Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

Plant genome assembly and annotation

Todd P. Michael[1,2,3,4]

Plant genome biology is entering a new era defined by fully phased, chromosome-scale, telomere-to-telomere assemblies, enabled by the convergence of long-read sequencing technologies, improved assembly algorithms, and powerful scaffolding strategies. Gapless, haplotype-resolved genomes are now feasible even for polyploid species, shifting the bottleneck from assembly to annotation and interpretation. Genome annotation remains one of the greatest opportunities and challenges in plant biology. While *ab initio* methods still form the backbone of structural prediction, evidence-based frameworks that integrate RNA sequencing, chromatin accessibility, methylation, and 3D genome data are rapidly advancing the field. At the same time, artificial intelligence–driven protein-coding gene predictors are redefining *ab initio* gene finding, and large-scale orthology networks continue to improve functional inference. The next frontier is extending annotation beyond protein-coding genes into regulatory and structural dimensions, a goal increasingly enabled by single-cell and multi-omic technologies. Looking forward, the integration of AI, multi-omics, and large language models promises to standardize and automate workflows from DNA isolation to functional annotation. These innovations will accelerate fundamental plant biology discovery, enable next-generation biodiversity conservation, and transform strategies for crop improvement and biotechnology.

Addresses
[1] The Plant Molecular and Cellular Laboratory, Salk Institute for Biological Studies, La Jolla, CA, 92037, USA
[2] Department of Cell and Developmental Biology, School of Biological Sciences, University of California, San Diego, La Jolla, CA, 92093, USA
[3] Center for Marine Biotechnology and Biomedicine, Scripps Institute of Oceanography, University of California, San Diego, La Jolla, CA, 92093, USA
[4] Department of Science and Conservation, San Diego Botanical Garden, Encinitas, CA, 92024, USA

Corresponding author: Michael, Todd P. (toddpmichael@gmail.com)

Current Opinion in Plant Biology 2026, **90**:102859

This review comes from a themed issue on **Genome studies and molecular genetics_2026**

Edited by **Eriko Sasaki** and **Arturo Marí-Ordóñez**

For complete overview of the section, please refer the article collection - Genome studies and molecular genetics_2026

Available online 11 February 2026

https://doi.org/10.1016/j.pbi.2026.102859

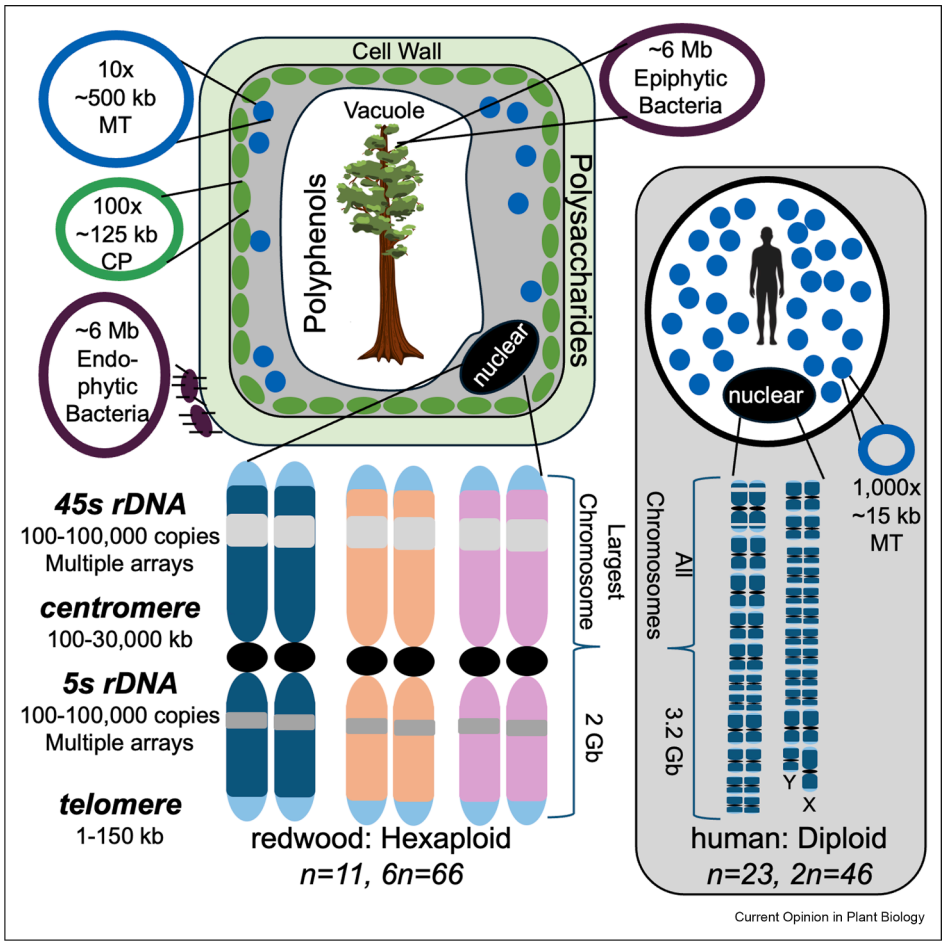
Introduction

Plant *de novo* genome assembly has entered a new era, defined by fully phased, chromosome-scale, haplotype-resolved, telomere-to-telomere (T2T) genomes [1]. Driven by breakthroughs in sequencing technologies, plant genomics has achieved levels of accuracy, throughput, and accessibility that were unimaginable a decade ago [2–4]. Yet plant genomes remain among the most challenging in nature: they are often exceptionally large, dominated by repetitive elements such as centromeres, telomeres, ribosomal DNA (rDNA) and transposable elements (TEs) [5] (Figure 1). Recurrent rounds of polyploidy further complicate assembly, blurring haplotype boundaries and amplifying complexity [6,7]. Despite these obstacles, innovations in sequencing, scaffolding, and computational algorithms have reshaped the landscape of plant genomics [8].

