I. Title page

The weak driver conundrum: data archiving and biological phenomena impact macrogenetic findings

Authors: Ivo Colmonero-Costeira[1] and Deborah M. Leigh*[1, 2,]

*corresponding author: deborah.leigh@senckenberg.de

Ivo Colmonero-Costeira

ivo.costeira@senckenberg.de

[1] Senckenberg Research Institute, Frankfurt, Germany

ORCID: 0000-0001-9914-0713

Deborah M. Leigh

deborah.leigh@senckenberg.de

[1] Senckenberg Research Institute, Frankfurt, Germany

[2] Institute of Ecology, Evolution, and Diversity, Faculty of Biosciences, Goethe University Frankfurt, Frankfurt, Germany

ORCID: 0000-0003-3902-2568

Keywords:

Open-data bias; Haplotype; Genetic Variation; Nucleotide diversity.

II. Abstract

Macrogenetics seeks to identify the global drivers and patterns in intraspecific genetic diversity, yet many reported patterns are weak or inconsistent. To achieve multispecies global inference, many macrogenetic studies leverage open sequencing data that can suffer from archiving biases. It remains unclear if macrogenetic inconsistencies are innate genetic phenomena, or are the product of open data limitations. Using three widely available genetic markers from the mitochondrion (*cytb*, *co1*) and nuclear (TLR4) genomes archived as haplotypes, here we demonstrate archiving biases are powerful enough to distort nucleotide diversity estimates and patterns. Distortion is worsened in analysis using geographic gridded cells, where archiving efforts both outweigh and interact with ecological predictors. Nevertheless, previously described incongruences in drivers of nuclear and mitochondrial diversity appear to be biologically meaningful, indicating some inconsistencies are innate to genetic data.

III. Main text

Introduction

Macrogenetics seeks to explain the ecological and evolutionary drivers that generate, maintain or degrade genetic diversity by aggregating and re-analysing previously collected genetic data at broad taxonomic and spatial-temporal scales (Blanchet *et al.* 2017; Leigh *et al.* 2021). A key goal of macrogenetic studies is to describe and map global genetic diversity patterns to improve biodiversity understanding (Csilléry *et al.* 2025; De Kort *et al.* 2021; Leigh *et al.* 2021; Miraldo *et al.* 2016; Theodoridis *et al.* 2020; Yiming *et al.* 2021).

Despite the potential for macrogenetics to address gaps in biodiversity understanding, many published multispecies genetic relationships are weak and sensitive to the underlying dataset used (discussed in Leigh *et al.* 2021). Most notably, there have been several investigations into the hypothesised latitudinal gradient in intraspecific genetic diversity, which have revealed dramatically different results depending on the genetic marker chosen. Specifically, while mitochondrial markers can show a negative trend in genetic diversity with increasing latitude (Miraldo *et al.* 2016; Pelletier & Carstens 2018) nuclear markers often show a lack of or even opposing trends (De Kort *et al.* 2021; Lawrence & Fraser 2020; Schmidt *et al.* 2022). Further inconsistencies can even be seen within the same study, for example, a recently identified relationship between genetic diversity of habitat forming species (e.g. corals and seagrass) and ecoregion species diversity was dependent on data included (Figuerola-Ferrando *et al.* 2023).

It remains unclear if weak or inconsistent macrogenetic trends are biologically meaningful, perhaps due to the inescapable random nature of genetic drift or complexity of driver interactions, or if they are the product of limitations innate to open data (hereafter the 'weak driver conundrum'). Regardless, it is clear that the limitation must be explored head on to ensure the accurate interpretation of results and thorough evolutionary understanding.

Not all macrogenetic studies are affected equally by innate limitations of open data. Those that extract, curate and re-analyze genetic haplotype data archived in public repositories ("Class III" see Leigh et al. 2021) are particularly vulnerable to inconsistencies in data archiving practices. As noted in Paz-Vinas et al. (2021) approximately half of a selected number of mitochondrial haplotype

sequence datasets deposited in GenBank are unique exemplars, meaning researchers have not comprehensively archived sequence data for all sampled individuals. This practice removes information on the relative frequency of the haplotypes, which is essential for estimating frequency-based estimates of genetic diversity, such as nucleotide diversity. Nucleotide diversity or 'pi' (π) is a commonly used metric in macrogenetic studies (Manel *et al.* 2020; Miraldo *et al.* 2016; Theodoridis *et al.* 2020; Yiming *et al.* 2021) and as a result, archiving comprehensiveness could distort estimates of genetic diversity and obscure relationships with ecological and evolutionary drivers. The role this distortion plays in the "weak driver conundrum" remains unclear, and effective correction methods have not been unexplored.

Here, we aimed at testing how variation in sequence archiving practices influences macrogenetic patterns using three widely available haplotype sequence-based datasets focusing on mammalian *cytb*, mammalian *co1* sequences as well as avian TLR4 sequences. Mitochondrial haplotypes (e.g., *cytb* and *co1*) are commonly repurposed in sequence-based macrogenetic research due to their availability and easy access. However, focusing on the mitochondrion alone may confound the impact of archiving practices with characteristics innate to the mitochondrion. Thus, we also focused on nuclear loci archived as haplotypes, specifically toll-like receptor genes, TLRs, which are often similarly sequenced using degenerate primers (Grueber et al. 2012) and deposited as haplotype sequence data.

To begin to disentangle the weak driver conundrum, we examined the extent archiving biases altered or changed species-level and intraspecific patterns, and

whether downstream data-processing approaches (e.g., haplotype collapsing) and sampling biases (e.g., rarefaction, Leight et al., 2021) interact with these effects.

Material and methods

Data retrieval

The mitochondrial sequencing database was the publicly available mammalian *cytb* and *co1* datasets curated by Theodoridis *et al.,* (2020). The dataset was downloaded from GitHub (https://github.com/spyrostheodoridis/Genetic-geography-of-terrestrial-mammals) on the 28th of July 2025. The original *cytb* dataset comprised of 24,395 geo-referenced sequences across 1,690 species and the *co1* dataset 22,570 sequences for 1,153 species. This dataset was filtered to remove species represented by a single sequence. We also re-assigned the aligned sequence data to the authors' 385.9km×385.9km global area grid cells.

