Peer Community In Genomics

Assessing a novel sequencing-based approach for population genomics in non-model species

Thomas Derrien *and* ***Sebastian Ernesto Ramos-Onsins*** *based on peer reviews by* ***Valentin Wucher*** *and 1 anonymous reviewer*

Emeline Deleury, Thomas Guillemaud, Aurélie Blin & Eric Lombaert (2020) An evaluation of pool-sequencing transcriptome-based exon capture for population genomics in non-model species. Missing preprint_server, ver. Missing article_version, peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/583534

Submitted: 26 February 2020, Recommended: 09 October 2020

Cite this recommendation as:
Derrien, T. and Ramos-Onsins, S. (2020) Assessing a novel sequencing-based approach for population genomics in non-model species. *Peer Community in Genomics*, 100002. 10.24072/pci.genomics.100002

Published: 09 October 2020

Developing new sequencing and bioinformatic strategies for non-model species is of great interest in many applications, such as phylogenetic studies of diverse related species, but also for studies in population genomics, where a relatively large number of individuals is necessary. Different approaches have been developed and used in these last two decades, such as RAD-Seq (e.g., Miller et al. 2007), exome sequencing (e.g., Teer and Mullikin 2010) and other genome reduced representation methods that avoid the use of a good reference and well annotated genome (reviewed at Davey et al. 2011). However, population genomics studies require the analysis of numerous individuals, which makes the studies still expensive. Pooling samples was thought as an inexpensive strategy to obtain estimates of variability and other related to the frequency spectrum, thus allowing the study of variability at population level (e.g., Van Tassell et al. 2008), although the major drawback was the loss of information related to the linkage of the variants. In addition, population analysis using all these sequencing strategies require statistical and empirical validations that are not always fully performed. A number of studies aiming to obtain unbiased estimates of variability using reduced representation libraries and/or with pooled data have been performed (e.g., Futschik and Schlötterer 2010, Gautier et al. 2013, Ferretti et al. 2013, Lynch et al. 2014), as well as validation of new sequencing methods for population genetic analyses (e.g., Gautier et al. 2013, Nevado et al. 2014). Nevertheless, empirical validation using both pooled and individual experimental approaches combined with different bioinformatic methods has not been always performed. Here, Deleury et al. (2020) proposed an efficient and elegant way of quantifying the single-nucleotide polymorphisms (SNPs) of exon-derived sequences in a non-model species (i.e. for which no reference genome sequence is available) at the population level scale. They also designed a new procedure to capture exon-derived sequences based on a reference transcriptome. In addition, they were able to make predictions of intron-exon boundaries for de novo transcripts based on the decay of read depth at the ends of the coding regions. Based on theoretical predictions (Gautier et al. 2013), Deleury et al. (2020) designed a procedure to test the accuracy of variant allele frequencies (AFs) with pooled samples, in a reduced genome-sequence library made with transcriptome regions, and additionally testing the effects of new bioinformatic methods in contrast to standardized methods. They applied their strategy on the non-model species Asian ladybird (*Harmonia axyridis*), for which a draft genome is available, thereby allowing them to benchmark their method with regard to a traditional mapping-based approach. Based on species-specific *de novo* transcriptomes, they designed capture probes which are then used to call SNPx and then compared the resulting SNP AFs at the individual (multiplexed) versus population (pooled) levels. Interestingly, they showed that SNP AFs in the pool sequencing strategy nicely correlate with the individual ones but obviously in a cost-effective way. Studies of population genomics for non-model species have usually limited budgets. The number of individuals required for population genomics analysis multiply the costs of the project, making pooling samples an interesting option. Furthermore, the use of pool sequencing is not always a choice, as many organisms are too small and/or individuals are too sticked each other to be individually sequenced (e.g., Choquet et al. 2019, Kurland et al. 2019). In addition, the study of a reduced section of the genome is cheaper and often sufficient for a number of population genetic questions, such as the understanding of general demographic events, or the estimation of the effects of positive and/or negative selection at functional coding regions. Studies on population genomics of non-model species have many applications in related fields, such as conservation genetics, control of invasive species, etc. The work of Deleury et al. (2020) is an elegant contribution to the assessment and validation of new methodologies used for the analysis of genome variations at the intra-population variability level, highlighting straight bioinformatic and reliable sequencing methods for population genomics studies.

References:

[1] Choquet et al. (2019). Towards population genomics in non-model species with large genomes: a case study of the marine zooplankton Calanus finmarchicus. Royal Society open science, 6(2), 180608. doi: https://doi.org/10.1098/rsos.180608

[2] Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M. and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics, 12(7), 499-510. doi: https://doi.org/10.1038/nrg3012

[3] Deleury, E., Guillemaud, T., Blin, A. and Lombaert, E. (2020) An evaluation of pool-sequencing transcriptome-based exon capture for population genomics in non-model species. bioRxiv, 10.1101/583534, ver. 7 peer-reviewed and recommended by PCI Genomics. https://doi.org/10.1101/583534

[4] Ferretti, L., Ramos-Onsins, S. E. and Pérez-Enciso, M. (2013). Population genomics from pool sequencing. Molecular ecology, 22(22), 5561-5576. doi: https://doi.org/10.1111/mec.12522

[5] Futschik, A. and Schlötterer, C. (2010). Massively parallel sequencing of pooled DNA samples—the next generation of molecular markers. Genetics, 186 (1), 207-218. doi: https://doi.org/10.1534/genetics.110.114397

[6] Gautier et al. (2013). Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Molecular Ecology, 22(14), 3766-3779. doi: https://doi.org/10.1111/mec.12360

[7] Kurland et al. (2019). Exploring a Pool-seq-only approach for gaining population genomic insights in nonmodel species. Ecology and evolution, 9(19), 11448-11463. doi: https://doi.org/10.1002/ece3.5646

[8] Lynch, M., Bost, D., Wilson, S., Maruki, T. and Harrison, S. (2014). Population-genetic inference from pooled-sequencing data. Genome biology and evolution, 6(5), 1210-1218. doi: https://doi.org/10.1093/gbe/evu085

[9] Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A. and Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome research, 17(2), 240-248. doi: https://doi.org/10.1101%2Fgr.5681207

[10] Nevado, B., Ramos-Onsins, S. E. and Perez-Enciso, M. (2014). Resequencing studies of nonmodel organisms using closely related reference genomes: optimal experimental designs and bioinformatics approaches for population genomics. Molecular ecology, 23(7), 1764-1779. doi: https://doi.org/10.1111/mec.12693

[11] Teer, J. K. and Mullikin, J. C. (2010). Exome sequencing: the sweet spot before whole genomes. Human molecular genetics, 19(R2), R145-R151. doi: https://doi.org/10.1093/hmg/ddq333

[12] Van Tassell et al. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature methods, 5(3), 247-252. doi: https://doi.org/10.1038/nmeth.1185

Reviews

Evaluation round #1

DOI or URL of the preprint: https://doi.org/10.1101/583534

Authors' reply, 15 September 2020

Download author's reply

Decision by Thomas Derrien, posted 07 April 2020

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

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.

The main points raised are the following:

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.

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?

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...)? - 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?

- One section of the discussion seems to have been duplicated.
- The references are presented twice. Overall, the manuscript is well written and report a very interesting and cost effective strategy to estimate allele frequencies in non-model organisms at the population level, therefore we are looking forward to seeing a revised version.

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." All the best,

Thomas DERRIEN

Reviewed by Valentin Wucher, 02 April 2020

Download the review

Reviewed by anonymous reviewer 1, 31 March 2020

Download the review