The history of plant genome assembly mirrors this rapid technological progress. The first reference genome, *Arabidopsis thaliana* in 2000, inaugurated the BAC-by-BAC era, during which genomes were assembled by cloning large DNA fragments into bacterial artificial chromosomes (BAC), sequencing and assembling each clone separately, and then stitching them together, a labor-intensive process that took years and cost millions of dollars, which ultimately limited high-quality genome assemblies to a few model organisms [2]. The rise of short-read sequencing transformed genome sequencing into a computational challenge, producing hundreds of assemblies with expanded gene-space coverage but often fragmented and incomplete genomes [4]. A new phase began with long-read sequencing and advanced scaffolding methods, which enabled the first round of highly contiguous, almost complete T2T genomes [3]. Recent T2T assemblies of *Arabidopsis* now provide the first complete views of entire centromeres, illuminating previous dark matter driving the evolution of plant genomes [9].

Today, *de novo* strategies routinely produce chromosome-scale, haplotype-resolved genomes, spanning both model species and key crops. These resources are fueling innovations in annotation frameworks, including pangenomes, machine learning, and multi-omic integration, which now reveal functional elements long hidden by fragmented, low-contiguity plant genomes. Looking forward, continued progress promises deeper resolution of polyploid complexity and more consistent gene predictions, with artificial intelligence and large-

Figure 1



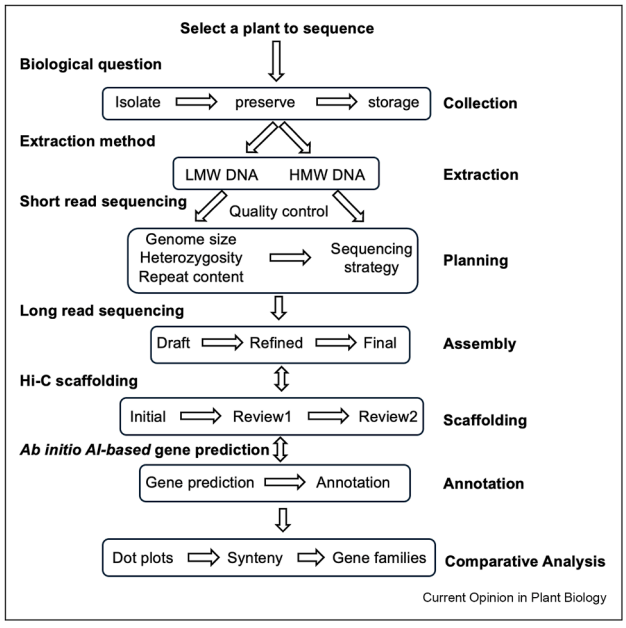
Plant and human cells differ in key cellular and genomic features that impact genome assembly complexity. Plant cell (left): Features a rigid cell wall composed of polysaccharides and large central vacuole that accumulates secondary metabolites such as polyphenols, both of which can interfere with DNA extraction. Plant cells also harbor endophytic and epiphytic bacteria (~6 Mb genomes). Organellar genomes include multiple copies of mitochondrial genomes (~10 per cell, ~500 kb each; MT) and chloroplast genomes (~100 per cell, ~125 kb each; CP). The redwood nuclear genome is hexaploid (*n = 11, 6n = 66*) with very large chromosomes, some exceeding 2 Gb in size, and includes repetitive features such as rDNA (45 S, 5 S), centromeres, and telomeres that are difficult features to assemble in telomere to telomere (T2T) efforts. Human cell (right): Surrounded by a flexible plasma membrane, human cells contain hundreds to thousands of mitochondria, each with multiple ~15 kb genomes (~1000 copies per cell). The nuclear genome is diploid (*n = 23, 2n = 46*) and totals ~3.2 Gb, with chromosome sizes ranging from ~249 Mb (chromosome 1) to ~48 Mb (chromosome 21). Although human genomes also contain challenging regions such as complex centromeres and large ribosomal DNA arrays that complicate complete T2T assembly, their overall genome size, ploidy, and structural complexity are substantially lower than those of many plant genomes.

scale machine learning poised to standardize assembly, polishing, and annotation pipelines. Yet critical challenges remain, as sample selection, preservation, and extraction of high-molecular-weight (HMW) DNA continue to represent major bottlenecks. While prior reviews have focused on technical aspects of sequencing and assembly [1,5,8,10], this article emphasizes the end-to-end process of building high-quality plant genomes from tissue collection and DNA isolation through assembly, scaffolding, and annotation highlighting both current best practices and the opportunities on the horizon (Figure 2).

Collection, preservation and DNA isolation critical to plant *de novo* assembly

A hallmark of the current state of plant genome assembly is its broad accessibility across laboratories and projects, which has elevated tissue collection, preservation, and DNA isolation to some of the most critical determinants of success. A persistent limitation in *de novo* plant genome sequencing remains the quality of input DNA used to generate long-read data [11]. Unlike microbial or animal systems, plant tissues are notoriously difficult substrates, often containing abundant secondary metabolites, polysaccharides, and polyphenols that

