Macrogenetic studies must not ignore limitations of genetic markers and scale

Ivan Paz-Vinas[1,2], Evelyn L Jensen[3], Laura D Bertola[4], Martin F Breed[5], Brian K Hand[6], Margaret E Hunter[7], Francine Kershaw[8], Deborah M Leigh[9], Gordon Luikart[6], Joachim Mergeay[10,11], Joshua M Miller[12], Charles B Van Rees[6], Gernot Segelbacher[13], and Sean Hoban[14]

[1]Institut de Recherche pour le Développement, Laboratoire Evolution & Diversité Biologique, UMR 5174, Centre National pour la Recherche Scientifique, Université de Toulouse, UPS, CNRS
[2]Laboratoire Ecologie Fonctionnelle et Environnement, UMR 5245, Université de Toulouse, UPS, CNRS, INP
[3]Department of Ecology and Evolutionary Biology, Yale University
[4]City College of New York
[5]College of Science and Engineering, Flinders University
[6]Flathead Lake Biological Station
[7]U.S. Geological Survey, Wetland and Aquatic Research Center
[8]Natural Resources Defense Council
[9]WSL Swiss Federal Research Institute
[10]Research Institute for Nature and Forest
[11]Aquatic Ecology, Evolution and Conservation, KULeuven
[12]Department of Biological Sciences, University of Alberta
[13]University Freiburg
[14]The Morton Arboretum

February 22, 2021

Abstract

Millette *et al.* (Ecology Letters, 2020, 23:55-67) reported no consistent worldwide anthropogenic effects on animal genetic diversity using repurposed mitochondrial sequences. We describe limitations to this study, some of which are common to other macrogenetic studies, that may lead to misinterpretations and unintended consequences for conservation.

Hosted file

`Suppl_table_S1_Appendix_A1.docx` available at https://authorea.com/users/395915/articles/509937-macrogenetic-studies-must-not-ignore-limitations-of-genetic-markers-and-scale

Macro-genetic studies must not ignore limitations of genetic markers and scale

Ivan PAZ-VINAS[1,2*+++], Evelyn L. JENSEN[3*++], Laura D. BERTOLA[4*ˆ], Martin F. BREED[5*ˆ], Brian K. HAND[6*], Margaret E. HUNTER[7*ˆ], Francine KERSHAW[8*], Deborah M. LEIGH[9*], Gordon LUIKART[6*ˆ], Joachim MERGEAY[10,11*ˆ], Joshua M. MILLER[12*], Charles B. van REES[6§], Gernot SEGELBACHER[13*ˆ], Sean HOBAN[14*ˆ]

1 Laboratoire Evolution & Diversité Biologique, Centre National pour la Recherche Scientifique, Institut de Recherche pour le Développement, Université de Toulouse, UPS, CNRS, IRD, UMR 5174, 118 route de Narbonne, Toulouse, 31062, France

2 Laboratoire Ecologie Fonctionnelle et Environnement, Université de Toulouse, UPS, CNRS, INP, UMR 5245, 118 route de Narbonne, Toulouse, 31062, France

3 Department of Ecology and Evolutionary Biology, Yale University, 21 Sachem St, New Haven, CT, 06520, USA

4 City College of New York, 160 Convent Ave., New York, NY, 10031, USA

5 College of Science and Engineering, Flinders University, Bedford Park, SA 5042, Australia

6 Flathead Lake Biological Station, 32125 Bio Station Ln, Polson, MT, 59860, USA

7 U.S. Geological Survey, Wetland and Aquatic Research Center, 7920 NW 71st St, Gainesville, FL,

32653, USA

8 Natural Resources Defense Council, 40 West 20th Street, New York, NY, USA

9 WSL Swiss Federal Research Institute, Zürcherstrasse 111, 8903 Birmensdorf, Switzerland

10 Research Institute for Nature and Forest, Gaverstraat 4, 9500 Geraardsbergen, Belgium

11 Aquatic Ecology, Evolution and Conservation, KULeuven, Charles Deberiotstraat 32, box 2439, 3000 Leuven, Belgium

12 Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

13 Chair of Wildlife Ecology and Management, University Freiburg, Tennenbacher Str. 4, D-79106

Freiburg, Germany

14 Center for Tree Science, The Morton Arboretum, 4100 Illinois Rt 53, Lisle, 60532, USA

* Group on Earth Observation Biodiversity Observation Network (GEO BON) Genetic Composition Working Group

