Classification: Biological Sciences. Ecology

Title: Estimation of Species Abundance Based on the Number of Segregating Sites using Environmental DNA (eDNA)

Short Title: Quantifying Species Abundance Using eDNA

Authors: Qiaoyun Ai*, Hao Yuan†,‡, Ying Wang§, Chenhong Li+

Affiliations:

[1] Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Shanghai Ocean University, Shanghai 201306, China

[2] Engineering Research Center of Environmental DNA and Ecological Water Health Assessment, Shanghai Ocean University, Shanghai 201306, China

Current address: *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48823, USA, †Genetics and Genome Sciences Program, Michigan State University, East Lansing, MI 48823, USA, Ecology, ‡Evolution and Behavior Program, Michigan State University, East Lansing, MI 48823, USA, §Amoydx, Xiamen, China

+**Correspondence to:** Chenhong Li, chli@shou.edu.cn

Keywords: Environmental DNA (eDNA), Biodiversity Monitoring, Genetic Variation, Aquatic Organisms

Abstract: The advance of environmental DNA (eDNA) has enabled rapid and non-invasive species detection in aquatic environments. Although most studies focus on species detections, some recent studies explored the potential of using eDNA concentration to quantify species abundance. However, the differential individual DNA contribution to eDNA samples could easily obscure the concentration-species abundance relationship. We propose using the number of segregating sites as a proxy for estimating species abundance. Since segregating sites reflects the genetic diversity of the population, which is less sensitive to differential individual DNA contribution than eDNA concentration. We examined the relationship between the number of segregating sites and species abundance in silico, in vitro, and in situ using two brackish goby species, *Acanthogobius hasta* and *Tridentiger bifasciatus*. Analyses of the simulated data and in vitro data with DNA mixed from a known number of individuals showed a strong correlation between the number of segregating sites and species abundance ($R^2 > 0.9$; $P < 0.01$). Results from the in situ experiment further validated the correlation ($R^2 = 0.70$, $P < 0.01$), and such correlation was not affected by biotic factors, including body size and feeding behavior ($P > 0.05$). Results of the cross-validation test also showed that the number of segregating sites predicted species abundance with less bias and variability than the eDNA concentration. Overall, the number of segregating sites correlates stronger with species abundance and also provides a better estimate than eDNA concentration. This advancement can significantly enhance the quantitative capabilities of eDNA technology.

INTRODUCTION

The environmental DNA (eDNA) technique has emerged as a prevalent approach in ecological studies due to its cost-effectiveness, enhanced efficiency, and non-invasive nature compared to traditional approaches (Furlan and Gleeson 2016; Hosler 2017; Sigsgaard et al. 2015). By

41 sampling environment samples from various ecosystems, such as tropical freshwater (Robson et
42 al. 2016), intermittent river (Bylemans et al. 2016), brackish water (Forsström and Vasemägi
43 2016), kelp forest (Port et al. 2016), and even terrestrial habitat (Bohmann and Lynggaard 2023),
44 the eDNA technique can capture ambient DNA released by organisms, which is subsequently
45 analyzed to detect species. The eDNA approach has been widely applied in biomonitoring,
46 including endangered (Fukumoto, Ushimaru, and Minamoto 2015) and invasive species (Uchii,
47 Doi, and Minamoto 2016), allowing for monitoring single or multiple species simultaneously
48 (DiBattista et al. 2017). Additionally, the eDNA technique has also been expanded into
49 population studies, which provide a practical and nondestructive approach to determining species
50 composition (Madduppa et al. 2021), symbiotic relationships (Shinzato et al. 2018), complex
51 species structures (DiBattista et al. 2017) and breeding habits (Sakata et al. 2017).

52 Another promising frontier in eDNA research is the estimation of species abundance.
53 This application, while challenging, holds significant value for conservation ecology and
54 population genetics. eDNA offers a non-invasive method to monitor changes in species
55 abundance, particularly for invasive species, thereby informing more effective management
56 strategies (Matthews and Whittaker 2015). Furthermore, species abundance is a key parameter in
57 understanding population dynamics. Fluctuations in abundance are closely related to changes in
58 various population attributes, including genetic diversity (Hughes et al. 2008) and selection
59 pressures (Lankau and Strauss 2011). Several studies have explored the potential of using eDNA
60 concentration as a proxy by relating it to species abundance (Chambert et al. 2018; Dunn et al.
61 2017:201; Hänfling et al. 2016; Yusishen et al. 2020). Some studies reported strong positive
62 relationships between eDNA concentration and species abundance (Karlsson et al. 2022; Yates,
63 Fraser, and Derry 2019), while others showed non-significant relationships (Brys et al. 2021).