Figure 2



Plant de novo genome assembly workflow. The major stages of plant genome sequencing are shown, from project conception to comparative analysis. Following definition of the biological question, plant material is collected, preserved, and stored (Collection). DNA is extracted using protocols optimized for either low-molecular-weight (LMW) or high-molecular-weight (HMW) DNA (Extraction). Short-read sequencing and quality control are then used to estimate genome size, heterozygosity, and repeat content, informing sequencing strategy and depth (Planning). Long-read sequencing produces genome assemblies that progress from draft to refined to final stages (Assembly). Hi-C scaffolding further organizes contigs into chromosome-scale assemblies through iterative inspection and correction (Scaffolding). Gene prediction and annotation follow, typically beginning with *ab initio*, AI-based approaches and refined through functional annotation; empirical data may be incorporated to train models, validate gene structures, and identify untranslated regions, promoters, enhancers, and other non-coding features (Annotation). Assembly, scaffolding, and gene prediction are inherently iterative processes (double-headed arrows), as annotation metrics can reveal gaps or misassemblies requiring revision. Finally, comparative analyses, including dot plots, synteny assessment, and gene family characterization, place the genome in an evolutionary context (Comparative Analysis).

interfere with DNA extraction, library preparation, and downstream sequencing (Figure 1). For high-quality assemblies, particularly those aiming for chromosome-scale, T2T resolution, the availability of HMW DNA of exceptional purity and integrity is essential [12,13].

The earliest stages of a project, sample collection and preservation, are decisive in determining DNA integrity [14,15] (Figure 2). Factors such as tissue type, developmental stage, and storage method (e.g., immediate flash-freezing, silica gel desiccation, ethanol fixation, or liquid nitrogen) all strongly influence the success of downstream DNA isolation [16,17]. Even subtle degradation reduces read lengths and compromises assembly contiguity, making careful planning and preservation protocols as critical as sequencing itself.

The importance of DNA quality has only increased with the latest long-read sequencing technologies. Platforms such as Oxford Nanopore Technologies (ONT), with their portability and steadily improving accuracy, and Pacific Biosciences (PacBio) HiFi chemistry, with its exceptional per-read fidelity, now enable high-quality *de novo* assemblies that were once restricted to specialized genome centers (Figure 3a and b). Portable sequencing has created opportunities for field-based genome projects, allowing reference-quality data to be generated directly from fresh collections in remote environments [18].

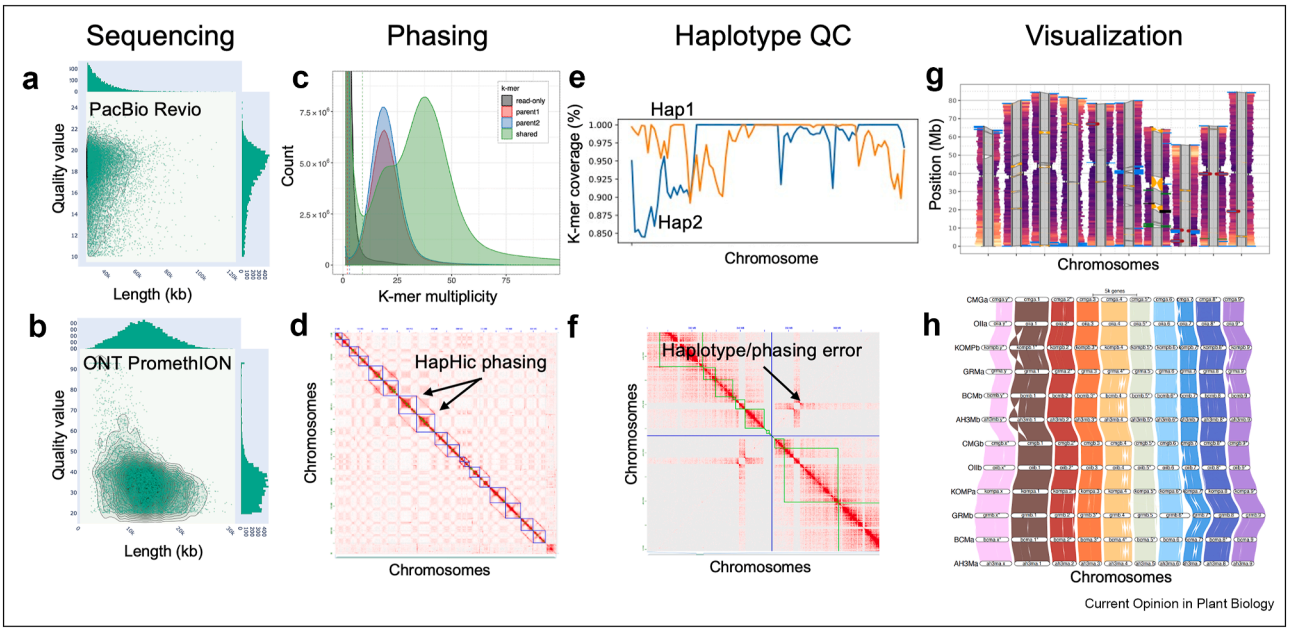
These innovations are transforming the accessibility and impact of plant genomics. Conservation and ecological restoration programs can now rapidly assemble genomes of endangered or rare species to guide management decisions. Botanical gardens and herbaria are increasingly adopting these technologies to build genomic catalogs of their collections, creating enduring resources for biodiversity preservation. Such efforts are also informing breeding programs, adaptation studies, and climate resilience strategies by linking high-quality genomic resources with applied goals [19].

Fully-phased, chromosome-scale, haplotype-resolved, telomere-to-telomere, plant genomes

The field of plant genomics has entered a new era defined by fully-phased, chromosome-scale, haplotype-resolved, T2T assemblies. This transformation has been propelled by the convergence of high-accuracy long-read sequencing technologies from PacBio and ONT coupled to next-generation assembly algorithms and scaffolding strategies capable of resolving some of the largest and most complex polyploid genomes (Figure 3). Recent advances now demonstrate that gapless T2T assemblies can be achieved with ONT ultra-long reads alone [20] or through gap filling enabled by ONT adaptive sampling [21]. Despite these recent advances, most genome assemblies described as T2T still fail to fully resolve highly repetitive centromeric and ribosomal DNA (rDNA) regions, and robust polyploid genome assembly remains beyond the reach of even the best assembly tools. Nevertheless, these developments underscore the remarkable speed with which plant genome assembly has progressed from a technically formidable endeavor to an increasingly routine practice.

