RESPONSE LETTER TO THE COMMENTS OF THE RECOMMENDER AND REVIEWERS

Resubmission MS Title:

An evaluation of pool-sequencing transcriptome-based exon capture for population genomics in non-model species

Available at: https://doi.org/10.1101/583534

Authors:

Emeline Deleury

Thomas Guillemaud

Aurélie Blin

Eric Lombaert

Submitted to: PCI Genomics

Dear Recommender,

We are deeply sorry for the long delay in response. This was due to a number of factors, some of which were beyond our control.

We found the reviewer's comments and criticisms extremely helpful and cogent. We have incorporated most of their suggestions and feel our manuscript is much improved. We hope that you will find the MS appropriate for recommendation in PCI Genomics.

Please find below our detailed responses (in italic characters) to your comments as well as those of the two reviewers (in bold characters).

Best wishes,

Emeline Deleury on behalf of all authors

Recommender's comments:

COMMENTS TO THE AUTHOR:

Dear Authors,

The two reviewers have now responded positively to your manuscript. Although they were impressed by the quantity of work that is described here, they have also made constructive comments and suggestions to clarify the manuscript.

<u>The main points raised are the following:</u>

1/ Illustration of the analysis workflow: Given the rather consequent analyses reported throughout the study (i.e. pool versus individual; SNP calling within exon versus at exon-intron junctions/borders (IEB); CDS mapping versus genome mapping…), it would be recommended to illustrate the workflow with a schema to guide the reader (e.g something like Transcriptome > CDS > probes > sequencing (pool vs individual) > SNP calling/mapping (CDS vs genome)). This would certainly add considerable value/directness to the described strategies and may also emphasize the contribution of the pooling strategy in the correct estimation of VAF as compared to indexed individuals. In addition, it could be interesting to define and use acronyms for the different methods for a better readability.

We added an illustration to graphically summarize the analysis workflow (Figure 1).

2/ Filtering: They are various filters used along both the method and result sections: CDS selection, SNPs calling, read coverage, CDS genome mapping. One could ask if (and how) they may influence/impact the effectiveness of the strategy. In the same lines, are the " ~5 Mb of randomly chosen" transcripts really random given that they were filtered based on their N-content, size, GC content?

We hope that our detailed responses to the reviewer #1 (see below) will give you a satisfying answer to these different points.

Briefly, we believe that the filters we used for the target selection and for the SNP calling are usual, and allow obtaining good quality result (e.g. avoiding sequencing errors and paralogues), as it is the aim in most NGS-based study. We used tools widely used in the field such as VarScan2 or Bowtie2.

About the random choice: after applying a number of quality filters, we then chose 5.5Mb at random from filtered CDS, corresponding to 5,736 CDS. In other words, the random choice was done after the filtering. We understand the misunderstanding because we have sometimes misused the term

"random" when dealing with our final set of targets, which was not entirely randomly chosen. We have corrected the text accordingly (e.g. Lines 27 and 162). Additionally, we explain in the method section why we performed this step (Lines 138-139): "We have chosen to work on an exome subset, particularly with the future goal of working simultaneously on a large number of population samples".

Minor points:

- Although this is not the main point of the study, would it possible to give more details about the de novo transcript annotation (initial numbers, method for reconstruction, sequenced tissues/stages…)?

We have not been clear enough in the manuscript on this point: we actually did not produce and annotate these data ourselves. We only searched putative peptide-coding sequences (CDS) and calculated the values for the filters used (GC%, N, size) for each of them. Information about the de novo transcripts of H. axyridis *are available in other studies, such as Vilcinskas et al. (2013; https://doi.org/10.1098/rspb.2012.2113) and Vogel et al. (2017; https://doi.org/10.1016/j.dci.2016.09.015). In consequence, we have not added any additional information in the manuscript, but we have modified the text to make the origin of these data more obvious to the reader (Lines 112 and 133-134).*

- line 443 : "the allele frequency estimates obtained with the two mapping methods were highly correlated both for the pool (r=0.998; Fig. 2C) and for the individuals (r=0.998)." It seems that the correlations of AF between the 2 mapping strategies (CDS vs genome) is slightly different for lower AF values (<0.2), with the mapping onto CDS slightly overestimating AF as compared to mapping onto genome (Fig 2C). Would it be interesting to do the correlations by bins/intervals of AFs?