ˆ Conservation Genetics Specialist Group, International Union for Conservation of Nature (IUCN), 1196 Gland, Switzerland

§ Group on Earth Observation Biodiversity Observation Network (GEO BON) Species Populations and Freshwater BONs

++ **These authors contributed equally to the work**

+ **Corresponding author**:

Ivan PAZ-VINAS

Laboratoire Evolution et Diversite Biologique (EDB UMR 5174), CNRS, IRD, UPS, Universite de Toulouse, 118 route de Narbonne, 31062 Toulouse, France.

E-mail: ivanpaz23@gmail.com

Phone number: +33 5 34 32 39 46

Authors email addresses and ORCID IDs:

Ivan Paz-Vinas: ivanpaz23@gmail.com, ORCID ID: 0000-0002-0043-9289

Evelyn L. Jensen: evelyn.jensen@yale.edu, ORCID ID: 0000-0002-1819-3117

Laura D. Bertola: laura.bertola@gmail.com, ORCID: 0000-0002-3445-0355

Martin F. Breed: martin.breed@flinders.edu.au, ORCID ID: 0000-0001-7810-9696

Brian K. Hand: brian.hand@umontana.edu, ORCID ID: 0000-0003-1152-665X

Margaret E. Hunter: mhunter@usgs.gov, ORCID ID: 0000-0002-4760-9302

Francine Kershaw: fkershaw@nrdc.org

Deborah M. Leigh: deborahmleigh.research@gmail.com, ORCID ID: 0000-0003-3902-2568

Gordon Luikart: gordon.luikart@mso.umt.edu, ORCID ID: 0000-00018697-0582

Joachim Mergeay: joachim.mergeay@inbo.be, ORCID ID: 0000-0002-6504-0551

Joshua M. Miller: jmm1@ualberta.ca, ORCID ID: 000-0002-4019-7675

Charles B. van Rees: charles.vanrees@flbs.umt.edu

Gernot Segelbacher: gernot.segelbacher@wildlife.uni-freiburg.de, ORCID ID: 0000-0002-8024-7008

Sean Hoban: shoban@mortonarb.org

Statement of authorship: All authors contributed to the brainstorming of ideas and to the development of the manuscript. IPV and ELJ wrote the drafts of the manuscript, and all authors contributed substantially to revisions. IPV made the figures and GIS analyses, and ELJ conducted the literature survey. SH initiated the project, and IPV and SH supervised the project.

Data availability statement: No new data has been used in this article. Mitochondrial COI sequence data for fish time-series were made available online by Millette *et al.* (2020) and can be found here: https://doi.org/10.5061/dryad.29rt7n0.

Running title: Markers and scale issues must not be ignored

Keywords (up to 10): macro-genetics, population genetics, macro-ecology, genetic data archiving, genetic diversity patterns, conservation, Genbank, anthropogenic impacts, COI, mitochondrial sequences

Type of article: Technical Comment (addressed article: Millette *et al.* Ecology Letters, 2020, 23:55-67)

Number of words in the abstract: 46

Number of words in the main text: 821

Number of references: 10

Number of figures and tables: 1 Figure

ABSTRACT

Millette *et al.* (Ecology Letters, 2020, 23:55-67) reported no consistent worldwide anthropogenic effects on animal genetic diversity using repurposed mitochondrial sequences. We describe limitations to this study, some of which are common to other macro-genetic studies, that may lead to misinterpretations and unintended consequences for conservation.

INTRODUCTION

Macro-ecology and conservation biology now include "macro-genetic" studies that repurpose genetic data from public databases (e.g. Manel et al. 2020) to explore patterns and drivers of intraspecific genetic diversity (IGD) for multiple taxa across large spatial and/or temporal scales (Blanchet et al. 2017). The macro-genetic study by Millette et al. (2020) sought to elucidate relationships between human population density and recent land-use change on animal IGD, but we explain here that technical limitations of the approach may prevent the detection of anthropogenic effects on IGD. Although the authors acknowledged several constraints, and presented their results with more nuance than previous macro-genetic studies (e.g. Miraldo et al. 2016), issues remain that cannot be resolved or adequately addressed by tempering the interpretation.

