Kraków, February 15, 2024

REWIEW

of the doctoral thesis of Mr. Mergi Daba Dinka, MSc

titled "*Building a Model: Developing Genomic Resources for Ferula communis (Apiaceae), a Traditional Medicinal Plant*"

Introduction

With the dynamic development of high-throughput DNA sequencing technologies, whole genome sequencing and assembly of reference genomes has become a routine strategy, paving a foundation for further analyses in structural, functional and comparative genomics. Availability of thousands of plant genomes in the public domain opens unprecedented possibilities in elucidating genetic basis of organismal biology as well as molecular mechanisms of evolutionary and adaptive processes. The results reported by Mr. Mergi Daba Dinka, focusing on the assembly and analysis of nuclear, plastid and mitochondrial genomes of *Ferula communis*, add a new element to that picture, especially interesting for those investigating the Apiaceae family. Thus, in mere twenty years, *de novo* assembly of a complex plant genome has become feasible as a PhD project, rather than a challenging effort requiring collaboration of several laboratories. Mr. Dinka presented results on a draft assembly of the nuclear genome, a complete assembly of the plastid genome and scaffolds of the mitochondrial genome of *F. communis*, utilizing a combination of short reads from Illumina and long reads from Oxford Nanopore.

Structure of the dissertation

The dissertation is presented in the form of a monograph comprising 156 pages of body text accompanied by the funding statement (one page), abstracts in English and in Polish (two and three pages, respectively), acknowledgements (three pages), table of contents (four pages), list of figures (two pages), and list of tables (one page). The main body of the dissertation includes introduction (27 pages; divided into subchapters), objectives of the study (one page), materials and methods (23 pages), results (28 pages), discussion (22 pages, including one page of conclusions), bibliography (49 pages), and two appendices (six pages). Additional supplementary files were provided on CD-ROM. The overall structure of the dissertation is complete and typical for PhD thesis in the form of monograph. No parts of the presented research have been published to date.

The documentation is complete, it complies with legal requirements and allows substantive evaluation of scientific achievements of the PhD candidate.

Evaluation of the scientific merit of the dissertation

In the first section of the Introduction, the Author highlighted the significance of whole genome assemblies to progress in plant science. He described general properties of genomes of flowering plants, pointing at the importance of repetitive DNA to the overall genome size structure and briefly presenting ways by which transposable elements (TEs) may affect gene function. He also characterized the structure of plant extranuclear genomes and finally provided a brief overview of genomic resources in Apiaceae, the family comprising *F. communis* among many economically important species of vegetables, condiments and medicinal plants. In the second section of the Introduction, he broadly characterized the Apiaceae family and subsequently focused on the tribe Scandiceae, which includes the model species for Apiaceae and the most economically important vegetable, i.e. carrot (*Daucus carota* L.). He argued that Scandiceae was an interesting taxon to study with respect to evolutionary and adaptive mechanisms, hence it was important to develop novel genomic resources for a range of species from that group, and the present effort of assembling the genome of *F. communis* was well-justified and highly welcome. It is why the major objectives defined by the Author of the present PhD dissertation were to assemble and annotate nuclear, plastid and mitochondrial genomes of *F. communis* and to perform comparative analyses with other available genome assemblies of Apiaceae.

The Materials and Methods section starts with a description of the origin and sampling of the *F. communis* specimen used for WGS, the respective protocols for Illumina and ONT sequencing, followed by a description of strategies used to build the assemblies of the nuclear, plastid and mitochondrial genomes. The provided diagram (Figure 4) helped depicting the complexity of the applied bioinformatic analyses. Subsequently, methods used to annotate the nuclear genome and to perform comparative analysis among Apiaceae were presented. Finally, strategies used to assemble and analyze organellar genomes of *F. communis* were described.