Achieving T2T resolution in plants has long been a challenge, largely due to the prevalence of highly repetitive genomic features, high levels of heterozygosity

Figure 3

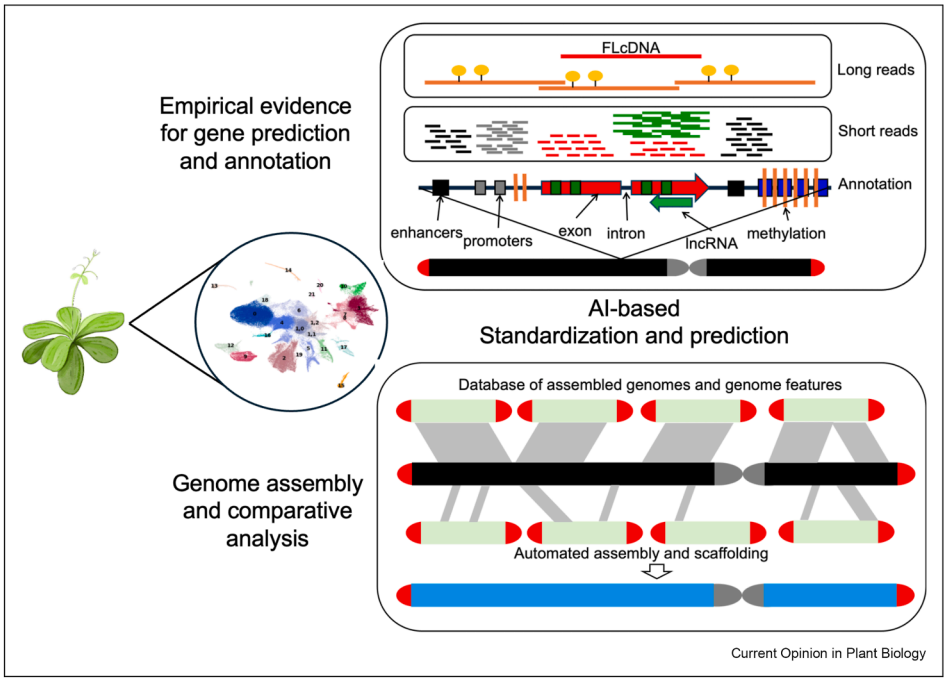


Plant genome assembly advanced by several key innovations from sequencing to visualization. Tools and visualizations are shown that collectively enable chromosome-scale, haplotype-resolved plant genome assembly and analysis, progressing from data generation to phasing, quality control, and comparative interpretation. (a–b) Long-read sequencing technologies form the foundation of modern assemblies. (a) PacBio Revio generates highly accurate HiFi reads, shown as read length versus quality distributions, which support precise base-level assembly, repeat resolution, and error correction. (b) Oxford Nanopore Technologies (ONT) PromethION produces ultra-long reads, visible as extended read-length distributions, that span centromeres, ribosomal DNA arrays, and other long repetitive regions, enabling improved contiguity and assembly across structurally complex loci. (c–d) Haplotype phasing integrates complementary sequence- and contact-based evidence. (c) k-mer multiplicity distributions separate haplotype-specific k-mers from shared sequence content, allowing estimation of heterozygosity and haplotype balance and enabling assignment of reads to parental or haplotypic origins, as used in trio-binning–based phasing. (d) Hi-C contact maps display chromatin interaction frequencies between genomic regions (brighter red indicating higher contact density), which tools such as HapHiC exploit to phase and scaffold contigs into chromosome-scale haplotypes and to identify phase blocks and large-scale structural consistency. (e–f) Haplotype quality control (QC) identifies phasing errors and structural inconsistencies. (e) k-mer–based haplotype consistency analysis (e.g., PanKmer) tracks haplotype-specific k-mer coverage along chromosomes, where abrupt shifts in signal indicate haplotype switching, collapse, or over-separation. (f) Hi-C contact maps provide an orthogonal QC signal, in which disrupted diagonal patterns, unexpected inter-haplotype contacts, or contact discontinuities reveal mis-joins, orientation errors, or phasing errors that often require manual inspection and correction. (g–h) Visualization tools enable interpretation of complex plant genomes. (g) Genome-wide visualization integrates multiple features across chromosomes, including gene density (purple to yellow color scale), GC content (track width), haplotype synteny (grey connecting lines), and structural variation (orange triangles), facilitating assessment of assembly completeness, structural integrity, and haplotype divergence. (h) Comparative visualization frameworks such as GENESPACE enable cross-species analyses of synteny and genome evolution, with conserved syntenic blocks shown as color-matched regions and structural variation or rearrangements highlighted by disruptions or gaps, placing newly assembled genomes in an evolutionary context.

and polyploidy [22]. Centromeres, telomeres, rDNA, and large blocks of transposable elements (TEs), especially long terminal repeat retrotransposons (LTR-RTs), posed persistent obstacles to contiguity and accuracy (Figure 1). Ultra-long ONT reads, with lengths exceeding hundreds of kilobases, provide the continuity to span centromeric regions, while PacBio HiFi reads supply the base accuracy needed to minimize assembly errors (Figure 3a and b). Complementary scaffolding technologies, including optical maps and high throughput chromatin conformation capture (Hi-C), have further anchored contigs into complete chromosomes [22].