The TLR4 database consisted of all available sequences from the NCBI GeneBank (https://www.ncbi.nlm.nih.gov). Sequences were downloaded with the Entrez Utilities Unix Command Line implemented in *Biopython* (Cock et al., 2009) on 2nd of April 2025. We restricted our queries to tetrapod species and excluded human sequences "(((TLR4*[Title]) OR (toll-like receptor 4*[Title])) AND tetrapoda[Organism]) NOT homo[Organism]". A total of 3,156 GeneBank records were retrieved. We retained only DNA sequences, and filtered the dataset for TLR4 interactor genes and pseudogenes, and sequences with less than 500 bp of length. We restricted our dataset to avian species as this was the most data-rich taxonomic group (60.30% of total number of DNA sequences, n = 1,364). Sequences from domestic

116 species including the domestic chicken *Gallus gallus* and the common pigeon
117 *Columba livia* were removed. After filtering, the avian TLR4 dataset consisted of
118 1,088 sequences over 62 species. Species-specific alignments were generated
119 using MAFFT sequence alignment algorithm for Unix (Katoh et al., 2002).

120

121 **Calculating genetic diversity**

122 Genetic diversity was defined as nucleotide diversity (π), which is the number
123 of nucleotide differences in a pairwise sequence comparison. Genetic diversity was
124 estimated at the two hierarchical commonly found in macrogenetics, (i) species level
125 averages which are typically used to explore taxon-related patterns (e.g., Lowe et
126 al. 2018; Stoffel et al. 2018); and, (ii) intraspecific spatially gridded scale, which
127 follows standard macrogenetic approaches that map fine-scale geographic
128 variation in genetic diversity (Miraldo *et al.* 2016; Theodoridis *et al.* 2020; Yiming *et*
129 *al.* 2021).

130 Species level nucleotide diversity was estimated as the average number of
131 differences across all possible pairwise sequence comparisons in a species-specific
132 sequence alignment (Nei and Li, 1979). For the intraspecific dataset, we obtained
133 estimates of genetic diversity for each grid cell by averaging the nucleotide diversity
134 across all species within it (following Theodoridis et al., 2020). Species only
135 represented by a single sequence within a grid cell were excluded.

136

137 **Data filtering for cryptic taxa or erroneously archived sequences**

For all datasets prior to the downstream analyses, we filtered and removed species for which average nucleotide diversity exceeded 0.05. These species contained highly divergent sequences which by far exceed some of the commonly applied intraspecific genetic divergence threshold (e.g., > 2 % for BOLD BIN system; Ratnasingham & Hebert 2013), and likely represent cryptic taxa or erroneously archived sequences, known to be present in open-data repositories (van der Burg and Vieites, 2023).

Standardizing by haplotype collapsing

To explore the interplay between data handling and filtering choices and archiving practices on macrogenetic signals, we collapsed all comprehensively archived sequence data into only unique haplotypes, to ensure genetic diversity estimates were equally biased across species. Nucleotide positions containing missing data or alignment gaps were masked and not considered when extracting unique haplotypes.

Sampling effort bias

To explore the impact of sampling effort, we rarefied the dataset by bootstrapping sequences with replacement. To maintain maximum taxonomic coverage, we randomly chose two samples per species per interaction to estimate genetic diversity at both the species-level and for each grid cell (intraspecific level). Because estimate precision may increase with sample size, we increased the number of rarified sequences up to the median number of samples per species

(*cytb* median = 6; *co1* median = 8, TLR4 median = 5). This represents the point where 50% of the taxonomic coverage of the datasets was lost. Estimates of genetic diversity were then taken based on 999 random bootstrapping replicates.

Predictor variables of genetic diversity

To understand the effect of the archiving practices on the relationship between genetic diversity and key drivers, we first searched for bio-climatic predictors with repeatable or important cross-study relationships. This search identified absolute Latitude (i.e., distance from the equator identified in Miraldo et al. 2016; Pelletier & Carstens 2018; Yiming et al. 2021) and climatic stability since the Last Glacial Maximum (identified in (De Kort *et al.* 2021; Theodoridis *et al.* 2020) as key drivers with potentially inconsistent relationships.

Latitude was taken as either the geographic centroid of all the individual geo-referenced sequences (curated by Theodoridis et al. 2020), or the geographic centroid of the 385.9km×385.9km grid cell. When collection coordinates were missing from the GeneBank records, we georeferenced the sequences using GeoNames API (https://www.geonames.org). The TLR4 dataset was particularly poor in coordinate information and thus latitude was only used at the species level.

Climatic stability was calculated by estimating the standardized differences between the current and Last Glacial Maximum mean annual temperatures obtained from CHELSA (as used in De Kort et al., 2021; Karger et al., 2017)

Finally, archiving effort or practice was coded as a dichotomous categorical variable, where 0 = sequence data were comprised of unique haplotypes; and 1 =

184 comprehensive sequence data. Data was considered to be unique haplotypes when
185 the total number of samples and haplotypes were equal, and comprehensive if the
186 number of unique haplotypes was lower than the number of samples. Archiving
187 effort was also estimated at the intraspecific level for each area grid cell, as the
188 proportion of species containing comprehensive sequence data.

