The manuscript by Zhao et al entitled "genome-wide identification and expression analysis of diacylglycerol acyltransferase genes in soybean (*Glycine max)*, reports, for the first time, the identification, in silico studies about genomic and proteomic features of DGAT enzymes in a very important agricultural plan soybean, although similar studies have been done in other plants. Some wet experiments about organ- and age/stage-specific expression of *DGAT* genes have been performed.

The manuscript is not novel/ transformative but definitively expansive since done in soybean for the first time as per the authors. Overall the manuscript is concise and easy to follow but there are a lot of places with typos / grammatical mistakes (see my minor comments below). There are many major lapses in the manuscript but I believe could be improved (see my major comments below). Despite my rigorous efforts to correct grammar and language throughout the text to enhance the readability of the manuscript, there might be still some undetected errors. Hence, I strongly encourage authors to check the grammar and language again. Sufficient statistical analyses have been done. I do not have any serious dissatisfaction with the scientific content and experimental designs. I am positive about the manuscript given that all the major and minor comments that I have made below are satisfactorily addressed and edited accordingly.

Major comments:

1. This type of study is already done in many plants. Is this study reporting for the first time genome-wide identification and expression analysis of DGATs in soybean? If yes, then it could be a point for why undertaking this work and it should be included in the abstract (very shortly) and introduction section (in detail in the last paragraph). I can see that it is included in the discussion section.

2. The resolution/ readability of fig 1(a) is very low. It should at least be like Fig 1b. Fig 1 a i, Fig 1a ii, and Fig 1a iii should be referenced in the result section. In the scales of fig 1a ii and fig 1a iii, mention whether it is an amino acid (aa) or nucleotide (nt). Enlarge Fig 1ai and if applicable, show the scale or phylogenetic distance marker. To remove confusion, I highly encourage the authors to group DGAT protein-related data (aa phylogenetic tree, predicted motifs, etc) and genes (intron and exon) as separate subfigures. In the current fig 1b (conserved protein domain), demarcate the conserved/ signature DGAT (pf03982) and WS/DGAT (pf03007 and pf06974) protein domains also, in the same figure if applicable, if not then as a separate subfigure. Make necessary additions/ edits in the result section.

3. Regarding fig. 3, in the current form the heat map is telling about the frequency of various promoter elements in the -2kb region of DGAT genes, not the position. Hence rephrase “position” as “frequency”. Since the relative position of elements with respect to transcriptional/ translational start site is also very

important for gene regulation, I highly encourage the authors to include an additional subfigure as fig 3b showing the location/ position of elements in the -2kb region. Make necessary additions/ edits in the result section.

4. Regarding figure 3 (promoter analysis), the result section mentions 5 divisions but figure 3 is showing only 3 divisions. Surprisingly, even the names of divisions are different in the result section and figure 3.

5. Rephrase the title of table 1 as ”Characteristics of soybean diacylglycerol acyltransferase genes and proteins” in all places as the table contains information about genes as well as proteins. I highly encourage authors to include an additional column about localization prediction. Also, cite the method or software for localization prediction in the methods section.

5. Regarding the raw data excel sheet about qPCR of *DGAT* genes in various organs (PeerJ-78736-pcr), I am surprised to see that *WS/DGATb2* raw Ct values are lower than the control gene. It is not common. It means *WS/DGATb2* gene has higher expression than *ACTIN* in flowers. I suspect some technical mistakes during the experiment or data processing. I highly recommend the authors reconfirm/ recheck the calculations.

6. Line 213, mentions a reference to figure 2b (about collinearity analysis), but there is no figure 2b. Figure 2 in its current form has only one figure.

7. Line 228, mentions 37 paralogous gene pairs but in table 2, but I can see only 35 gene pairs. I strongly recommend authors cross-check the consistency between the actual figures and explanations in the result section for all figures thoroughly.

8. Regarding Table 2 and the result section “selective pressure on DGAT genes”, I could not find any raw data about the calculations. I do see the nucleotide sequence word file. I highly recommend including an excel file showing raw calculations of how you get Ka, Ks values, and other analysis, for example, a chi-squared test to distinguish between a null model (Ka/Ks=1) and the observed results.

9. Line 259-290, Regarding results “*DGAT* gene-expression patterns”. Rephrase the title as “Spatial *DGAT* genes-expression patterns” or similar. Line 272, regarding the sentence “Given its function, DGAT2a4 may play a key role in this subfamily”. Explain why? Is it only due to its higher level of expression in late seed development? If so that is not convincing, although lower, there is an appreciable level of expression of other members also. If it cannot be justified with other reasons, the sentence could be deleted. The authors should not use the term “main gene” in the family depending on the higher or lower expression level. “important/ dominant” or a similar word would be proper. In line 276, “root nodules and roots in the mature stages of seed formation, respectively” is extremely confusing since figure 4 does not contain these terms.