Those contradictory findings likely resulted from the influences of various biotic and abiotic factors on eDNA dynamics - shedding, decaying, and transporting, which collectively affect eDNA concentration (Barnes and Turner 2016). For example, Biological characteristics such as species (Ficetola et al. 2008; Jerde et al. 2011), tissue source (Merkes et al. 2014), life stage (Maruyama et al. 2014), and weight (Mizumoto et al. 2018) could affect the shedding rate of eDNA. Environmental conditions including temperature (Strickler, Fremier, and Goldberg 2015), UV intensity (Strickler et al. 2015), and microbial community (Poté, Ackermann, and Wildi 2009) play a role in determining the decaying rate of eDNA. Additionally, physical processes like water flow (Foppen et al. 2011) and precipitation/resuspension (Turner, Uy, and Everhart 2015) contribute to the transport of eDNA in the ecosystem. These factors combined lead to the uneven contribution of DNA from individuals among the population. Some individuals significantly influence the measured concentration, potentially leading to overestimation of their numbers, while others contribute less to the overall eDNA concentration and potentially are underrepresented in species abundance estimates. Thus, the differential contribution of individuals to the collected eDNA samples skews the species abundance estimates from eDNA concentration (Andres et al. 2023). Consequently, developing a metric resilient to differential individual contribution to eDNA samples is imperative for achieving an accurate and precise estimate of species abundance.

To establish a resilient metric against the differential individual contribution to eDNA samples, we propose using the number of segregating sites - polymorphic sites within the population - as a proxy for species abundance. As an indicator of genetic variation, the number of segregating sites rises with increasing sample size in equilibrium populations (Tajima 1989; Takahashi and Tajima 2017). The inherent redundancy nature of segregating sites ensures that

87 each individual is accessed only once regardless of the number of copies of DNA detected in the
88 samples. We investigated the relationship between the number of segregating sites and species
89 abundance through a comprehensive, three-tiered experimental design. Our study began with an
90 in silico analysis, where we tested the relationship using completely simulated data generated at
91 various mutation rates (Fig. 1a). The in silico study was then complemented by an in vitro
92 analysis, in which we simulated eDNA samples by subsampling sequences from tissues of two
93 brackish water goby species, *Acanthogobius hasta* and *Tridentiger bifasciatus* (Fig. 1b). Finally,
94 to evaluate our findings in a more natural context, we conducted in situ (mesocosm) experiments
95 to analyze eDNA collected from aquariums (Fig. 1c). Results revealed that the number of
96 segregating sites extracted from enriched sequences showed a stronger correlation and a better
97 estimate of species abundance than eDNA concentration.

98 **2 MATERIAL AND METHODS**

99 **2.1 In Silico Experiment**

100 We first assessed the relationship between the number of segregating sites and species
101 abundance by entirely simulated sequences. The length of simulated sequences was set at 17,000
102 bp. The number of simulated sequences/individuals was 1,000, and sequences were generated at
103 the mutation rate of 10^{-6} /bp/gen. To account for mutation rate variation among different species,
104 we also generated another two datasets at the mutation rate of 10^{-7} /bp/gen and 10^{-8} /bp/gen.
105 These mutation rates represent a biologically meaningful range for vertebrates (Allio et al. 2017).
106 All data were generated using the software Fastsimcoal2 (Excoffier and Foll 2011). A subset of
107 sequences were randomly chosen from the simulated data, ranging from 20 to 980 sequences
108 with intervals of 20. Selected sequences were aligned using MUSCLE v1.0 (Edgar 2004), and
109 then the number of segregating sites was counted from alignments. The simulation process was

110 repeated three times at each specified number of sequences. The correlation between the number
111 of segregating sites and the number of individuals/sequences was estimated by regression
112 analysis using Microsoft Excel, in which the number of segregating sites was the dependent
113 variable (y), the number of individuals/sequences was the independent variable (x) and the
114 significance of the correlation was estimated using R^2 and p-value. The simulated sequences are
115 available in Dryad at https://doi.org/10.5061/dryad.w3r2280zz.

