PCI Genomics Review

Dear Chiara Bortoluzzi,
Your article, entitled **Trends in genome diversity of small populations under a conservation program: a case study of two French chicken breeds**, has now been reviewed.
The referees' comments and the recommender's decision are shown below. As you can see, the recommender found your article very interesting but suggested certain revisions.
We shall, in principle, be happy to recommend your article as soon as it has been revised in response to the points raised by the referees.
When revising your article, we remind you that your article must contain the following sections (see our Guide for Authors in the Help section of the PCI Genomics website):

1) **Data, script and code availability (if applicable)**
   - **Data, statistical scripts, command lines and simulation code must be made available to readers**. They should either be included in the article or deposited **in** an open repository such as Zenodo **with a DOI**. A perennial URL can be provided if no DOI is available; please note that GitHub URL are not perennial.
   - **If deposited in** an open repository, a reference to **Data, statistical scripts, command lines and simulation code**, with a DOI or a perennial URL, must be provided in the reference list and in the "Data, script and code availability" section
   - The "Data, script and code availability" section must clearly indicate **where and how** data can be accessed.
   - Wherever possible, data, scripts and code should be provided in machine-readable formats. Avoid PDFs other than for textual supplementary information.
   - Metadata should accompany the data, to make the data understandable and reusable by the reader.

2) **Supplementary information (if applicable)**
   - Supplementary information (text, tables, figures, videos, etc.) can be referred to in the article. It must be available in an open repository (such as Zenodo, Dryad, OSF, Figshare, Morphobank, Morphosource, Github, MorphoMuseuM, Phenome10k, etc. or any institutional repository, etc...) with a DOI. A perennial URL can be provided if no DOI is available.
   - A reference to the supplementary information, with a DOI or a perennial URL, must be provided in the reference list and in the "Supplementary information" section.
   - List all documents attached to the manuscript as Supplementary Information in the "Supplementary Information" section.

3) **Funding (mandatory)**
   - All sources of funding must be listed in a separate "Funding section". The absence of funding must be clearly indicated in this section.

4) **Conflict of interest disclosure (mandatory)**

- Authors should declare any potential non-financial conflict of interest (financial conflicts of interest are forbidden, see the PCI code of conduct).
- In the absence of competing interests, the authors should add the following sentence to the “Conflict of interest disclosure” section: “The authors declare they have no conflict of interest relating to the content of this article.” If appropriate, this disclosure may be completed by a sentence indicating that some of the authors are PCI recommenders: “XXX is a recommender for PCI XX.”

5) **Materials and methods (mandatory)**

- Details of experimental procedures and quantitative analyses must be made **fully available** to readers, in the text, as appendices, or as Supplementary Information deposited in an open repository, such as Zenodo, Dryad or institutional repositories with a DOI.
- For specimen-based studies, **complete repository information** should be provided and institutional abbreviations should be listed in a dedicated subsection (if applicable). Specimens on which conclusions are based **must be deposited in an accessible and permanent repository**.

When your revised article is ready, please:

1) Upload the new version of your manuscript onto your favourite open archive and **wait until it appears online**;

2) Follow this link https://genomics.peercommunityin.org/user/my_articles or logging onto the PCI Genomics website and go to 'For Contributors -> Your submitted preprints' in the top menu and **click on the blue ‘VIEW/EDIT' button at the right end of the line** referring to the preprint in question.

3) Click on the black ‘EDIT YOUR ARTICLE DATA’ button (mandatory step). You can then edit the title, authors, DOI, abstract, keywords, disciplines, and DOI/URL of data, scripts and code. Do not forget to save your modifications by clicking on the green button.

4) Click on the blue ‘EDIT YOUR REPLY TO THE RECOMMENDER’ button (mandatory step). You could then write or paste your text, upload your reply as a PDF file, and upload a document with the modifications marked in TrackChange mode. If you are submitting the final formatted version ready to be recommended, you should only add a sentence indicating that you posted the final version on the preprint server. Do not forget to **save your modifications by clicking on the green button**.

5) Click on the green ‘SEND RESUBMISSION’ button. This will result in your submission being sent to the recommender.

Once the recommender has read the revised version, they may decide to recommend it directly, in which case the editorial correspondence (reviews, recommender’s decisions, authors’ replies) and a recommendation text will be published by PCI Genomics under the license CC-BY.

Alternatively, other rounds of reviews may be needed before the recommender reaches a favorable conclusion. They may also reject your article, in which case the reviews and decision will be sent to you, but they will not be published or publicly released by PCI Genomics. They will be safely stored in our database, to which only the Managing Board has access. You will be notified by e-mail at each stage in the procedure.

