

INTRODUCTION

Making sense of complexity: Advances in bioinformatics for plant biology

Coined by Dutch theoretical biologists in the 1970s, the term bioinformatics originally denoted a broad concept relating to the study of information processing in biological systems, such as ecosystem interaction, neuronal messaging, and transfer of genetic information (Hogeweg, 2011). Subsequently co-opted to describe the sequencing and analysis of molecules (from nucleic acids to proteins), bioinformatics has diverse applications including the analysis, visualization, storage, and generation of data relating to living organisms and the molecular information they carry. Plant biology has reaped dividends from the development and maturation of bioinformatics; it has not only extended our understanding of model plant species such as *Arabidopsis thaliana* (Cantó-Pastor et al., 2021) but also driven innovative solutions to characterize non-model species (Nevado et al., 2014). Both avenues of discovery contribute to key objectives in improving food security, conservation, and biotechnology.

The size and complexity of many plant genomes has historically made their analysis financially and computationally difficult. Frequent polyploidy and repeat element expansion make the elucidation of plant genome sequences challenging (Soltis et al., 2015). Furthermore, high heterozygosity in wild populations, pervasive hybridization, and a lack of inbred lines present roadblocks to analyses such as read mapping and assembly (Kajitani et al., 2019). Long-read technologies have become ever more accessible in recent years, and algorithmic advances have accommodated sequential updates to error models, read lengths, and library types (Michael and VanBuren, 2020). Moreover, novel methods to scaffold contigs and obtain long-range interaction information have driven impressive improvements in genome assembly quality, making telomere-to-telomere genome sequencing projects an achievable goal for many labs (Kress et al., 2022).

Long-read technologies paired with novel mapping algorithms have fueled discovery of new transposable element (TE) dynamics, and there has been an associated resurgence of interest in their role in adaptive trait evolution and phenotypic variation (Schrader and Schmitz, 2019; Pimpinelli and Piacentini, 2020). Bioinformatics developments in this field have led to vast improvements in our ability to detect complex TE mobilization patterns such as

nested insertions and structural variants (Bree et al., 2022; Lemay et al., 2022). Despite these advancements, characterization and annotation of genomic features such as genes and repetitive elements remain challenging due to species-specific genomic configurations, taxonomically patchy reference databases, and a lack of robust benchmarking and quality control. While structural and functional annotation methods still have significant obstacles to overcome, many important contributions have been made to improve the comparison and optimization of these approaches (Caballero and Wegrzyn, 2019). Moreover, the extension and aggregation of existing gene, variant, and repeat annotation software is beginning to allow researchers to combine and curate different algorithmic approaches and databases (Nelson et al., 2017; Kirsche et al., 2023).

The scale of plant diversity to be characterized remains a challenge, however, and incorporating samples from preserved, non-model, or difficult-to-access material requires innovative wet lab and bioinformatics solutions (Lang et al., 2020). Reduced representation sequencing (RRS) methods represent a crucial tool for the study of non-model plants; this adaptation of emerging sequencing technologies has allowed for cost-effective population studies, analyses of historical diversity using herbarium specimens, and phylogenomic explorations on a large scale (Kersey, 2019; One Thousand Plant Transcriptomes Initiative, 2019). Limitations associated with RRS such as paralogous genes, different selection landscapes of coding and non-coding sequences, and missing data are increasingly accounted for with the continuous improvement of software and methodology (Johnson et al., 2016), and integration of -omics data for non-model taxa in online portals creates an ever more accessible environment for researchers to characterize the world's flora (Goodstein et al., 2012).

Bioinformatics, since its inception in biological applications, has been a field in constant flux, with a high turnover of technologies, sequencing platforms, algorithms, and techniques, and the current landscape of bioinformatics in plant sciences is no different. This special issue of *Applications in Plant Sciences* presents five papers that explore bioinformatics approaches to address issues in plant biology, such as genome assembly, reduced representation sequencing, and structural and functional annotation. We summarize these papers here.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Applications in Plant Sciences* published by Wiley Periodicals LLC on behalf of Botanical Society of America.

