REVIEW

REVISED White paper: standards for handling and analyzing plant pan-genomes

[version 2; peer review: 2 approved, 2 approved with reservations]

Marc C. Heuermann[1], Pedro Barros[2], Sebastian Beier[3], Heidrun Gundlach[4], Jorge Alvarez-Jarreta[5], Keywan Hassani-Pak[6], Patrick König[1], Anne Fiebig[1], Tim Godec[7], Kristina Gruden[7], Nadja Nolte[7], Marko Petek[7], Uwe Scholz[1], Maja Zagorščak[7], Klaas Vandepoele[8], Michiel Van Bel[8]

[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Saxony-Anhalt, 06466, Germany
[2]Universidade Nova de Lisboa Instituto de Tecnologia Quimica e Biologica, Oeiras, Lisbon, Portugal
[3]Forschungszentrum Jülich GmbH Institute of Bio- and Geosciences, Jülich, North Rhine-Westphalia, 52425, Germany
[4]Helmholtz Zentrum München, Neuherberg, 85764, Germany
[5]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
[6]Rothamsted Research, Harpenden, England, AL52JQ, UK
[7]National Institute of Biology, Večna pot 111, Ljubljana, 1000, Slovenia
[8]Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 71, Ghent, 9052, Belgium

v2 **First published:** 28 Jul 2025, **14**:739
https://doi.org/10.12688/f1000research.166538.1

Latest published: 18 Nov 2025, **14**:739
https://doi.org/10.12688/f1000research.166538.2

Abstract

Plant pan-genomes, which aggregate genomic sequences and annotations from multiple individuals of a species, have emerged as transformative tools for understanding genetic diversity, adaptation, and evolutionary dynamics. Super-pan-genomes, extending across species boundaries, further enable comparative analyses of clades or genera, bridging breeding applications with evolutionary insights (Shang et al., 2022; Li et al., 2023a). However, the absence of standardized practices for data generation, analysis, and sharing hinders reproducibility and interoperability. This white paper presents a harmonized framework developed by the ELIXIR E-PAN consortium, addressing nomenclature, quality control (QC), data formats, visualization, and community practices. By adopting these guidelines, researchers can enhance FAIR (Findable, Accessible, Interoperable, Reusable) compliance, foster collaboration, and accelerate translational applications in crop improvement and evolutionary biology.

Keywords

plant pan-genome, white paper, standards, quality control

Open Peer Review

Approval Status ✓ ✓ ? ?

	1	2	3	4
version 2 (revision) 18 Nov 2025	✓ view	✓ view	? view	? view
	↑	↑		
version 1 28 Jul 2025	? view	? view		

1. **Sunil Kumar Sahu**, State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen, China

2. **Xiaoming Xie**, China Agricultural University College of Agronomy and Biotechnology (Ringgold ID: 200630), Beijing, China

3. **Rutwik Barmukh**, Murdoch University, Murdoch, Australia

4. **Jianping Xu**, McMaster University, Hamilton, Canada

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the ELIXIR gateway.

This article is included in the Plant Science gateway.

This article is included in the Genomics and Genetics gateway.

Corresponding author: Marc C. Heuermann (marcheuermann@ipk-gatersleben.de)

Author roles: Heuermann MC: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Barros P**: Writing – Review & Editing; **Beier S**: Writing – Review & Editing; **Gundlach H**: Writing – Review & Editing; **Alvarez-Jarreta J**: Writing – Review & Editing; **Hassani-Pak K**: Writing – Review & Editing; **König P**: Writing – Review & Editing; **Fiebig A**: Writing – Review & Editing; **Godec T**: Writing – Review & Editing; **Gruden K**: Writing – Review & Editing; **Nolte N**: Writing – Review & Editing; **Petek M**: Writing – Review & Editing; **Scholz U**: Writing – Review & Editing; **Zagorščak M**: Writing – Review & Editing; **Vandepoele K**: Project Administration, Writing – Review & Editing; **Van Bel M**: Project Administration, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This study has received funding from ELIXIR under the call for proposals on Biodiversity, Food Safety and Pathogens (2024-SCIENCE-BFSP) as part of WP2 E-PAN: Enhancing pan-genome analysis in plants. MCH received funding from ELIXIR-DE, which was supported by the Federal Ministry of Education and Research BMBF within the framework of de.NBI/ELIXIR-DE (W-de.NBI-009).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Heuermann MC, Barros P, Beier S *et al.* **White paper: standards for handling and analyzing plant pan-genomes [version 2; peer review: 2 approved, 2 approved with reservations]** F1000Research 2025, **14**:739 https://doi.org/10.12688/f1000research.166538.2

First published: 28 Jul 2025, **14**:739 https://doi.org/10.12688/f1000research.166538.1

REVISED **Amendments from Version 1**

The reviewers' comments have helped us significantly refine the white paper. The new figure provides readers with a clear, visualized, and summarized overview of the essential steps required to successfully conduct a pan-genome analysis. These steps encompass quality control, annotation, pan-genome construction and analysis, and visualization throughout a pan-genome project, with exemplary tools highlighted for each stage. Furthermore, we have substantially expanded the scope of the chapters on quality control standards, data formats and sharing, visualization and analysis guidelines, and case studies.

Any further responses from the reviewers can be found at the end of the article

1. Introduction

Pan-genomes capture both core genomic elements (shared across individuals) and accessory components (variable or unique to subsets), offering unprecedented resolution for studying traits such as disease resistance, environmental adaptation, and domestication (Qin et al., 2021; Zhou et al., 2022). Super-pan-genomes, which span multiple species, provide evolutionary context for gene family dynamics and speciation events, as demonstrated in clades like Brassicaceae (Jiao & Schneeberger, 2020) and Solanaceae (Alonge et al., 2022). In plant genomics, pan-genomes are vital for understanding genetic diversity, adaptation, and evolutionary dynamics, particularly given the extensive variation observed in plant species (Schreiber et al., 2024). Despite their potential, inconsistencies in data management—such as ad hoc naming conventions, variable QC practices, and fragmented repository use—limit cross-study comparisons and data reuse.

The ELIXIR E-PAN consortium synthesizes insights from foundational studies on barley (*Hordeum vulgare*), rice (*Oryza sativa*), tomato (*Solanum lycopersicum*), and Arabidopsis (*Arabidopsis thaliana*) to propose actionable standards. These guidelines aim to unify the plant genomics community, ensuring robust, interoperable resources for breeding and evolutionary research.

2. Naming conventions and ontologies

2.1 Accession and assembly identifiers

Accession naming should adhere to MIAPPE (Minimum Information About Plant Phenotyping Experiments) standards. The Biological Material ID should incorporate institutional identifiers, followed by the accession number from germplasm catalogue or common name of the plant source/variety (e.g., IPK-Gatersleben:HOR_13170 for barley accession "Barke") to ensure traceability (MIAPPE v1.1, Papoutsoglou et al., 2020). When complementary data regarding a specific accession is also available at external sources (e.g. Biosamples), a link to a Biological material external ID should be provided in the metadata.

Genome assembly identifiers should contain at least 4 fields—species, variety/line, project group, assembly version — separated by period ('.'), with an optional fifth field for additional information (Cannon et al., 2025). For example, drOrySati.Nipponbare.RicePan.1.0, which refers to the assembly of *Oryza sativa*, Nipponbare cultivar, RicePan project, version 1.0 (ToLID identifier, https://id.tol.sanger.ac.uk/, Darwin Tree of Life Consortium, 2023).