Equally transformative is the ability to generate fully phased, haplotype-resolved assemblies (Figure 3c and d). Unlike haploid consensus genomes, phased assemblies capture heterozygosity and structural variation, providing critical insights into outcrossing species, hybrids, and polyploids. Several complementary strategies are used for haplotype resolution, with Hi-C–based phasing emerging as the standard approach due to its broad applicability and scalability [24]. Hi-C leverages chromatin contact maps to phase haplotypes along chromosome-scale scaffolds; however, this approach often requires extensive manual inspection and correction of contact maps. Mis-joins, orientation errors, and

Figure 4



Future directions in plant genome assembly and annotation. Plant genome research is moving toward a future defined by the seamless integration of empirical data, comparative genomics, and artificial intelligence. Left: *Arabidopsis*, shown as an example of a model system, illustrates how comprehensive datasets spanning whole-plant and developmental stages at single-cell resolution will serve as the foundation for predictive models that refine gene prediction, functional annotation, and even genome assembly. Top right: Expanding empirical datasets, including full-length cDNA (FLcDNA), long- and short-read transcriptomes, and epigenomic layers such as DNA methylation, provide high-resolution evidence to accurately define coding and noncoding features, regulatory elements, and chromatin states. Bottom right: AI-driven approaches are poised to harness the rapidly growing databases of assembled genomes and annotated features to deliver standardized, automated assembly and annotation pipelines, ultimately enabling predictive and interpretable models of genome function across diverse plant species.

haplotype switching are frequent, particularly in polyploid genomes and those enriched for long repetitive elements, necessitating expert-driven quality-control workflows. More specialized strategies include trio binning, which uses parental sequence data to assign offspring reads to haplotypes [23], and pollen sequencing, which exploits haploid gametophytes to directly phase haplotypes [25]. Together, these approaches enable genome reconstructions that reveal allele-specific expression, structural rearrangements, and evolutionary histories often obscured in haploid assemblies [26].

Despite these advances, polyploid genome assembly remains challenging. Most current long-read assemblers are fundamentally designed around diploid assumptions, modeling assembly graphs as two-haplotype bubbles. In polyploid genomes, particularly autopolyploids or recently diverged allopolyploids, this frequently results in collapse of similar homeologs, retention of redundant contigs, or haplotype inflation driven by sequencing and assembly errors. Distinguishing true structural variation from assembly artifacts also remains difficult in repetitive regions. Although high-quality polyploid plant genomes can be produced through careful parameter tuning, coverage management, and auxiliary data integration, robust automated polyploid-aware assembly remains an unresolved challenge.

As assemblies approach T2T and fully phased status, quality control (QC) has necessarily evolved beyond traditional benchmarks [27]. While Benchmarking Universal Single-Copy Orthologs (BUSCO) remains widely used for measuring gene content, its reliance on a limited set of orthologs (~3000) only provides one measure of completeness [28]. Additional metrics now include the LTR Assembly Index (LAI), which quantifies retrotransposon assembly quality [29]. Merqury, which assesses base-level accuracy via k-mer analysis; and switch error rates, which evaluate phasing fidelity [30]. Standard QC measures such as coverage uniformity and read mapping rates remain indispensable [31]. Beyond single-genome assessment, pangenome frameworks are now being leveraged to identify structural inconsistencies, haplotype switching, and assembly gaps across species and families [32] (Figure 3e and f).

Increasingly, specialized visualization tools including GeneSpace for synteny, haplotype divergence, and structural variation [33] are being complemented by new strategies for exploring repetitive regions [34] and novel frameworks for visualizing haplotypes [35] (Figure 3g and h).

Plant genome annotation developing from an art to a science

Although fully phased, chromosome-scale, haplotype-resolved, T2T plant genomes are becoming increasingly routine, genome annotation remains one of the most persistent challenges, and greatest opportunities for advancing discovery in plant biology, agriculture, and biotechnology [36] (Figure 4). Annotation encompasses not only the accurate prediction of genomic features such as protein-coding gene models, but also the reliable assignment of biological functions to those genes. Furthermore, the rapid expansion of multi-omic datasets, enabled by innovative sequencing applications that effectively position sequencing as the "new microscope," has broadened the scope of annotation well beyond protein-coding genes [37].

The field is steadily moving beyond traditional *ab initio* prediction toward evidence-based annotation frameworks. A growing emphasis is placed on generating comprehensive RNA sequencing (RNA-seq) datasets across diverse tissues, developmental stages, and environmental conditions to improve gene model prediction and capture isoform diversity [38]. In parallel, multi-omic integration combining assay for transposase-accessible chromatin sequencing (ATAC-seq), chromatin immunoprecipitation sequencing (ChIP-seq), DNA methylation profiling, and Hi-C provide complementary and orthogonal evidence [39]. Together, these datasets not only enhance the training and identification of features such as promoters, untranslated regions (UTRs), enhancers, insulators, activation domains, and topologically associating domains (TADs), but also extend annotation into the regulatory and structural dimensions of the genome that have historically been underexplored (Figure 4).

Concurrently, artificial intelligence (AI)-driven gene predictors have emerged that outperform many traditional *ab initio* tools in both accuracy and efficiency, even without empirical evidence (Figure 4). Helixer, for instance, combines convolutional and recurrent neural networks with a Hidden Markov Model decoder to predict gene structures, and has been shown to improve exon-intron boundary accuracy compared to classic predictors such as AUGUSTUS [36,40]. Tiberius advances this approach by integrating convolutional layers, bidirectional long short-term memory networks, and a differentiable Hidden Markov Model (HMM) into a single end-to-end framework, enabling it to learn both sequence features and structural rules directly from DNA [41]. As with other *ab initio* gene predictors, current AI-based methods face challenges in defining exon-intron boundaries, UTRs, and authentic single-exon genes in repeat-rich plant genomes, and typically do not model alternative splicing, instead reporting primary gene models. Looking forward, these AI-based predictors could be further enhanced by training on rapidly expanding multi-omic datasets, allowing them to capture subtle genomic signals and regulatory complexity that evidence-only pipelines may overlook [42].