Figure 3C (formerly 2C) involves a very large number of points (174,307), and the difference mentioned here is mainly a visual artefact due to a few points. Indeed, when focusing on frequency below 0.2 (154,322 SNPs), the correlation between allele frequency estimates remains very high: 0.984 for both pool and individuals measurements. Rather than displaying correlation by intervals, we chose to follow reviewer #1's advice, and we computed Bland Altman average bias with its standard deviation (lines 284-288). Overall, the allele frequency estimates obtained with the two mapping methods displayed a good level of agreement, both for the pool (average bias = $6.8E10^{-4}$; bias SD = 0.013) and for the individuals (average bias = $1.0E10^{-3}$; bias SD = 0.012) (lines 454-458). For information, biases were only marginally higher when focusing on frequency below 0.2 (average bias = $1.1E10^{-3}$ and bias SD = 0.007 for the pool; average bias = $1.4E10^{-3}$ and bias SD = 0.007 and for the individuals).

- One section of the discussion seems to have been duplicated.

This was indeed a mistake, and we now removed one of the paragraphs.

- The references are presented twice.

This has been corrected.

Additional requirements of the managing board:

As indicated in the 'How does it work?' section and in the code of conduct, please make sure that:

-Data are available to readers, either in the text or through an open data repository such as Zenodo (free), Dryad or some other institutional repository. Data must be reusable, thus metadata or accompanying text must carefully describe the data.

-Details on quantitative analyses (e.g., data treatment and statistical scripts in R, bioinformatic pipeline scripts, etc.) and details concerning simulations (scripts, codes) are available to readers in the text, as appendices, or through an open data repository, such as Zenodo, Dryad or some other institutional repository. The scripts or codes must be carefully described so that they can be reused.

-Details on experimental procedures are available to readers in the text or as appendices.

-Authors have no financial conflict of interest relating to the article. The article must contain a "Conflict of interest disclosure" paragraph before the reference section containing this sentence: "The authors of this preprint declare that they have no financial conflict of interest with 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 one of the PCI XXX recommenders."

We have carefully followed the "How does it work?" section and code of conduct of PCI Genomics. In particular:

- *the exome sequence capture data reported in the manuscript have been deposited in the European Nucleotide Archive (accession no. PRJEB31592).*
- *The CDS sequences targeted and the description of the exon positions on the subset of transcripts have been deposited at Zenodo, https://doi.org/10.5281/zenodo.2598388*
- *The HaxR v1.0 genome sequence used in this study is available from http://bipaa.genouest.org/sp/harmonia_axyridis/*
- *Scripts and tool for intron-exon boundary prediction are available from https://github.com/edeleury/IEB-finder*
- *The "Conflict of interest disclosure" paragraph has been updated.*

Reviewer #1's comments

General remarks

Overall, this paper contains a lot of work and an interesting method to find intron-exon boundaries. In general, it was for me rather difficult to follow the reasoning behind some steps, especially those sections related to the SNPs and the filters used at various steps in the materials and methods. Furthermore, I do not entirely understand what the exact aim is of the paper: exome sequencing or "random subset of the exome" sequencing or sequencing of orthologous targets (because a lot of references are made to phylogenetic studies, even the first sentence immediately refers to that), … To me, the main idea seemed to be the random subset targeted sequencing of coding regions. If that is indeed the general idea and if it thus is not to be used for phylogenetics, neither for sequencing the entire exome, I would more focus the writing of the paper on that: why do you want to sequence a subset, what is the rationale for that, for what can it be used. If the general idea is that sequencing a pool has a limited effect on allelic frequencies, this should also be quantified in more detail. Overall, I do like the general concept of the paper, however, I think it can be refined more.

See below, in the "in more details" section, for detailed answers about the points raised here.