2.2 Gene identifiers

Gene identifiers must balance stability with biological relevance, as outlined by Cannon et al. (2025), keeping track of the annotation version, chromosome and gene ID. Their framework proposes human- and machine-readable identifiers, including the assembly names (e.g. drOrySati.Nipponbare.RicePan.1.0) with the addition of gene models like drOrySati.Nipponbare.RicePan.1.0.1.01.g000100 (assembly version 1.0, annotation version 1, chromosome 01, gene 100). To enhance this for pan-genomics, the "group" field can denote pan-genome projects (e.g., RicePan), linking multiple assemblies, while optional fields like "Hap1" or metadata tags distinguish haplotypes or accession types (e.g., wild vs. cultivated). Pangenes, representing orthologous gene clusters, can be assigned identifiers like drOrySati.RicePan.pan00001, with metadata linking to specific gene models across assemblies. Cannon et al. (2025) advocate preserving legacy identifiers via cross-references to ensure stability, avoiding disruptive renaming as new accessions are added.

2.3 Metadata and ontologies

A core metadata schema is critical for interoperability. Required fields to properly annotate pan-genome studies include species details such as name (TaxonID), pedigree, geographic origin, ploidy and chromosome number, as well as sequencing technology used (e.g. PacBio HiFi, Oxford Nanopore, Hi-C, Illumina), assembly pipelines (e.g., Flye (Kolmogorov et al., 2020), hifiasm (Cheng et al., 2024), Canu (Koren et al., 2017), …), and assembly QC metrics (e.g., BUSCO scores (Manni et al., 2021)). Existing ontologies such as the Sequence Ontology (SO) should be extended to include pan-genome-specific terms that describe the layouts and structures of pan-genomes (Eilbeck et al., 2005). These can be categorized as core, shell and cloud genome genes, but these terms may depend on the number of genomes and genotypes selected (Jayakodi et al., 2024). Any downstream comparative analysis requires open and transparent reporting on the thresholds used, so that these must be included in the metadata. Collaboration with the AgBioData Nomenclature Working Group and the Genomics Standards Consortium (https://www.gensc.org/) ensures alignment with broader genomic standards (Cannon et al., 2025).

2.4 Generalized feature identification

As pan-genome graphs grow to encompass not just core and variable genes but a full spectrum of genomic elements, we need a unified identification system. Current annotation often focuses on genes, leaving features like transposable elements, SSRs, non-coding RNAs, and regulatory motifs with inconsistent or tool-specific labels. We propose the development of a **generalized feature identifier (GFI)**. This system would provide a stable, queryable, and standardized format for any annotated feature, independent of its type or the discovery tool used. A GFI would be important for pan-genome-scale association studies and for functionally characterizing the entire "dark matter" of the genome, ensuring that a SNP in a long terminal repeat or a copy number variation in a novel ncRNA can be cataloged and compared with the same rigor as in a protein-coding gene.

3. Quality Control (QC) standards

3.1 Sequencing and assembly QC

Quality control in genome assembly workflows begins with sequencing QC, where tools like FASTQC assess raw read integrity, including base quality, GC content, and adapter contamination (Figure 1A). K-mer plots, generated via Jellyfish (Marçais et al., 2011) paired with GenomeScope 2 (Ranallo-Benavidez et al., 2020), provide insights into genome complexity, such as ploidy, heterozygosity, and repetitive element profiles (Figure 1A). For individual assembly QC, QUAST (Mikheenko et al., 2023) is recommended for evaluating contiguity metrics (e.g., N50, L50) and is particularly effective for comparing multiple assemblies of diploid genomes, while CRAQ (Li et al., 2023) excels in assessing consensus accuracy and structural errors in polyploid genomes due to its sensitivity to haplotype-specific misassemblies (Figure 1A). When results from QUAST and CRAQ conflict (e.g., differing contig counts due to haplotype collapsing), users should prioritize CRAQ for polyploid assemblies and cross-validate with raw read alignments (e.g., using Minimap2) to resolve discrepancies. Merqury (Rhie et al., 2020) further validates haplotype resolution in polyploid or heterozygous genomes (e.g., wheat, potato) by comparing k-mer spectra between raw reads and assemblies, offering a robust check for completeness and phasing errors (Figure 1A). For repeat quality control, the LTR Assembly Index (LAI; Ou et al., 2018) assesses the completeness of long terminal repeat retrotransposons, while *tidk* (Brown et al., 2025) detects telomeric motifs to evaluate chromosomal end-to-end integrity (Figure 1A). When results from these tools conflict, LAI generally provides a more reliable indicator of assembly quality. Even in high-quality plant genomes assembled from long reads, some chromosome ends may still lack detectable telomeric repeats.

3.2 Annotation QC

Annotation pipelines must be documented alongside assembly strategies. These may include gene model integration pipelines like MAKER2 (Holt & Yandell, 2011), PASA (Haas et al., 2003) or BRAKER3 (Gabriel et al., 2024), while Helixer (Stiehler et al., 2020) is recommended for ab initio prediction in non-model organisms due to its deep learning-based approach (Figure 1B). Liftoff (Shumate & Salzberg, 2021) is ideal for annotation transfer between closely related species and should be part of a standard annotation pipeline (Figure 1B). Use versioned workflows (e.g., Snakemake (Köster & Rahmann, 2012), or Nextflow (Di Tommaso et al., 2017)) to ensure reproducibility, provenance tracking, and portability. Transcriptomic data (RNA-Seq) from multiple tissues (e.g., roots, shoots) and stress conditions (e.g., drought, disease) with sufficient read coverage validates gene models, especially for accessory genes lacking orthologs (Qin et al., 2021). Long-read RNA sequencing technologies are recommended to recover full-length transcripts and accurately characterize alternative isoforms. For structural annotation QC, BUSCO (Manni et al., 2021) assesses gene space completeness using lineage-specific datasets that can be adjusted for polyploid genomes (Figure 1B). Mercator4 (Bolger et al., 2021) assigns functional categories based on the MapMan bin system and is useful for identifying missing functions in a single genome (Figure 1B). PSAURON (Sommer et al., 2025) validates structural annotations, and OMArk (Nevers et al., 2025) detects contamination via evolutionary consistency checks (Figure 1B). In cases where evaluation tools disagree (e.g., BUSCO reports missing genes but PSAURON suggests completeness), integrating RNA-Seq support and orthology evidence provides a more reliable basis for resolving such discrepancies.

3.3 Pan-genome-specific QC

Pan-genome completeness requires saturation analysis, where gene accumulation curves assess whether additional accessions contribute novel genes (Tettelin et al., 2005). For species with varying ploidy levels (e.g., diploid vs. polyploid barley), a minimum of 10–20 accessions is typically required for diploid species to approach saturation, while polyploid species may need 30–50 accessions due to increased gene content complexity (Jayakodi et al., 2024). Users should plot

