Comments to the Ph.D. thesis by Magdalena Lučanová

The thesis consist of (i) a brief review of the genome size issue in plants, later at the end of the thesis appended with a conclusion chapter and references, (ii) text of applicant´s paper draft handling intraspecific genome size variation in one model species, (iii) the text of five (of 21 existing) papers co-authored by the applicant and already published in international impacted journals.

The review part (11 pages) is written in understandable way and covers full spectrum of topics related to the current genome size research: genome size variation among and within species, its reasons and evolutionary and/or ecological consequences, and methodical aspects of genome size measurements.

Criticism of the review part

- "Plants" in the thesis are mainly represented by angiosperms and other plant groups are reflected really to a very minor extent in the text. In this respect the title of the work seems excessively broad and instead of "in plants" it should be used "in flowering plants" (in angiosperms).
- It´s a pity the review part does not refer (or even intentionally exclude) the applicant´s published papers which are presented in detail later in the thesis. This may evoke questions about the completeness and topicality of the provided review and prevents natural evaluation of the applicant´s scientific contribution, which must be done separately in another part of the thesis. It will be also fine if the concluding paragraph and future perspectives will be placed directly after the review part and not somewhere at the end of the work in between several pages of references.
- References are sometimes lacking by some important statements (e.g., by indication conversion of pg to Mbp or the genome size variation in protists in the first paragraph of the review). Applicant also tend to use referencing to reviews rather than to the original, primary papers (namely in the second half of the review). This seem effective for writing but prevents rigorous evaluation of how deep applicant really navigates in the discussed problems.
- Towards the end, the review starts to be slightly worse structured and argued compared to the initial parts (e.g., parts handling genome size and phylogeny; broken sentence at the end of page 18 and start of page 19).

The five published papers presented in the applicants´ Ph.D. thesis were published in recognized international journals. In a paper handling genome size of invasive plants in the Preslia journal, the applicant serves as the first author (corresponding author is not indicated in this paper). In any of the four presented or the total of 20 co-authored papers, the applicant does not serve neither as a first (major) or a corresponding author. From the provided "contribution statement" it seem applicant was mainly responsible for plant sampling and flow cytometry measurements.

Due to the performed conversions the current text of papers contains numerous concatenated words, sometimes parts of the original text are missing (such as citation to Hollister et al. 2012 in the paper from Bot. J. Linn. Soc. 2016; or a dash between 5 and 8 °C in page 96, resulting in 58 °C a significantly changing the provided information) or information-carrying formatting is lost (such as bolding of text in Table 1 of the paper about invasive plants, which should marked species analyzed for the first time).

The draft of the paper is the only paper where the applicant clearly serves both as the first and corresponding author. The paper examines heritability of genome size in intraspecific crosses of *Taraxacum stenocephaum* (Asteraceae), a species with very high intraspecific variation reported by the applicant´s supervisor few years ago (Trávníček et al. 2013). It also test possible link between genome size and several fitness parameters of the F1 progeny. The positives of the paper are the properly selected model species, large number of examined plants, rigorously tested sexual system, and the attempt to explain the existing genome size variation by detailed karyological observations. The paper, however, also contains numerous problems which must be carefully addressed before the paper would be suitable for publication. In the current form I would suggest "reject with resubmission welcomed", namely due to the problems with statistical analysis and the possible extensive changes of the related text. Detailed review of the paper for the purpose of its future revision is appended following the "questions" section.