I have a couple general remarks:

- **Reporting of the results**
- **The possibility of an upwards bias of the results**
- **The filters used between line 217 - 232**
- **The use of the word "exome"**
- **The statistical analyses in general**
- **The bias towards the ends (line 33, line 606)**

See below for responses to these remarks.

In more detail:

Reporting of the results: this paper contains an enormous amounts of comparison in general sets, subsets, with further subdivisions and so on. It would improve the understanding of several results if results would be reported as XX% (number/total number). Maybe a nice figure detailing the subsets would also be nice and make it more easy to follow.

We have (i) reported throughout the text as much results as meaningful XX%, in order to make the manuscript easier to read (e.g. Lines 244, 270, 341), and (ii) added an illustration (Figure 1) to

summarize the analyses corresponding to the main objectives of the study, and we believe that this helps a lot the general understanding of our work. We have.

There are several reasons why I think the results presented here are biased upwards:

Line 133 – 136: While I understand why you used these filters, it also results in omitting regions that are typically difficult to sequence (see Broeckx et al., 2015). If you talk about capturing exomes (line 107, 500, …), you will also have to capture these regions. By omitting them, your results are likely biased upwards towards more easily sequenced regions.

This point is related to the general question of the aim of our paper that may not have been clear enough, according to reviewer #1. The goal is not to make a "whole-exome sequencing" but to capture a good-quality subset of the exome. This may be useful in a number of studies which aim at comparing various samples at the inter-specific or intra-specific scale (e.g. phylogenetics, genome-scan, comparative studies on genetic load), giving access to a large, although not full, section of the genome under natural selection. We have made some changes to the text so that the confusion between "exome" (which may be understand as "whole-exome") and "exome subset" is avoided as much as possible (e.g. Lines 111, 147, 268, 520 and 546).

Here, the reviewer refers to the missing data filter (i.e. N content), the GC% filter and the CDS length filter that we used for the target choice and probe design. These filters have been used in a number of studies to which we compare our results in the discussion section. We agree that the global efficiency of the capture will likely be sensitive to the precise number of filters, and to the threshold values of each filter, which may be quite different depending on the considered study. We deliberately chose to apply the filters recommended by the probe manufacturer to select our CDS in order to maximize the theoretical capture efficiency, and to have the necessary data to achieve our main objective (i.e. to properly assess SNPs allelic frequencies by combining capture and pooling). Moreover, it is worth noting that the coverages obtained in our experiment are heterogeneous between targets and along a target, i.e. capture efficiency is heterogeneous but reproducible between samples: this suggests that it may be possible to obtain usable data for targets that are less easy to capture.

Line 150: I have checked later on, you seem to compare your results with what was predicted to be sequenced by Roche Nimblegen (5717 CDS), not what you aimed to sequence (5736 CDS). This also will bias your results upwards.

The decrease from 5,736 CDS to 5,717 CDS is related to a series of bioinformatics filters carried out by Nimblegen, and briefly described in the paragraph lines 149-159, eg removing mitochondrial CDS). Therefore, we did not aim at sequencing 5,736 CDS but 5,717. It is now clearer in this MS. We now say

at the beginning of the paragraph: "The probes based on the selected CDS were designed and produced by NimbleGen". And at the end of the paragraph: "The final probe set contained 6,400 regions of overlapping probes, corresponding to a targeted exome size of 5,347,461 bases distributed over 5,717 of our selected CDS (Zenodo https://doi.org/10.5281/zenodo.2598388). This final probe set was used for the preparation of a SeqCap EZ Developer assay by Roche NimbleGen Technologies, which synthesized the probes as biotinylated DNA oligomers".

Line 352: Covered by one read says something, this is of course not really useful. A base that is covered once will not allow you to make a reliable variant call. The set of bases that is captured and sequenced at a sequencing depth that is sufficiently high for variant calling will be far lower.

In this section, we want here to provide statistics about the capture itself, not genotyping or variant calling. These are very different information: the capture efficiency on one side, and the quality of the genotyping on the other. About the quality of the genotyping (and thus variant call), we provide and discuss many data (especially about sequencing depth) throughout the whole manuscript, elsewhere.