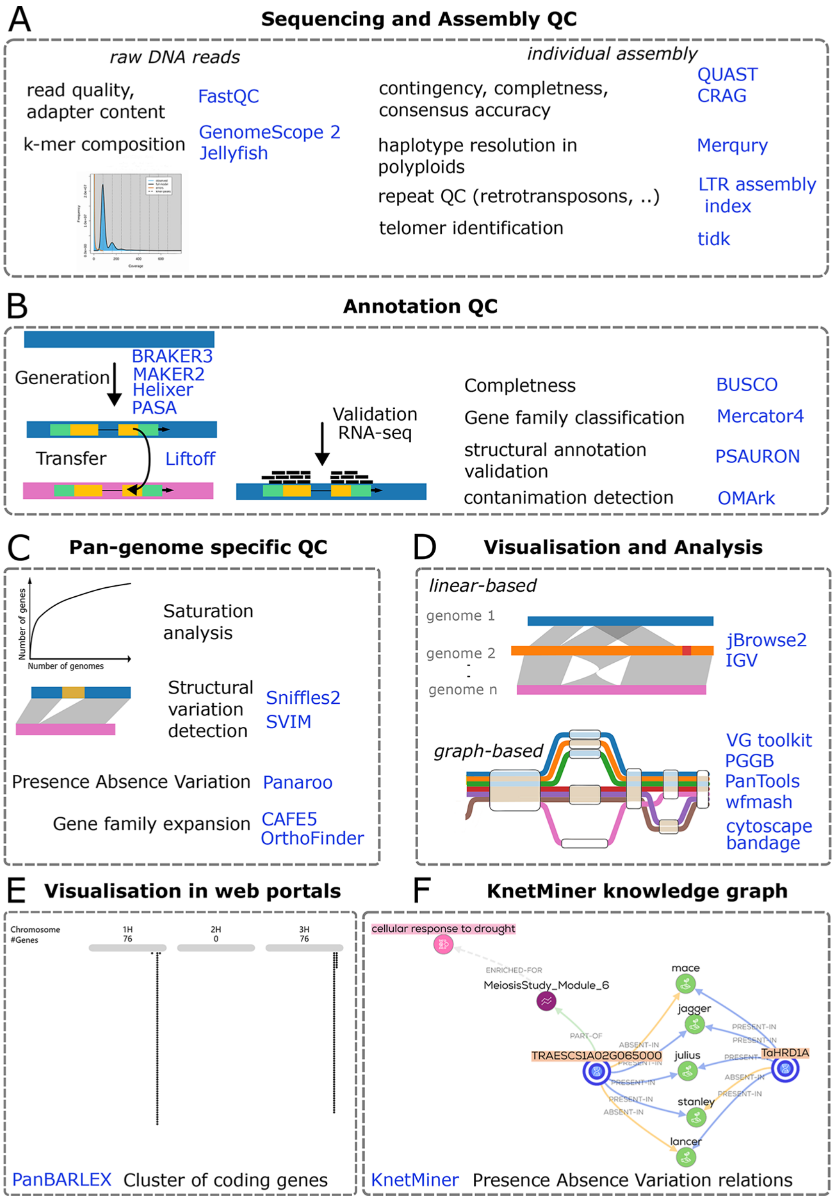

Figure 1. Overview of quality control, annotation, pan-genome analysis, and visualization steps across a -pan-genome project, with example tools highlighted. A, Sequencing and assembly QC. Raw DNA reads are screened for base quality, adapter contamination, and k-mer composition using **FastQC**, **GenomeScope 2**, and **Jellyfish**. Individual assemblies are evaluated for contiguity, completeness, and consensus accuracy with **QUAST** and **CRAG**; haplotype resolution in polyploids with **Merqury**; repeat content and assembly of long terminal repeat retrotransposons with **LTR assembly index**; and telomere identification with **tidk**. **B**, Annotation QC. Gene models are generated and refined with **BRAKER3**, **MAKER2**, **Helixer**, and **PASA**, and can be transferred between assemblies using Liftoff. Validation incorporates RNA-seq support and summary metrics including gene set completeness with **BUSCO**, gene family classification with **Mercator4**, structural annotation validation with **PSAURON**, and contamination detection with **OMArk**. **C**, Pan-genome–specific QC and discovery. Across multiple genomes, analyses include gene accumulation and saturation behavior, detection of structural variants with **Sniffles2** and **SVIM**, assessment of presence–absence variation with **Panaroo**, and tests for gene family expansion or contraction with **CAFE5** and **OrthoFinder**. **D**, Visualization and comparative analysis. Linear genome browsers support side-by-side inspection of assemblies and annotations (**jBrowse2**, **IGV**). Graph-based frameworks represent shared and alternative haplotypes and enable mapping and variant interrogation across many genomes (**VG toolkit**, **PGGB**, **PanTools**, **wfmash**), complemented by network and assembly graph viewers (**cytoscape**, **bandage**). **E**, Pre-rendered web portals. Project-specific portals provide searchable tracks and summary plots for community access, exemplified by **PanBARLEX**, (https://panbarlex.ipk-gatersleben.de/#seqcluster/BarleyCDS90_02985). **F**, Presence absence variation (PAV) relations shown in knowledge graphs produced by **KnetMiner**. Dashed boxes delineate workflow stages; icons are schematic. The listed software represents commonly used options and is not exhaustive. Abbreviations: QC, quality control; RNA-seq, RNA sequencing; PAV, presence–absence variation.

accumulation curves using tools like Panaroo and evaluate saturation by fitting models (e.g., Heap's Law) to confirm diminishing returns in gene discovery (Figure 1C). For species like barley, benchmark datasets of 100+ conserved genes enable orthology tool validation (Jayakodi et al., 2024). OrthoFinder (Emms & Kelly, 2019) and CAFE5 (Mendes et al., 2020) facilitate gene family expansion and contraction analyses, providing insights into evolutionary dynamics (Figure 1C). Structural variant detection, using Sniffles2 (Smolka et al., 2024) for long-read data or SVIM (Heller & Vingron, 2019) for short-read data, quantifies indels and inversions (Qin et al., 2021) (Figure 1C). When tools like Sniffles2 and SVIM yield conflicting variant calls, users should integrate multi-platform data (e.g., combining long- and short-read alignments) and prioritize calls supported by higher read depth or mapping quality. Presence-absence variation (PAV) detection via Panaroo or PAV-specific pipelines is critical for identifying variable gene content tied to phenotypic diversity (Tonkin-Hill et al., 2020).

4. Data formats and sharing

4.1 File formats

- **Raw data**: Assemblies must be submitted in FASTA format with headers containing unique sequence identifiers (e.g., >chr01, >chr02). Annotations must be provided in GFF3 or GTF format (compliant with Sequence Ontology), with the sequence IDs in the first column exactly matching the sequence identifiers used in the FASTA headers.

- **Derived data**: Structural variants in VCF/BCF, orthogroups in TSV (cluster ID + member gene), and graph-based representations (GFA format) for complex pan-genomes (Li et al., 2020).

4.2 Repositories

Centralized repositories would archive versioned datasets (e.g., Barley v2, Rice v1.5) with DOI-based identifiers (DataCite). Public deposition in INSDC (raw reads and assembly, https://www.insdc.org/) and Ensembl (annotations, see documentation of Ensembl, 2025, https://beta.ensembl.org/) ensures global accessibility (ENA Documentation, 2025).

4.3 Metadata requirements

Mandatory metadata fields include sequencing technology and coverage (e.g., PacBio HiFi, Oxford Nanopore), assembly method (e.g., Flye, Hifiasm), accession provenance (BioSample IDs), and software versioning of all software and pipelines used. Missing metadata, as observed in early barley submissions, must be addressed via enforced submission guidelines (Jayakodi et al., 2024).

For pangenome datasets, additional metadata fields are critical to ensure traceability and interoperability across studies. These should include the species name and NCBI Taxonomy ID, pangenome version and build date, and a complete list of constituent genomes with corresponding assembly accessions, strain names, and versions. Furthermore, metadata should describe the methods and parameters used to construct the pangenome.

Capturing this information in structured formats such as JSON-LD or RO-Crate (Peroni et al., 2022) would align pangenome submissions with broader FAIR data principles and facilitate integration with knowledge graphs and comparative genomics resources.

5. Visualization and analysis guidelines

5.1 Visualization tools

Plant pan-genomes capture a species' full genomic diversity, constructed using either linear-based or graph-based methods, each with distinct strengths and limitations. To provide a clearer comparison, linear-based approaches are divided into two distinct categories: sequence-based and gene-based analyses.