Questions to the review part

- In the review I lack notice about (cyto)skeletal DNA theory of genome size evolution (Cavalier-Smith in J. Cell Sci. 1978 and Ann. Bot. 2005). Could you explain why you prefer nucleotype theory and the large genome constraint hypothesis over this one to explain existing genome size variation in plants?
- Large scale genome constraint hypothesis (Knight et al. Ann. Bot. 2005) arguing for overall disadvantageous effect of large genome size, and frequently mentioned in the applicant´s review, is mainly formulated based on studies from flowering plants (angiosperms). Isn´t large genome size typical of non-flowering plants, such as ferns and gymnosperms successfully surviving Mesozoic and Cenozoic global climatic changes, on the contrary, an evidence for the long-term evolutionary superiority of large genome size?
- In the review you repeatedly comment correlation of genome size with rate of photosynthesis. Could you explain backgrounds of such a relationships (i.e. how it could be theoretically justified)?
- In page 20 you claim that "association between genome size and invasiveness is based on assumption that small genomes have faster division and therefore faster growth". Is this assumption right and has it been somewhere seriously tested that plants with small genomes really grow faster? What about the situation in spring geophytes?
- Could you explain the statement on page 21 that "it is more difficult for small genomes to become and stay large and easier for large genomes to became and stay small"?
- I disagree with your claim that polyploidy is major factor in genome size evolution (appearing several times in the review) noting for example *Arabidopsis thaliana* as a counterevidence. Which facts led you to this conclusion?
- You claim that mixed ploidy populations are very important to study polyploid evolution and in another part of the review that "polyploids often have different ecological niches and can occupy different regions". I am somehow surprised the results from mixed ploidy populations, intensively studied in your home institution, were not used to support this claim. Could you summary what the research of the mixed ploidy populations in your institute tell to this question (ecological and geographical differentiation of diploids and polyploids)?

Questions to the paper handling genome size in invasive plants (70% contribution declared by the applicant)

- The paper has good and attractive idea and I like its precise description of statistical methods and critical comments of the results in the discussion chapter. The major weakness I see in the inconsistent and ambiguous usage of "invasiveness" categories: (i) term "non-invasive species" is used both as the opposite to neophytes (i.e. for "native species") and simultaneously also for a category within neophytes, (ii) 93 measured neophytes are called "naturalized aliens" in Fig. 2, and (iii I am also not sure what is the difference between "naturalized aliens" and "naturalized plants" in the third paragraph of the discussion. I am therefore afraid that many authors perhaps understand and cite this paper incorrectly. Could applicant explain the used terminology?
- Why authors did use data from the Plant C-value database only for non-invasive (native) taxa (>4000 species) but not also for the invasive ones, particularly those missing in their study? Could applicant also explain how they selected the data from the Plant C-value database when two or more data were available for one species?
- The argumentation refuting meaningfulness of phylogenetic corrections in the discussion I consider incorrect. Could applicant show an example when phylogenetic correction can bring significant results even if normal regression will fail to detect any relationship?
- We have recently measured genome sizes in majority of Czech vascular plants. They show good correspondence with the genome size of invasive plants in the applicant´s paper but there are also few exceptions (all values recalculated to Prague genome size standards), such as *Hesperis matronalis* (7.61 pg Praha vs. 16.03 pg Brno), *Sedum hispanicum* (5.39 pg Praha vs. 0.865 pg Brno), *Silene dichotoma* (5.89 pg Praha vs. 2.92 pg Brno), *Trifolium hybridum* (1.9 pg Praha vs. 1.14 pg Brno; genome size form Praha refers rather to the morphologically similar *T. repens*). Could applicant check identification of these species in the stored herbarium vouchers?

Questions to the presented paper draft (applicant is the first and corresponding author)

- You claim that mean genome size in F1 offspring match well the expected parental mean. However, looking to the Fig. 2, it is clear that mean genome size of offspring is always higher than the parental mean. How it can happen?
- Did polynomial (quadratic) relationships remained anywhere in the data (particularly Fig. 6, 9, 10), when analyzed with adequate statistical methods?
- What was the difference between samples measured individually using PI dye and the same samples when measured jointly with DAPI (i.e. in double peaks)? Why didn´t you use consistently DAPI when it provides better signal resolution?
- According to Fig. S1, leaves started to dramatically reduce their length after ±150 days of cultivation. Is it realistic?
- You counted chromosomes and observed size of their satellites in some plants but did not measure their gnome size. Wasn´t there really any piece of leaf or root to measure its genome size with FCM?
- By measuring genome size of the used genome size standard, the distance of peaks from the primary standard was nearly 4-fold, i.e. clearly much more than recommended by you former supervisor. Is this genome size value still used in your lab for this standard? How do you check for signal linearity?

