Received: 23 August 2021 | Revised: 21 September 2021 | Accepted: 11 October 2021

DOI: 10.1111/1755-0998.13531

MOLECULAR ECOLOGY RESOURCES WILEY

NEWS AND VIEWS

Perspective

Good from far, but far from good: The impact of a reference genome on evolutionary inference

Blair P. Bentley[1] | **Ellie E. Armstrong**[2]

[1]Department of Environmental Conservation, University of Massachusetts, Amherst, Massachusetts, USA

[2]Stanford University Department of Biology, Stanford, California, USA

Correspondence
Blair P. Bentley, Department of Environmental Conservation, University of Massachusetts, Amherst, MA, USA.
Email: bbentley@umass.edu

Genomic diversity and past population histories are key considerations in the fields of conservation and evolutionary biology. In this issue of *Molecular Ecology Resources*, Prasad et al. (*Mol. Ecol. Resour.*, 2021) examine how the quality and phylogenetic divergence of reference genomes influences the outcomes of downstream analyses such as diversity and demographic history inference. Using the beluga whale and rowi kiwi as examples (Figure 1), they systematically estimate heterozygosity, runs of homozygosity (ROH), and demographic history (PSMC) using reference genomes of varying quality and phylogenetic divergence from the target species. They show that demographic history analyses are impacted by phylogenetic distance, although this is not pronounced until divergence exceeds 3% from the target species. Similarly, their results imply that heterozygosity estimates are dependent on phylogenetic distance and the method used to perform the estimates, and ROHs are potentially undetectable when a nonconspecific reference is used. This investigation into the role of divergence and quality of reference genomes highlights the impact and potential biases generated by genome selection on downstream analyses, and provides a possible alternative in cross-species scaffolding in instances where a conspecific reference genome is not available.

Demographic history and genomic diversity of individuals, populations, and species are key features in both evolutionary and conservation biology. In the absence of a conspecific reference genome, closely related species are often used to align the raw reads of a target species for the purpose of downstream analyses. Despite the widespread use of non-conspecific reference genomes, few studies have directly addressed how phylogenetic divergence or assembly quality of the reference may bias or limit downstream analyses. The article by Prasad et al. (2021) in the current issue of *Molecular Ecology Resources* attempts to address this gap in knowledge using two data sets from distinct vertebrate lineages. In their paper, they find that in general, demographic trajectories, but not absolute values are reliable, irrespective of the phylogenetic divergence and quality of the reference genome. Heterozygosity estimates display a more complicated pattern and increase incrementally as phylogenetic distance from the target increases, unless a parameter to adjust for multiple mismatches is included in the analysis pipeline, which results in decreasing heterozygosity with distance. The authors also reveal that phylogenetic distance influences ROH detection, with any non-conspecific genome unable to resolve ROHs for the rowi kiwi. As no ROHs were detected in the beluga whale when aligned to any reference genome (including conspecific), the effect of reference genome divergence on ROHs warrants further investigation (Figure 1).

Genome-wide heterozygosity is a key component of many downstream population genetic analyses, as estimates of demographic history, runs of homozygosity, and genetic load (to name a few) are directly dependent on the ability to accurately and reliably call heterozygous alleles. In this study, Prasad et al. (2021) are able to effectively demonstrate that the divergence time between the reference genome species and the target species substantially

FIGURE 1 Prasad et al. (2021) use the rowi kiwi and beluga whale, pictured here, to assess the impact of phylogenetic divergence and quality of reference genome on a suite of down stream analyses

impacts estimates of diversity. Heterozygosity was shown to deviate as divergence increased, raising concerns for previous studies that have used nonconspecific species for their alignments, particularly those that have used the same genome to compare species with variable phylogenetic distances from the reference. Moreover, the study revealed additional discrepancies in heterozygosity estimates when using different analysis tools and parameters for genotyping.

Previous studies have demonstrated that mapping bias between the species of interest and a nonconspecific reference genome leads to an increase in homozygous calls (Brandt et al., 2015). Prasad et al. (2021) demonstrate that heterozygosity can also be overestimated using divergent reference genomes, yet when they included parameters to adjust for multiple mismatches heterozygosity was underestimated. These results indicate that heterozygosity estimates are directly dependent on not only the software implemented for genotype calling, but also the specific parameters used. This outcome, in conjunction with the observation that increased fragmentation of the genome also impacts heterozygosity estimates, should serve as a cautionary tale to the many recent studies which attempt to compare heterozygosity from various species with different reference genome qualities, those that make heterozygosity estimates using different data types, and those that compare results obtained from different analysis methods.

While the authors in this study were unable to resolve any ROHs in the beluga whale genome, their results suggest that phylogenetic divergence plays an integral role in the outcomes of these commonly employed analyses when using the rowi kiwi genome. The ROH analysis on the rowi kiwi data followed the expected trend, showing that more contiguous assemblies were more reliable at capturing ROH segments, although noting that there are potential issues with the conspecific assembly, suggesting further investigation is warranted. Contrastingly, cross-species scaffolded (CSS) assemblies, and the use of the closest nonconspecific assembly captured at least some proportion of the genome in ROH, but this was substantially lower than the conspecific assembly, and any further distance between the rowi kiwi and the reference genome species resulted in no ROHs being detected.