Sequence-based linear analysis involves aligning multiple genomes to a single reference or consensus sequence to identify sequence-level variations, such as single-nucleotide polymorphisms (SNPs) and insertions/deletions (indels). This process typically employs variant callers like GATK (McKenna et al., 2010) or freebayes (Garrison & Marth, 2012) to detect SNPs and indels from whole-genome alignments. These methods are computationally efficient and compatible with visualization tools like JBrowse2 (Diesh et al., 2023) or IGV (Robinson et al., 2023) for synteny and variant visualization (Figure 1D). Web-portals such as PanBARLEX (PanBARLEX - Barley Pangenome Explorer) enable pan-genome research by providing searchable and pre-rendered visualizations (Figure 1E). However, reference bias in sequence-based linear approaches can limit their ability to capture complex structural variations, particularly in repetitive or polyploid plant genomes.

Gene-based linear analysis focuses on inferring orthology and identifying gene-level presence/absence variations (PAVs) using tools like OrthoFinder (Emms & Kelly, 2019) or Ensembl Compara (Dyer et al., 2025). These tools analyze annotated gene sets to determine the pan-gene repertoire, identifying core and accessory genes across a species. While effective for gene-level PAV detection, these methods do not directly address sequence-level variations like SNPs or indels, requiring separate workflows for comprehensive analysis. Orthology inference tools must be benchmarked using inflation value sweeps to minimize false positives (Emms & Kelly, 2019). Visualization of gene-level PAVs can be achieved through UpSet plots or as presence/absence relationships in KnetMiner knowledge graphs (Hassani-Pak et al., 2021) (Figure 1F).

In contrast, **graph-based approaches** model genomes as interconnected nodes (shared regions) and edges (SNPs, indels, and structural variants) using tools like VG Toolkit (Hickey et al., 2020), PGGB (Garrison et al., 2024), PanTools (Jonkheer et al., 2022), or wfmash (Guarracino et al., 2021) (Figure 1D). These methods integrate both sequence-level and structural variations in a single framework, offering an unbiased, comprehensive view of genomic diversity. They are particularly suited for complex genomes, such as tomato (Zhou et al., 2022). Visualization tools like Bandage (Wick et al., 2015) or Cytoscape (Shannon et al., 2003) are used to represent structural complexity, though these approaches are computationally intensive and require specialized expertise (Figure 1D).

In summary, sequence-based linear methods excel in rapid SNP and indel detection but are limited by reference bias, while gene-based linear methods are ideal for pan-gene analysis but require separate homology-based workflows. Graph-based approaches unify both gene-level and structural variation analyses, offering greater flexibility for complex genomes despite higher computational demands. As computational resources and tools advance, graph-based methods are becoming more accessible, enhancing plant pan-genome studies as demonstrated in rice (Qin et al., 2021).

5.2 Integrative analysis best practices

The integration of pangenomic information into crop improvement remains challenging, despite its potential to illuminate the genetic basis of agronomic traits. Pangenomes reveal extensive structural polymorphisms and gene content diversity across accessions, yet these findings often remain siloed from other key data sources such as GWAS and QTL mappings, gene expression profiles, gene regulation, functional annotations, and published literature. Without coherent integration, researchers face difficulties in linking genomic variation to phenotype and in distinguishing biologically meaningful signals from background noise. Bridging these data types requires frameworks capable of harmonizing heterogeneous evidence, tracking provenance, and enabling transparent reasoning across molecular, phenotypic, and bibliographic domains.

Platforms such as KnetMiner (Hassani-Pak et al., 2021, https://knetminer.com) address these challenges by synthesizing pangenomic, association, omics, and literature-derived evidence within a unified knowledge graph. This integrative approach allows relationships among genes, traits, and pathways to be explored in context, supporting AI-assisted hypothesis generation and candidate gene prioritization. By providing explainable connections between diverse evidence sources, KnetMiner exemplifies how knowledge graph technologies can transform FAIR yet fragmented genomic data into a coherent foundation for evidence-based crop breeding.

6. Case studies

Barley Pan-genome (Jayakodi et al., 2024): The IPK barley pan-genome, encompassing 76 accessions, faced significant challenges in diploid genome assembly due to the crop's complex genetic structure. The adoption of automated quality control (QC) pipelines, implemented via Snakemake (Köster & Rahmann, 2012) and, alongside validation gene sets, was critical to ensuring reproducibility and accuracy. These standardized tools mitigated errors from manual curation, which previously led to inconsistent gene annotations across accessions. By streamlining QC processes, the project achieved robust assembly outcomes, enabling reliable downstream analyses for barley breeding programs. Without such standards, the project risked fragmented datasets, highlighting the necessity of automation for handling complexity.

Rice Pan-genome (Qin et al., 2021): Analysis of 31 rice accessions using Sniffles revealed hidden structural variations critical for understanding genetic diversity. However, the absence of standardized QC metrics initially led to discrepancies in variant calling, complicating comparisons across accessions. The project's success in identifying novel variations was enhanced by post-hoc implementation of rigorous QC protocols, which improved variant validation and reproducibility. This case underscores the need for predefined, community-wide QC standards to ensure consistency in pan-genome analyses, as their absence delayed insights into rice diversity and potential breeding applications.

Tomato Super-Pan-genome (Zhou et al., 2022): The tomato super-pan-genome, comprising 838 genomes, utilized a graph-based representation to resolve complex structural variants, directly informing breeding strategies for disease resistance. The adoption of standardized graph-based assembly tools ensured accurate representation of genetic diversity, overcoming limitations of linear reference genomes. This standardized approach facilitated the identification of novel resistance genes, significantly advancing breeding outcomes. Without such standards, the project could have faced misassembled variants, reducing its utility for applied breeding. This case exemplifies how standardized frameworks enhance the resolution of complex genomic data for practical applications.

Arabidopsis (Jiao & Schneeberger 2020; Zhong et al., 2025): Annotation gene naming and transfer across Arabidopsis MAGIC founders using Liftoff achieved cross-accession consistency. The use of standardized annotation pipelines ensured accurate gene mapping, enabling robust multi-omic and pan-genomic comparisons. This standardization was pivotal in identifying functional genomic variations within the population, supporting downstream genetic studies. In contrast, earlier Arabidopsis pan-genome efforts lacking such standardized tools faced annotation inconsistencies, which hindered comparative analyses. This case highlights how standardized naming conventions and annotation transfer tools like Liftoff are essential for ensuring reliable and reproducible pan-genomic insights.

7. Future directions

- **Artificial Intelligence**: Tools like DeepVariant (Poplin et al., 2018) will enhance variant calling in polyploid genomes. Detection of other genomic features, such as repeat elements, regulatory elements, and binding sites, will be enabled and refined using foundational models, as demonstrated in recent high-impact studies. For instance, BigRNA predicts tissue-specific RNA expression and identifies regulatory elements like microRNA and protein binding sites with high accuracy (Celaj et al., 2023). Similarly, Evo 2 detects transcription factor binding sites and exon-intron boundaries across diverse genomes (Brixi et al., 2025), while models like DNABERT (Ji et al., 2021) and Enformer (Avsec et al., 2021) excel in promoter prediction and variant effect analysis (Li et al., 2024). These advancements highlight the transformative potential of foundational models in refining genomic feature detection, particularly for complex polyploid genomes.

- **Cross-species standards**: Develop clade-wide frameworks (e.g., Brassicaceae) to unify super-pan-genome analyses.

