Prof. dr hab. Jakub Sawicki

Department of Botany and Evolutionary Ecology

Faculty of Biology and Biotechnology

University of Warmia and Mazury in Olsztyn

Plac Łódzki 1

10-719 Olsztyn

email:jakub.sawicki@uwm.edu.pl

tel: (89) 523 39 80

Review of

PhD thesis of Dergi Daba Dinka realized in the Department of Ecology and Biogeography, Faculty of Biological and Veterinary Sciences, Nicolaus Copernicus University in Toruń under supervision of Prof. dr hab. Krzysztof Szpila and dr Marcin Piwczyński, entitled "Building a Model: Developing Genomic Resources for *Ferula communis* (Apiaceae), a Traditional Medicinal Plant"

The review was prepared at the request of the Scientific Council of the Discipline of Biological Sciences, Faculty of Biological and Veterinary Sciences, Nicolaus Copernicus University in Toruń, in accordance with Article 190(2) of July 20, 2018 (Journal of Laws of 2023, item 742 with amendments).

The doctoral thesis of Mr. Mergi Daba Dinka was prepared in the form of a single-author manuscript consisting of 156 pages. The thesis explores the assembly and analysis of the genome of *Ferula communis*, a member of the Apiaceae family, to understand the evolutionary dynamics and genomic architecture underlying the diversity of angiosperms. Utilizing next-generation sequencing technologies, including Illumina and Oxford Nanopore Technologies (ONT), the study successfully assembled the nuclear, plastid, and mitochondrial genomes of *F. communis*, revealing insights into the plant's genome evolution, gene family expansions, and the roles of transposable elements.

The work begins with an extensive, 28-page introduction that delves into the evolutionary significance of whole genomes in angiosperms, the pivotal role of whole-genome duplication (WGD) events, the characteristics of plant genomes, and the genomic resources of angiosperms. It highlights angiosperms' vast diversity, their dominance in various ecosystems, and the crucial insights provided by next-generation sequencing technologies into their evolutionary dynamics. The discussion underscores that all flowering plants are polyploids, having experienced multiple WGD events throughout their evolution. These duplications are linked to significant diversification and adaptation processes, although the direct relationship between WGD and evolutionary success remains complex and varied across different lineages and timescales.

Further elaboration on plant genomes reveals a surprising balance between the occurrence of WGD events and a relatively small average genome size in angiosperms compared to other plant groups. This balance is maintained through mechanisms collectively termed post-polyploidization diploidization, which include DNA loss, chromosomal rearrangements, and gene loss, contributing to angiosperms' unique genomic plasticity and their ability to diversify into new ecological niches.

The introduction also addresses the structure and organization of organellar genomes in angiosperms, including plastids and mitochondria, which have originated from endosymbiotic events. The plastid genome is described as a uniparentally inherited, circular DNA molecule with a quadripartite structure, while the mitochondrial genome is noted for its wide size range, complex structure, and high level of RNA editing. These organellar genomes exhibit significant variation and play essential roles in plant cellular energy production and photosynthesis.

Lastly, the introduction discusses the current state of genomic resources in angiosperms, emphasizing the limited but growing number of sequenced genomes. Despite the sequencing of genomes from several key species, the vast diversity of angiosperms means that much remains to be understood about their genomic organization, evolution, and the link between genomic diversity and plant morphology, chemistry, and ecology. The need for further genomic studies, especially in economically important plant families like the Apiaceae, is highlighted as essential for advancing our understanding of plant biology and for addressing broader ecological and societal challenges.

The introduction effectively establishes the thesis as a significant contribution to the field of genomics and evolution of Apiaceae. However, the objectives of this study listed in chapter 1.4 are rather disappointing. In the context of this part, the conducted research is not hypothesis-driven with clearly pointed out objectives, but rather a list of to-do tasks like: assembly the genome, plastome, mitogenome etc. Even based on the results section, more ambitious objectives could be formulated.