189

190 **Multi-model inference of predictors of genetic diversity**

191 Due to the high skewness of nucleotide diversity estimates towards zero, we
192 implemented a general linear mixed model approach (glmm, R package *glmmTMB*
193 v1.1.12, McGillycuddy et al., 2025) with a Gamma distribution and a log link
194 function. A small constant was added (1^{-4}) as a Gamma distribution is always
195 positive. For all models, we also weighted the residuals by the sample size with the
196 assumption that higher sample sizes generate more accurate estimates of genetic
197 diversity (De Kort et al., 2021):

198 $$weights = 1/\sqrt{log\ (sample\ size)}$$
199

200 All possible predictor combinations of the full models (i.e., Genetic Diversity ~
201 |Latitude| + Climatic Stability + Archiving Effort) were tested using multi-model
202 inference based on the Akaike's information criterion (AICc) using R package
203 *MuMIn* v1.48.11 (Barton, 2025). To explore a potential relationship between
204 archiving efforts and sampling location, the |Latitude| * Archiving Effort interaction
205 was included in all models. We retained models with deltaAICc < 4 compared to the
206 top performing model (De Kort et al., 2021). Because we were interested in
207 estimating the relative importance of each of the predictors, the retained models

were averaged and used to estimate the sum of Akaike weights (SW_+). We considered predictor variables with $SW \geq 0.5$ and a p-value < 0.05 to be significant. The model assumptions and model fits biases were evaluated visually using R package *performance* v0.15.1 (Lüdecke et al., 2021). The co-linearity between predictors was evaluated using the Variance Inflation Factor using R package *car* v3.1-3 (Fox and Weisberg, 2019). All R packages were run in R environment v4.5.1 (R Core Team, 2025) coupled with RStudio v2025.05.1+513 (Posit team, 2025).

Results

Dataset structure

After filtering, the mammalian *cytb* dataset contained 23,394 sequences across 919 species with a per species average of 25.46 (± 74.28 SD) sequences and approximately 62 % of data comprehensively archived. The mammalian *co1* dataset contained the 22,003 sequences over 729 species, with a species average of 30.18 (± 75.12 SD) sequences and approximately 82 % comprehensively archived. The avian TLR4 dataset contained 1,088 sequences over 62 species, per species average of 17.55 (± 47.89 SD) sequences and only approximately 11 % comprehensively archived.

The impact of archiving is strongest at the intraspecific level

At the species level a relationship with latitude was identified in both the *cytb* and *co1*, with genetic diversity decreasing with distance from the equator (complete dataset - *cytb* latitude: SW = 1.00, p <0.01, slope = -7.59E-03; *co1* latitude: SW =

231 1.00, $p < 0.05$, slope = -9.66E-03; Fig. 1A, Table 1). Climatic stability and archiving
232 effort were not significant predictors for either mitochondrial marker. For TLR4 in
233 contrast, no variables were significant though climatic stability was marginally
234 informative (complete dataset - TLR4 climatic stability: SW = 0.50, $p = 0.13$, slope = -
235 1.35E+01; Fig. 1A, Table 1).

236 At the intraspecific gridded level, archiving effort became the most important
237 predictor of mitochondrial genetic diversity (complete dataset - *cytb* archiving effort:
238 SW = 1.00, $p < 0.01$, slope = -1.04E+00; *co1* archiving effort: SW = 1.00, $p < 0.05$,
239 slope = -7.38E-01; Fig. 1A, Table 2). The higher the proportion of species within a
240 grid cell with comprehensive sequence data, the lower the estimated genetic
241 diversity (approximately 52% decrease in average genetic diversity in cells with
242 archiving effort of 1, i.e., containing only comprehensively archived sequences).
243 While absolute latitude remained an important predictor of *cytb* intraspecific genetic
244 diversity (complete dataset - *cytb* latitude: SW = 1.00, $p < 0.01$, slope = -9.48E-03), it
245 was no longer important for *co1* intraspecific diversity. Instead, the interaction
246 between latitude and archiving efforts was (complete dataset - *co1* latitude: SW =
247 1.00, $p = 0.18$, slope = -6.50E-03; latitude * archiving effort: SW = 0.86, $p < 0.05$,
248 slope = -1.67E-02), with the impact of archiving effort highest at higher latitudes (Fig.
249 1A, C, Table 2).

250

251 **Using only unique haplotypes does not remove all archiving biases**

252 At the species level, collapsing sequences to only unique haplotypes resulted
253 in substantial inflation of nucleotide diversity for species that were originally

254 comprehensively archived (Suppl. Fig. 1). Several species then exceeded the
255 commonly accepted upper limit for intraspecific diversity (> 0.05), even though they
256 had passed the initial filtering.

257 Despite data being more uniform in this dataset, archiving effort had a
258 significant effect on nucleotide diversity (unique haplotypes dataset - *cytb* archiving
259 effort: SW = 1.00, p <0.001, slope = -7.38E-01; *co1* archiving effort: SW = 1.00, p <
260 0.001, slope = 4.27E-01; Table 1). The set of bio-climatic predictors that were
261 significant in the complete dataset remained significant after collapsing with
262 reduction in effect size compared to the complete datasets (unique haplotype
263 dataset - *cytb* latitude: SW = 1.00, p <0.01, slope = -6.88E-03; *co1* latitude SW =
264 0.93, p < 0.05, slope = -7.92E-03; Table 1). For the TLR4 dataset, no variables were
265 significant, however, latitude replaced climatic stability as the marginally significant
266 variable (unique haplotype dataset - TLR4 latitude: SW = 0.51, p = 0.39, slope =
267 5.99E-03; TLR4 climatic stability: SW = 0.33, p = 0.12, slope = -1.35E+01; Table 1).

268 At the intraspecific level, when considering only unique haplotypes the impact
269 of archiving practices became non-significant for both *cytb* and *co1* markers (unique
270 haplotypes dataset - *cytb* archiving effort: SW = 0.74, p = 0.12, slope = -3.04E-01;
271 *co1* archiving effort: SW = 0.81, p = 0.27, slope = 3.07E-01; Table 1). The
272 relationship between intraspecific genetic diversity and latitude remained for *cytb* but
273 with lower effect size (unique haplotypes dataset - *cytb* latitude: SW = 1.00; p<0.001,
274 slope = -8.67E-03; Table 1). Similarly, for *co1*, the interaction between latitude and
275 archiving efforts weakened but remained significant (unique haplotypes dataset - *co1*
276 latitude * archiving effort: SW = 0.71, p < 0.05, slope = -1.36E-02).