We thank you in advance for submitting your revised version.
Yours sincerely,
The Managing Board of PCI Genomics

Revision round #1
Decision for round #1 : *Revision needed*
Revision

Dear authors,
The manuscript **Trends in genome diversity of small populations under a conservation program: a case study of two French chicken breeds** has been examined by two expert scientists in population genetics. Although the two reviewers found merit in this study and recognized the quality of the associated paper, they raised a number of concerns that should be addressed before any decision could be rendered. I enclosed below detailed evaluation points. If you think you are able to provide a detailed answer to the different points, I encourage you to respond point by point and submit a new version of the preprint.
Thank you for submitting to PCI Genomics.
Best regards,
Claudia Kasper
*by **Claudia Kasper**, 26 Apr 2024 11:00*

We would like to thank the recommender, Claudia Kasper, for giving us the possibility to resubmit a new version of our preprint. We would also like to thank the two reviewers, Markus Neuditschko and Claudia Fontsere Alemany, for their comments, which we believe have enhanced the quality and relevance of our work. We would like to highlight in this rebuttal letter that, even though the reviewers had no remark on the chicken reference genome used in our submitted preprint, we took the initiative to re-run all our analyses to the latest chicken reference genome (GenBank assembly accession: GCA_016699485.1) generated by the Vertebrate Genomes Project to ensure that our research was conducted with the most up-to-date genomic resources. We also changed some steps of the mapping pipeline, to ensure that the latest softwares and programmes were used for reproducibility purposes. Although our methods have slightly changed in this new version of the preprint, the main conclusions have not changed.

Manuscript: **https://doi.org/10.1101/2024.02.22.581528**
version: 1

Review by Markus Neuditschko, 26 Apr 2024 10:48

Bortoluzzi et al., assessed various genetic diversity parameters of two local French Chicken breeds taking advantage of whole-genome sequencing information. The study is well-written and easy to follow. However, I have some major concerns about the small sample size, as they only analysed 15 samples for each time period (2003 and 2013). To better assess the selected sample size, authors should also provide more information about the breeds (e.g. the number of registered animals in the pedigree). Furthermore, the authors did not provide any information, how the 15 animals were selected. Based on pedigree information, it is possible to select most informative individuals by assessing their marginal gene contribution.
We clarified in the discussion that, although our sample size is a major limitation, such sample size is not uncommon in studies on local livestock breeds and endangered species. Regarding the question on the sampling, actually the initial number of founders was in the range of 15 and we also took 15 for the comparison 10 generations later. Indeed, the animals were sampled from different sire families as much as possible to make them the most informative. Given the size of the reproductive nucleus, there were about 10 sire families, and thus we sampled at least one individual per family and various dams.

Besides, that I have also identified some minor issues:
L27: replace still in existence with current livestock populations
Changed

L43: Livestock breeders instead of keepers
Changed

L45: routinely implemented
Changed

L97: robust and generally
Changed

Figure 1b: The PCA visualisation is not quite informative, as the points are coloured according to the time point and not to breed origin. To increase the visualisation of the PCA, I suggest using different shapes (time point) and colours (breeds). Also the variance explained by the first two components is rather low, hence it might also be informative to explore additional components.
This is indeed a great suggestion. Breeds are now coloured differently while the two time points are identified by different shapes. We provided an additional PCA in the supplementary material showing PC3 (see Figure S2), while still presenting PC1 vs PC2 as part of Figure 1, because these two components are the most informative ones based on our PCA analysis. In addition, having such a low inertia explained by the two first axes is not statistically surprising giving the initial number of variables (i.e. the number of SNP), and thus having more than 15% of inertia explained by only the 2 first axes is really high (relative to 1/nb SNP, the expected value for equally informative variable in scaled PCA).

L459: The authors mentioned that the breeds were previously genotyped on 57K SNP chip data and simultaneously highlight the added value of genome sequencing data. To do so, I would suggest downsizing sequencing data to 57K Data, to confirm the arguments posted in this section.
Unfortunately, we cannot downsize the sequencing data to the 57K SNP chip data and re-run our analyses on that small subset, because our analyses are tailored to whole-genome sequencing data and this proof was not the purpose of the present study. Nonetheless, we expanded on this aspect in our Discussion.

Review by Claudia Fontsere Alemany, 16 Apr 2024 12:11
Comments to authors

Bortoluzzi et al present a study on the effects of two different conservation programs on the genomes of two French chicken breeds.
Overall, I find the paper easy to follow and the analysis done appropriate for the research questions they had.
Title and abstract represent the content and the main findings of the study.
Introduction is well written and clearly presents the background of the study.