- **Community engagement**: ELIXIR hackathons will refine workflows and ontology terms, ensuring adaptability to technological advances.

8. Conclusion

This white paper establishes a community-driven framework for plant pan-genome research. By adopting these guidelines, researchers can ensure data interoperability, reproducibility, and translational impact. The E-PAN consortium calls for global collaboration to iteratively refine these standards, fostering innovation in plant genomics and breeding.

Endorsed by ELIXIR Nodes: DE, BE, PT, SI, UK.

Contact: elixir-epan@elixir-europe.org

Data availability

No data is associated with this article.

Acknowledgments

The E-PAN consortium acknowledges contributions from researchers at ELIXIR nodes and foundational studies in rice, barley, tomato, and Arabidopsis. AI tools (DeepSeek R1, Qwen QwQ 32B) hosted on https://chat-ai.academiccloud.de helped create the draft from multiple meeting notes, with thorough human oversight ensuring scientific accuracy.

For updates, visit the *ELIXIR Plant Sciences Community Portal*.

References

Alonge M, *et al.*: **Major impacts of widespread structural variation on gene expression and crop improvement in tomato.** *Nature.* 2022; **606**: 527–534.
PubMed Abstract | Publisher Full Text | Free Full Text

Avsec Ž, *et al.*: **Effective gene expression prediction from sequence by integrating long-range interactions.** *Nat. Methods.* 2021; **18**: 1196–1203.
PubMed Abstract | Publisher Full Text | Free Full Text

Bolger M, *et al.*: **MapMan Visualization of RNA-Seq Data Using Mercator4 Functional Annotations.** Dobnik D, Gruden K, Ramšak Ž, *et al.*, editors. *Solanum tuberosum. Methods in Molecular Biology.* New York, NY: Humana; 2021; vol. **2354**.
Publisher Full Text

Brixi G, *et al.*: **Genome modeling and design across all domains of life with Evo 2.** *bioRxiv.* 2025.
Publisher Full Text

Brown MR, *et al.*: **tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets Open Access.** *Bioinformatics.* February 2025; **41**(2): btaf049.
PubMed Abstract | Publisher Full Text | Free Full Text

Cannon EKS, *et al.*: **Guidelines for gene and genome assembly nomenclature.** *Genetics.* 2025; **229**(3).
PubMed Abstract | Publisher Full Text | Free Full Text

Celaj A, *et al.*: **An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics.** *bioRxiv.* 2023.
Publisher Full Text

Cheng H, *et al.*: **Scalable telomere-to-telomere assembly for diploid and polyploid genomes with double graph.** *Nat. Methods.* 2024; **21**: 967–970.
PubMed Abstract | Publisher Full Text | Free Full Text

Darwin Tree of Life Consortium: **The Darwin Tree of Life project: Sequencing all life in Britain and Ireland.** 2023.
Reference Source

Diesh C, *et al.*: **JBrowse2: A modular genome browser with next-generation data support.** *Genome Biol.* 2023; **24**(74): 74.
PubMed Abstract | Publisher Full Text | Free Full Text

Di Tommaso P, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat. Biotechnol.* 2017; **35**: 316–319.
PubMed Abstract | Publisher Full Text

Dyer S, *et al.*: **Ensembl 2025.** *Nucleic Acids Res.* 6 January 2025; **53**(D1): D948–D957.
PubMed Abstract | Publisher Full Text | Free Full Text

Eilbeck K, *et al.*: **The Sequence Ontology: A tool for the unification of genome annotations.** *Genome Biol.* 2005; **6**(R44): R44.
PubMed Abstract | Publisher Full Text | Free Full Text

Emms DM, Kelly S: **OrthoFinder: Phylogenetic orthology inference for comparative genomics.** *Genome Biol.* 2019; **20**(238): 238.
PubMed Abstract | Publisher Full Text | Free Full Text

Gabriel L, *et al.*: **BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA.** *Genome Res.* 2024; **34**(34): 769–777.
PubMed Abstract | Publisher Full Text | Free Full Text

Garrison E, *et al.*: **Building pangenome graphs.** *Nat. Methods.* 2024; **21**: 2008–2012.
PubMed Abstract | Publisher Full Text

Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** arXiv preprint arXiv:1207.3907 [q-bio.GN] 2012. 2012.
Publisher Full Text

Guarracino A, *et al.*: *wfmash: whole-chromosome pairwise alignment using the hierarchical wavefront algorithm.* GitHub; 2021.
Reference Source

Haas B-J, *et al.*: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Res.* 1 October 2003; **31**(19): 5654–5666.
PubMed Abstract | Publisher Full Text | Free Full Text

Hassani-Pak K, *et al.*: **KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species.** *Plant Biotechnol. J.* 2021; **19**: 1670–1678.
PubMed Abstract | Publisher Full Text | Free Full Text

Heller D, Vingron M: **SVIM: structural variant identification using mapped long reads Open Access.** *Bioinformatics.* September 2019; **35**(17): 2907–2915.
PubMed Abstract | Publisher Full Text | Free Full Text

Hickey G, *et al.*: **Genotyping structural variants in pangenome graphs using the vg toolkit.** *Genome Biol.* 2020; **21**: 35.
PubMed Abstract | Publisher Full Text | Free Full Text

Holt C, Yandell M: **MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.** *BMC Bioinformatics.* 2011; **12**: Article number: 491.
PubMed Abstract | Publisher Full Text | Free Full Text

Jayakodi M, *et al.*: **The barley pan-genome reveals genomic diversity across wild and cultivated accessions.** *Nature.* 2024; **636**: 654–662.
PubMed Abstract | Publisher Full Text | Free Full Text

Ji Y, *et al.*: **DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome Free.** *Bioinformatics.* August 2021; **37**(15): 2112–2120.
PubMed Abstract | Publisher Full Text | Free Full Text

Jiao W-B, Schneeberger K: **Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics.** *Nat. Commun.* 2020; **11**(989): 989.
PubMed Abstract | Publisher Full Text | Free Full Text

Jonkheer EM, *et al.*: **PanTools v3: Functional analysis of prokaryotic pangenomes.** *Bioinformatics.* 2022; **38**(18): 4403–4405.
PubMed Abstract | Publisher Full Text | Free Full Text

Kolmogorov M, *et al.*: **metaFlye: scalable long-read metagenome assembly using repeat graphs.** *Nat. Methods.* 2020; **17**: 1103–1110.
PubMed Abstract | Publisher Full Text | Free Full Text

Koren S, *et al.*: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.** *Genome Res.* 2017; **27**: 722–736.
PubMed Abstract | Publisher Full Text | Free Full Text

Köster J, Rahmann S: **Snakemake—A scalable bioinformatics workflow engine.** *Bioinformatics.* 2012; **28**(19): 2520–2522.
PubMed Abstract | Publisher Full Text

Li H, *et al.*: **The design and construction of reference pangenome graphs.** *Genome Biol.* 2020; **21**(265): 265.
PubMed Abstract | Publisher Full Text | Free Full Text

Li N, *et al.*: **Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species.** *Nat. Genet.* 2023a; **55**(8): 852–860.
PubMed Abstract | Publisher Full Text | Free Full Text

Li K, *et al.*: **Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement.** *Nat. Commun.* 2023b; **14**: 6556.
PubMed Abstract | Publisher Full Text | Free Full Text

Li Q, *et al.*: **Progress and opportunities of foundation models in bioinformatics.** *Brief. Bioinform.* 2024; **25**(6).
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, *et al.*: **BUSCO: Assessing genome assembly and annotation completeness.** *Current Protocols.* 2021; **1**(7): e323.
PubMed Abstract | Publisher Full Text