277

Sampling and reporting biases likely impact macrogenetic driver identification

In the species level rarified dataset, estimates of genetic diversity were characterised by large coefficients of variation (25 – 700%) and substantial variability in species estimates of genetic diversity (Suppl. Figure 2).

Relative to the full datasets, rarefication to the mode of the number of samples (n = 2), resulted in little to no change to the significance of the predictors of genetic diversity, and generally resulted in decreased effect sizes with the exception of latitude for *co1* (rarefaction n = 2 dataset - *co1* latitude SW = 1.00, $p < 0.05$, slope = -1.11E-02; Table 1). For *cytb* and *co1*, archiving efforts remained a significant predictor for the rarefied (n = 2) datasets, in both species and intraspecific analyses (Table 1 and Table 2, respectively). For TLR4, we did not find significant changes to model composition in the rarefied (n = 2) data compared to the un-rarefied dataset. Across all genetic markers, rarefaction to higher sample sizes minimized the coefficient of variation in nucleotide diversity estimates (Suppl. Fig. 2). However, across the rarified datasets with increasing sample sizes, we detected highly unstable models with the relative importance, significance and effect sizes of all predictors varying depending on the dataset used (Table 1).

Discussion

Here we explored how inconsistent archiving practices of haplotype sequence data may impact global genetic diversity patterns and contribute to the weak driver conundrum in macrogenetics.

When analysing genetic diversity species level (i.e. without gridding data), we found that predictors and archiving effects were highly marker specific. While a relationship was found between mitochondrial genetic diversity and latitude when all data were considered (*cytb* and *co1,* as shown in Miraldo et al. 2016; Pelletier & Carstens 2018; Yiming et al. 2021), there was none for nuclear genetic diversity at TLR4. Despite re-identifying a previously published latitudinal clines in mitochondrial genetic diversity Miraldo et al. 2016; Pelletier & Carstens 2018), we note that these datasets were not consistent because they were composed of both unique haplotypes and comprehensive datasets. When collapsing the datasets to a uniform format (i.e. unique haplotypes) patterns changed for the mitochondrial markers, and archiving effort became a significant predictor of genetic diversity signalling an important bias that must be addressed.

At the intraspecific (gridded) scale, which is representative of the more common macrogenetic gridded scale, archiving practices was one of the strongest predictors of nucleotide diversity for both mitochondrial markers. We speculate that this is because the grid cell level captures local differences in data-archiving standards that amplify regional nucleotide diversity. This is nicely shown in West Africa where the archiving of unique haplotypes appears to be standard and there is subsequently high regional *co1* nucleotide diversity. This relationship is lost when collapsing the datasets to unique haplotypes, likely reflecting the removal of this local bias. However, such collapsing removes the within species focus central to Macrogenetics, signalling it is not a suitable correction for this bias.

Focusing on the macrogenetic relationship between genetic diversity and latitude, while *cytb* displayed a latitudinal cline in mitochondrial genetic diversity at

the intraspecific level, this relationship was not seen for *co1*. For *co1*, archiving comprehensiveness interacted with latitude and, for TLR4*,* there was no relationship with latitude. Despite the small sample size, the lack of a cline in nuclear genetic diversity follows our expectations based on previous nuclear studies, signalling that biological differences between the marker types may drive some of the inconsistent marogenetic patterns reported (Clark and Pinsky, 2024; Schmidt et al., 2022).

Based on our results, we cannot discount that a genuine relationship is present between mitochondrial genetic diversity and latitude. What drives this is unclear and while macrogenetic studies have only considered broad ecological theories, simple evolutionary explanations should not be overlooked. For example, the mitochondrion is involved in thermoregulation (Lajbner et al. 2018), which could have led to temperature-based selection and thus a mitochondrial genetic diversity cline with latitude (Balloux et al. 2009; Lajbner et al. 2018). Such temperature mediated clines have been detected in drosophila and human mitochondrion haplotypes (Balloux et al. 2009; Lajbner et al. 2018), supporting that this may be the root of the multispecies mitochondrial macrogenetic pattern.

A genuine cline in mitochondrial genetic diversity would also explain the interaction seen in *co1*, because nucleotide diversity is a frequency-based statistic any incomplete archiving would inflate genetic diversity and, this inflation is worse in low diversity populations (Nei & Li 1979). This could drive an amplified effect of archiving bias at higher latitudes because species genuinely have lower mitochondrial genetic diversity.

Looking closer into other biases that could contribute to macrogenetic inconsistencies, we found that when standardising the species and intraspecific

gridded datasets to unique haplotypes we unexpectedly detected several erroneous sequences within the data that had not been successfully filtered and exceeded accepted intraspecific diversity levels (Ratnasingham and Hebert, 2013). These are likely cryptic species or misidentified haplotypes which are known to be present in public databases (van der Burg and Vieites, 2023). Although such sequences have little influence in comprehensively archived datasets due to their low frequency, they will add noise to patterns detected and may contribute to the weak driver phenomenon by overall reduction of effect sizes. Whilst time consuming, future haplotype based macrogenetic studies will need to screen individual haplotypes to remove these, through initially collapsing all haplotypes to unique sequences or through haplotype networks or other biologically supported analyses.

Many important questions surrounding genetic diversity patterns remain and it is important that all potential data corrections and limitations are explored. Rarefaction, commonly used to test sampling bias, partially captures the random nature of “which” haplotypes are reported “where” and may interact with archiving biases in unexpected ways. Our rarefaction analyses showed that randomly sampling haplotypes generates incredibly high variance in estimated genetic diversity (25 – 700 % coefficient of variation), indicating reliable estimates of nucleotide diversity cannot be obtained from rarefied data. Rarefaction to higher sample sizes was accompanied by significant instability in the importance and weakening effect sizes of bio-climatic and artifactual macrogenetic predictors alike. This is in part due to reduced taxonomic coverage and subsequent reduction in statistical power, and further signals it cannot be universally used as a data correction. Indeed, previous mitochondrial studies (e.g., Millette et al. 2019) reported

372 that rarefaction analyses were not equally robust across taxonomic groups.
373 Worryingly this analysis suggests that the random nature of where a haplotype is first
374 reported in unique sequence datasets likely also contributes to unstable
375 macrogenetic findings and ultimately, the weak driver conundrum.