To summarize the global efficiency of the capture itself, and to be able to compare capture efficiency between different studies (Puritz & Lotterhos, 2018), we provide the capture sensitivity (by definition, the percentage of targets that are covered by at least one read (e.g. Bi et al. 2012, Portik et al 2016, Puritz & Lotterhos, 2018)) and the capture specificity (i.e. the percentage of cleaned reads that are aligned to target sequences).

The filters used in lines 217 – 232: I am not entirely sure why these filters are applied and if they are applied, whether they bias the results or not. For instance, why restrict the analysis to ensure target size is the same. If one of the approaches succeeds in sequencing more, it seems like something you like and not something that should be diminished. The reasoning for 3 reads is something I understand, albeit that that means nearly no supporting evidence for a variant. The 15 is something I do not understand entirely however. Furthermore, why only variants called in 20 individuals? Can you clarify these filters and explain why they are safe to use, i.e. do not cause a bias.

Our main objective is to compare the allelic frequencies of SNPs obtained from the pool with those obtained from individuals (considered as the reference here), i.e. do we find the same SNPs identified by these two approaches, and if so do they have the same frequencies? To compare in a coherent way the SNPs resulting from the two captures (pool vs. individuals), it was necessary to consider the same size of covered exome for the pool and for the individuals. We have chosen the minimum sequencing depth parameters (one of the SNP calling parameters) for the individuals (40X) and for the pool (250X) which allow us to consider about the same size of covered targets (90-93%). As we have shown that

the capture efficiency was reproducible between the two captures (e.g. Fig. S5), we assumed that the same regions of the targets were considered for the two captures. We have thus placed ourselves in relevant conditions of comparison. Note that 6% of the targets (base level, cf. table 1) are not captured by any library, so we do not substantially reduce our initial dataset by applying these filters.

*The filters applied here are rather "classical" filters in the context of variant calling. They aim at avoiding as much as possible sequencing errors by discarding positions that lack sequencing depth, as well as variants supported by too low number of reads. The minimum number of reads to call a variant was then defined in relation to the minimum coverage defined for the pool/individual comparison (see above). In the manuscript, we say: "given the pool's minimum coverage of 250X, we expected to have ~3 reads per haplotype if the 72 haplotypes in the pool were homogeneously captured" (Lines 229-231). The rational is the same for the individuals: the individual's minimum coverage is 40X, and we thus expect a minimum of ~20 reads per haplotype (*Harmonia axyridis *being a diploid species). We arbitrarily chose 15 reads to take into account the variance between both haplotypes sequencing depths (we added precisions to the text, lines 228-229). For the pool it was difficult to account for this variance, and it is likely that we missed some SNPs that were poorly represented in this sample, but it was difficult to go below 3 at the risk of getting sequencing errors.*

Variant called was not performed only on 20 individuals. The variant called was performed on at least 20 individuals (knowing that the maximum number of individuals is 23). We did not retain the positions which were genotyped in less than 20 individuals in order to provide a good and sharp estimation of allele frequencies. We added precisions to the text (lines 239-240).

In general, I am confused by the usage of the word "exome". Is the goal to ultimately use this technique to sequence the entire exome for a large number of individuals in pool OR to sequence a random subset of the genome to obtain frequency estimates? If the goal is to sequence a random subset, it is also more OK to use the 5717 CDS instead of the 5736 CDS (see earlier remark). The third remark (in the section of biases) still remains at that moment however. In general however, I do have the feeling that you put the subset and the exome at the same level, as also stated in the discussion (line 500) and that biases the results.

We agree that we have not been careful enough in the text with regard to the use of the term "exome". Our goal was not to capture the full exome. The purpose of our article is to show the feasibility of a transcriptome-based exome capture on pool data of a non-model species. The idea is therefore to capture a good-quality subset (potentially quite large) of the exome, which may be useful in a number of studies which aim at comparing various population samples at the inter-specific and intra-specific scales (e.g. phylogenetics, genome-scan, comparative studies on genetic load), giving access to a large, although not full, section of the genome under natural selection.