Reduced representation sequencing methods such as target capture, RAD sequencing, and genome skimming provide powerful tools for phylogenomic studies, especially in cases where whole genome analyses are infeasible or many non-model organisms must be sampled cost efficiently. Bioinformatic methods such as probe design and resolution of paralogous sequences have critical impacts on downstream analyses and interpretations; therefore, clear guidelines and accessible implementation are important to ensure that maximum benefits are reaped by the scientific community. Two papers in this issue discuss aspects of RRS.

Despite recent advances in whole genome sequencing, RRS approaches continue to be of great importance in biodiversity and evolutionary studies, particularly in situations where obtaining fresh plant material is not feasible or the number of samples is very large. In their contribution, Pezzini et al. (2023) provide a comprehensive review of genome skimming and target capture, two techniques used commonly for the study of non-model organisms and difficult material such as herbarium specimens. This review is timely, because while the design of target capture probes (i.e., bait sets) for specific taxa has historically been hindered by the limited availability of genomic resources for non-model organisms, this is likely to change in the next few years thanks to ambitious whole genome sequencing efforts such as the Earth Biogenome Project (Lewin et al., 2022). The rapid growth in the number and taxonomic resolution of bait sets is making analysis of non-model plant species easier by using probes that are universal or cover larger clades. Pezzini and co-authors discuss a variety of approaches utilizing existing resources such as combining universal and taxon-specific bait sets for use in non-model organisms, or combining new results with legacy data to enable broader taxon sampling. Considerations for genome skimming and target capture have similarities; however, the untargeted technique used by genome skimming results in sequence data that are highly dependent on copy number, favoring more frequently represented regions such as those in chloroplasts and mitochondria. Including both project planning and downstream analysis considerations, the authors review the merits and drawbacks of both target capture and genome skimming approaches, providing a valuable resource for researchers who may have a variety of data, taxa, and tissue types at hand.

In their contribution to this issue, Jackson et al. (2023) build on the existing bioinformatic pipelines HybPiper (Johnson et al., 2016) and ParaGone (Yang and Smith, 2014), providing a streamlined version of both pipelines within a Singularity container, vastly simplifying dependency installation and implementation. These two pipelines perform target capture read assembly and paralogy resolution, respectively, and the use of both is a common workflow employed by phylogeneticists prior to species tree inference. Within the containerized pipeline, the authors implement two Nextflow workflows, *hybpiper-nf* and *paragone-nf*, which include improved sample handling and methodological improvements. *Hybpiper-nf* addresses

organization and tractability of large sample sizes, automatically detecting sequence types in BLAST (Altschul et al., 1990) and Diamond (Buchfink et al., 2015) runs and parsing sequence names from read files. Additional improvements over the previous standalone implementations of HybPiper include additional options to manipulate the resolution of chimeric locus assemblies, giving the user greater insight and control over the processing of target capture data. The process of phylogenomic inference is streamlined by the production of correctly formatted files from *hybpiper-nf* that are directly compatible with *paragone-nf*, where four different paralog inference algorithms are implemented (originally described in Yang and Smith [2014]). The authors test their workflow using the Angiosperms353 and Compositae1061 bait sets applied to data sets including Asteraceae and Orchidaceae, demonstrating greatly improved usability and streamlining of the target capture workflow. This new, containerized workflow will provide the non-model plant biology community with more accessible bioinformatic tools to analyze RRS data and greatly streamline new phylogenomic projects.

Transposable elements are a ubiquitous feature of plant genomes, and the revival of interest in TEs and their role in genome dynamics, trait evolution, and evolutionary trajectories has coincided with the emergence of long-read sequencing technologies, which can allow researchers to capture 5' and 3' insertion sites in a single read, a feat not previously possible with short reads. Popular TE annotation software, however, remains computationally inaccessible for some researchers due to long run times and high computational demands. Gonzalez-García et al. (2023) leverage algorithmic advances in long-read mapping techniques to annotate TEs, using a computationally efficient homology-based method employing minimizers. The comparatively high error rate of long reads is a useful proxy for the imperfect sequence conservation between members of TE families, and the authors build on the long-read alignment method used by *Minimap2* (Li, 2018) to reduce run time from hours to minutes, marking an improvement of orders of magnitude in computational efficiency. Moreover, the authors make use of alternatives to commonly used *de novo* TE annotation pipelines (Orozco-Arias et al., 2023), broadening the diversity of bioinformatic resources for TE annotation, a field which, despite its age, still presents significant challenges in model and non-model organisms alike.

