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.

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)

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.

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.
- 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.
- 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.

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.

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.

Statistical remarks:

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

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?

Some smaller remarks:

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

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.

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).

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?

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)

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

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"

Line 41: represents => is

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.

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

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* ...

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.

Line 66: through => towards

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

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

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).

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.

Line 160: is PIF an abbreviation?

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

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?

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

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

Line 531: "a random subset" of the exome

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

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

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)

Line 602: SNP polymorphismS

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

Line 613: the actual deviations were not measured.