Major critics and weaknesses of the paper draft (to be considered during paper revision)

- Authors claim that mean genome size in F1 offspring match well the expected parental mean and document this by very high correlation between both parameters. However, looking to the Fig. 2 with the data, it is clear that mean genome size of offspring is consistently higher than the parental mean. This fact must be corrected in the results and the reasons rigorously explained in the discussion. By testing the match of the data, authors should correctly test the x=y dependence in the data, not a simple existence of a correlation.
- By testing several relationships authors show polynomial (quadratic) relationship in the data where "best" or "highest" values of dependent variable (leaf length, leaf number, aboveground biomass) are typical for mid-sized values of the genome size (independent variable). These relationships may easily originate as a "data density sampling artifacts" and must be tested in a different way (e.g., by comparing with a null model). Simply if a dependent variable (y) is limited from the left (number of leaves, days to germination) and therefore has right skewed distribution with some extremes, its highest values will most probably occur in the most densely sampled interval of the independent variable (x), such as around the mean genome sizes in the authors´ genome size dataset. Similar figures (like Fig. 6, 10) with the quadratic like dependencies may be easily generated by combining randomly generated data from normal and lognormal distribution (in R for example using: plot(rnorm(100,0,3),rlnorm(100,0,1)) which hardly be claimed to reflect any actual quadratic relationship. In the same line I acknowledge authors used Poisson error structure by analyzing leaf numbers. However, such approach must be applied also in all other non-negative variables which necessarily have asymmetrical distribution errors (leaf length, number of days to flowering, germination, or aboveground biomass; Figs. 6, 9, 10). These cannot be simply analyzed with simple linear regression.
- By measuring genome size of the used genome size standard, the distance of peaks from the primary standard was nearly 4-fold. This is much larger than recommended by the "best practice" of flow cytometry as described by the applicant´s former supervisor (i.e. two-fold; Doležel et al. Nat. Protocols 2007). This will be perhaps fine if this will be one of numerous other samples but if this is for the reference standard this step should be done very accurately as it influence all estimates. How authors checked for signal linearity during the measurements to eliminate this potential error?
- Supplementary tables (1, 2, 3) reporting the results of individual measurements are completely missing. I also lack data to the success of the crossing experiments.
- In the same line I lack the data and results of analyses handling proportions of viable fruits and fruits dimensions, which are discussed in the discussion (page 40). Seems that some analyses were removed from results but not simultaneously from the discussion.
- Samples were measured by PI dye but the representative double peak were made with DAPI dye. How much differences from individual measurements and those from the double peaks (peak ratios) matched? Best show a 2D graph with regression of "PI difference=DAPI double peak difference" (x=y) relationship.
- According to Fig S1, leaves start to dramatically reduce their length after ±150 days of cultivation. This seem unrealistic and pointing to some methodical problem. Has the same data been used for the statistical analyses? How this could affect the results and their interpretations?
- Authors counted chromosomes and observed size of their satellites, however, they provide no direct link with these observations because they were unable to measure genome size in counted seedlings. This is somehow surprising (and suspicious), given the flow cytometry allow measurements of a very

small amounts of material and that at least roots must have been available to authors for the purpose of such measurements.

- It would be interesting to know if there was some maternal effect on the offspring fitness and flowering and to exclude it eventually from the analyses.
- Regarding the offspring fitness it may be useful to test not only dependence on the absolute DNA content but also the dependence on the difference from parental mean (expecting lower fitness of largely deviating plants irrespective of their absolute genome size).
- The information about existence of 1.262-fold difference in all 775 measured seedlings does not appear neither in the results, nor the abstract. Could this information be documented by a double peak?
- In the results you indicate “moderate” correlation between fruit weight and genome size. However, in the discussion (page 42) you indicate this correlation was “strong”. This is improper.
- Identification of differences in percentages is sometimes confusing – best use expression using X-fold through the whole paper.
- In Fig. 1 I lack details to peaks, namely the peak ratio. Samples show should be identified to make possible to find their genome sizes in Table S1.
- In Fig. 4. write legend in normal font and explain according what criterion samples were sorted in the figure. Note that parents and offspring cannot be seriously distinguished in black and white figure.
- In legend to Fig. 5 write “2C of pollen donor plant”.

Petr Šmarda Ph.D.
Department of Botany and Zoology
Masaryk University, Brno
5 September 2018