The annotation of gene features is a fundamental step in ascribing context to genomic data sets, paving the way for further studies such as expression assays, comparative genomics, and population dynamics. Despite advances in genome assembly methods, genome annotation remains one of the most challenging bottlenecks facing plant genome science, with intron length variation, divergent TE dynamics, and low sequence conservation hampering the annotation efforts of non-model genome projects. In their contribution, Vuruptoor et al. (2023) address the need to improve quantification of structural genome annotation methods, employing a mixture of existing and emerging metrics to

benchmark genome annotation methods. They approach the issue in a robust manner by using a broad diversity of taxa with challenging genomic features such as variable ploidy, high TE content, and large genomes. As well as commonly used metrics such as BUSCO, the authors draw attention to equally informative measures of annotation quality such as the ratio of mono-exonic to multi-exonic genes to detect unlikely gene models and false positive genes resulting from incomplete repeat masking. That the problem of genome annotation is not solved, even in model plant species, is testament to the importance of benchmarking studies such as this, and the inclusion of challenging taxa during software design is vital to ensure non-model plant species can equally benefit from bioinformatic innovations.

Upstream bioinformatics analyses frequently produce an extensive list of genes of interest, for example, transcripts that are differentially expressed between control and perturbed conditions, genes that show signals of accelerated rates of evolution, or particularly duplication-rich gene families. In order to make these results statistically meaningful and human readable, further contextualization is required through categorizing the genes employing the widely used system of gene ontology (GO). In GO, hierarchical structures of molecular functions, cellular locations, or biological processes are arranged from the general to the specific, and these categories represent a universal way to describe gene function. Gene Ontology annotation results in a large amount of data that is difficult to synthesize manually, precluding quick insights into the results of upstream applications. Here, Sessa et al. (2023) describe and test GOgetter, an easy-to-use pipeline for the summarization and visualization of GO annotations from a set of FASTA files and a GO slim mapping file as input. GOgetter combines functionalities for transferring annotations via homology searchers, calculating summaries for every data set, and producing publication-ready graphs. GOgetter is flexible, allowing users to apply different quality and similarity filters as well as use different reference databases to accommodate non-model organisms. Three case studies demonstrate GOgetter's flexibility, wide applicability from bryophytes to angiosperms, and robustness. We anticipate that this software will facilitate the rapid exploration of new transcriptomes and genomes by streamlining the GO annotation process.

Bioinformatics has revolutionized plant biology, enabling researchers to harness analytical advancements and reveal the enormous complexity of plant genomes, relationships, and biology. As technological innovations promise to provide us with ever greater insights, our bioinformatic analyses of novel data types must keep pace by supporting techniques to further our understanding of plant biology, benchmarking methods for complex bioinformatic operations such as genome annotation, and contextualizing biological data in functional or structural terms. This special issue reflects the diversity of approaches to new and old problems in plant biology, showcasing the wide range of applications of bioinformatics in plant biology, and we hope that it will support the continuing


development of bioinformatics tools and methods for a new generation of technological advance.

AUTHOR CONTRIBUTIONS

K.E. prepared the initial draft of the manuscript. All authors contributed to article summaries, reviewed and edited subsequent drafts, and approved the final version of the manuscript.

ACKNOWLEDGMENTS

We thank the authors for their contributions to this special issue, and the reviewers for their thoughtful comments and suggestions. We also are very grateful to the *Applications in Plant Sciences* editor-in-chief Dr. Briana L. Gross and managing editor Beth Parada for their help and guidance throughout the editorial process.