Marçais G, *et al.*: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers Free.** *Bioinformatics.* March 2011; **27**(6): 764–770.
PubMed Abstract | Publisher Full Text | Free Full Text

McKenna A, *et al.*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**: 1297–1303.
PubMed Abstract | Publisher Full Text | Free Full Text

Mendes FK, *et al.*: **CAFE 5 models variation in evolutionary rates among gene families.** *Bioinformatics.* December 2020; **36**(22-23): 5516–5518.
PubMed Abstract | Publisher Full Text

Mikheenko A, *et al.*: **WebQUAST: online evaluation of genome assemblies Open Access.** *Nucleic Acids Res.* 5 July 2023; **51**(W1): W601–W606.
Publisher Full Text

Naithani S, *et al.*: **Exploring Pan-Genomes: An Overview of Resources and Tools for Unraveling Structure, Function, and Evolution of Crop Genes and Genomes.** *Biomolecules.* 2023 Sep 17; **13**(9): 1403.
PubMed Abstract | Publisher Full Text | Free Full Text

Nevers Y, *et al.*: **OMArk: Genome assembly quality assessment using evolutionary signals.** *Nat. Biotechnol.* 2025; **43**: 124–133.
PubMed Abstract | Publisher Full Text | Free Full Text

Papoutsoglou EA, *et al.*: **Enabling reusability of plant phenomic datasets with MIAPPE 1.1.** *New Phytol.* 2020; **227**: 260–273.
PubMed Abstract | Publisher Full Text | Free Full Text

Peroni S, *et al.*: **Packaging research artefacts with RO-Crate.** *Data Sci.* 2022; **5**(2): 97–138.
Publisher Full Text

Poplin R, *et al.*: **A universal SNP and small-indel variant caller using deep neural networks.** *Nat. Biotechnol.* 2018; **36**(10): 983–987.
PubMed Abstract | Publisher Full Text

Qin P, *et al.*: **Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations.** *Cell.* 2021; **184**(13): 3542–3558.e16.
PubMed Abstract | Publisher Full Text

Ranallo-Benavidez TR, *et al.*: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat. Commun.* 2020; **11**:

1432.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, *et al.*: **Merqury: Reference-free quality assessment of genome assemblies.** *Genome Biol.* 2020; **21**(245): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Robinson J-T, *et al.*: **igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV).** *Bioinformatics.* January 2023; **39**(1): btac830.
PubMed Abstract | Publisher Full Text | Free Full Text

Schreiber M, *et al.*: **Plant pangenomes for crop improvement, biodiversity and evolution.** *Nat. Rev. Genet.* 2024; **25**: 563–577.
PubMed Abstract | Publisher Full Text | Free Full Text

Shang L, *et al.*: **A super pan-genomic landscape of rice.** *Cell Res.* 2022; **32**: 878–896.
PubMed Abstract | Publisher Full Text | Free Full Text

Shannon P, *et al.*: **Cytoscape: A software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003; **13**(11): 2498–2504.
PubMed Abstract | Publisher Full Text | Free Full Text

Shumate A, Salzberg SL: **Liftoff: Accurate alignment-based annotation transfer in phylogenomics.** *Bioinformatics.* 2021; **37**(12): 1639–1643.
PubMed Abstract | Publisher Full Text | Free Full Text

Smolka M, *et al.*: **(2024). Detection of mosaic and population-level structural variants with Sniffles2.** *Nat. Biotechnol.* 2024; **42**: 1571–1580.
PubMed Abstract | Publisher Full Text | Free Full Text

Sommer MJ, *et al.*: **PSAURON: A tool for structural annotation quality assessment.** *NAR Genomics and Bioinformatics.* 2025; **7**(1).
PubMed Abstract | Publisher Full Text | Free Full Text

Stiehler F, *et al.*: **Helixer: Cross-species gene annotation with deep learning.** *Bioinformatics.* 2020; **36**(22-23): 5291–5298.
PubMed Abstract | Publisher Full Text | Free Full Text

Tettelin H, *et al.*: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae.** *Proc. Natl. Acad. Sci.* 2005; **102**(39): 13950–13955.
PubMed Abstract | Publisher Full Text | Free Full Text

Tonkin-Hill G, *et al.*: **Panaroo: Pangenome analysis pipeline for microbial genomes.** *Genome Biol.* 2020; **21**(180): 180.
PubMed Abstract | Publisher Full Text | Free Full Text

Ou S, *et al.*: **Assessing genome assembly quality using the LTR Assembly Index (LAI) Open Access.** *Nucleic Acids Res.* 30 November 2018; **46**(21): e126.
PubMed Abstract | Publisher Full Text | Free Full Text

Wick RR, *et al.*: **Bandage: Interactive visualization of de novo genome assemblies.** *Bioinformatics.* 2015; **31**(20): 3350–3352.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhong Z, *et al.*: **The distinct roles of genome, methylation, transcription, and translation on protein expression in *Arabidopsis thaliana* resolve the Central Dogma's information flow.** *Genome Biol.* 2025; **26**(1): 319.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhou Y, *et al.*: **Graph pangenome captures missing heritability and empowers tomato breeding.** *Nature.* 2022; **606**: 527–534.
PubMed Abstract | Publisher Full Text | Free Full Text

Additional Resources

AgBioData Nomenclature Working Group: **GitHub repository.** 2025.
Reference Source

ENA Metadata Standards: **European Nucleotide Archive.** 2025.
Reference Source

ENSEMBL: **Genome data & annotation.** 2025.
Reference Source

FAIR Cookbook: **ELIXIR Europe.** 2025.
Reference Source

INSDC: **The International Nucleotide Sequence Database Collaboration.** 2025.
Reference Source

Merqury Documentation: **GitHub.** 2020.
Reference Source

PanBARLEX: 2025.
Reference Source

Open Peer Review

Current Peer Review Status: ✓ ✓ ? ?

Version 2

Reviewer Report 30 December 2025

https://doi.org/10.5256/f1000research.191128.r436510

?

Jianping Xu
Department of Biology, McMaster University, Hamilton, Canada

This is an excellent set of recommendations for handling and analyzing plant pan-genomes. I have only three minor comments/questions for authors' considerations.
1. Should there be a clearly stated minimum standard for a genome to be included in plant pan-genome analyses? In the examples/case studies provided, what were the inclusion criteria and is there an emerging consensus?
2. Even though the abstract included "Super-pan-genomes", aside from the tomato case study example at the very end, very little attention was paid to this topic in the main text. Should there be additional criteria for super-pan-genome studies on top of what's recommended for pan-genome analyses?
3. This White Paper listed INSDC and Ensembl as the suggested data repositories. Given the increasing importance of genomic and pan-genomic data from China and the depositions of such data in the Chinese National Genomic Data Center (https://ngdc.cncb.ac.cn/), I think it's important to include that database as a suggested repository for pan-genome datasets.
A minor editorial comment: two citations were included in the abstract of this paper. Can the abstract in this journal include citations?

Is the topic of the review discussed comprehensively in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Yes

Is the review written in accessible language?
Yes

Are the conclusions drawn appropriate in the context of the current research literature?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: population genetics and genomics, with a focus on fungi

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 30 December 2025

https://doi.org/10.5256/f1000research.191128.r434365

? **Rutwik Barmukh**
Centre for Crop and Food Innovation, Murdoch University, Murdoch, Western Australia, Australia

This white paper makes an impressive effort to harmonize methodological practices for plant pan-genome construction, analysis, and sharing, as pan-genomic resources are rapidly increasing across different crop species. The manuscript is highly relevant and addresses a clear need within the plant genomics community. However, to function effectively as a widely adopted standards document, the manuscript would benefit from stronger justification of recommended practices and better consideration of downstream use cases and implementation challenges. The manuscript's clarity, usability, and long-term impact will significantly improve after addressing some issues highlighted below.