376 It is important to note that not all macrogenetic studies are equally impacted
377 by archiving bias and the issues discussed here. Macrogenetic analysis of nuclear
378 non-haplotypic markers, will be less sensitive to choices in archiving practices. When
379 archiving non-haplotype data (e.g. microsatellite genotypes) researchers upload the
380 complete set of multilocus genotypes. Consequently, these datasets inherently retain
381 individual frequencies. This fundamental difference in the data archiving standards,
382 coupled with real biological differences, likely also contributes to why macrogenetic
383 patterns between mitochondrial (nucleotide diversity, sequence-based) and nuclear
384 (heterozygosity, individual-based) markers are often inconsistent or even
385 contradictory (Leigh *et al.* 2021; Paz-Vinas *et al.* 2021). However, we note that such
386 nuclear datasets also have their own limitations arising from data archiving gaps and
387 poor metadata (Leigh et al., 2024).

388 To ensure we can comprehensively understand genetic diversity, frequency-
389 based diversity estimates from haplotypic data require careful treatment in future.
390 This includes filtering for non-comprehensive sequence datasets, estimating
391 archiving efforts for its inclusion as a random effect on multi-model inference
392 analyses and accessing the impacts of downstream data processing (e.g.,
393 standardization by haplotype collapsing or rarefication) on predictor effect sizes. We
394 stress that transparent reporting by data-generating authors of archiving
395 completeness, data-processing steps, and marker-specific limitations is also

essential (Leigh et al., 2024). We hope that this and similar methodological studies on limitations and challenges of macrogenetics may advance the field towards achieving the potential to support global biodiversity understanding and real-world conservation efforts.

Acknowledgments

All data analyses were conducted on the Delta Computing Cluster of the Senckenberg Gesellschaft für Naturforschung.

Data accessibility statement:

All sequence data, associated metadata and scripts employed in the study will be deposited GitHub, freely available in (https://github.com/deborahmleigh/Archiving-in-macrogenetics).

Statement of authorship:

Conceptualization: ICC and DML; data assembly and curation: ICC; data analyses and interpretation: ICC and DML; Writing: ICC and DML; Project supervision: DML

References

Balloux, F., Lawson Handley, L. J., Jombart, T., Liu, H., & Manica, A. (2009). Climate shaped the worldwide distribution of human mitochondrial DNA sequence

variation. *Proceedings of the Royal Society B: Biological Sciences*, 276, 3447–3455. https://doi.org/10.1098/RSPB.2009.0752.

Blanchet, S., Prunier, J.G. & De Kort, H. (2017). Time to Go Bigger: Emerging Patterns in Macrogenetics. *Trends in Genetics*, 33, 579–580. https://doi.org/10.1016/J.TIG.2017.06.007.

Clark, R. D., & Pinsky, M. L. (2024). Global patterns of nuclear and mitochondrial genetic diversity in marine fishes. Ecology and Evolution, 14, e11365. https://doi.org/10.1002/ece3.11365

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25, 1422–1423. https://doi.org/10.1093/BIOINFORMATICS/BTP163.

Csilléry, K., Yang, H., Hung, T.H., Rodríguez-Rodríguez, P., Miller, J., Mantovani, V., *et al.* (2025). GenDivRange: A global dataset of geo-referenced population genetic diversity across species ranges. *Scientific Data*, 12, 980. https://doi.org/10.1038/s41597-025-05303-2.

De Kort, H., Prunier, J. G., Ducatez, S., Honnay, O., Baguette, M., Stevens, V. M., & Blanchet, S. (2021). Life history, climate and biogeography interactively affect worldwide genetic diversity of plant and animal populations. *Nature Communications*, 12, 1–11. https://doi.org/10.1038/s41467-021-20958-2.

Figuerola-Ferrando, L., Barreiro, A., Montero-Serra, I., Pagès-Escolà, M., Garrabou, J., Linares, C., *et al.* (2023). Global patterns and drivers of genetic diversity

among marine habitat-clarckecies. *Global Ecology and Biogeography*, 32, 1218–1229. https://doi.org/10.1111/GEB.13685.

Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). SAGE Publications, Inc.

Grueber, C. E., Wallis, G. P., King, T. M., & Jamieson, I. G. (2012). Variation at Innate Immunity Toll-Like Receptor Genes in a Bottlenecked Population of a New Zealand Robin. *PLOS ONE, 7,* e45011. https://doi.org/10.1371/JOURNAL.PONE.0045011.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific Data*, 4, 1–20. https://doi.org/10.1038/sdata.2017.122.

Katoh, K., Misawa, K., Kuma, K. I., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research, 30,* 3059–3066. https://doi.org/10.1093/NAR/GKF436.

Lajbner, Z., Pnini, R., Camus, M. F., Miller, J., & Dowling, D. K. (2018). Experimental evidence that thermal selection shapes mitochondrial genome evolution. *Scientific Reports*, 8, 9500-. https://doi.org/10.1038/s41598-018-27805-3.

Lawrence, E.R. & Fraser, D.J. (2020). Latitudinal biodiversity gradients at three levels: Linking species richness, population richness and genetic diversity. *Global Ecology and Biogeography*, 29, 770–788. https://doi.org/10.1111/GEB.13075.

Leigh, D. M., Vandergast, A. G., Hunter, M. E., Crandall, E. D., Funk, W. C., Garroway, C. J., Hoban, S., Oyler-McCance, S. J., Rellstab, C., Segelbacher, G., Schmidt, C., Vázquez-Domínguez, E., & Paz-Vinas, I. (2024). Best practices for genetic

and genomic data archiving. *Nature Ecology & Evolution*, 8, 1224–1232. https://doi.org/10.1038/s41559-024-02423-7.