Line 288 mentions seed development stages S7, S8, and S9 but in figure 4, I cannot see S. I can see seed.R1 to R8. There is not even R9. It is extremely confusing to follow /understand. Make all necessary corrections making sure that there is consistency between text and figures.

10. Regarding figure 5 and the corresponding result section " Expression of *DGAT* genes", the title seems to be similar to the preceding title. You could differentiate as public vs experimental or other suitable ways. Start the paragraph mentioning why you decide to qPCR by yourself since public data is already available (fig 4). Why you chose particularly 9 genes? If randomly or without any bias, then mention that. Or choosing the representative among the 3 families, etc could be one of the possible reasons. Describe any matching or unmatching between figs 4 and 5. Lines 302 to 305 are not fitting and hence should be moved to suitable places in the discussion section.

Minor Comments:

1. Line 21, delete "in"

2. Line 63, delete "it"

3. Line 264, rephrase "stem" as "stems"

4. Line 279, rephrase "tissues was" as "tissues were"

5. Line 299, rephrase " gene expressed" as "gene is expressed"

6. Lines 40 and 41, rephrase the last sentence of the first paragraph of the introduction section as "Specifically, DGAT catalyzes the formation of an ester linkage between a fatty acyl CoA and the free hydroxyl group of diacylglycerol to form TAGs"

7. Line 46, rephrase "Further" as "further"

8. Line 49, delete "were" and "and".

9. Lines 110 to 113, regarding the section "Estimating Ka/Ks ratios for duplicated gene pair, if applicable mention the cutoff ratio to determine whether it is undergoing purifying selection or not.

10. 1Line 135, Do not italicize all results subheadings as it creates confusion about whether genes (italicized) are being referred to or proteins (not italicized). Rephrase the title as " Identification of DGAT proteins in soybean (*G. max*).

11. 1Line s 136 to 138, there is confusion about whether it is about genes or proteins, and hence for clarity, rephrase the sentence as "Thirteen and twelve genes containing the signature DGAT (pf03982) and WS/DGAT (pf03007 and pf06974) protein domains, respectively, were identified…"

12. 1Line 150. After "(Table 1)", add a sentence " The amino acid (aa) length of DGATs varied from XXX (XXX) to XX Xaa (XXX), the isoelectric point (pI) ranged from X.XX (XXX) to XX.XX (XXX).

3. Lines 150-152, the sentence " There were clear basic properties that were similar among members within one group but distinct among groups" is not true as per Table 1. There is actually an overlapping range of properties like pI, aa length, CDS length, MW, etc in every group. Hence the sentence should be rephrased as "There were similarities in basic gene and protein properties within as well as among groups." or other sentences with a similar sense.

14. 1Lines, 153 to 167, these sentences/ paragraphs are not suitable in the results section. Move them to a fitting place in the discussion section.

15. Line 178, rephrase as "DGAT genes" as "DGAT proteins"

16. Line 189-190, rephrase the sentence as "In contrast, cytoDGAT do not contain any of the conserved domains but contains the unique 2Fe-2S_Thioredx Damian".

17. Line 209, rephrase as "*DGAT*". Make sure that genes are capitalized and italicized while proteins are capitalized and not italicized throughout the whole text.

18. Line 213, provide reference/s that tells collinearity analysis is done to understand the evolution and amplification of genes in a family.

19. 2Line 223, add one or two sentences telling what is means or conclusion you get or the implication from these collinearity results.

22. Line 226, for clarity, mention what Ka and Ks mean.

21. 2Line 237, rephrase "expands" as "expanded"

22. Line 254, add a sentence like "Interestingly, although *cytoDGAT* does not share protein domains, it does share promoter features with other *DGAT* genes.

23. 2Line 261, delete "Thus,"

24. 2Line 264, start a new sentence as "*WS/DGAT* subfamily members are expressed….."

25. Regarding lines 84-85 and 136-137, for consistency, either use lowercase or uppercase, whichever is the correct form, in both places and throughout the manuscript.

26. Regarding fig. 5 legend, add information about the reference gene as well as mention whether the expression level is relative to the reference gene or specific organ. An example could be " The transcript

levels were normalized and expressed relative to the reference XXX gene." or other sentences with a similar sense.

27. In figures 3 and 4 legends, explain the scales whether they are simple fold, absolute value or log2 fold, etc.

28. In the figure 4 legend, mention the source of data, public or form your wet experiments.

29. Regarding raw data (Peerj-78736 word file), at the start mention the source as well as the meaning of differential colorings.

30. In the table 2 legend, include full forms of all abbreviations like in Table 1.

31. While referring to the finds from this study in the discussion section, provide the references like (figure X)

32. Line 331, rephrase as "which may be due to whole genome duplication during evolution". Also, provide reference/s supporting it.

33. Line 338, provide reference/s that shows a relation between light-responsive elements and seed oil content.

34. Line 338, rephrase "presented" as "contained"

35. Regarding raw data (excel sheet), it is hard to follow. Mention each data belongs to which figure and provide the relevant units