Katie Emelianova¹ 

Diego Mauricio Riaño-Pachón²

Maria Fernanda Torres Jimenez³

¹Department of Botany and Biodiversity Research,
University of Vienna, Vienna, Austria

²Center for Nuclear Energy in Agriculture,
University of São Paulo, São Paulo, Brazil

³Life Sciences Center,
Institute of Biosciences, Vilnius University,
Vilnius, Lithuania

Correspondence

Katie Emelianova, Department of Botany and
Biodiversity Research, Rennweg 14,
A-1030 Vienna, Austria.

Email: katherine.emelianova@univie.ac.at

This article is part of the special issue “Bioinformatics for
Plant Biology.”

ORCID

Katie Emelianova  <http://orcid.org/0000-0002-5981-4442>

REFERENCES

- Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3): 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
- van Bree, E. J., R. L. F. P. Guimarães, M. Lundberg, E. R. Blujdea, J. L. Rosenkrantz, F. T. G. White, J. Poppinga, et al. 2022. A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci. *Genome Research* 32(4): 656–670. <https://doi.org/10.1101/gr.275515.121>
- Buchfink, B., C. Xie, and D. H. Huson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12(1): 59–60. <https://doi.org/10.1038/nmeth.3176>
- Caballero, M., and J. Wegrzyn. 2019. gFACs: Gene Filtering, Analysis, and Conversion to unify genome annotations across alignment and gene prediction frameworks. *Genomics, Proteomics & Bioinformatics* 17(3): 305–310. <https://doi.org/10.1016/j.gpb.2019.04.002>
- Cantó-Pastor, A., G. A. Mason, S. M. Brady, and N. J. Provart. 2021. Arabidopsis bioinformatics: Tools and strategies. *The Plant Journal* 108(6): 1585–1596. <https://doi.org/10.1111/tpj.15547>

