ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ
Faculty of Biology
Institute of Molecular Biology and Biotechnology

Zofia Szweykowska-Kulińska, prof. dr hab.
Department of Gene expression

Poznań, 25.01.2024

Opinion on PhD dissertation of MSc Mergi Daba Dinka entitled "Building a Model: Developing Genomic Resources for *Ferula communis* (Apiaceae), a Traditional Medicinal Plant."

The PhD dissertation written by MSc Mergi Daba Dinka was prepared in the Department of Ecology and Biogeography at the Faculty of Biological and Veterinary Sciences of Nicolaus Copernicus University. This work was carried out under the supervision of Prof. Dr. Hab. Krzysztof Szpila and assistant supervisor Dr Marcin Piwczyński. This work is focused on providing new genomic data on plant species from Apiaceae family – *Ferula communis*. Apiaceae represents one of the largest families in seed plants consisting of more than 3800 species. Within this family there are economically important species used as vegetables, herbs and spices like *Daucus carrota* (carrot), *Apium graveolens* (celery), *Petroselinum crispum* (parsley), *Foeniculum vulgare* (fennel), *Corriandrum sativum* (coriander). There are also very well known toxic plants within this family like *Cicuta virosa* (water hemlock). From this short description it is clear that members of this family are extremely reach in phytochemicals important in food and pharmacological industry. Thus it is important to learn about genomic resources within this family, both from the applicative as well as from the basic science point of views. The last one is important mainly to understand plant genome evolution processes within Apiaceae family.

Ferula communis belongs to Apioideae subfamily, Scandiceae tribe which is known to contain species living in various habitats, exhibiting various floral morphology an all types of

ul. Uniwersytetu Poznańskiego 6, Collegium Biologicum, 61-614 Poznań, Poland
tel. +48 61 829 59 50
ibmib@amu.edu.pl
www.ibmib.amu.edu.pl

fruit morphology. The most economically important Apiaceae family species belong to this tribe. *F.communis* was recently located in the subtribe Ferulinae with sister groups Daucinae and Torilidinae. This species is diploid and contains 22 chromosomes (2n=22). In *Ferula* genus substantial differences in genome size between species have been observed. Moreover, in the case of *F.communis* incongruences between organellar and nuclear markers have been reported. These data suggest that homoploid hybridization could take place in the *F.communis* species. All these data prompted the Candidate to perform Illumina and Nanopre sequencing experiments (short and long reads) in order to learn about nuclear, chloroplast and mitochondrial genomes of *F.communis*. The Candidate decided to use DNAseq data to assemble nuclear genome of *F.communis*, to annotate protein coding genes and transposons, to perform comparative studies in the assembled genome with other Apiaceae genomes found in different repositories and on the basis of this comparison to draw conclusions on the studied species history and its evolution, and finally to assemble, annotate and characterize chloroplast and mitochondrial genomes of *F.communis*. According to me all goals of the presented PhD dissertation have been fulfilled.

Having DNA seq data the Candidate tested various algorithms for nuclear genome assembly. Illumina platform generated more than 10^9 trimmed and filtered pair-end high quality reads of 150bp in length while ONT-MinION platform – more than $4{,}5 \times 10^6$ clean reads with the average length 7419 bp and N50 read length – 14266 bp. Different methods and algorithms estimated *F.communis* genome size within the range of 1.7Mb to 3,1Mb. After the genome assembly using other algorithms it seems that the genome size estimated for 2,7-2,9 MB should reflect the real size of *F.communis* genome. I would like to listen to the opinion of the Candidate in this matter. Surprisingly the heterozygosity seems to be rather low within the

ul. Uniwersytetu Poznańskiego 6, Collegium Biologicum, 61-614 Poznań, Poland
tel. +48 61 829 59 50
ibmib@amu.edu.pl
www.ibmib.amu.edu.pl

plants ranging from1-3%. Plants were collected from Menorca – a small island in the Mediterranean region – could it be the reason for the low heterozygosity values?

Various parameters were assessed when a couple of available algorithms for genome assembly were used. One of them was the percentage of complete BUSCO genes that allows for the identification of complete, duplicated and fragmented genes, contig numbers and contig N50. Of all assemblers used Flye and DBG2OLC that assembles long-reads were chosen for subsequent analysis. After genome assembly quality assessment it turned out that Flye outperformed DBG2OLC in several parameters. That is why Flye assembler was used for further analyses. Interestingly, more than 600 Kb of bacterial and 500 kb of viral sequences have been identified forming roughly 20 contigs each. These were removed from the further analyses.

Final characterization of *F.communis* nuclear genome assembly revealed its length to be estimated for 2,772 Mb, N50 value of 0,17Mb, L50 548 and 59178 scaffolds. The number of scaffolds is very high accounting that *F.communis* has 22 chromosomes. I understand that ideal number of scaffolds should be 22. What could be the reason, what kind of sequences are missing for assembling the genome into longer scaffolds? I would like to listen to the Candidate opinion in this matter. GC content was estimated to be 34% while repetitive sequences account for 87% of the genome. I wonder what is the GC content in other Apioideae species? Is it similar? Using two different algorithms for gene prediction and using *Arabidopsis thaliana* genome as a reference the number of genes was estimated between 65k to more than 80 k and these numbers are twice than that of *A.thaliana*. Average number of exons in *F.communis* when compared to *A.thaliana* is smaller (4,2 in contrast to 9,7) while the number of exons is similar in both species. These discrepancies are reasonably explained by the Candidate: there are no data on alternative splicing events in *F.communis* primarily

ul. Uniwersytetu Poznańskiego 6, Collegium Biologicum, 61-614 Poznań, Poland
tel. +48 61 829 59 50
ibmib@amu.edu.pl
www.ibmib.amu.edu.pl

because of the absence of tissue-specific transcriptomic data while in the case of *A.thaliana* extensive studies on AS had been carried out. Transposable elements represent the majority of the genome and represent both retrotransposons and transposons. In the first group LTR RTE (*Copia* and *Gipsy*) are mainly presented while DNA TE represent a small part consisting mainly of helitron and CACTA transposable elements. There is an intriguing information in the Candidate's thesis – the presence of unknown LTR elements that represent 18.51% of the whole genome. Are they also absent in *D.carrota* or *Corriandrum sativum* and other Scandiceae or even Apaiceae species? For sure this a story that one should work further on.