Are COI sequences the most appropriate data?

Millette *et al* . (2020) used 175,247 mitochondrial cytochrome c oxidase subunit 1 (COI) sequences from 17,082 vertebrate species deposited in BOLD and GenBank. COI became a popular marker for species molecular barcoding due to its low within-species and high between-species variation. However, these characteristics make COI inappropriate for measuring IGD, as Millette *et al* . acknowledge, in addition to potential discordance with nuclear variation. Despite these well-known issues, the large availability of COI sequences has, nevertheless, resulted in its continued use to represent IGD in macro-genetic studies (e.g. Miraldo et al. 2016; Millette et al. 2020; Theodoridis et al. 2020; Manel et al. 2020).

Even if COI could provide a useful IGD measure, we have identified a subtle -yet serious- constraint of repurposing publicly-available data due to inconsistent archiving practices. Specifically, it is common for only unique or newly-discovered haplotypes to be deposited in repositories, and not the study's full dataset. As an example, we screened 18 Molecular Ecology issues (Table S1): of 40 papers that deposited mitochondrial sequences in GenBank, 22 deposited all sequences generated, while 18 deposited only novel haplotypes (sequences detected for the first time) or exemplars of each haplotype. Therefore, deposited data may more accurately represent haplotype accumulation curves across space and time; databases consequently do not allow comparable snapshots of genetic diversity at different times. This bias compromises attempts to quantify temporal trends in IGD using GenBank, as done in Millette et al. (2020), and is a potential issue in many spatial macro-genetic studies. Macro-genetic studies should extract metadata regarding sample sizes and complete haplotype (or allele) frequencies from the original manuscripts (as done by Lawrence et al. 2019) to avoid bias from inconsistently-archived data.

Are the spatial and temporal scales biologically meaningful?

Millette *et al* . (2020) examined IGD temporal trends across 909 animal species where COI sequences were available for [?]4 years. Sequences were grouped across [?]1,000 km to avoid "*conflating spatial and temporal effects"* . This scale far exceeds the dispersal capabilities of many included species, the scale of habitat change affecting them, and thus the scale at which population genetic processes influencing IGD operate. Additionally, the clustering algorithm used can 'daisy chain' locations together so that sites >1,000 km apart are grouped together (Appendix A1). Grouping sequences into biologically-implausible "populations" likely obscures any anthropogenic effects on IGD, especially when combined with the small sample sizes (<10 sequences/year for 77% of time-series overall) and the large number of locations sampled yearly (e.g. mean of 3 locations sampled *per* year for fish; Appendix A1). We reexamined 104 *Inland and Coastal Bony Fish* time-series from Millette et al. (2020), and found that the sequences included were sampled in inland freshwaters for most time-series (96/104), with sequences from multiple demographically-independent locations (i.e. from disconnected drainage basins; median 3 [range 0-15], see Figure 1 and Appendix A1 for examples) erroneously pooled into "populations". By pooling sequences from independent locations, changes in IGD attributable to anthropogenic pressures would be lost in the noise, with uneven sampling across space and time compounding the issue.

Additionally, the median span of included time-series is only seven years overall, and represents an average of just 2.2 generations for fish (Appendix A1). Thus for many taxa, the data covers an insufficient time-span for most measurable evolutionary changes in IGD. Of course, such analysis also neglects any pre-1980s impacts. DNA from museums, herbaria and fossil archives are needed for this.

Conclusions

We support the goals of Millette et al. (2020) and recognize that some of the flaws outlined are not unique to their study, although their temporal focus presents novel issues. Combined, these constraints increase the risk of overinterpretation of macro-genetic studies' conclusions (e.g. no or no consistent anthropogenically-driven IGD changes), which could misinform important conservation policy decisions. Macro-geneticists must not continue to merely acknowledge such limitations and carry on with their studies regardless, especially when meta-analyses using appropriate molecular markers consistently show anthropogenically-driven changes in IGD (e.g. due to harvest, habitat loss and fragmentation; Aguilar et al. 2008; Pinsky and Palumbi, 2014; Schlaepfer et al. 2019; Gonzalez et al. 2020). Macro-geneticists must accurately study variables of interest by using the most appropriate data rather than the most abundant data.