1. Although the manuscript highlights the importance of the proposed standards for improving FAIR compliance, the practical aspects of implementing these standards in real-world scenarios can be explored further. For instance, challenges related to community-wide adoption, such as maintaining consistent metadata across projects and the limited support for certain recommended formats (e.g. GFA, JSON-LD) in existing repositories, deserve more explicit discussion. In addition, suggestions for handling legacy datasets that lack complete or standardized metadata would be beneficial. Providing concrete examples of common pitfalls (e.g., interoperability failures caused by inconsistent metadata) along with potential mitigation approaches would further strengthen the paper.

2. While the manuscript precisely captures genomic and computational standards, it currently lacks guidance on how pan-genome outputs should interface with downstream applications that are most relevant to plant breeders and geneticists (e.g., GWAS, QTL mapping, selection decisions, etc.). A subsection linking pan-genome data to common downstream analyses can be added. Also, example schemas or pipelines showing how standardized pan-genome representations improve trait mapping or selection decisions in practice can be included.

3. The Case Studies section highlights several key projects, but it reads as descriptive rather than analytical. It would be more convincing if this section contained quantitative comparisons showing

how applying the proposed standards increased reproducibility, efficiency, or interpretability over previous, unstandardized approaches.

Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Is the review written in accessible language?

Partly

Are the conclusions drawn appropriate in the context of the current research literature?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Crop genomics, bioinformatics, molecular breeding

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 24 December 2025

https://doi.org/10.5256/f1000research.191128.r433712

✓ **Xiaoming Xie**

Wheat Genetics and Genomics Center, China Agricultural University College of Agronomy and Biotechnology (Ringgold ID: 200630), Beijing, Beijing, China

I have no further comments to make.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: wheat pangenome, gene-based pangenome, comparative genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 25 November 2025

https://doi.org/10.5256/f1000research.191128.r433713

✓ **Sunil Kumar Sahu**
State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen, China

I am happy with the author's thorough revision and detailed reponse. I have no further comments

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant genomics and evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 04 September 2025

https://doi.org/10.5256/f1000research.183537.r404154

? **Xiaoming Xie**
Wheat Genetics and Genomics Center, China Agricultural University College of Agronomy and Biotechnology (Ringgold ID: 200630), Beijing, Beijing, China

This white paper by Heuermann *et al.* presents a timely and comprehensive framework aiming to establish community-wide standards for plant pan-genome analysis. The authors cover critical aspects from nomenclature and quality control to data sharing and visualization. This work represents a valuable and necessary initiative to promote FAIR principles in a rapidly evolving field. However, several major revisions are required to enhance its practical utility, scientific rigor, and overall impact before it can be endorsed as a foundational guide for the community.

Major comments

1. Lack of Practical Guidance and Actionable Workflows. While the paper provides an exhaustive list of tools and standards, it currently functions more as a catalogue than a practical guide. A

researcher new to the field would struggle to navigate the options and select the most appropriate methodology for their specific project. To address this, the authors should incorporate one or more decision-tree figures or summary tables that guide users based on their specific research context (e.g., species ploidy, data type, biological question). Such a resource would transform this document from a simple list into an indispensable, actionable guide.

2. Understated Importance of the Generalized Feature Identifier (GFI) Concept. The proposal of a 'Generalized Feature Identifier' (GFI) in the 'Future Directions' section is a highly innovative and critical idea. It elegantly addresses a major bottleneck in functionally annotating the non-genic 'dark matter' of pan-genomes, which is often overlooked. However, its placement as a future thought diminishes its significance. This concept should be introduced much earlier in the manuscript, possibly in the nomenclature section, and framed as a core recommendation of this white paper to highlight its forward-thinking contribution.

3. Imprecise Comparison of Linear- vs. Graph-Based Approaches. The distinction between linear- and graph-based pan-genomes in Section 5.1 is crucial, but the current description of linear approaches could be refined for greater accuracy and clarity. The manuscript conflates two distinct types of 'linear-based' analyses. It states that these approaches identify sequence-level variations (SNPs, indels) as well as gene-level presence/absence variations (PAVs). However, the tools cited as examples, such as OrthoFinder and Ensembl Compara, are primarily used for inferring orthology and identifying gene-level PAVs. They are not the primary tools for calling SNPs and small indels from whole-genome alignments (which typically involves a separate workflow with variant callers like GATK against a linear reference). This conflation creates an imprecise comparison with graph-based approaches, which are fundamentally designed to model sequence-level variation directly. To improve this section, we recommend the authors explicitly distinguish between: (a) Sequence-based linear analysis: Aligning multiple genomes to a single linear reference to call SNPs and indels. (b) Gene-based linear analysis: Using orthology inference tools on annotated gene sets to determine the pan-gene repertoire and gene-level PAVs. By separating these two concepts, the manuscript can provide a more accurate and nuanced comparison, highlighting how graph-based pan-genomes aim to integrate both types of variation in a way that requires distinct workflows in a traditional linear framework.

Minor comments

1. The authors should adopt a more authoritative tone appropriate for a standards paper. Phrases like "probably more suited" (Section 2.3) should be replaced with definitive recommendations (e.g., "we recommend the use of...") to provide clear guidance.

2. In Section 4.3 (Metadata requirements), the list of mandatory fields should be expanded to include the specific versions of all software and pipelines used. This is essential for ensuring full reproducibility of the analyses.

3. The case studies in Section 6 are too brief to be impactful. Each case should be slightly expanded to illustrate how the application (or lack thereof) of the proposed standards directly impacted the project's outcomes, challenges, or successes. This would provide concrete evidence for the importance of the proposed standards.

4. Regarding the classification of genes into "core, shell, and cloud" (Section 2.3), it is crucial to

state that the specific percentage thresholds used for these definitions must be explicitly reported in the metadata, as they can significantly impact downstream comparative analyses.

Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Is the review written in accessible language?

Partly

Are the conclusions drawn appropriate in the context of the current research literature?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: wheat pangenome, gene-based pangenome, comparative genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 10 Nov 2025

Marc Christian Heuermann

Reviewer 2

Xiaoming Xie (https://orcid.org/0000-0002-7925-4964), Wheat Genetics and Genomics Center, China Agricultural University College of Agronomy and Biotechnology (Ringgold ID: 200630), Beijing, Beijing, China

Approved with Reservations

This white paper by Heuermann *et al.* presents a timely and comprehensive framework aiming to establish community-wide standards for plant pan-genome analysis. The authors cover critical aspects from nomenclature and quality control to data sharing and visualization. This work represents a valuable and necessary initiative to promote FAIR principles in a rapidly evolving field. However, several major revisions are required to enhance its practical utility, scientific rigor, and overall impact before it can be endorsed as a foundational guide for the community.

Major comments

1. Lack of Practical Guidance and Actionable Workflows. While the paper provides an exhaustive list of tools and standards, it currently functions more as a catalogue than a practical guide. A researcher new to the field would struggle to navigate the options and

select the most appropriate methodology for their specific project. To address this, the authors should incorporate one or more decision-tree figures or summary tables that guide users based on their specific research context (e.g., species ploidy, data type, biological question). Such a resource would transform this document from a simple list into an indispensable, actionable guide.