Regarding the results I have a few suggestions that, in my opinion, can enhance the quality of the paper (which I already enjoyed reading):

In all the boxplot comparisons between 2003 and 2013/15, there should be a statistical test to account for significance. Even if the authors describe a trend it would be good to add if it is statistically significant or not. I understand that the sample size might be a limitation, but this is something the authors can discuss in the paper. Also, when using boxplots, it is good practice to also include the individual dots, to get a perception of how they are distributed.
We changed all our boxplots following the reviewer's suggestions, including a statistical test in the figure and main test (see comment below) and individual dots to better show how data points are distributed. Regarding the limitation of the sample size, reviewer 1 raised a similar remark, so we have now included a short paragraph at the beginning of our discussion (Line 486 - 493), where we highlighted that such small sample size is not uncommon for studies on local livestock breeds and endangered species, putting our study into perspective.

Why did the authors decide to include a semen sample given it is a different sample type? Have the authors detected any difference in the number of genotypes detected in this sample compared to the others? Just to check that sample type is not adding more noise.

We decided to include a semen sample to investigate the usefulness of this sample type in population genomic studies, as the project was conducted under the umbrella of the IMAGE project, whose objective was to demonstrate gene banks' usefulness to retrospectively assess genetic changes in livestock breeds. Based on our analyses and previous experiences in sampling and re-sequencing from sperm and blood, we do not report any unexpected behaviour in the semen sample. For instance, the number of reads and mapped reads are within our range, as well as the average mapping quality, average depth, and % of missing sites (Table S1). This confirms our expectations, as we did not expect any effect for re-sequencing data.

Why is sample 7218 so different from the rest of 2013/15 Gasconne samples in the PCA? Do you see a different pattern in the other analysis as well? How many of the genotypes are private to this sample? I think this deserves an explanation both in the text (results/discussion) and in the Figure 1 caption.
We didn't observe any allele private to the Gasconne sample 7218. Although this sample had a lower genome-wide heterozygosity and higher number of ROH (resulting in a high Fped: 0.12500), these values were not so far off to make us consider this sample an outlier in all our downstream analyses. It is possible that this sample belongs to a close family or that it is just the result of a single related mating. However, we didn't have such information to make such a conclusion.

Regarding the ROH detection, how does your method account for having a window with heterozygous calls inside a ROH that would break it down? Do you account for genotyping errors breaking the (otherwise) long ROH? How do you deal with this? I would like to see a bit more explanations of the method.
Yes, our method accounts for alignment errors, which are often expressed in whole-genome sequencing data as peaks in heterozygous sites. As stated in the methodology, we splitted the candidate ROH into smaller chunks only if - by including the peak in heterozygosity - the overall heterozygosity within the ROH is larger than 0.20 the average genome-wide heterozygosity. However, if the peak in heterozygosity does not inflate the heterozygosity within the candidate ROH, the peak is retained to avoid breaking a likely long ROH into smaller chunks.

Also, in a more theoretical point of view: Does it make sense to calculate ROHs that are <= 100kb? In my previous experience I found them not to be much informative. Is there any reason why the authors decided to include them? On the other hand, I would consider adding another classification of > 10Mb to account for very long ROHs (if in fact they do exist).

We agree with the reviewer that considering ROH <= 100Kb is not very informative. We redefined the different ROH classes as follows: short (100 Kb - 1 Mb), medium (1 - 3 Mb) and long (>= 3 Mb). Within the class of long ROHs we followed the reviewer's suggestion by including an additional class of ROH >10 Mb. However, we didn't include these ROHs in Figure 3, because these ROHs are present only in a few individuals. We discussed them nonetheless in the main text (Line 397 - 399).

Regarding genome-wide heterozygosity, given that the authors already have the regions that are ROHs, I was wondering if they could also plot the heterozygosity outside ROH vs global heterozygosity, to compare both breeds with their historical genome-wide diversity. Of course, I am adding this suggestion only if it adds value to the analysis and discussion.
Great suggestion. We generated an additional boxplot on the heterozygosity outside ROH to compare this with the genome-wide heterozygosity (see Figure S5). The plot has been moved to the SM.

When discussing over the phenotypic data: could the authors hypothesize why even though there is a loss of heterozygosity and increase of inbreeding, it seems that the Barbezieux chickens are more productive? Would it be possible to link this to a reduction of genetic load? I would like to see a bit more discussion on this, even if it is just ideas/hypothesis.
We have now added a new small paragraph about this (Line 505 - 509).