References:

Aguilar R, Quesada M, Ashworth L, Herrerias-Diego Y, Lobo J. (2008). Genetic consequences of habitat fragmentation in plant populations: susceptible signals in plant traits and methodological approaches. *Mol Ecol.* 17, 5177-5188.

Blanchet, S., Prunier, J.G. & De Kort, H. (2017). Time to go bigger: Emerging patterns in macrogenetics. *Trends Genet.*, 33, 579–580.

Denys, G.P.J., Geiger, M., Persat, H., Keith, P. & Dettai, A. (2015). Invalidity of *Gasterosteus gymnurus* (Cuvier, 1829) (Actinopterygii, Gasterosteidae) according to integrative taxonomy. *Cybium*, 39, 37–45.

Gonzalez, A.V., Gomez-Silva, V., Ramirez, M.J. and Fonturbel, F.E. (2020). Meta-analysis of the differential effects of habitat fragmentation and degradation on plant genetic diversity. *Cons Biol.*, 34, 711-720.

Lawrence, E.R., Benavente, J.N., Matte, J.-M., Marin, K., Wells, Z.R.R., Bernos, T.A., *et al.* (2019). Geo-referenced population-specific microsatellite data across American continents, the MacroPopGen Database. *Sci. Data*, 6, 14

Manel, S., Guerin, P.-E., Mouillot, D., Blanchet, S., Velez, L., Albouy, C., *et al.* (2020). Global determinants of freshwater and marine fish genetic diversity. *Nat. Commun.*, 11, 692

Millette, K.L., Fugere, V., Debyser, C., Greiner, A., Chain, F.J.J. & Gonzalez, A. (2020). No consistent effects of humans on animal genetic diversity worldwide. Ecol. Lett., 23, 55–67.

Pinsky, M.L. and Palumbi, S.R. (2014). Meta-analysis reveals lower genetic diversity in overfished populations. *Mol Ecol.*, 23, 29-39.

Schlaepfer, D. R., Braschler, B., Rusterholz, H.-P., and Baur, B. (2018). Genetic effects of anthropogenic habitat fragmentation on remnant animal and plant populations: a meta-analysis. *Ecosphere* 9, e02488.

Theodoridis, S., Fordham, D.A., Brown, S.C., Li, S., Rahbek, C. & Nogues-Bravo, D. (2020). Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. *Nat. Commun.*, 11, 2557

Acknowledgements: We acknowledge the support of the GEO BON Genetic Composition Working Group in the development of this manuscript. We also thank Arne Mooers and the reviewers for their comments. I.P-V works in a laboratory supported by the 'Laboratoire d'Excellence' (LABEX) entitled TULIP (ANR-10-LABX-41).

Competing interests: The authors declare no competing interests.

Disclaimer: Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Prepublication disclaimer: This draft manuscript is distributed solely for purposes of scientific peer review. Its content is deliberative and predecisional, so it must not be disclosed or released by reviewers. Because the manuscript has not yet been approved for publication by the U.S. Geological Survey (USGS), it does not represent any official USGS finding or policy.

FIGURE 1: Map showing the grouping of sequences from the fish species *Gasterosteus gymnurus* (Cuvier, 1829; a junior synonym of*Gasterosteus aculeatus* , Linnaeus, 1758) into a single "population" to measure change in intraspecific genetic diversity. This is one of the 909 time-series datasets in Millette et al. (2020). This time-series consists of 53 mitochondrial cytochrome c oxidase subunit 1 (COI) sequences collected at 24 different sampling sites (colored dots). The sampling sites are all within the 1,000 km distance threshold set by Millette et al. for being pooled into a population, despite being located in nine watersheds from six of the major hydrographical regions in France. Sample sizes are highly uneven across the time series, with just three sequences from a single site each in 2004, 2007 and 2009, and then 44 sequences from 21 sites in 2013. Millette et al. (2020) analyzed the trend in nucleotide diversity across these temporal points, despite the 2013 sample consisting of sequences pooled across many different regions, while the other years had a single site, in different regions.