We have made some changes to the text so that this confusion between "exome" and "exome subset" is avoided (e.g. Lines 111, 147, 268, 520 and 546).

Statistical remarks:

- Which correlation coefficient was used? (e.g. line 375-376)

We used the Pearson correlation coefficient. We have added this information throughout the manuscript.

- A more general remark is the question: what is the aim? Demonstrating that allelic frequencies are rather similar in both approaches and an accurate representation of true population allelic frequencies? In general, obtaining a (rather) high correlation coefficient is not that unexpected, especially given the fact that a large number of samples is shared. Much more informative is to get an idea on how divergent the estimates are, e.g. by using Bland Altman plots and calculating SDs for the difference. At that moment, you have an idea to what extent the allelic frequencies differ and whether that is acceptable or not.

Demonstrating that allelic frequencies can be accurately estimated with pool data mapped directly onto CDS (thus without a reference genome) is the core objective of our work. As suggested by reviewer #1, we now use the Bland Altman approach: we have added for all the main allelic frequencies comparisons (i.e. pool vs. individuals; CDS mapping vs genome mapping) the average bias with its standard deviation (lines 284-288, 428-429, 449-451 and 456-458).

The bias towards the end: you state in the abstract and line 606 that there is a bigger bias towards the end when it comes to estimating allele frequencies. This is to be clarified more. From the paper, I get the feeling you mean that this refers to more variants that are not in HWE. Bias in allelic frequency to me refers in this case to more widely diverging estimates of the true allele frequency, not to deviations of HWE. At line 610, there is also a statement of 296,736 SNPs. This is the only time I found this number in the manuscript. Where does it come from?

We do not refer to a bias in the estimation of allele frequencies, but to the prediction of false SNP. We have shown that the direct mapping onto CDS can produce false SNPs near the IEBs (due to the similarity in sequence between the beginning of the intron and the beginning of the next exon). Figure S6 illustrates this with an example and Figure S2B explains how this is possible (alignment of the beginning of the intron).

Where does the number 296,736 come from? Since our IEB predictor allows us to locate these regions, we propose in the discussion to remove all SNPs found near the predicted IEBs and in short

regions - framed by 2 predicted IEBs. By doing this, we remove 3,300 SNPs and end up with 296,736 SNPS (300,036-3,300=296,736 SNPs). This is now clearer in the MS (lines 609-610).

Of the 3,300 SNPs removed, a large proportion (408 out of 3,300 = 12.3%) are not at HW. Only 0.3% (1,007 SNPs) of the 300,036 SNPs displayed significant deviation from Hardy-Weinberg equilibrium (line 416) and a large part of the SNPs that are not at HW equilibrium (408 out of 1,007 = 40.5%) are located next to IEBs. This is, if not a proof, a good additional indication of the existence of false SNPs at the vicinity of IEBs when mapping directly onto CDS.

We have modified the text to make it clearer: "A large proportion of the 3,300 excluded SNPs (408 SNPs, i.e. 12.36%) were not at Hardy Weinberg equilibrium in individuals, which is to be expected if many of them are false SNPs" (lines 610-612). Also, we do not talk anymore about "biases", but only about "false SNPs" (lines 606-607).

Some smaller remarks:

Line 137 – 138: how did you do the random selection 5.5Mb?

Using a homemade script, targets were randomly drawn from the filtered CDS until the desired exome size was reached. More precisely: the index (1 to 12,739) of the CDS was drawn from a discrete uniform distribution U[1;12,739] without replacement until the desired subset exome size is reached. We have added these precisions in the manuscript (lines 146-148).

Line 147: why are probes that match other species omitted? This also implies that you have to have access to reference genomes of closely related species? If this is the case, best to mention this as a limitation.