Leigh, D.M., van Rees, C.B., Millette, K.L., Breed, M.F., Schmidt, C., Bertola, L.D., *et al.* (2021). Opportunities and challenges of macrogenetic studies. *Nat Rev Genet*., 22. https://doi.org/0.1038/s41576-021-00394-0.

Lowe, A.J., Breed, M.F., Caron, H., Colpaert, N., Dick, C., Finegan, B., *et al.* (2018). Standardized genetic diversity-life history correlates for improved genetic resource management of Neotropical trees. *Divers Distrib*, 24, 730–741. https://doi.org/10.1111/DDI.12716.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, 6, 3139. https://doi.org/10.21105/JOSS.03139.

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. *Nature Communications*, 11, 692. https://doi.org/10.1038/s41467-020-14409-7.

Marton, K. (2025). *MuMIn: Multi-Model Inference R package.* (1.48.11). https://doi.org/10.32614/CRAN.package.MuMIn.

McGillycuddy, M., Popovic, G., Bolker, B. M., & Warton, D. I. (2025). Parsimoniously Fitting Large Multivariate Random Effects in glmmTMB. *Journal of Statistical Software*, 112, 1–19. https://doi.org/10.18637/JSS.V112.I01.

Miraldo, A., Li, S., Borregaard, M.K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., *et al.* (2016). An Anthropocene map of genetic diversity. *Science*, 353, 1532–1535. https://doi.org/10.1126/science.aaf438.

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences*, 76, 5269–5273. https://doi.org/10.1073/PNAS.76.10.5269.

Paz-Vinas, I., Jensen, E.L., Bertola, L.D., Breed, M.F., Hand, B.K., Hunter, M.E., *et al.* (2021). Macrogenetic studies must not ignore limitations of genetic markers and scale. *Ecol Lett*, 24, 1282–1284. https://doi.org/10.1111/ELE.13732.

Pelletier, T.A. & Carstens, B.C. (2018). Geographical range size and latitude predict population genetic structure in a global survey. *Biol Lett*, 14. https://doi.org/10.1098/RSBL.2017.0566/50522.

Posit team. (2025). *RStudio: Integrated Development for R.* (v2025.05.1+513). Posit Software, PBC. https://posit.co/download/rstudio-desktop/.

R CoreTeam. (2025). *R: A language and environment for statistical computing.* (4.5.1). Foundation for Statistical Computing. https://www.r-project.org.

Ratnasingham, S. & Hebert, P.D.N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS One*, 8, e66213. https://doi.org/10.1371/JOURNAL.PONE.0066213.

Schmidt, C., Dray, S. & Garroway, C.J. (2022). Genetic and species-level biodiversity patterns are linked by demography and ecological opportunity. *Evolution*, 76, 86–100. https://doi.org/10.1111/EVO.14407.

Stoffel, M.A., Humble, E., Paijmans, A.J., Acevedo-Whitehouse, K., Chilvers, B.L., Dickerson, B., *et al.* (2018). Demographic histories and genetic diversity across pinnipeds are shaped by human exploitation, ecology and life-history. *Nature Communications*, 9, *4836*. https://doi.org/10.1038/s41467-018-06695-z.

512 Theodoridis, S., Fordham, D. A., Brown, S. C., Li, S., Rahbek, C., & Nogues-Bravo,
513 D. (2020). Evolutionary history and past climate change shape the distribution
514 of genetic diversity in terrestrial mammals. *Nature Communications*,11, 1–11.
515 https://doi.org/10.1038/s41467-020-16449-5.
516 van den Burg, M. P., & Vieites, D. R. (2023). Bird genetic databases need improved
517 curation and error reporting to NCBI. *Ibis*, 165, 472–481.
518 https://doi.org/10.1111/IBI.13143.
519 Yiming, L., Siqi, W., Chaoyuan, C., Jiaqi, Z., Supen, W., Xianglei, H., *et al.* (2021).
520 Latitudinal gradients in genetic diversity and natural selection at a highly
521 adaptive gene in terrestrial mammals. *Ecography*, 44, 206–218.
522 https://doi.org/10.1111/ecog.05082.
523

