Reviewer Report: Article genes-510672

Summary:
In the article entitled “Exploring molecular signs of sex in the marine diatom *Skeletonema marinoi*”, the authors compared the transcriptomes of sexually induced strains of the diatom with uninduced strains. The authors aimed to investigate the genes involved in sexual development in this species with an effort to potentially identify genes that could be used as markers to indicated sexual reproduction in metagenomic and metatranscriptomic datasets. The results include the identification of a number of genes related to sexual processes, such as meiosis as well as those involved in flagellar formation. Additionally, the authors define a set of unannotated genes as “sexually induced genes” due to their differential regulation during sexual development. They were further able to observe differential expression of a subset of these genes. They thus conclude that this study has expanded the set of genes that can be used as a marker for sexual reproduction, particularly for diatom species.

I believe this manuscript fits well within the scope of Genes, particularly with regards to the study of gene expression, genomics and transcriptomics. The paper also has the potential to make a significant contribution to this specific field, particularly if the research is presented in a more understandable and coherent manner. My overall impression of this manuscript is that although the science is mostly sound and that the analyses conducted are appropriate for transcriptomic data, a number of improvements could be made to make it more suitable for publication purposes.

Broad Comments:
1. **Major concern 1- The transcriptomic comparisons:**
   - The authors present the transcriptomes of three different cell types, 1) Small cells with salinity stress (sexualized cells), 2) Large cells with salinity stress (non-sexualized cells) and 3) Large cells without salinity stress (control). It would also have been appropriate to include a fourth cell type- small cells without salinity stress, ie: small cells that are not sexually competent. The manuscript’s comparison between the sexualized and non-sexualized cells allows for genes involved in cell size to be detected as differentially expressed between the two cell types, which then may be incorrectly considered as sexually-induced genes.
   - It is fairly difficult to keep track of the various comparisons made between the numerous datasets. For example, in text (line 140), it states that “four pairwise comparisons” were conducted, however, in table 1, there are five listed comparisons. Perhaps Figure 1 could be improved and added to by including the various sample types that were collected at T1 and T2, and what comparisons were done. Alternatively, an entirely new figure could be generated.
• The abbreviations given to each comparison are not intuitive and perhaps are not entirely necessary. Furthermore, using the abbreviations “SST1” and “SST2” to refer to the salinity stress timepoints 1 and 2 is confusing given that “SST” is used as the abbreviation for “sexualization size threshold” earlier in the manuscript (lines 45). Perhaps use something like “SS_T1”.

2. **Major concern 2 - Discussion involving the flagella genes:**
   - *This concern is also linked to the comment 9 below regarding the time points selected.* It is originally made clear (line 189) that swimming, assumingly flagellate, gametes were only observed more than 62 hours after transferal to the salinity stress media. However, later in the manuscript (line 260), the authors state that the differential expression of flagella genes is consistent with the finding of flagellate gametes at T2. T2 is significantly earlier than the 62 hours after which flagellate gametes are reported. This statement is made again in the discussion (line 326). This needs urgent clarification.
   - In both the abstract (lines 21 – 22) and the discussion (lines 313 - 317), the fact that genes involved in flagellar synthesis were not found to be expressed in *S. marinoi* despite having been found in *L. danicus* was specifically referenced. Why is this particular result focused on so much? Especially considering genes involved in flagellar synthesis in *E. huxlei* were found to be differentially expressed in *S. marinoi*.

3. **Major concern 3 - Section 2.6:**
   - This section is very difficult to follow and not written in a logical manner. I think the entire section needs to be rewritten in an effort to better this manuscript. It appears that slightly different steps were taken for genes from *L. danicus* vs genes from *P. multistriata*, this may well be appropriate, but it could be better explained. The last paragraph of this section also needs some clarity as the steps of the method are difficult to follow.

4. There are too many tables and not enough figures in this manuscript. While transcriptomic data often calls for the use of tables and is appropriate in certain cases, some of the data presented here could be better visualized in figure form. One such example is Table 1. Perhaps including a figure that compares sexualized and non-sexualized cells (could be indicated by small and large cells, respectively) at the two time points with the number of up- and down-regulated genes indicated on the figure. Additionally, the GO term enrichment analyses performed provide perfect data for figure-based visualization.