We did not omit "probes that match other species". As explained in the manuscript (lines 151-154), we only omitted probes with more than one close match in the H. axyridis *de novo transcriptome or in the draft genome of* A. glabripennis*.*

Our method does not require the use of reference genomes of closely related species. However, we believe that the use of the genomes of outgroup species improves the quality of the probe design. This is not a limitation because using easily available genomic data of model species is actually achievable in the probe design process of any non-model species. In our case, it is important to emphasize that Tribolium castaneum *and* Anoplophora glabripennis *are not "related species" of* H. Axyridis*: the estimated divergence time of both species with* H. axyridis *is approximately 240 MYA (see http://www.timetree.org/).*

Line 238 – 241: why was the Hardy-Weinberg equilibrium test performed? If it is used as a proxy for genotyping error, I do not entirely agree with the concept… It has been shown that HWE testing will not achieve this. Some references:

Leal, S. M. Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. Genet. Epidemiol. 29, 204–214 (2005).

Zou, G. Y. & Donner, A. The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: A cautionary note. Ann. Hum. Genet. 70, 923–933 (2006).

Teo, Y. Y., Fry, A. E., Clark, T. G., Tai, E. S. & Seielstad, M. On the usage of HWE for identifying genotyping errors. Ann. Hum. Genet. 71, 701–703 (2007).

Fardo, D. W., Becker, K. D., Bertram, L., Tanzi, R. E. & Lange, C. Recovering unused information in genome-wide association studies: the benefit of analyzing SNPs out of Hardy-Weinberg equilibrium. Eur. J. Hum. Genet. 17, 1676–82 (2009).

We only used the HW test as an indicator of a possible bias and of possible SNP calling errors, especially those associated with false-positive SNPs (pseudo-SNPs) due to paralogs or the presence of an IEB region. However, we never interpret these HW test results as conclusive evidences, but only as interesting clues. Thus, we did not use the HW test to select/remove SNPs in our workflow.

For example, we found that the proportion of SNPs not at HW equilibrium decreased when the "maximum depth" filter was applied (lines 416-418). This result is consistent with the hypothesis of false-positive SNPs associated to paralogs, but we do not state in the manuscript that the HW test is conclusive evidence (lines 571-575).

Also, we observed that the proportion of SNPs in disequilibrium close to the IEBs was much higher than that observed on the rest of our targets. This result is consistent with the false-positive SNPs in the vicinity of IEBs as described in Figure S6, but we do not state in the manuscript that the HW test is conclusive evidence (lines 610-612).

Line 470 – 472 and 479 – 481: I do not entirely understand the clarifications in these sentences. Can you clarify more? Regions of 7 and 8 bp are explained in the methods but less not. Where does the cut-off of 10 bp come from and where can we find the 153 regions?

This question actually refers to two different points:

About former Lines 470-472: Theoretically, and as explained in the Methods section, regions predicted as IEBs have a minimum size of 7 bp if the region is (i) fully covered and (ii) not at the end of the CDS. In some cases, it is possible that one of the 2 exon ends are not covered by reads (for example, small exon at the end of a CDS). In this case, this region is considered not covered at all, and the deviation of the signal found on the second exon end covered cannot deport to the uncovered region.

In this case, we obtain a signal (1 base) and an extension of the signal in one direction only (3 bases), i.e. a region "+" of only 4 bases. We added a sentence in the Methods section, lines 328-330.
About former Lines 479-481: On figure S2B, we have illustrated a signal deviation when the 3 bases of the beginning of the intron are identical to the 3 bases of the beginning of the next exon. If the signal has a greater deviation than expected by the predictor (n=3), then the tool predicts, for a true IEB, two predicted IEB regions very close one to another (i.e. two predicted IEB regions separated by a region - very short). We thus counted the number of short regions -, lower than 10bp (arbitrary threshold here) obtained by considering two values of signal offset (n=3 and n=5). We counted 153 and 51 regions for n=3 and n=5, respectively. By increasing the length of the signal offset, fewer short - regions are predicted, i.e. fewer false predictions. We modified a sentence in the Methods section to make it clearer, lines 492-494.