Answer: Thank you for your insightful comments. Both reviewers raised this point, prompting us to create Figure 1, which clearly illustrates how each tool fits into the pan-genome analysis workflow. We have now cross-referenced the figure in both Section 3 and Section 5.

2. Understated Importance of the Generalized Feature Identifier (GFI) Concept. The proposal of a 'Generalized Feature Identifier' (GFI) in the 'Future Directions' section is a highly innovative and critical idea. It elegantly addresses a major bottleneck in functionally annotating the non-genic 'dark matter' of pan-genomes, which is often overlooked. However, its placement as a future thought diminishes its significance. This concept should be introduced much earlier in the manuscript, possibly in the nomenclature section, and framed as a core recommendation of this white paper to highlight its forward-thinking contribution.

Answer: We do agree with the assessment and positioned the GFI concept now as paragraph 2.4.

3. Imprecise Comparison of Linear- vs. Graph-Based Approaches. The distinction between linear- and graph-based pan-genomes in Section 5.1 is crucial, but the current description of linear approaches could be refined for greater accuracy and clarity. The manuscript conflates two distinct types of 'linear-based' analyses. It states that these approaches identify sequence-level variations (SNPs, indels) as well as gene-level presence/absence variations (PAVs). However, the tools cited as examples, such as OrthoFinder and Ensembl Compara, are primarily used for inferring orthology and identifying gene-level PAVs. They are not the primary tools for calling SNPs and small indels from whole-genome alignments (which typically involves a separate workflow with variant callers like GATK against a linear reference). This conflation creates an imprecise comparison with graph-based approaches, which are fundamentally designed to model sequence-level variation directly. To improve this section, we recommend the authors explicitly distinguish between: (a) Sequence-based linear analysis: Aligning multiple genomes to a single linear reference to call SNPs and indels. (b) Gene-based linear analysis: Using orthology inference tools on annotated gene sets to determine the pan-gene repertoire and gene-level PAVs. By separating these two concepts, the manuscript can provide a more accurate and nuanced comparison, highlighting how graph-based pan-genomes aim to integrate both types of variation in a way that requires distinct workflows in a traditional linear framework.

Answer: Thank you for this specific and helpful comment. We have reworked and updated our paragraph 5 accordingly.

Minor comments

1. The authors should adopt a more authoritative tone appropriate for a standards paper. Phrases like "probably more suited" (Section 2.3) should be replaced with definitive recommendations (e.g., "we recommend the use of...") to provide clear guidance.

Answer: Thank you for the suggestion. We rephrased the paragraph.

2. In Section 4.3 (Metadata requirements), the list of mandatory fields should be expanded to include the specific versions of all software and pipelines used. This is essential for ensuring full reproducibility of the analyses.

Answer: Thanks, we have now addressed this important point and extended the paragraph to elaborate more on the importance of metadata requirements.

3. The case studies in Section 6 are too brief to be impactful. Each case should be slightly expanded to illustrate how the application (or lack thereof) of the proposed standards directly impacted the project's outcomes, challenges, or successes. This would provide concrete evidence for the importance of the proposed standards.

Answer: Agreed. We updated the paragraph and extended the description of the case studies and their relevance.

4. Regarding the classification of genes into "core, shell, and cloud" (Section 2.3), it is crucial to state that the specific percentage thresholds used for these definitions must be explicitly reported in the metadata, as they can significantly impact downstream comparative analyses.

Answer: We added a sentence to be more precise regarding this point.

- Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

- Are all factual statements correct and adequately supported by citations?

Yes

- Is the review written in accessible language?

Partly

- Are the conclusions drawn appropriate in the context of the current research literature?

Yes

Competing Interests: No competing interests.

Reviewer Report 03 September 2025

https://doi.org/10.5256/f1000research.183537.r404153

?

Sunil Kumar Sahu

State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen, China

This article presents a very comprehensive and thoughtful set of recommendations, covering a wide spectrum from naming conventions to quality control and data sharing. I enjoyed reading it and have a few suggestions that I believe could strengthen its impact and practicality for the community.

My main thought is that the sheer breadth of recommendations, while excellent, might feel daunting for some labs, especially those with limited resources. To make the framework more accessible, it would be incredibly helpful if the authors could more clearly distinguish between what they consider essential "minimum standards" and what are "aspirational best practices." For instance, while the framework is well-described, I found myself wishing for a more concrete, practical roadmap. The QC section, for example, lists many excellent tools (FastQC, GenomeScope, QUAST, etc.), but a visual workflow diagram would be immensely valuable. A figure illustrating the step-by-step process from raw data QC, through assembly and annotation QC, to pan-genome QC would really help readers visualize how to integrate these tools into their own standardized processes.

Finally, on the topic of quality control, the article does a great job listing the available tools but could go further in guiding users on how to apply them. For example, some guidance on tool selection would be useful, such as which aspects of QUAST or CRAQ are best for evaluating polyploid assemblies. It would also be helpful to address how to interpret conflicting results from different tools or databases. Furthermore, in the pan-genome section, the concept of "saturation analysis" is mentioned. It would strengthen this part to include some discussion on the sample size required to confidently claim saturation, particularly for species with different ploidy levels.

These are all meant as constructive feedback to enhance what is already a very valuable and needed framework. I hope my comments are helpful.

Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?

Yes

Is the review written in accessible language?

Yes

Are the conclusions drawn appropriate in the context of the current research literature?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant genomics and evolution

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 10 Nov 2025

Marc Christian Heuermann

Reviewer 1
Sunil Kumar Sahu (https://orcid.org/0000-0002-4742-9870), State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, Shenzhen, China

Approved with Reservations
This article presents a very comprehensive and thoughtful set of recommendations, covering a wide spectrum from naming conventions to quality control and data sharing. I enjoyed reading it and have a few suggestions that I believe could strengthen its impact and practicality for the community.
My main thought is that the sheer breadth of recommendations, while excellent, might feel daunting for some labs, especially those with limited resources. To make the framework more accessible, it would be incredibly helpful if the authors could more clearly distinguish between what they consider essential "minimum standards" and what are "aspirational best practices." For instance, while the framework is well-described, I found myself wishing for a more concrete, practical roadmap. The QC section, for example, lists many excellent tools (FastQC, GenomeScope, QUAST, etc.), but a visual workflow diagram would be immensely valuable. A figure illustrating the step-by-step process from raw data QC, through assembly and annotation QC, to pan-genome QC would really help readers visualize how to integrate these tools into their own standardized processes.

Answer: Thank you for your constructive feedback. We have added Figure 1 to the manuscript to clearly illustrate the recommended tools for each step of the pan-genome analysis workflow. The figure is now cross-referenced in both Section 3 and Section 5 for better integration with the discussed content.

Finally, on the topic of quality control, the article does a great job listing the available tools but could go further in guiding users on how to apply them. For example, some guidance on tool selection would be useful, such as which aspects of QUAST or CRAQ are best for evaluating polyploid assemblies. It would also be helpful to address how to interpret conflicting results from different tools or databases. Furthermore, in the pan-genome section, the concept of "saturation analysis" is mentioned. It would strengthen this part to include some discussion on the sample size required to confidently claim saturation, particularly for species with different ploidy levels.
These are all meant as constructive feedback to enhance what is already a very valuable and needed framework. I hope my comments are helpful.

Answer: We have revised section 3. Quality Control (QC) Standards to fully incorporate your suggestions.

- ○ Is the topic of the review discussed comprehensively in the context of the current literature?

Yes

- ○ Are all factual statements correct and adequately supported by citations?

Yes

- ○ Is the review written in accessible language?

Yes

- ○ Are the conclusions drawn appropriate in the context of the current research literature?

Yes

Competing Interests: No competing interests.