Comparative orthology studies using several species from Apioideae subfamily and *Vitis vinifera*, *Brassica carinata* and *Lactuca sativa* (outgroup species) allowed to divide more than 400K protein sequences to 40K orthogroups. *F.communis* was found to have the highest number of species-specific orthologs (more than 3000 gene clusters). GO analysis revealed that the most abundant group (more than 800) is associated with DNA integration gene family, the next in row were genes associated with DNA recombination and RNA-dependent DNA polymerase-activity. Thinking about these groups I think that they are mainly associated with mobile elements activity that occupy more than 87% of *F.communis* genome. I would like to have a comment form the Candidate on this matter.

Comparison of the number of gene clusters and protein sequences within Apiaceae species revealed the highest shared number of gene clusters or protein coding genes between *F.communis* and *Corriandrum sativum*. However, 2003 of these gene clusters lack annotation. The rest was found and annotated in GO or Swiss-prot databases.

Analysis of gene expansion and contraction revealed that F.*communis* and *C.sativum* have the highest number of gene duplications within Apiaceae family. Conversely, *D.carrota* and

Apium graveolens have the highest number of gene contraction events. If there is any evolutionary explanation for these observations I would like to ask the Candidate to comment on this matter.

Final part of the Results section describes the assembly of chloroplast and mitochondrial genomes. In the case of the first one small changes have been observed when this genome was compared to other known seed plant genomes. The most interesting discovery is the presence of two chloroplast haplotypes with the reverse orientation of the SSC region (which most probably is a result of intramolecular recombination events between two IRs) and the lack of *ycf15* gene (of unknown function) in comparison to other Ferula species. The ratio of non-synonymous to synonymous mutations (dN/dS) was calculated for cp genome and showed strong positive selection in the case *ccsA* gene which encodes a protein required for heme attachment to c-type cytochromes.

Mitochondrial genome assembly revealed the presence of 16 scaffolds instead of single circular mitochondrial genome. I wonder what was found in the case of other Apiaceae/Scandiceae mitochondrial genomes? Do they represent a set of master and mini circles or have also complex, unsolved structure?

Describing gene structure of cp- and mt-genomes the Candidate is writing only about introns. I would like to know whether these are (more likely) introns from group II and some are from group I? Would be interesting to know in comparison to other cp- and mt-genomes.

The last part of Results section is devoted to mitochondrial editing. A very interesting observation concerns editing of mt-genes. All of them seem to introduce non-synonymous changes (70-69% in the second codon position and 29-30% in the first codon position).

ul. Uniwersytetu Poznańskiego 6, Collegium Biologicum, 61-614 Poznań, Poland
tel. +48 61 829 59 50
ibmib@amu.edu.pl
www.ibmib.amu.edu.pl

ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ
Faculty of Biology
Institute of Molecular Biology and Biotechnology

Where there also some editing sites introducing synonymous amino acid? Another question deals with the observed efficiency of editing. When observing editing – do you see also unedited forms? If yes, what is the percentage of edited to non-editied sites in mt- mRNAs?

Finally, I found intriguing that mtDNA can be found in cp-genome as an effect of introgression. I would like the Candidate to introduce this topic in a more complete form during the PhD defense.

The whole PhD dissertation is written in a classical form having introduction, aim of the study, materials and methods, results and discussion. Many supplementary data are provided on the CD plate that unfortunately, I was not able to open because I have no device anymore to open it.

All parts of the PhD dissertation are logically written in a friendly way. I really appreciate the form since being not a bioinformatician I was warried to encounter problems upon reading. But this was not the case. I regard this PhD achievement as an important step towards understanding plant genome evolution and plasticity. *F.communis* genome draft is quite informative and already suggests many interesting evolutionary trends that can be observed within various Apiaceae species. I noticed that Dr Piwczyński and Daba Dinka have already one paper published together in Molecular Phylogenetics and Evolution however, on other studies, not related to PhD dissertation results. I hope that the results presented in the dissertation will be published soon.

The reviewed doctoral dissertation presents a thorough scientific thesis containing novel discoveries and consequently fulfills all the requirements set for doctoral theses. All questions and comments I wrote in my revision do not lower the value of this dissertation. I

ul. Uniwersytetu Poznańskiego 6, Collegium Biologicum, 61-614 Poznań, Poland
tel. +48 61 829 59 50
ibmib@amu.edu.pl
www.ibmib.amu.edu.pl

ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ
Faculty of Biology
Institute of Molecular Biology and Biotechnology

just expect the Candidate to answer theses questions/address comments during the PhD defense.

In light of the above, I recommend the Council of the Discipline of Biological Sciences at Nicolaus Copernicus University in Toruń to accept the doctoral thesis of MSc Mergi Daba Dinka and to allow the doctoral Candidate to proceed to the next stages of the doctoral procedure.

Zofia Szweykowska-Kulińska

ul. Uniwersytetu Poznańskiego 6, Collegium Biologicum, 61-614 Poznań, Poland
tel. +48 61 829 59 50
ibmib@amu.edu.pl
www.ibmib.amu.edu.pl