other breeds. *PLOS One* 13: e0194282. https://doi.org/10.1371/journal.pone.0194282
81. Tang S, Li D, Lu Y, et al. (2026) Integrative advances in pig genomics: from reference assemblies and evolutionary history to the mechanistic dissection of key traits. *Biology* 15: 447. https://doi.org/10.3390/biology15050447



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)