5. The introduction is not comprehensive enough if the authors wish to appeal to the broader audience of this journal. A more in depth description of sexual reproduction in diatoms would be helpful, perhaps including a figure to show the various stages of sex in these species. For example, the unique manner of the progressive size reduction and “the rigid fustule” the authors refer to are not as clear as they could be. Additionally, a short description of the difference between pennate and central
diatoms would give good context for some of the comparisons to other species which are made later in the manuscript.

6. The paragraphs beginning on lines 48 and 57 could be combined and rewritten to form a more logical “story”. Alternatively, two independent paragraphs can be written with this information as the two paragraphs have a mixture of information regarding chemical sensing and genes involved in sexual reproduction.

7. The final paragraph of the introduction, where the aim of the study is discussed, it packed with unnecessary detail that takes away from the main message. This paragraphs should be written more concisely. The extra information could either be moved to somewhere in the introduction, methods and material or discussion, where appropriate.

8. In two cases, the databases used were accessed a number of years ago. In line 133, for example, the UniRef90 database was accessed mid-2016 while in line 166 the NCBI NR database was accessed early in 2017. It would be appropriate to relook at these databases and determine whether new data has been published in these databases that may be of use.

9. Under section 2.1 Experimental set up, the authors state they do a pilot test to determine the correct time points for the transcriptomic study, by looking at timepoints 20, 30, 40 and 62 hours post transfer (line 92). Shortly thereafter (line 100), the authors state that Timepoint 2 will be at 52 hours- why? There is no apparent justification for the choice of this time point.

10. In lines 205 – 208, the authors state that they expected to observe the induction of certain genes that were detected in a similar study in P. multistriata. Why are the other comparative species not mentioned here?

11. Late in the manuscript, the authors refer to the genome sequence of this species. If such data exists, why was a de novo transcriptome assembled? Why did the authors not simply align the RNA reads to the genome in order to determine differential expression etc?

12. The results section from lines 224 – 233 is insufficient. The authors refer to “significantly enriched GO terms” but do not state what these are. A figure would be more descriptive and informative. Furthermore, GO terms can be broken down into various categories- cellular component, molecular function and biological process. Indicating which of these categories are enriched would also provide better information and more understandable results.

13. In table 2, the query IDs are from Arabidopsis and Saccharomyces- why is this the case? Are there not more appropriate, closer relatives to use?

14. In lines 248 – 253, the absence of the GEX1 and HAP2 genes are indicated. This is briefly brought up again in the discussion, but I think it needs to be discussed more, given the importance of these genes to other eukaryotes and the suggestion that these genes be used as the markers for eukaryotic sex.
15. The paragraphs starting on lines 365 and 373 seem somewhat irrelevant to the overall discussion of this manuscript. I’d suggest either integrating this information with the data collected here or excluding it entirely.

16. An interesting finding in this manuscript is the general down-regulation seen in genes associated with the chloroplast during sexual reproduction. The paragraph beginning on line 412 very briefly discusses this. I suggest going into more detail discussing this result and providing some kind of plausible explanation.

Specific Comments:
1. **Consistency:**
   - In section 2.2 RNA extraction and RNA-seq, the single-end and pair-end libraries are abbreviated to SE and PE (lines 107 and 111, respectively). This abbreviation, however, is not used consistently throughout the rest of the document. This should be fixed (eg: lines 120, 123, and 129).
   - When referring to E-values, both capital and lowercase E’s are used to denote the scientific numbering format (eg: Line 158: E-value of 1E-02 vs Line 240: (E-value < 1e-5). Please ensure this is consistent throughout the document.
   - It is general notation to write gene names in italics (eg: Msh4) and protein names in capital letters and standard font (eg: MSH4). This should be standardized throughout the manuscript.
   - Both “up-regulated, down-regulated” and “upregulated, downregulated” are used. Please standardize throughout the manuscript.
   - Numbers in the excess of 1000 are referred to in the manuscript either as “1000” or “1,000”. Please standardize throughout the manuscript.