Prasad et al. (2021) also investigated the impact of divergence and reference genome fragmentation on the pairwise sequential markovian coalescence (PSMC) method of demographic history modelling. This method was originally introduced by Li and Durbin (2011) to estimate population size changes over evolutionary time using coalescence in a single diploid genome. Like many other downstream analyses, the PSMC method is directly reliant on the ability to accurately call heterozygous sites within alignments, with long stretches of homozygous sites indicative of more recent coalescence events. This single genome strategy is theoretically ideal for species of high conservation concern, where data generation is limited and difficult to acquire. However, despite suggestions from the original author that alternative software may be more suitable in many circumstances (Li, 2014), PSMC has continued to dominate studies of demographic history within the literature. In this study, for both the beluga whale and the rowi kiwi, reference genomes that were more divergent from the target species showed similar trajectories to demographic results generated with conspecific references, but effective population size estimates were slightly lower. The authors suggest that using nonconspecific assemblies with a divergence over a 3% threshold are less reliable.

The results here, like others before them (Beichman et al., 2018) show an inability to reliably estimate demographic history. Although methods such as PSMC provide us with a general idea of what a species' demographic history looks like, this analysis is fraught with additional confounding factors such as population structure, the need for accurate mutation and recombination rate estimates, and fluctuating or unknown generation times, which are often ignored by the studies employing them. While this does not mean that such analyses are obsolete—and are sometimes the best that we can do with the data we have—we must be measured in the conclusions we draw from them.

While this paper raises concerns about the applicability of non-conspecific reference genomes for downstream analyses, it does suggest an alternative strategy, which the authors describe as "cross-species scaffolding". Cross-species scaffolding involves creating a "hybrid genome" using high-coverage short read data to create a de novo assembly, then subsequently generating in silico mate-pair libraries from the genome of a closely related species. Other strategies such as iterative mapping (Sarver et al., 2017) are also viable alternatives to de novo genome assembly, though with an increasing number of genomes being developed, this may become a nonissue in the near future (Rhie et al., 2021). The paper shows that CSS genomes generate demographic history and diversity estimates more concordant with the conspecific reference when compared to highly divergent genomes, an important note if there is no closely related genome or if the reference genome is highly fragmented. However, the species analysed here are members of groups (mammals and birds) that, for the most part, have comparatively stable karyotypes within clades and for which there often exist closely related species with minimal genetic divergence. Results from other groups, such as reptiles, insects, or fish, are unlikely to perform as well with the alternatives suggested here (cross-species scaffolding), since large differences in karyotype will strongly impact the utility of in silico mate-pair libraries on a cross-species scaffolded assembly. Using CSS the authors also show that there may be an issue with the published reference genome of the rowi kiwi, despite this being the only genome for which long ROH were detectable. However, the cause of these results and the utility of CSS for detecting genome misassemblies are unclear.

Although the main objective of this study was to evaluate the role of reference genome choice in population genetic analyses, it also highlights an issue which should be of greater concern to the non-model genetics community, specifically, the impact of software choice on genotype calling (specifically heterozygous sites) and therefore all downstream analyses that are reliant on the accuracy of those calls. This is particularly salient for recent meta-analyses which seek to compare heterozygosity and ROH data without accounting for the data source or the software used. It is tedious and time consuming for each paper to re-estimate these parameters with their own pipelines for the purpose of comparison, as such, a more thorough benchmarking of methods is clearly needed to allow for reliable comparisons between species. Although only one ROH calling software was used for comparisons here, this is a prime subject for follow-up investigations.

With next-generation sequencing data becoming more abundant than ever, meta-analyses will undoubtedly give us unprecedented insight into the patterns of diversity within and between species. However, as shown here, the choice of tool and the quality of the data available can have substantial impacts on the metrics of interest. It is clear that there is a need for additional studies which use real data to examine the biases introduced by these tools.

ACKNOWLEDGEMENTS

The authors thank J. Kelley for the invitation to provide this perspective piece, and the reviewer who provided important clarification and comments.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed.

ORCID

Blair P. Bentley https://orcid.org/0000-0002-9606-6770
Ellie E. Armstrong https://orcid.org/0000-0001-7107-6318

REFERENCES

Beichman, A. C., Huerta-Sanchez, E., & Lohmueller, K. E. (2018). Using genomic data to infer historic population dynamics of nonmodel organisms. *Annual Review of Ecology, Evolution, and Systematics*, *49*, 433–456.

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., & Meyer, D. (2015). Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project Phase I Data. *G3: Genes, Genomes, Genetics*, *5*(5), 931–941.

Li, H. (2014). *Alternatives to PSMC*. https://lh3.github.io/2014/07/20/alternatives-to-psmc

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496.

Prasad, A., Lorenzen, E. D., & Westbury, M. V. (2021) Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Molecular Ecology Resources*, *22*, 45–55. https://doi.org/10.1111/1755-0998.13457

Rhie, A., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737–746.

Sarver, B. A. J., et al. (2017). Phylogenomic insights into mouse evolution using a pseudoreference approach. *Genome Biology and Evolution*, *9*(3), 726–739.

How to cite this article: Bentley, B. P., & Armstrong, E. E. (2022). Good from far, but far from good: The impact of a reference genome on evolutionary inference. *Molecular Ecology Resources*, 22, 12–14. https://doi.org/10.1111/1755-0998.13531