Figures

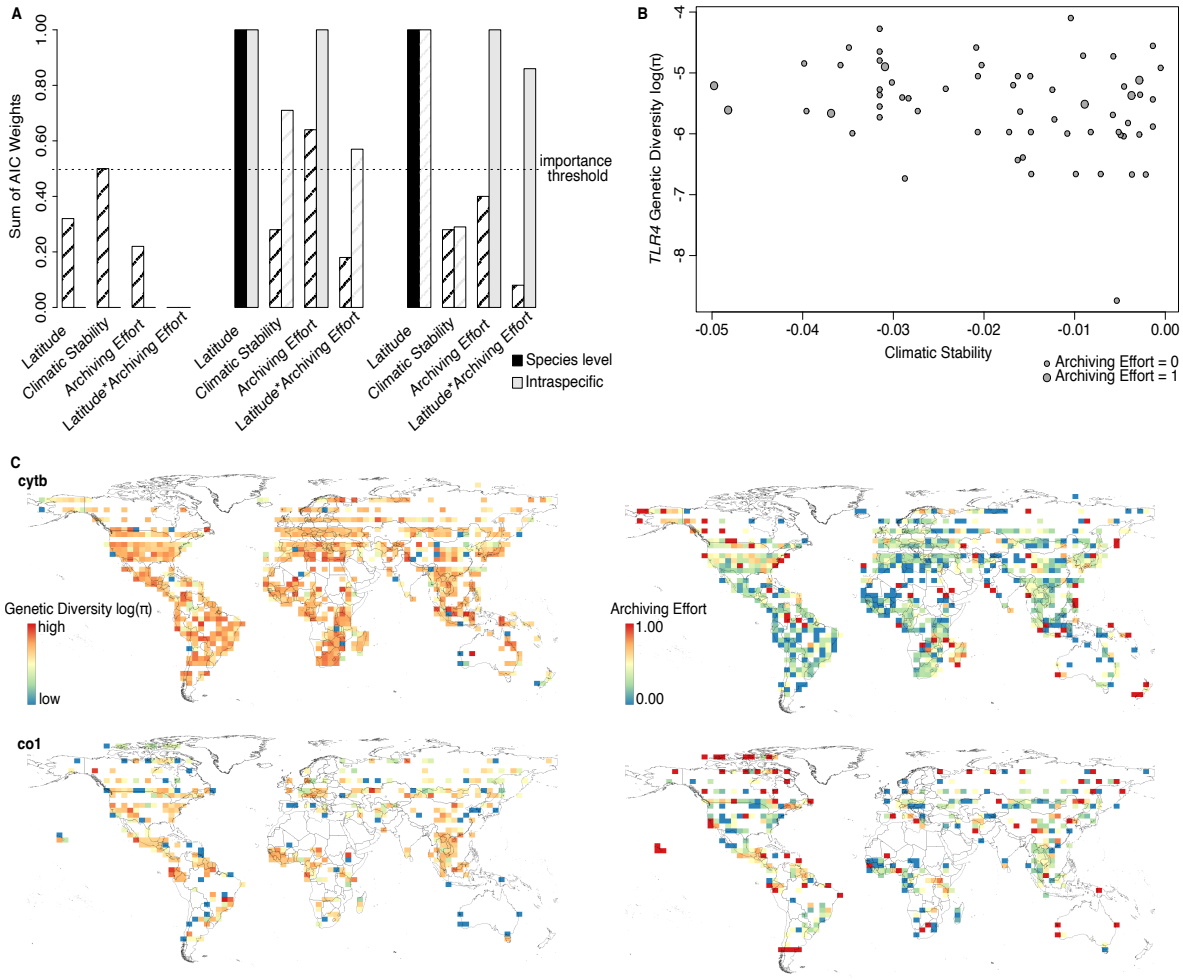

Figure 1.

Figure Legends

Figure 1. Archiving efforts as drivers of global avian TLR4, mammalian *cytb* and mammalian *co1* genetic diversity patterns. **A)** Predictor variable significance at both species and intraspecific (i.e. gridded) level. Predictors with relative importance higher than 50% (i.e. the relative sum of Akaike weights for the set of best models with $\Delta AIC < 4$) are considered as significant contributors to genetic diversity. Solid bars represent significant predictors in multi-model inference analysis ($p < 0.05$), black colours or lines represent the species level and grey the intraspecific. **B)** Species level relationship between TLR4 genetic diversity, climatic stability and archiving efforts. Archiving effort for each species is represented as bubbles with increasing size. At this level, archiving effort = 0 signifies that only unique haplotypes were archived for the species, while an archiving effort = 1 signifies that the species sequence data was archived comprehensively. **C)** Global distribution of mammalian *cytb* and *co1* genetic diversity and sequence archiving efforts. Archiving effort at the intraspecific (gridded) level was estimated as the proportion of species whose sequence data were archived comprehensively within a cell.

Tables

Table 1. Multi-model inference of predictors of genetic diversity at the species level.

Genetic Marker	Model	Predictor Variables	Slope (± SE)	Sum of Akaike Weights	*p-value*
TLR4	*Complete*	Latitude	5.79E-03 ± 6.95E-03	0.32	0.41
		Climatic Stability	-1.35E+01 ± 8.66E+00	0.50	0.13
		Archiving Effort	2.38E-02 ± 3.83E-01	0.22	0.95
		Latitude*Archiving Effort	–	–	–
	Unique Haplotypes	Latitude	5.99E-03 ± 6.82E-03	0.51	0.39
		Climatic Stability	-1.35E+01 ± 8.51E+00	0.33	0.12
		Archiving Effort	1.15E-01 ± 3.77E-01	0.23	0.77
		Latitude*Archiving Effort	–	–	–
	Rarefaction (n = 2)	Latitude	2.84E-03 ± 8.41E-03	0.31	0.74
		Climatic Stability	-1.54E+01 ± 9.15E+00	0.58	0.10
		Archiving Effort	4.08E-01 ± 4.13E-01	0.36	0.33
		Latitude*Archiving Effort	–	–	–
	Rarefaction (n = 3)	Latitude	2.99E-03 ± 5.03E-03	0.16	0.57

		Climatic Stability	-5.54E+00 ± 6.48E+00	0.25	0.41
		Archiving Effort	1.31E-01 ± 2.71E-01	0.21	0.64
		Latitude*Archiving Effort	–	–	–
	Rarefaction (n = 5)	Latitude	3.56E-04 ± 5.85E-03	0.15	0.95
		Climatic Stability	-3.68E+00 ± 7.08E+00	0.17	0.62
		Archiving Effort	9.82E-02 ± 2.74E-01	0.16	0.73
		Latitude*Archiving Effort	–	–	–
cytb	*Complete*	**Latitude**	**-7.76E-03 ± 3.01E-03**	**1.00**	**<0.01**
		Climatic Stability	-2.40E+00 ± 7.54E+00	0.28	0.75
		Archiving Effort	-1.21E-01 ± 1.03E-01	0.64	0.24
		Latitude*Archiving Effort	-1.56E-03 ± 5.13E-03	0.18	0.76
	Unique Haplotypes	**Latitude**	**-6.88E-03 ± 2.42E-03**	**1.00**	**<0.01**
		Climatic Stability	-1.83E+00 ± 5.95E+00	0.28	0.76
		Archiving Effort	**2.83E-01 ± 8.11E-02**	**1.00**	**<0.001**
		Latitude*Archiving Effort	-5.46E-04 ± 4.08E-03	0.27	0.89
	Rarefaction (n = 2)	**Latitude**	**-6.35E-03 ± 3.13E-03**	**1.00**	**<0.05**
		Climatic Stability	-1.80E+00 ± 7.69E+00	0.25	0.82