2. **Line 36:** “a box and its lid” is an unnecessary comparison.

3. **Lines 40 – 42:** The definition of homothallic mating is not particularly descriptive of the system, especially in comparison to heterothallic mating. This definition should include a clear indication that homothallic mating occurs in the absence of mating partners. Alternatively, given that the type of mating occurring is irrelevant to the overall aims and results of this paper- these definitions could be removed.

4. **Line 86 and other:** The abbreviation for sexualization size threshold (SST) is indicated in line 45. In other places where this comes up, please use the abbreviation instead of the entire term.

5. **Lines 107 – 110 and 111 – 112:** The detail provided for the preparation of the SE libraries was perfect. However, the same level of detail is not provided for the preparation of the PE library. This should be supplemented. Furthermore, it is clear why the SE libraries were prepared, but why were the PE libraries made as well? Perhaps a sentence starting “In order to **, we also sequenced a PE library...”.

6. **Line 112:** Why was a different S. marinoi strain used for the PE library? The origin of this isolate should also be indicated in the experimental set up.

7. **Line 114:** What is a “summary control”?
8. **Line 114 and others**: There are many instances in the manuscript where particular software packages are used but not referenced. Please add references to these where appropriate. (eg: FastQC software is used and should be referenced as: “Andrews, Simon. FastQC: a quality control tool for high throughput sequence data. (2010)”)

9. **Line 127 – 130**: A single sentence spans these lines and is confusing to understand. Please break it up into two or more independent sentences for better understanding.

10. **Line 130**: Please provide an explanation for “positive contig impact score” and explain in more detail why these contigs were also added to the final analyses.

11. **Line 128**: Instead of providing the URL for the genome sequence used, instead provide the two references that relevant for this genome (Bowler et al. 2008 & Armbrust et al. 2004).

12. **Line 136**: The authors refer to the biological replicate. It is not made clear beforehand what represses a biological control, even in Figure S1.

13. **Line 139**: “... at least above 1 CPM...”. What is CPM in this context and how is it calculated?

14. **Line 140**: “... was used to identify differentially expressed genes for each...”- This is the first reference to differentially expressed genes and the DEG abbreviation should be used here so that it can be used elsewhere (eg: line 229).

15. **Lines 141 and 142**: “...T1 and T2: i) large cells...”- What does the i) refer to? There is no follow up ii) so it seems unnecessary.

16. **Lines 143, 144 and others**: The logFC changes are referred to as “-1<logFC<1” throughout the manuscript. This is not correct. It should be indicated as “-1>logFC>1” as significant differences are usually considered if the fold change is BELOW -1 or ABOVE 1.

17. **Lines 154 and 155**: “The list of meiotic-related genes was taken from [7] and the one for flagella [9] and [10].”- Please indicate from which species these genes were taken, it will provide better context that simply providing the references.

18. **Line 165**: “...*S. marinoi* were blasted (BLASTx)...”- One cannot use “BLAST” as a verb as in this sentence, as in “blasted”. Please rephrase so that this is not the case.

19. **Line 176**: The authors simply state “*Bicoecoid*” despite the fact the sentence is referring to a particular species. Which species is specifically being used?

20. **Line 188**: The authors refer to the gametangia as “elongated cells with a bent shape”. Instead of this description, could the authors state “gametangia, which represent an early sexual tissue” or something similar?

21. **Line 189**: The authors refer to the auxospores as “round structures still attached to the gametangial cells”. Instead of this description, could the authors state “auxospores, which are the sexual cells that result after cell fusion” or something similar?

22. **Line 202 onwards**: The first paragraph of this section of results has information better suited to the methods and materials section. It should be rewritten with only the results.
23. **Lines 207, 263, 269, 280 and others**: There are many instances in the manuscript where species names have not been italicized.

24. **Line 209**: “We sequenced 14 samples...” - Although these 14 samples are explained in Figure S1, it should be made clearer in the main text. Furthermore, why are there only 2 replicates for the Large cells (both salt and non-salt), but 3 replicates for the Small cells? Transcriptomics projects that are not followed up with a confirmation experiment (like RT-qPCR) usually have a minimum of 3 repeats per condition.