In the Results, standard statistics of the raw and filtered sequencing data obtained from Illumina and ONT platforms were presented. Notably, different strategies used to assemble the nuclear genome yielded strikingly different genome size estimates, mostly exceeding the expected genome size of ca. 1.7 Gb. Other measures, e.g. N50 and BUSCO score, also differed widely among the applied tools and parameters. The Author decided to use the Flye assembly, spanning 2.8 Gb, for downstream analyses. It was annotated using *in silico* methods (the annotation was not supported by transcriptome data), which yielded a rather large number of 65 to 79 thousands of genes. The provided information, coupled with the fact that the analyses indicated an unexpectedly low level of heterozygosity, makes me wondering if in fact the donor plant could have been extremely heterozygous and some scaffolds represented alternative variants derived from each parental genome. It would explain the larger than expected assembly size and gene count, as well as the reported expansion of certain gene families (presented in more detail in the Discussion; the expansion was the most pronounced in rapidly evolving and frequently clustered gene families, e.g. pathogenesis-related and governing terpene synthesis). It would certainly help if the DNA content was measured by flow cytometry in the somatic cells of the donor plant to provide an estimate of the actual size of the nuclear genome.

The subsequent sections of Results describe the structure of completely assembled plastid genome and the scaffold-scale assembly of the mitochondrial genome. With respect to the latter, the Author proposed a linear organization of mtDNA, which to me seems rather unlikely. I believe a standard master circle still exists, while obviously a number of sub-genomic variants resulting from recombination can be present and those more frequent ones are likely represented by the 14 reported scaffolds. Interestingly, the Author highlights presence of a 58 kb-long intron in *nad1* while the largest scaffold is only 40 kb-long (see Figure 11). While it seems to be an artifact as the intron would span more than one fifth of the whole mt genome, it implies that the higher level of organization, i.e. the master circle, was actually used to annotate mt genes.

I have few more detailed questions and comments to the M&M and Results chapters:

- were there any particular criteria to select the plant for sequencing or was it a random choice?
- what are the 5,391 ‘structural variants’ reported on p. 56 and in Table 6? As I see it, structural variants can only be called when different genomes are compared.
- what does the ‘number of transcripts’ refer to (Table 8)? The numbers differ from the number of genes (more than genes for Braker2, less than genes for Funnotate). As no transcriptomic data were reported, it seems confusing.
- In Table 9, orders (bolded) and superfamilies are specified, the title of the respective column (‘Class’) is misleading. In the column labeled ‘Count’ numbers of hits are provided and it should not be confused with numbers of TE copies.
- I suggest to draw scaffolds in Figure 11 up to scale as their sizes differ significantly.

All the reported results were thoroughly discussed in the Discussion chapter, it also included sections (e.g. that related to gene family expansions) which could have been placed in the Results chapter. The first section of the Discussion addressed technical limitations of the applied assembly strategy. The Author concluded that the current assembly should be classified as ‘draft’ and additional methods should be employed to make it more contiguous – I fully agree with his opinion. Subsequently, he compared the genome of *F. communis* to that of *A. thaliana* highlighting their plasticity and pointed at the significance of TEs in shaping plant genomes. He also discussed gene content and gene expansion in *F. communis* (see my comments above), compared orthologs among Apiaceae and characterized genes he classified as ‘unique’ to *F. communis*, i.e. those having ‘no apparent homologs’ among phylogenetically close relatives (page 91). However, no criteria for such ‘apparent homology’ were provided. The results on the characterization of mitochondrial and plastid genomes were also adequately discussed and all findings were summarized in the final paragraph titled ‘Conclusions’.

Conclusion

The PhD thesis of Mr. Mergi Daba Dinka fulfills requirements of a doctoral dissertation. The reported results are of high quality and novelty, they provide basis for further research on Apiaceae genomics. The theoretical section of the thesis shows knowledge of the doctoral candidate on the subject of investigation. The research was largely performed *in silico* and the PhD candidate is a competent researcher capable of utilizing tools essential in plant structural genomics. The study was well designed and competently conducted and adequately discussed.

The dissertation fulfills all requirements indicated in the currently binding regulations (art. 187 Ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce, tekst jednolity: Dz.U. z 2021 r., poz. 478). In particular, it provides an original solution of a scientific problem, it confirms the general theoretical knowledge of Mr. Mergi Daba Dinka in the discipline of biological sciences and his ability to conduct research.

Thus, I put forward a motion to the Scientific Board of the discipline of biological sciences, Nicolaus Copernicus University in Toruń, to admit the PhD candidate, Mr. Mergi Daba Dinka, to further stages of the procedure.

prof. dr hab. inż. Dariusz Grzebelus