- Gonzalez-García, L. N., D. Lozano-Arce, J. P. Londoño, R. Guyot, and J. Duitama. 2023. Efficient homology-based annotation of transposable elements using minimizers. *Applications in Plant Sciences* 11(4): e11520.
- Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, et al. 2012. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research* 40(D1): D1178–D1186. <https://doi.org/10.1093/nar/gkr944>
- Hogeweg, P. 2011. The roots of bioinformatics in theoretical biology. *PLoS Computational Biology* 7(3): e1002021. <https://doi.org/10.1371/journal.pcbi.1002021>
- Jackson, C., T. McLay, and A. N. Schmidt-Lebuhn. 2023. hybpipec-nf and paragone-nf: Containerization and additional options for target capture assembly and paralog resolution. *Applications in Plant Sciences* 11(4): e11532.
- Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J. Wickett. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4(7): e1600016. <https://doi.org/10.3732/apps.1600016>
- Kajitani, R., D. Yoshimura, M. Okuno, Y. Minakuchi, H. Kagoshima, A. Fujiyama, K. Kubokawa, et al. 2019. Platanus-alley is a de novo haplotype assembler enabling a comprehensive access to divergent heterozygous regions. *Nature Communications* 10(1): 1702. <https://doi.org/10.1038/s41467-019-09575-2>
- Kersey, P. J. 2019. Plant genome sequences: Past, present, future. *Current Opinion in Plant Biology* 48: 1–8. <https://doi.org/10.1016/j.pbi.2018.11.001>
- Kirsche, M., G. Prabhu, R. Sherman, B. Ni, A. Battle, S. Aganezov, and M. C. Schatz. 2023. Jasmine and Iris: Population-scale structural variant comparison and analysis. *Nature Methods* 20(3): 408–417. <https://doi.org/10.1038/s41592-022-01753-3>
- Kress, W. J., D. E. Soltis, P. J. Kersey, J. L. Wegrzyn, J. H. Leebens-Mack, M. R. Gostel, X. Liu, and P. S. Soltis. 2022. Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences, USA* 119(4): e2115640118. <https://doi.org/10.1073/pnas.2115640118>
- Lang, P. L. M., C. L. Weiß, S. Kersten, S. M. Latorre, S. Nagel, B. Nickel, M. Meyer, and H. A. Burbano. 2020. Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. *Molecular Ecology Resources* 20(5): 1228–1247. <https://doi.org/10.1111/1755-0998.13168>
- Lemay, M.-A., J. A. Sibbesen, D. Torkamaneh, J. Hamel, R. C. Levesque, and F. Belzile. 2022. Combined use of Oxford Nanopore and Illumina sequencing yields insights into soybean structural variation biology. *BMC Biology* 20(1): 53. <https://doi.org/10.1186/s12915-022-01255-w>
- Lewin, H. A., S. Richards, E. Lieberman Aiden, M. L. Allende, J. M. Archibald, M. Bálint, K. B. Barker, et al. 2022. The Earth BioGenome Project 2020: Starting the clock. *Proceedings of the National Academy of Sciences, USA* 119(4): e2115635118. <https://doi.org/10.1073/pnas.2115635118>
- Li, H. 2018. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* 34(18): 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Michael, T. P., and R. VanBuren. 2020. Building near-complete plant genomes. *Current Opinion in Plant Biology* 54: 26–33. <https://doi.org/10.1016/j.pbi.2019.12.009>
- Nelson, M. G., R. S. Linheiro, and C. M. Bergman. 2017. McClintock: An integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3: Genes, Genomes, Genetics* 7(8): 2763–2778. <https://doi.org/10.1534/g3.117.043893>
- Nevado, B., S. E. Ramos-Onsins, and M. Perez-Enciso. 2014. Resequencing studies of nonmodel organisms using closely related reference genomes: Optimal experimental designs and bioinformatics approaches for population genomics. *Molecular Ecology* 23(7): 1764–1779. <https://doi.org/10.1111/mec.12693>
- One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574(7780): 7780. <https://doi.org/10.1038/s41586-019-1693-2>
- Orozco-Arias, S., L. Humberto Lopez-Murillo, M. S. Candamil-Cortés, M. Arias, P. A. Jaimes, A. Rossi Paschoal, R. Tabares-Soto, et al. 2023. Inpactor2: A software based on deep learning to identify and classify LTR-retrotransposons in plant genomes. *Briefings in Bioinformatics* 24(1): bbac511. <https://doi.org/10.1093/bib/bbac511>
- Pezzini, F. F., G. Ferrari, L. L. Forrest, M. L. Hart, K. Nishii, and C. A. Kidner. 2023. Target capture and genome skimming for plant diversity studies. *Applications in Plant Sciences* 11(4): e11537.
- Pimpinelli, S., and L. Piacentini. 2020. Environmental change and the evolution of genomes: Transposable elements as translators of phenotypic plasticity into genotypic variability. *Functional Ecology* 34(2): 428–441. <https://doi.org/10.1111/1365-2435.13497>
- Schrader, L., and J. Schmitz. 2019. The impact of transposable elements in adaptive evolution. *Molecular Ecology* 28(6): 1537–1549. <https://doi.org/10.1111/mec.14794>
- Sessa, E. B., R. R. Masalia, N. Arrigo, M. S. Barker, and J. A. Pelosi. 2023. GOgetter: A pipeline for summarizing and visualizing GO slim annotations for plant genetic data. *Applications in Plant Sciences* 11(4): e11536.
- Soltis, P. S., D. B. Marchant, Y. Van de Peer, and D. E. Soltis. 2015. Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development* 35: 119–125. <https://doi.org/10.1016/j.gde.2015.11.003>
- Vuruputoor, V. S., D. Monyak, K. C. Fetter, C. Webster, A. Bhattarai, B. Shrestha, S. Zaman, et al. 2023. Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. *Applications in Plant Sciences* 11(4): e11533.
- Yang, Y., and S. A. Smith. 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* 31(11): 3081–3092. <https://doi.org/10.1093/molbev/msu245>

How to cite this article: Emelianova, K., D. M. Riaño-Pachón, and M. F. Torres Jimenez. 2023. Making sense of complexity: Advances in bioinformatics for plant biology. *Applications in Plant Sciences* 11(4): e11538. <https://doi.org/10.1002/aps3.11538>