Line 501 – 502: actually, you did use a genome from a different species or at least I have the feeling that you did (line 147)

Yes we used the genome of two outgroup species, namely Tribolium castaneum *and* Anoplophora glabripennis*, to improve the quality of the probe design. However, our method does not mandatorily require the genome of a related species (*Tribolium castaneum *and* Anoplophora glabripennis *are not closely related to the focus species* Harmonia axyridis: *the estimated divergence time of both species with* H. axyridis *is approximately 240 MYA (see http://www.timetree.org/)). Using easily available genomic data of model species distantly related to the focus species is actually achievable in the probe design process of any non-model species.*

Line 548 – 574: a big part is repeated in this section. Something went wrong here?

This was indeed a mistake, and we consequently removed one of the paragraphs.

Detailed remarks:

Line 38: I was a bit confused by first sentence. I would suggest to rephrase the part of "reliable set of orthologous loci" to "obtaining genotype calls for a set of orthologous loci"

This has been modified (lines 38-39).

Line 41: represents => is

This has been corrected (line 41).

Line 42 - 44: Hybridization capture … DNA fragments => Hybridization capture is one of these reduced representation methods that allows the enrichment of a preselected set of hundreds to thousands of genes or DNA fragments from the genomic DNA.

This has been modified (lines 42-44).

Line 52: given the tendency of functional elements to be conserved even ~~in~~ after

This has been corrected (line 52).

Line 59-60: "An alternative … to capture probe design" => An alternative approach for non-model species involves designing DNA capture probes based on a de novo ...

This has been corrected (lines 59-60).

Line 60: can you also add a little bit more information on the technique? I was confused for a second about whether the aim was to target DNA or RNA.

We have clarified that section: "An alternative approach for non-model species involves designing DNA capture probes on the basis of sequences obtained from a de novo *assembled transcriptome for the species studied or a related species (e.g. Bi et al., 2012; Neves et al., 2013)" (lines 59-61).*

Further explanation can be found in the "methods" section.

Line 66: through => towards

This has been corrected (line 67).

Line 68: Even => In addition, even …. have <u>still</u> been found to

This has been corrected (Line 70).

Line 113: the same individuals gave me the feeling that exactly the same set of individuals were sequenced, which is not the case.

We changed "individuals" to "population sample" to avoid any confusion at this point of the MS.

Line 126: This sentence ("We designed … of H. axyridis.") is confusing here. I would omit it (especially as a similar sentence with more information is available at line 137 – 138).

We removed this sentence, and replaced "the species" with "H. axyridis" in the following one (line 134)

Line 138 – 141: I think this step is purely a step that explains what the results of the random selection is? Can you maybe add that these are "out of curiosity" results because it made me doubt whether this was used further for anything downstream.

These results were actually not used in the following analyses. Therefore, we added "To get additional information on our unannotated exome subset, […]" at the beginning of the sentence (lines 148-149).

Line 160: is PIF an abbreviation?

PIFs is the acronym of "Pool Individual Frequency SNP". We added this information in the manuscript (line 167).

Line 383: "not therefore" => "therefore not"

This has been corrected (line 400).

Line 387 – 392: can you add the total number of SNPs as well here (i.e. the 409,328) to make the calculations more clear?

The total number of SNPs has been added to the paragraph (line 406).

Figure 1. The individuals only box is not visible, probably due to the low number? Best to say something about it.

There are only 86 positions upon 300,036 (0.03%) that are polymorphic only in individuals. This is why they were not visible in Figure 2 (formerly 1). We decided to remove those SNPs from Figure 2, and mention them in the caption.

Line 517: orthology, due to … => "orthology. In both cases, this is due to the …"

This has been modified (line 531).

Line 531: "a random subset" of the exome

This has been modified (lines 545-546).

Line 534: estimation of allelic frequencies => the actual impact is not detailed on, only correlation coefficients are used.

We agree with that comment, and we have revised that sentence to soften our statement: "The study of the subset of CDS with complete genomic sequence matches further supported the idea that direct mapping onto the CDS has probably no major impact on SNP identification or on the estimation of allele frequencies relative to mapping onto the genome" (lines 546-547). Note that, as detailed above, we

have now calculated and added the Bland-Altman average bias with its standard deviation for every comparisons of allele frequencies (lines 284-288, 428-429, 449-451 and 456-458).