Beyond structural prediction, functional annotation is benefiting from the exponential growth of plant genomic resources. More than 5000 plant genomes are now deposited in NCBI, of which approximately 1500 meet the criteria for draft genome assemblies. Although many of these genomes lack high-quality annotation, the recent advances in AI-based gene prediction enable the rapid generation of standardized gene models. These uniform annotations facilitate comparative analyses, such as OrthoFinder, that leverage dense orthology networks to produce increasingly accurate cross-species gene function predictions [43,44]. At the same time, large language models (LLMs) and graph-based AI frameworks are being developed to synthesize heterogeneous data sources including sequence homology, gene expression, protein−protein interactions, and evolutionary signals to annotate genes that lack well-characterized homologs [45]. Despite these advances, annotation remains an iterative and interpretive process, requiring expert oversight to distinguish biologically meaningful features from technical artifacts.

Prospects for *de novo* plant genome assembly and annotation with AI

Future *de novo* plant genome assemblies will increasingly depend on automated, end-to-end pipelines that minimize manual intervention while improving consistency, scalability, and reproducibility (Figure 4). As sequencing throughput continues to increase and costs decline, next-generation frameworks will integrate long- and short-read data, optical maps, and Hi-C signals into near-complete assemblies in a largely automated fashion [46]. Continuous benchmarking through standardized pipelines will further harmonize quality control, reducing variability across projects and ensuring reliable comparisons [47].

LLMs represent a particularly promising new direction for genome interpretation that has been largely untapped to date [48]. Adapted for biological sequences and genome graphs, LLMs offer the potential to transform phasing and haplotyping of complex plant genomes, especially polyploids and aneuploids where traditional methods often fall short [46]. By learning to recognize recurring sequence and structural patterns,

LLMs may help disentangle haplotypes across highly repetitive and centromeric regions that remain challenging even for current T2T assemblies.

In parallel, the integration of single-cell genomics and other multi-omic datasets will refine both assemblies and annotations [49]. Single-cell and single-nucleus RNA-seq (sc/snRNA-seq) together with ATAC-seq provide unprecedented resolution of expression and chromatin accessibility, enabling cell-type- and stage-specific functional gene annotation. When coupled with proteomic, metabolomic, and methylome data, these modalities will power AI-driven models capable of detecting subtle regulatory features such as enhancers, insulators, and TADs [50]. The outcome will be functionally enriched, context-aware plant genome annotations that bridge structure with function.

Looking ahead, the convergence of automated assembly pipelines, LLM-guided phasing, and multi-omic integration points toward a future where generating a fully-phased, chromosome-scale, haplotype-resolved, and functionally annotated plant genome becomes routine and possibly most important, reproducible (Figure 4). Such standardization will usher in a new era of genome exploration, with transformative impacts on biodiversity conservation, ecological restoration, precision agriculture, and synthetic biology. Ultimately, this paradigm will make possible a seamless transition from raw sequence data to functionally interpretable genomes, accelerating both fundamental discovery and applied innovation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Todd Michael reports financial support was provided by Bill and Melinda Gates Foundation. Todd Michael reports a relationship with Cqesta that includes: equity or stocks. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is dedicated to Joanne Chory, a trailblazer in plant biology who championed the sequencing of the first plant genome, *Arabidopsis thaliana*, and whose vision recognized plants as a vital solution for the challenges facing our world. I am deeply grateful to Nolan Hartwick, Nicholas Allsing, and Allen Mamerto of the Michael Lab genome assembly team for their insightful discussions and contributions. This work was funded by a Bill and Melinda Gates Foundation (INV-040541) to T.P.M.

Data availability

No data was used for the research described in the article.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. * Garg V, Bohra A, Mascher M, Spannagl M, Xu X, Bevan MW, Bennetzen JL, Varshney RK: **Unlocking plant genetics with telomere-to-telomere genome assemblies**. *Nat Genet* 2024, **56**:1788–1799.
This review synthesizes recent advances in T2T plant genome assemblies, highlighting how complete reference genomes are reshaping our understanding of genome structure, evolution, and function. The authors emphasize challenges unique to plants, including polyploidy and repetitive content, and outline future opportunities in crop improvement and comparative genomics made possible by fully resolved assemblies.

2. Michael TP, Jackson S: **The first 50 plant genomes**. *Plant Genome* 2013, **6**.

3. Michael TP, VanBuren R: **Building near-complete plant genomes**. *Curr Opin Plant Biol* 2020, **54**:26–33.

4. Michael TP, VanBuren R: **Progress, challenges and the future of crop genomes**. *Curr Opin Plant Biol* 2015, **24**:71–81.

5. Xie L, Gong X, Yang K, Huang Y, Zhang S, Shen L, Sun Y, Wu D, Ye C, Zhu Q-H, *et al.*: **Technology-enabled great leap in deciphering plant genomes**. *Nat Plants* 2024, **10**:551–566.

6. Zhao T, Schranz ME: **Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes**. *Proc Natl Acad Sci U S A* 2019, **116**:2165–2174.

7. Zhao T, Zwaenepoel A, Xue J-Y, Kao S-M, Li Z, Schranz ME, Van de Peer Y: **Whole-genome microsynteny-based phylogeny of angiosperms**. *Nat Commun* 2021, **12**:3498.

8. Zhao Z, Shi T: **Haplotype-resolved assembly in polyploid plants: methods, challenges, and implications for evolutionary and breeding research**. *Genes (Basel)* 2025, **16**.

9. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmücker A, Mandáková T, Jamge B, Lambing C, Kuo P, *et al.*: **The genetic and epigenetic landscape of the Arabidopsis centromeres**. *Science* 2021, **374**:eabi7489.