The Materials and Methods section of the thesis, chapters 2.1 to 2.6 with subchapters, provides detailed insights into the methodologies employed for sample collection, DNA extraction, sequencing, assembly, and annotation of the *Ferula communis* genome, as well as

the comparative genomics analyses conducted. The sequences for further processing were generated using both short-reads and long-reads technologies using Illumina and Oxford Nanopore platforms. The assembly of the nuclear genome involved estimating genome size and heterozygosity using k-mer analysis, followed by multiple de novo assembly strategies for both short and long reads, as well as hybrid assembly methods. Tools such as SOAPdenovo2, Meraculous-2D, NECAT, Flye, and DBG2OLC were employed, each contributing to a comprehensive understanding of the *F. communis* genome structure. Both plastid and mitochondrial genomes were assembled using GetOrganelle and NOVOPlasty, with annotation performed through web server-based pipelines such as CPGAVAS2 and GESeq. This process included identifying and annotating protein-coding genes, tRNAs, and repetitive sequences, providing a complete picture of the organellar genomes. Comparative genomics analyses were carried out to understand the evolutionary relationships and gene family expansions or contractions among *Ferula communis* and other related species. Gene orthology was analyzed using OrthoFinder and OrthoVenn3, revealing shared and unique gene clusters across species. Additionally, transposable elements within the *F. communis* genome were identified and categorized, highlighting the significant proportion of the genome they occupy.

Throughout these sections, a combination of bioinformatics tools and custom scripts were utilized to handle and analyze the sequencing data. This comprehensive approach ensured a detailed exploration of the genomic landscape of *Ferula communis*, contributing valuable insights into its genetic makeup and evolutionary history.

This chapter is generally well written, however in my opinion, some of the methodological assumptions could be better chosen, at both wet-lab and bioinformatics level:

1. Why was RNA editing sites prediction used only for mitogenome since plastomes are edited too?
2. What about detecting structural heteroplasmy of plastomes using long reads? In most Angiosperms, at least 2 structural variants are present.
3. Why were the generated long-reads not used for mitogenome assembly? Especially in the context of medicore assembly results of this molecule.
4. The proper genome annotations can't be based exclusively on prediction methods and in the case of this thesis RNA-seq analysis wasn't performed. Therefore novel genes that could be specific for *Ferula* couldn't be identified.
5. And last but not least, the genome assembly approach lacks any wet-lab scaffolding method like optical mapping, Hi-C or PoreC sequencing. Nowadays it's rather a standard part of the plant genome assembly process required by every reputable journal.

Chapter 3 of the dissertation provides an in-depth analysis of the genomic characteristics of *Ferula communis*, revealing significant findings through various subsections, each focused on different aspects of genomic data. The chapter begins with the sequencing efforts, detailing the generation of a vast amount of data from both Illumina and Oxford Nanopore Technologies, setting a solid foundation for the complex assembly and analysis tasks that follow. The meticulous approach to quality assessment and data preparation underscores the rigorous methodology employed in this study.

The estimation of the nuclear genome size and heterozygosity provides essential insights into the complexity and diversity of the *F. communis* genome. The detailed exploration of various genome assembly strategies, from short-read assemblers to hybrid approaches, showcases the challenges and considerations in assembling large and complex plant genomes. The choice of

DBG2OLC for a comprehensive assembly, despite its fragmentation, reflects a strategic decision to capture the genome's extensive repetitive sequences and structural complexity.

The quality assessment of genome assemblies, particularly through the comparison of the Flye and DBG2OLC assemblies, demonstrates a careful and methodical approach to selecting the best possible assembly for further analysis. This section not only highlights the importance of assembly quality in genomic studies but also illustrates the use of various metrics and tools to evaluate assembly integrity and completeness.

The characterization of the *F. communis* genome reveals its substantial repetitive content, emphasizing the significant role of transposable elements in shaping genome architecture and evolution. The gene prediction and functional annotation efforts, comparing the *F. communis* genome with that of *Arabidopsis thaliana*, provide valuable insights into the gene structure and functional potential of this species, despite the challenges in direct comparison due to the absence of comprehensive transcriptomic data.

The analysis of transposable elements further enriches our understanding of the *F. communis* genome, highlighting the prevalence of retrotransposon elements and their contribution to genomic diversity and evolution. The orthology analysis, revealing extensive gene sharing among species and a high number of species-specific orthologs in *F. communis*, underscores the unique evolutionary trajectory of this species within the Apiaceae family.