25. **Line 237 and 238**: The authors refer specifically to genes BRCA1, SMC2 and SMC4. What are these genes, what kind of proteins do they encode and what role do they have in sexual reproduction?

26. **Lines 238 and 435**: The authors refer to genes being “overexpressed”. Do they mean “up-regulated”? These two terms mean slightly different things.

27. **Table 2**: Instead of using (1) and (2) to link to the footnotes, using a superscript a and b would be more readable and less confusing.

28. **Table 4**: Why are some of the LogFC changes in bold text? Why do some of the contigs not have annotations (if they are unknown proteins, indicate as such)? What does -::- mean under the *P. multistriata* annotation mean? Indicating the genes that show the same direction of regulation by “^” is not very clear and these genes are difficult to identify. Perhaps indicating their LogFC in bold or italics, or with a superscript would be clearer.

29. **Lines 261 – 269**: The authors discuss the retrieval of genes related to flagella synthesis and state that eight of the *L. danicus* genes are found in the *S. marinoi* transcriptome and a further two in the genome, bringing the total number of homologous genes to 10. This suggests that the first eight were not found in the genome- is this the case? If so, how?

30. **Line 272**: The comparison of sex-related genes from *T. weissflogii* is brought up for the first time here- perhaps this should be discussed earlier when the other comparative species are brought to attention.

31. **Line 284**: The identification of the SIG1, 2 and 3 genes is never brought up in the methods and materials section. How were these genes discovered in *S. marinoi* and what is their relevance?

32. **Lines 290 – 299**: These lines represent a single sentence which is difficult to follow and understand. Please rewrite into concise, independent sentences.

33. **Line 300**: “…can be found as highly regulated...” - What does highly regulated mean? Up-regulated? Significantly differentially expressed?

34. **Line 310**: the authors state that their analysis was able to identify genes associated with the switch to meiosis. This is not strictly true. Although meiosis is an important factor in sexual reproduction, many other pathways and processes are taking place during this time. A more appropriate remark would be “… associated with the switch from the vegetative lifestyle to sexual development”.
35. **Lines 327 – 328:** Delete “ie genes for which functions outside meiosis are not known” — this is implied by the description “specific for meiosis”.

36. **Lines 354 – 356:** The authors need to provide a reference for this sentence.

37. **Line 365:** The authors have been consistent in describing the type of species, such as the “pennate *P. multistriata*”. This is helpful and should also be done for *T. pseudonana* and other species as it provides context for each comparison.

38. **Lines 385 and 396:** The authors refer to the set of flagella genes as an “asset”. Do they perhaps mean “set”?

39. **Line 399:** The abbreviation MMESTP comes up in this sentence. This abbreviation needs clarification.

40. **Line 407:** The term “mastigoneme” is used. This word would only be understood by a very particular audience and should either be excluded or, if relevant, explained.

41. **Grammatical/typographical errors:**
   - Line 15: “increase” should be “increased”
   - Line 45: “cohorts” should be “cohort”
   - Line 46: “deriving” should be “derived”
   - Line 50: The sentence should read “Chemical cues also play a role…”
   - Line 53: “to identify” should be “the identification of”
   - Line 58: “case of” should be “case for”
   - Line 92: “transfer at higher” should be “transfer to media of a higher”
   - Line 97: “adjusted at” should be “adjusted to”
   - Line 102: It should be mentioned that it is DNA that is being extracted
   - Line 161 & 177: “blast” should be “BLAST”
   - Line 178: “an aligned” should be “and aligned”
   - Line 187: “below SST transferred” should be “below the SST and transferred”
   - Line 225: “a few” should be “few”
   - Line 256 and 258: The species name *E. huxleyi* is spelt incorrectly.
   - Line 302: “supposedly” should be “putatively”
   - Line 354: “its absence is some” should be “its absence in some”
   - Line 302: “Pennate diatoms...” should be “While pennate diatoms...” (to indicate comparison to centric diatoms)
   - Line 410: “will allow to reveal” should be “will reveal”
   - Line 411: “worth” should be “worthy”
   - Line 440: There is an extra “B” before the supplementary materials are listed.