116 **2.2 Target selection**

117 In our study, we employ target enrichment rather than metabarcoding to collect DNA from
118 eDNA samples. Target enrichment offers superior coverage to PCR, allowing us to extract more
119 segregating sites from wider genomic regions for species abundance estimates (Li et al. 2023).
120 Our goal is to design a single set of RNA baits for enrichment that can be applied to a wide range
121 of species. To achieve this versatility, we focus on identifying mtDNA regions that are highly
122 conserved across diverse species. This strategy ensures better enrichment across multiple species
123 while utilizing variable flanking regions to extract segregating sites.

124 We conducted a comprehensive alignment of mitochondrial genome sequences from a
125 broad spectrum of vertebrates, including *Danio rerio*, *Cyprinus carpio*, *Oryzias latipes*,
126 *Tetraodon nigroviridis*, *Gasterosteus aculeatus*, *Lepisosteus oculatus*, *Anguilla japonica*, *Gadus*
127 *morhua*, *Oreochromis niloticus*, *Homo sapiens*, *Xenopus tropicalis*, *Anolis carolinensis*,
128 *Callorhinchus milii*, and *Gallus gallus* (Table S1). We selected regions with an average pairwise
129 similarity above 0.8 and a minimum length of 120 bp to ensure sufficient conservation. This
130 rigorous selection process yielded 11 segments from the mitochondrial genome that met our
131 criteria. These segments were subsequently employed as target markers for both our in vitro and
132 in situ experiments.

133 **2.3 In Vitro Experiment**

134 Before using 11 markers to enrich targets from eDNA samples, we need to test the performance
135 of using those markers to enrich DNA extracted from tissue samples. Thus, we examined the
136 correlation between the number of segregating sites and species abundance using enriched
137 sequences to get a sense of correlation from actual populations. We selected two species of fish
138 for the in vitro experiment: *A. hasta* (92 samples) were bought from fishermen who caught fish
139 from the East China Sea in August 2016, and *T. bifasciatus* (78 samples) caught from Dishui
140 Lake in Shanghai. The DNA of two species was extracted from fin clips using an Ezup Column
141 Animal Genomic DNA Purification Kit (Sangon, Shanghai, China). The purified DNA was
142 quantified with a NanoDrop 3300 Fluorospectrometer (Thermo Fisher Scientific, Wilmington,
143 DE, U.S.A.) and visualized by agarose gel electrophoresis. The primers were designed according
144 to the mitochondrial genome sequence of *A. hasta* (NCBI accession: AY486321) and *T.*
145 *trigonocephalus* (NCBI accession: NC_029738) (Table S1). We evaluated the specificity of the
146 primers using the software Primer3.

147 RNA baits used for target enrichment were RT-PCR products of the 11 target segments.
148 TakaRa Ex Taq® (RR001 Takara) was used to amplify 11 target DNA segments using the DNA
149 from an *A. hasta* sample *(*CL1176_43) as a template. Eleven pairs of primers used to amplify
150 target segments are listed in Table S4. The PCR products were checked using agarose gel
151 electrophoresis and mixed as template transcription. Then, we followed the manual of
152 MEGAshortscript™ kit (Thermo Fisher Scientific, Wilmington, DE, U.S.A.) to transcribe PCR
153 products to RNA transcripts. The concentration of purified transcribed product was measured
154 and then stored at -80 °C until the gene enrichment.

155 The process of gene enrichment followed Li et al.'s protocol (Li et al. 2013). We sheared
156 the 1,000 ng purified DNA into around 250 bp using a Covaris M220 Focused-ultrasonicator
157 (Woburn, Massachusetts, U.S.A.). Sheared DNA was used to prepare libraries with two different
158 indexes incorporated in both ends. All indexed libraries were mixed into one tube equimolarly
159 for gene enrichment and Illumina sequencing on a HiSeq 2500 sequencer (Illumina, Inc, San
160 Diego, CA, U.S.A.).