		Archiving Effort	**2.14E-01 ± 1.06E-01**	**0.93**	**<0.05**
		Latitude*Archiving Effort	-1.69E-03 ± 5.22E-03	0.26	0.75
	Rarefaction (n = 4)	Latitude	-4.65E-03 ± 3.96E-03	0.60	0.24
		Climatic Stability	9.88E+00 ± 7.73E+00	0.47	0.20
		Archiving Effort	-1.08E-01 ± 1.33E-01	0.46	0.42
		Latitude*Archiving Effort	-3.98E-03 ± 6.61E-03	0.09	0.55
	Rarefaction (n = 6)	Latitude	-2.42E-03 ± 6.19E-03	0.58	0.70
		Climatic Stability	6.72E+00 ± 8.74E+00	0.35	0.44
		Archiving Effort	3.77E-03 ± 2.11E-01	0.39	0.99
		Latitude*Archiving Effort	-1.15E-02 ± 8.37E-03	0.13	0.17
co1	*Complete*	**Latitude**	**-1.01E-02 ± 3.92E-03**	**1.00**	**<0.05**
		Climatic Stability	-4.33E+00 ± 6.83E+00	0.28	0.53
		Archiving Effort	-1.08E-01 ± 1.45E-01	0.40	0.46
		Latitude*Archiving Effort	9.06E-05 ± 7.64E-03	0.08	0.99
	Unique Haplotypes	**Latitude**	**-7.92E-03 ± 3.78E-03**	**0.93**	**<0.05**
		Climatic Stability	2.92E+00 ± 5.74E+00	0.32	0.61
		Archiving Effort	**4.27E-01 ± 1.15E-01**	**1.00**	**<0.001**

	Latitude*Archiving Effort	3.43E-03 ± 5.61E-03	0.28	0.54
Rarefaction (n = 2)	**Latitude**	**-1.11E-02 ± 4.68E-03**	**1.00**	**<0.05**
	Climatic Stability	-4.40E+00 ± 6.88E+00	0.31	0.52
	Archiving Effort	2.54E-01 ± 1.53E-01	0.80	0.10
	Latitude*Archiving Effort	2.32E-03 ± 7.72E-03	0.22	0.76
Rarefaction (n = 5)	Latitude	-7.15E-03 ± 6.42E-03	0.60	0.27
	Climatic Stability	1.10E+01 ± 6.43E+00	0.58	0.09
	Archiving Effort	6.29E-02 ± 2.77E-01	0.29	0.82
	Latitude*Archiving Effort	6.55E-03 ± 1.95E-02	0.04	0.74
Rarefaction (n = 8)	Latitude	-5.26E-03 ± 4.23E-03	0.47	0.22
	Climatic Stability	9.53E+00 ± 6.68E+00	0.51	0.16
	Archiving Effort	6.46E-02 ± 4.88E-01	0.27	0.90
	Latitude*Archiving Effort	–	–	–

All predictor variables identified in the top performing models (deltaAICc < 4) of every marker-standardizing scheme are presented. The effect sizes (slope), the sum of Akaike weights for each predictor variable and the respective p-value for the averaged conditional models are represented. Significant predictor variables for both the sum of Akaike weights (≥ 0.5) and p-values (< 0.05) are presented in bold.

548

Table 2. Multi-model inference of predictors of genetic diversity at the intraspecific level (gridded) level.

Genetic Marker	Dataset	Predictor Variables	Slope (± SE)	Sum of Akaike Weights	*p-value*
cytb	*Complete*	**Latitude**	**-9.48E-03 ± 2.96E-03**	**1.00**	**<0.01**
		Climatic Stability	-6.61E+00 ± 3.81E+00	0.63	0.08
		Archiving Effort	**-1.04E+00 ± 2.26E-01**	**1.00**	**<0.001**
		Latitude*Archiving Effort	9.18E-03 ± 5.95E-03	0.54	0.12
	Unique Haplotypes	**Latitude**	**-8.67E-03 ± 2.39E-03**	**1.00**	**<0.001**
		Climatic Stability	-5.61E+00 ± 3.07E+00	0.67	0.07
		Archiving Effort	-3.04E-01 ± 2.00E-01	0.74	0.12
		Latitude*Archiving Effort	8.33E-03 ± 5.07E-03	0.43	0.10
	Rarefaction (n = 2)	**Latitude**	**-9.27E-03 ± 3.09E-03**	**1.00**	**<0.01**
		Climatic Stability	-7.49E+00 ± 3.90E+00	0.71	0.06
		Archiving Effort	**-7.13E-01 ± 2.40E-01**	**1.00**	**<0.01**
		Latitude*Archiving Effort	9.99E-03 ± 6.19E-03	0.57	0.11
co1	*Complete*	Latitude	-6.50E-03 ± 4.86E-03	1.00	0.18
		Climatic Stability	-5.00E+00 ± 5.77E+00	0.29	0.39

	Archiving Effort	**-7.38E-01 ± 3.70E-01**	**1.00**	**<0.05**
	Latitude*Archiving Effort	**-1.67E-02 ± 7.53E-03**	**0.86**	**<0.05**
Unique Haplotypes	Latitude	-3.86E-03 ± 3.40E-03	1.00	0.26
	Climatic Stability	-7.02E-01 ± 3.85E+00	0.19	0.86
	Archiving Effort	3.07E-01 ± 2.80E-01	0.81	0.27
	Latitude*Archiving Effort	**-1.36E-02 ± 5.80E-03**	**0.71**	**<0.05**
Rarefaction (n = 2)	Latitude	-8.68E-03 ± 5.21E-03	1.00	0.10
	Climatic Stability	-6.84E+00 ± 5.72E+00	0.43	0.23
	Archiving Effort	-3.85E-01 ± 3.86E-01	1.00	0.32
	Latitude*Archiving Effort	**-1.55E-02 ± 7.60E-03**	**0.74**	**<0.05**

All predictor variables identified in the top performing models (deltaAICc < 4) of every marker-sampling scheme are presented. The effect sizes (slope), the sum of Akaike weights for each predictor variable and the respective p-value for the averaged conditional models are represented. Significant predictor variables for both the sum of Akaike weights (≥ 0.5) and p-values (< 0.05) are presented in bold.

549

550