10. * Li H, Durbin R: **Genome assembly in the telomere-to-telomere era**. *Nat Rev Genet* 2024, **25**:658–670.
This technical review by two of the pre-eminent leaders in genome assembly provides an authoritative overview of the algorithms, sequencing technologies, and scaffolding methods that have enabled T2T assemblies. Li and Durbin critically assess current methodological challenges, including repetitive sequences and polyploid genomes, and chart future directions for advancing assembly accuracy and completeness across diverse organisms.

11. ** Minich JJ, Moore ML, Allsing NA, Aylward A, Murray ER, Tran L, Michael TP: **Generating high-quality plant and fish reference genomes from field-collected specimens by optimizing preservation**. *Commun Biol* 2023, **6**:1246.
This study demonstrates how optimizing tissue collection and preservation protocols enables the recovery of ultra–high molecular weight DNA suitable for long-read sequencing from field-collected specimens. With a particular focus on plants, the work highlights strategies to overcome challenges posed by polysaccharides and polyphenols, providing a roadmap for generating reference-quality genomes for conservation and biodiversity genomics.

12. Kang M, Chanderbali A, Lee S, Soltis DE, Soltis PS, Kim S: **High-molecular-weight DNA extraction for long-read sequencing of plant genomes: an optimization of standard methods**. *Appl Plant Sci* 2023, **11**, e11528.

13. Nishii K, Möller M, Foster RG, Forrest LL, Kelso N, Barber S, Howard C, Hart ML: **A high quality, high molecular weight DNA extraction method for PacBio HiFi genome sequencing of recalcitrant plants**. *Plant Methods* 2023, **19**:41.

14. Messner E, Becker L, DeSanctis ML, Soranno EA, Pianka R, Pierce C, Johnson M, Poulin RF, Appiah-Madson HJ, Distel DL: **Perish the thawed? EDTA reduces DNA degradation during extraction from frozen tissue**. *PLoS One* 2025, **20**, e0321872.

15. Russo A, Mayjonade B, Frei D, Potente G, Kellenberger RT, Frachon L, Copetti D, Studer B, Frey JE, Grossniklaus U, *et al.*: **Low-input high-molecular-weight DNA extraction for long-read sequencing from plants of diverse families**. *Front Plant Sci* 2022, **13**, 883897.

16. Bein B, Chrysostomakis I, Arantes LS, Brown T, Gerheim C, Schell T, Schneider C, Leushkin E, Chen Z, Sigwart J, *et al.*: **Long-read sequencing and genome assembly of natural history collection samples and challenging specimens**. *Genome Biol* 2025, **26**:25.

17. Quatela A-S, Cangren P, Jafari F, Michel T, de Boer HJ, Oxelman B: **Retrieval of long DNA reads from herbarium specimens**. *AoB Plants* 2023, **15**:lad074.

18. Hotaling S, Wilcox ER, Heckenhauer J, Stewart RJ, Frandsen PB: **Highly accurate long reads are crucial for realizing the potential of biodiversity genomics**. *BMC Genom* 2023, **24**:117.

19. Theissinger K, Fernandes C, Formenti G, Bista I, Berg PR, Bleidorn C, Bombarely A, Crottini A, Gallo GR, Godoy JA, *et al.*: **How genomics can help biodiversity conservation**. *Trends Genet* 2023, **39**:545–559.

20. ** Koren S, Bao Z, Guarracino A, Ou S, Goodwin S, Jenike KM, Lucas J, McNulty B, Park J, Rautiainen M, *et al.*: **Gapless assembly of complete human and plant chromosomes using only nanopore sequencing**. *Genome Res* 2024, **34**: 1919–1930.
This landmark paper demonstrates that ultra-long Oxford Nanopore Technologies (ONT) reads alone are sufficient to achieve gapless, telomere-to-telomere assemblies of both human and plant chromosomes. The study establishes ONT as a standalone technology for resolving highly repetitive regions such as centromeres and rDNA arrays, marking a critical advance for plant genomics.

21. Lu D, Liu C, Ji W, Xia R, Li S, Liu Y, Liu N, Liu Y, Deng XW, Li B: **Nanopore ultra-long sequencing and adaptive sampling spur plant complete telomere-to-telomere genome assembly**. *Mol Plant* 2024, **17**:1773–1786.

22. Lin Y, Ye C, Li X, Chen Q, Wu Y, Zhang F, Pan R, Zhang S, Chen S, Wang X, *et al.*: **quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification**. *Hortic Res* 2023, **10**:uhad127.

23. Delorean EE, Youngblood RC, Simpson SA, Schoonmaker AN, Scheffler BE, Rutter WB, Hulse-Kemp AM: **Representing true plant genomes: haplotype-resolved hybrid pepper genome with trio-binning**. *Front Plant Sci* 2023, **14**, 1184112.

24. ** Zeng X, Yi Z, Zhang X, Du Y, Li Y, Zhou Z, Chen S, Zhao H, Yang S, Wang Y, Chen G: **Achieving de novo scaffolding of chromosome-level haplotypes using Hi-C data**. *Nat Plants* 2024, **10**:1157–1158.
This study introduces HapHiC, a scaffolding approach that harnesses Hi-C chromatin contact maps to resolve haplotypes at chromosome scale. By leveraging the long-range phasing power of Hi-C, HapHiC enables *de novo* haplotype separation without relying on parental data or external references, addressing a key challenge in assembling highly heterozygous and polyploid plant genomes.

25. Sun H, Jiao W-B, Krause K, Campoy JA, Goel M, Folz-Donahue K, Kukat C, Huettel B, Schneeberger K: **Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar**. *Nat Genet* 2022, **54**:342–348.