Gene expansion and contraction analysis provides a dynamic view of the genome, illustrating the evolutionary processes that have shaped the gene content and organization in *F. communis* and related species. The detailed examination of the plastid genome, including its assembly, gene content, and evidence of positive selection on specific genes, adds another layer to the understanding of *F. communis*, highlighting the functional and evolutionary significance of the plastid genome.

In summary, Chapter 3 presents a comprehensive and detailed examination of the *Ferula communis* genome, from sequencing and assembly to gene annotation and evolutionary analysis. The findings contribute significantly to our understanding of plant genomics, particularly within the Apiaceae family, providing a valuable resource for future research on plant evolution, diversity, and adaptation. Besides the limitation of used methods, I found only one minor issue within this chapter, I'm not recommend using the "scaffold" term in the case of assembled mitochondrial contings, since any scaffolding approach was used according to the Material and Methods section.

The main part of the work is crowned by a 21-page discussion referring to the latest discoveries in the addressed topic within the context of genomic evolution, gene family expansions, the impact of transposable elements (TEs), and the structural and evolutionary insights into the plastid and mitochondrial genomes. This chapter methodically discusses the role of TEs as a major factor in genome size variation, highlighting the substantial presence of TEs, particularly Gypsy and Copia elements, which are instrumental in genome expansion through dynamic processes involving amplification and recombination. The discussion extends to the substantial gene family expansions observed in *F. communis*, attributing this phenomenon to whole-genome duplications (WGD) events that have historically influenced gene diversity and chromosome numbers, enabling the species to adapt to diverse habitats.

Further, the chapter contrasts *F. communis* with closely related species, elucidating the evolutionary relationships and the unique genetic components contributing to its adaptive traits and metabolic diversity, especially in the production of secondary metabolites like terpenoids. This is complemented by an examination of unique genes in *F. communis* that are indicative of specialized functions such as disease resistance and stress response, suggesting a genetic basis for its adaptation to various environmental challenges.

The discussion on the structure and evolution of the plastid genome (pDNA) in *F. communis* explores its highly conserved nature, the presence of distinct haplotypes, and the implications of sequence divergence, particularly in non-coding regions which play a role in phylogenetic relationships within the Apiaceae family. It also addresses the phenomenon of positive selection on specific pDNA genes, implying adaptive evolution to particular habitats.

The mitochondrial genome's multipartite structure, characterized by non-circular scaffolds and the presence of intermediate repeats conducive to recombination, underscores the complexity and adaptability of plant organelle genomes. The significance of RNA editing in the mitochondrial genome is highlighted as a critical process for error correction, expanding the genetic code, and ensuring accurate protein synthesis, which is essential for the plant's adaptation and survival.

In conclusion, Chapter 4 synthesizes the comprehensive genomic analysis of *F. communis*, emphasizing the evolutionary processes, structural adaptations, and the functional diversification that underpin its ecological success. It underscores the importance of *F. communis* as a model system for exploring genomic and evolutionary questions, particularly in understanding how genetic elements influence genome structure, function, and adaptation to extreme environments.

Final conclusions

The entirety of the doctoral thesis presented to me for evaluation constitutes a valuable contribution to the understanding of genomic diversity of Apiaceae and provides valuable resources for further studies. Modern molecular biology and bioinformatics methods were applied for characterization of nuclear and organellar genomes of *Ferula communis*.

The reviewed doctoral dissertation gives the impression of a thorough scientific thesis and fulfills all the requirements set for doctoral theses according to the law of July 18, 2018 (Article 187) Journal of Laws of 2023, item 742.

In light of the above, I kindly request the High Scientific Council of the Discipline of Biological Sciences, Faculty of Biological and Veterinary Sciences, Nicolaus Copernicus University in Toruń to accept the doctoral thesis of Mr. Mergi Daba Dinka and to allow the doctoral candidate to proceed to the next stages of the doctoral process.

Sincerely,

Jakub Sawicki

Olsztyn, 18.02.2024