I want to comment as well that I appreciate the authors adding the positive selection scan section even if they do not find any signature.

I also have just a few more minor comments:

Introduction

Line 27 the authors could add that these 7745 still existing breeds are world-wide (making this fact explicit in the text). Also, is it known how many breeds have gone extinct already in recent years? If this information is known, could make it even a stronger point for the significance of paper.
We now added a few more sentences in the Introduction regarding the percentage of breeds that went extinct in recent years (Line 83 - 87).

There is formatting error in reference (**J** Fernández, Meuwissen, et al, 2011), in lane 48 (and other places where the reference is added, or the first author is J Fernández) where there is the J of the name in front of the surname.
We corrected the reference error throughout the preprint.

Material and methods

Line 116. Consider changing Sibs to Siblings wherever this appears.

Changed throughout the preprint and supplementary material.

Line 128. Add (if true) that this is a double-stranded library preparation. Also add citation of the method used.
Changed to double-stranded library.

Lines 136-143. I am curious why the authors decide to do this very conservative strategy (which is ok) of doing SNP calling with different callers and using the overlap. Did you see any weird behaviour of any of them? Which one is better?
We decided to go for a very conservative strategy to increase our confidence in the set of variants called by an old - and now outdated version - of GATK v3.7.0. However, for this new version of the preprint we decided to switch to the latest chicken reference genome generated by the Vertebrate Genomes Project and to switch to the latest version of GATK v4.2.4.0. Considering that our ability to confidently call variants has improved from GATK3 to GATK4, we decided for this new version of the preprint to rely solely on variants called by GATK4, while following the same GATK best practices presented in the previous version of the preprint.

Lines 145 to 150. Here the authors mention SNPrelate to do the PCA, could the authors add the parameters used or a link to GitHub with the script? Also, for pruning there should be a mention of which software and parameters were used.
We added a few more sentences to the PCA analysis explaining the parameters used. Moreover, we are now providing all codes used throughout the paper, as mentioned in our Data availability statement (Line 208 - 212).

Lines 157-162. For calculating heterozygosity, the authors cite the paper they have followed. However, in my opinion there should be a bit of explanation of the principles of the methods and/or which software they have used. Also consider adding the script to a GitHub page.
For coverage filtering the authors use 2 times the mean genome-wide coverage. However, in the “Sequencing, read processing and alignment” (see line 143) the authors mention a filtering of 2.5 times the individual mean genome-wide coverage. Probably there is a typo in one of them but just checking.
We did provide additional information on the correction performed on the heterozygosity analysis (Line 229 - 231). The custom python script used to perform the genome-wide heterozygosity and ROH analysis have been made available with the preprint (see Data availability). Also, we thank the reviewer for pointing out the typo regarding the mean genome-wide coverage. It is now corrected.

Line 167: “corrected number of genotypes”. Corrected how?

We added a few words explaining what the correction consists of.

Line 195. For phasing, the authors use a Ne of 100,000 but later they calculate Ne for its breed (lines 312-313) and the resulting Ne is much smaller. How will this impact your phasing accuracy? Have you considered using the calculated Ne?
We used a Ne of 100,000 individuals because we expect the ancestral effective population size of chicken to be very large and these ancestral Ne's are the ones relevant for the phasing step in Beagle. For this reason, we didn't use the estimated current Ne. We would also like to highlight that for this new preprint we changed the method used to estimate changes in effective population size, because the NB package previously used in R is now deprecated, meaning that our analysis was not reproducible. Thus we used the very recently developed currentNe package which is well adapted to the domestic populations, having small population and strong family structures.

Line 201. Pedigree inbreeding: which software did you use to calculate FPED?
We used the pedigree library in R. We added this specification to the M&M (Line 269 - 271).

Line 205-106. Genome-based inbreeding. Does your "actual length of the genome (Lauto) covered in our dataset" exclude complex regions (duplicates etc)? If so, specify this.
We did not filter the length of the genome (here made by chromosome 1 up to chromosome 39) for complex regions, as duplicates were removed as much as possible in the alignment/mapping step. We specified this in the main text to avoid any further misunderstanding. Furthermore, since we considered the same Lauto for all individuals it is still allowing for relative comparisons between individuals and populations.

Line 212. Add the link to the whole-genome alignment from Ensembl so it is findable.
Added.

Line 227. Add the link to the GERP score so it is findable.
Added.

Lines 214-222. Please elaborate a bit more on the parameters and method used for VEP and chCADD – maybe link to the script or just add the parameters used. Also, when describing the filtering criteria: which software and how did you select this?

