Responses to reviewer comments

Q: It was unclear to me what ‘scaffolding’ meant to the analysis. Why is it so important to note this?

A: Scaffolding is a process that connects contigs (segments of continuous A, C, T, and G nucleotides such as ACTGATCTAG) together but there may be some unknown base pairs. A scaffold is a segment of sequence that contains A's, T's, C's, G's, and N's where the N's denote unknown base pairs (such as ACTATCATCTACTATCNNNNNNNNNCAGGTATCGACT). The length of the N segment is determined in the scaffolding process and can be anything from a few base pairs to 10's of thousands. Any time you carry out an assembly process you perform scaffolding (because you may have unknown base pairs between two contigs where all the sequence is known. A good example of this is with the linked-read scaffolding where you know all sequences come from within 50kbp of one another but you may not know how far apart they are. The assembler will make an educated guess based on algorithms in the software.

Q: What does the significant about the forms of chromosome pairs? Why is it important to note that there were subtelocentric/acrocentric vs metacentric chromosome pairs?

A: There is no significance per se to the composition of chromosome pairs. The composition is only a descriptive metric that is commonly reported in karyotyping analysis. The chromosome composition did not affect our genome assembly. I have updated the first paragraph of the genome assembly discussion to reflect this: "Similar to other reports, we note a preponderance of subtelocentric/acrocentric chromosome pairs over metacentric chromosome pairs, as one might expect for closely related species. Chromosome composition is a descriptive metric, and our findings did not alter or affect our genome assembly process."

Q: I noticed there are figures referenced but did not see them in the document. Did I miss another document from you that would include those figures?

A: Thanks for finding this! There was some kind of error in converting the word document to a pdf and it appears none of the figures were included in the final report! No wonder some things were confusing. We are sorry about this and they are included in the new one! As a note there was a typo in the domestication selection section where a "Figure 1" was erroneously mentioned. We have corrected this error and it now correctly lists "Figure 11"

Q: The report notes that data from the Vertebrate Genome Project has not been returned. When will the data from the Vertebrate genome be returned? What will be the plan when you do get it? It’s mentioned that you have the genome so does that mean that not getting the data from the Vertebrate Genome Project will not impair the deliverable for this contract?

A: Despite consistent inquiries over the past 6 months, we have not received any useful correspondence regarding when we should expect the data from the Vertebrate Genome Project. Fortunately the female hi-c data that we received from Phase Genomics before the Covid shutdowns allowed us to perform scaffolding on both the male and female assemblies. The female hi-c data will scaffold effectively for all autosomal chromosome. Since, karyotype analysis did not show signs of a sex chromosome, we don't expect the male hi-c data to significantly alter the effectiveness of Step 3 in the assembly process. When we get the data, we will re-scaffold the male assembly and resubmit an updated male reference genome to the NCBI genome archive. Revisions to assemblies are quite common. Since the initial proposal only listed one genome and we have provided two chromosome scale genomes we believe this task has effectively been completed. If you would like, we can notify you when the revised genome is uploaded to the NCBI server.

Q: Could we get the Ne estimates with CI’s in a table as well as the figure for the report?

A: Yes. Due to the number of data points (630) we have included this as a Supplementary Table in a separate document (see Supplemental\_Table\_1-temporal\_Ne\_values\_all\_by\_all.csv)

Q: As part of the deliverables can you provide the data and the images that were taken as well as the R script?

A: Yes. Data: The total raw data is approximately 4.4Tb in size, do you have a method that you like to use for transferring large amounts of data? Alternatively, when we publish each of these papers we will be uploading the raw sequencing files to the SRA and the genome assemblies to NCBI.

Scripts: I am working on cleaning up the directory structure to open source all of the in-house perl, python, R, shell and slurm scripts we used in this analysis. They will be located on GitHub. Images: could you clarify what you mean by images? If they are the images for the karyotype, those in the word document are the highest resolution that we have been sent by the collaborators.

Q: It is noted that Ne went up yet it was considered an artifact of the analysis. How confident are you on concluding that Ne going down is not an ‘artifact’ of the analysis?

A: Ne did not increase, but before correcting for the technical changes in the lab protocol genetic diversity (theta) increased from 2018-2020 which was concordant with the labwork change. However, we have reanalyzed this data by trimming the last 50 base pairs from the longer sequences generated in 2018-2020 and have amended the report accordingly. As expected, genetic diversity continued to decline from 2018 to 2020. (See updated Figure 10). So as of today we are 100% confident that the previously observed uptick was a technical artifact.