Line 580: “instead identifying” => “instead it identifies”

This has been corrected (line 581).

Line 590: “level of coverage” => “coverage levels” (although coverage might be confusing in an article about targeted sequencing where you talk about how much of a region is covered; maybe sequencing depth is a better word throughout the manuscript)

We modified the text accordingly. Also, we agree with Reviewer 1, and we consequently modified the term "coverage" with the more appropriate term "sequencing depth" in all the relevant places throughout the manuscript.

Line 602: SNP polymorphismS

This has been corrected (line 602).

Line 610-612: this seems to suggest that deviations from HWE = genotyping error, which I have troubles with (see general comments)

Of course, there are many other reasons than genotyping errors to explain a deviation from HWE. Consequently, we revised the sentence to mitigate our statement, which was too affirmative in the previous version of the manuscript: “A large proportion of the 3,300 SNPs excluded (408 SNPs, i.e. 12.36%) were not at Hardy Weinberg equilibrium in individuals, which is an additional evidence of confirming the relevance of this filter” (lines 610-612).

Line 613: the actual deviations were not measured.

As detailed above, we have calculated and added the Bland-Altman average bias with its standard deviation for every comparisons of allele frequencies (lines 284-288, 428-429, 449-451 and 456-458).

Reviewer X's comments

In "An evaluation of pool-sequencing transcriptome-based exon capture for population genomics in non-model species", Deleury et al. propose a new method to generate a transcriptome-based exon capture suitable for large scale population genomics studies at the computational levels by mapping directly to the transcriptome and cost efficient by the use of pool-sequencing. They illustrate the different step on *Harmonia axyridis*. The also create a new method to identify intron-exon boundaries, method available through github.
I would recommend the paper after some revisions.

My comments are listed below.

General comments:

The authors performed a lot of different steps to generate the data use for the benchmark. Nonetheless, because of the amount of steps, the reader can be lost in the different filtering and the data used. A flowchart could help to understand and to follow the text, as well as the stating what are the data used in the paper, we can be lost between the draft genome of *Harmonia axyridis*, the de novo transcriptome of *H. axyridis* and the use of other species.
We added an illustration to graphically summarize the analysis workflow (Figure 1).

Specific comments:

line 52: 'even in after, the in should be remove.
This has been corrected (line 52).

Figure 1: We cannot see the 'Individuals only', probably because it is too small compare to the other two. Maybe another representation is needed, otherwise, we don't know where the 'Individuals only' are.
There are only 86 positions upon 300,036 (0.03%) that are polymorphic only in individuals. This is why the there were not visible in Figure 2 (formerly 1). We decided to remove those SNPs from Figure 2, and mention them in the caption.

Figure 2: The arrow in the text of the figure should be removed, they didn't add anything and are confusing. The figure is also a bit unclear and it is difficult to read and understand which text is referring to which panel. Maybe a figure with the panels on the left and the text on the right will be better.

Figure 3 (formerly 2), as well as its caption, was modified. Note that the addition of figure 1 should allow a better understanding of this figure.

line 446: Is the use of the word 'private' here mean specific? Is it classically used? If not, the word specific should be use instead, here and after.

The word "private" refers here to SNPs identified in a single mapping approach. We added this precision to the text (Lines 459-461).

line 459: IEB will be better written in full letters because it seems that it is the first time the acronym is used in the Result part.

This has been modified (line 473).

line 462: IEB will be better written in full letters in the title.

This has been modified (line 477).

line 548-574: The two paragraphs say the same thing. It is, I think a mistake. One should be choose.

This was indeed a mistake, and we consequently removed one of the paragraphs.

Comment on IEB_finder:

I was able to run the first step, i.e. Step 1 : collect_CDS_infos.pl, but for the second step (Mapping genomic reads on CDS sequences), there is no 'genomicReads.fq' file to test the tool.

The file "genomicReads.fq" was added to the github platform (https://github.com/edeleury/IEB-finder). The readme.md has also been updated.