26. Gao W, Wang S, Jiang T, Hu H, Gao R, Zhou M, Wang G: **Chromosome-scale and haplotype-resolved genome assembly of Populus trichocarpa**. *Hortic Res* 2025, **12**:uhaf012.

27. Wang P, Wang F: **A proposed metric set for evaluation of genome assembly quality**. *Trends Genet* 2023, **39**:175–186.

28. Tegenfeldt F, Kuznetsov D, Manni M, Berkeley M, Zdobnov EM, Kriventseva EV: **OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes**. *Nucleic Acids Res* 2025, **53**:D516–D522.

29. Mokhtar MM, Abd-Elhalim HM, El Allali A: **A large-scale assessment of the quality of plant genome assemblies using the LTR assembly index**. *AoB Plants* 2023, **15**:lad015.

30. Rhie A, Walenz BP, Koren S, Phillippy AM: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies**. *Genome Biol* 2020, **21**:245.

31. Li K, Xu P, Wang J, Yi X, Jiao Y: **Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement**. *Nat Commun* 2023, **14**:6556.

32. Aylward AJ, Petrus S, Mamerto A, Hartwick NT, Michael TP: **PanKmer: k-mer-based and reference-free pangenome analysis**. *Bioinformatics* 2023, **39**.

33. * Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J: **GENESPACE tracks regions of interest and gene copy number variation across multiple genomes**. *eLife* 2022, **11**.
This study introduces GENESPACE, a visualization and analysis platform for tracking synteny, regions of interest, and gene copy number variation across multiple genomes. Demonstrated extensively in plants, the tool enables exploration of genome evolution and structural diversity beyond linear assemblies. The work highlights the growing need for more powerful visualization and comparative frameworks to interpret the complexity of polyploid and repeat-rich plant genomes.

34. Vollger MR, Kerpedjiev P, Phillippy AM, Eichler EE: **StainedGlass: interactive visualization of massive tandem repeat structures with identity heatmaps**. *Bioinformatics* 2022, **38**: 2049–2051.

35. Lynch RC, Padgitt-Cobb LK, Garfinkel AR, Knaus BJ, Hartwick NT, Allsing N, Aylward A, Bentz PC, Carey SB, Mamerto A, *et al.*: **Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome**. *Nature* 2025, **643**:1001–1010.

36. Abbas Q, Wilhelm M, Kuster B, Poppenberger B, Frishman D: **Exploring crop genomes: assembly features, gene prediction accuracy, and implications for proteomics studies**. *BMC Genom* 2024, **25**:619.

37. Michael TP: **Can a plant biologist fix a thermostat?** *New Phytol* 2025, **245**:1403–1410.

38. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M: **BRAKER3: fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA**. *Genome Res* 2024, **34**:769–777.

39. Zhang X, Luo Z, Marand AP, Yan H, Jang H, Bang S, Mendieta JP, Minow MAA, Schmitz RJ: **A spatially resolved multi-omic single-cell atlas of soybean development**. *Cell* 2025, **188**:550–567.e19.

40. Stiehler F, Steinborn M, Scholz S, Dey D, Weber APM, Denton AK: **Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning**. *Bioinformatics* 2021, **36**:5291–5298.

41. Gabriel L, Becker F, Hoff KJ, Stanke M: **Tiberius: end-to-end deep learning with an HMM for gene prediction**. *Bioinformatics* 2024, **40**.

42. Zhai J, Gokaslan A, Schiff Y, Berthel A, Liu Z-Y, Lai W-Y, Miller ZR, Scheben A, Stitzer MC, Romay MC, *et al.*: **Cross-species modeling of plant genomes at single-nucleotide resolution using a pretrained DNA language model**. *Proc Natl Acad Sci U S A* 2025, **122**, e2421738122.

43. Emms DM, Liu Y, Belcher LJ, Holmes J, Kelly S: **OrthoFinder: scalable phylogenetic orthology inference for comparative genomics**. *bioRxiv* 2025, https://doi.org/10.1101/2025.07.15.664860.

44. Hartwick NT, Michael TP: **OrthoBrowser: gene family analysis and visualization**. *Bioinform Adv* 2025, **5**:vbaf009.

45. Sunil RS, Lim SC, Itharajula M, Mutwil M: **The gene function prediction challenge: large language models and knowledge graphs to the rescue**. *Curr Opin Plant Biol* 2024, **82**, 102665.

46. Jiang Z, Peng Z, Wei Z, Sun J, Luo Y, Bie L, Zhang G, Wang Y: **A deep learning-based method enables the automatic and accurate assembly of chromosome-level genomes**. *Nucleic Acids Res* 2024, **52**, e92.

47. Sullivan J, De Panis D, Galeone V, Mazzoni CJ: **Genome Evaluation Pipeline (GEP): a fully automated quality control tool for parallel evaluation of genome assemblies**. *Bioinform Adv* 2025, **5**:vbaf147.

48. ** Lam HYI, Ong XE, Mutwil M: **Large language models in plant biology**. *Trends Plant Sci* 2024, **29**:1145–1155.
This clear and forward-looking review outlines how large language models (LLMs) are poised to transform plant biology, from literature mining and data integration to functional annotation and hypothesis generation. The authors provide an accessible introduction to LLMs while charting a visionary roadmap for their application, making the article a touchstone for future work at the interface of AI and plant science.

49. Rosen Y, Brbić M, Roohani Y, Swanson K, Li Z, Leskovec J: **Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN**. *Nat Methods* 2024, **21**:1492–1500.

50. Liu G, Chen L, Wu Y, Han Y, Bao Y, Zhang T: **PDLLMs: a group of tailored DNA large language models for analyzing plant genomes**. *Mol Plant* 2025, **18**:175–178.