Q: Why was the individuals needed for the Ne estimates set to 35?

A: Years with fewer individuals would not have a sufficient number of gene copies to compare to years with greater than 50 individuals. Because there was one year (2010) with only 35 individuals we set the low end of the cutoff to 35. Below is Table 7 showing variance in individuals available for Ne estimation:

Table 7. Distribution of the total samples acquired for effective population size estimations.

|  |  |
| --- | --- |
| **Year** | **Number of Individuals** |
| 1993 | 9 |
| 1995 | 74 |
| 1996 | 111 |
| 1997 | 65 |
| 1998 | 126 |
| 1999 | 45 |
| 2000 | 5 |
| 2002 | 192 |
| 2004 | 192 |
| 2006 | 189 |
| 2007 | 1 |
| 2008 | 191 |
| 2009 | 183 |
| 2010 | 35 |
| 2011 | 192 |
| 2012 | 191 |
| 2013 | 189 |
| 2014 | 217 |
| 2015 | 108 |
| 2016 | 83 |
| 2017 | 167 |
| 2018 | 167 |
| 2019 | 147 |
| 2020 | 66 |

There were 3 years with <10 individuals and 21 years with >=35 individuals. Less than 10 is insufficient number, and for technical reasons, we wanted to use an equal number of individuals for the estimates. Thus, we did not make estimates for the low years, and randomly sampled 35 individuals from the other years.

Q: I was a bit lost on the sex gene determination. It is likely my lack of understanding but it is noted there are indicators of a sex gene on Chromosome 9 and that unique k-mers for males was also noted. Why are those not sufficient indicators for the sex gene? Do these factors merely narrow whether the sex gene exists and where the sex gene is located?

A: The indicator of a sex determining region came from a comparison of two individuals (one male and one female individual that we sequenced for our genome assembly work). However these candidate sequences were not supported by our RAD sequencing data (lower resolution sequencing method) where we had an additional 24 females and 24 males. So, although there are discrete differences between the sequences of the two (single male and single female) individuals, the extent to which those difference contribute to sex determination will remain ambiguous until whole genome sequencing is performed on a large number of male and female individuals.

Q: It is noted that Ne thresholds have various support in the literature. But it seems that you feel that these thresholds are premature and should not be used to inform management. Is that correct? Can you provide some examples of how these thresholds would not be useful to inform management? Also the values would not be used as the only line of evidence but as one of many when considering management options. Does that affect your conclusions about not supporting its use?

Answering this question is challenging without knowing what types of management actions would be taken based on the threshold, and which exact threshold would be used – the old 50/500 rule or the more recent 100/1000 rule. Furthermore, it’s important to note that our Ne estimates are exactly that, estimates. Although the declining trajectory is very clear and robust, there is potential for all Ne estimates to be systematically biased in one direction or the other due to technical issues as well as for Ne estimates from specific years to be anomalous due to technical issues. For example, due to potential technical issues, all our estimates could be systematically high (or low), and some specific years may be particularly inaccurate. Based on these and other considerations, it’s would be fairly arbitrary to take dramatic management actions based on the Ne estimate from a single year moving above or below some predetermined threshold.

Q: Why is it important to note for management that there are two domestication selection regions?

A: We added this text to the domestication selection section of the report:

“The management benefit of identifying regions under domestication selection is that it provides the foundation for future research to elucidate the genes and biological pathways involved in the process of adaptation to captivity. This knowledge can provide insight into, and thus the potential to modify environmental condition that are causing the domestication. For example, if we find that our candidate regions are involved in salinity regulation, the FCCL can modify the salinity condition in the hatchery. Or, if the genes are involved in immune response, modifying fish density may be warranted.”

Q: Why is the trend of Ne more important the discrete estimates of Ne?

As mentioned above, there is potential for all Ne estimates to be systematically biased in one direction or the other due to technical issues as well as for Ne estimates from specific years to be anomalous due to technical issues. However, given that consistent methods were used to estimate Ne across time, it would be virtually impossible for technical artifacts to produce the clear declining trend that we observed across the 25 year sample set. Thus, we consider the declining trench much more robust and useful than the point estimate from any single year.