We added a few lines clarifying the VEP and chCADD analysis. We also added an additional reference, being that of Derks et al. 2018, as the idea behind the filtering of genes 1:1 ortholog between chicken and zebra finch was first developed and applied by Derks and colleagues. We also added the link to the publicly available chicken CADD scores in the Data availability section. We would like to highlight here that, since we switched to the latest chicken reference genome, we had to lift over all chCADD scores to the latest reference genome. To do this, we had to generate a liftover chain file, as explained in the new section entitled 'Liftover chain file generation' (Line 279).

In formula (1) is written chCADDi = etc. Is the i in the chCADD***i***=... correct? I am not really used to mathematical notation, but this felt wrong.
We thank the reviewer for pointing out this mistake in the formula, which has now been corrected.

Results

Line 275-276. Here the authors put the % of decrease in genome-wide heterozygosity and then in parenthesis the pi value. Is this the pi or the DeltaPi value? Otherwise to which population does this pi value refer to (the pre or the post)? Please clarify this.
The pi value reported in parenthesis refers to the genome-wide heterozygosity of the Barbezieux and Gasconne breed in 2013/15 or in other words the genome-wide heterozygosity after the establishment of the conservation programme. We realized that this value was confusing and misleading, so we rephrased that sentence.

Line 286. Add the range of the genome covered with long ROHs for the Gasconne as you have done with the Barbezieux before. "1 to 20 long ROHs that covered up to 29% of the genome (X-X%)".
Changed.

Phenotypic data paragraph (lines 343-359). I would love to see supplementary table 5 to 8 as plots (only for supplementary) as well. It is more useful to understand the trends and the data itself.
While we retained Table 5 to Table 8 in the SM, we generated two additional figures for the SM for Table 5 and Table 6. We did not do this for Table 7 and Table 8, as we believe these are simpler to understand as a table than a figure.

Discussion

Line 377. The authors mention that the founding nucleus was sampled from fancy breeders, was there an attempt to avoid relatedness? In other words, is it known if the founders were related? In parallel, the authors could estimate relatedness from the 2003 population to see how related they were to start with.

Unfortunately we are unable to answer this question, because we don't know whether the founders were related or not. Since they came from fancy breeders, it is very possible that the founders were related, as fancy breeders often keep a small flock where inbreeding is not always avoided. However, since we do not have information on the founders, we cannot provide any further information to the reviewer.

Line 443. First time chCADD acronym appears, define it as chicken CADD as you do here. This happens the first time in line 216.
Changed.

Lines 458-459. The authors could compare the results of their paper with the previous 57k SNP genotyping. How similar/different are they? Stress better why we need genome-wide sequencing.
Please see our answer to the same comment of Reviewer 1.

Figure and tables

Table 1. I am not familiar with chicken morphology and probably others are not as well. So, it was unclear the value of the "Morphology" column. Is it to show they are "similar breeds"? If that is the case maybe more context could be added in the methods section.
We understand the confusion. We have decided to change "Morphology" with "Comb type, feather colour" as this column essentially describes these two phenotypes.

Figure 2. Add "genome-wide" heterozygosity in the first sentence of the caption. I would recommend the authors to do a statistical test in the heterozygosity levels.
a) What does the values on top of the boxplot represent? DeltaPi? The symbols are strange, and the values are not the same as the ones in the text.
We changed the caption as suggested. Furthermore, we now provide a statistical test for the heterozygosity analysis and all other temporal analyses. We decided to remove the value on top of the boxplots because we understand this might look confusing. Instead we replaced the value with a symbol to indicate whether changes in diversity are significant or not.

Figure 3.
a) It is difficult to see which individuals are from 2003 and 2013/15 given that bold and italics are complicated to differentiate. The same goes to Gasconne and Barbezieux as some IDs do not have the prefix. I propose to use a color legend for the Heterozygosity bar plot so then it is easier to see which ones are which and add a prefix to each name. Alternatively, you could use color for breed and empty/full bar plot for pre/post sampling.
b) Is the difference statistically significant? Also, add color legend.
c) Add the p-value and Pearson's r in the caption as well (not just in the text).
We changed Figure 3a and 3b. We hope the current figure will be more clear than the previous one.

Figure 4. Add a statistical test for the genetic load.
Added.

Supplementary Material

Figure S1. Make the figure bigger (one on top of the other).
As we changed our pipeline for this new version of the preprint (see comments above), Figure S1 has been removed from the SM.

I really enjoyed reading and reviewing the paper and looking forward to reading a revised version of it.