161 The enriched sequences were mapped to mtDNA reference sequences of *A. hasta* and *T.*
162 *bifasciatus* to get potential target sequences, which were further assembled using the software
163 Trinity2.06 (Grabherr et al. 2011). We evaluated the enrichment quality based on raw read
164 numbers, mapped read numbers, enriched loci numbers, and sequencing depth.

165 We constructed a phylogenetic tree using the assembled sequences to validate the species
166 identity of our collected samples. Assembled sequences of 92 *A. hasta* and 78 *T. bifasciatus* were
167 used to build maximum likelihood trees using GTRGAMMA model in RAxML v8.0.0
168 (Stamatakis 2014) with *Mugilogobius abei* (NCBI accession: NC_023353.1) as outgroup. We
169 used 92 sequences enriched from *A. hasta* and 78 sequences enriched from *T. bifasciatus* to
170 evaluate the correlation between the number of segregating sites and the number of individuals.

171 The sequences were randomly selected across diverse group sizes, ranging from 5 to 90
172 for *A. hasta*, and 5 to 70 for *T. bifasciatus*. We generated three repeats for each group size. The
173 sequences in each group were aligned using MUSCLE v1.0 (Edgar 2004), then the number of
174 segregating sites was counted using MEGA7 (Kumar, Stecher, and Tamura 2016). The
175 correlation between the number of segregating sites and the number of individuals was evaluated
176 by regression analysis using Microsoft Excel. The number of segregating sites was set as the

177 dependent variable (y), while the number of individuals was the independent variable (x) and the
178 significance of the correlation was estimated by R^2 and p-value.

179 **2.3 In Situ Experiment (Mesocosm experiment)**

180 Finally, we conducted a mesocosm experiment to evaluate the relationship between the number
181 segregating sites and species abundance in water samples. *A. hasta* used for the mesocosm
182 experiment were captured from the East China Sea in September 2017. The fish were kept in
183 dechlorinated water with 5‰ salinity and fed with frozen shrimp before the experiment. The
184 water in the aquaria was changed at least four times, and the dead fish were removed in a timely
185 manner until all the fish were accustomed to the aquaria. All equipment, including containers, air
186 stones, and pipes were sterilized one week before the experiment. To reduce the mortality of fish,
187 we did not weigh the individuals for size grouping, but divided them into four groups according
188 to the body length (< 10 cm, 10-12.5 cm, 12.5-15 cm, > 15 cm). The average weight of fish was
189 measured after the experiment (Table 1). Twenty 65 L plastic containers were prepared to keep
190 *A. hasta* (Table 1), and one aquarium without fish was set as negative control. We set five fish
191 number gradients, including 7, 14, 21, 28, 35. To account for the effect of feeding behavior in the
192 estimation of species abundance, *Penaeus vannamei* was cut into pieces to feed the fish, at
193 approximately 5% of the estimated total weight of fish in each aquarium. One liter of water in
194 every aquarium was collected after three days. The collected water was filtered using 47-mm
195 GF/F glass filters (pore size c. o.7 μm; Whatman International, Ltd), then the filters were soaked
196 in pure ethanol and stored at -80 °C until DNA extraction. DNA samples were extracted using an
197 Ezup Column Animal Genomic DNA Purification Kit (Sangon, Shanghai, China). The purified
198 eDNA was checked using agarose gel electrophoresis and stored at -20 °C until the target gene
199 enrichment and qPCR.

200 For better target enrichment efficiency, we redesigned primers for the in situ experiment
201 according to sequences we obtained from the in vitro experiment (Table S4). PCR products were
202 further transcribed and prepared as baits following instructions in section 2.3. By checking
203 agarose gel electrophoresis, we found that the maximum length of purified eDNA was too large
204 for next-generation sequencing. Thus, we sheared DNA to around 250 bp using a Covaris M220
205 Focused-ultrasonicator (Woburn, Massachusetts, U.S.A.). Libraries preparation, target gene
206 enrichment and sequencing followed the above procedure. After preprocessing raw reads
207 including trimming adaptors, removing low-quantity reads, and PCR duplicates, the reads of
208 each sample were mapped to the reference of *A. hasta* using BWA-MEM (Li and Durbin 2010).
209 The number of segregating sites was counted from the targeted regions as well as flanking
210 sequences, spanning 50 bp upstream and downstream of the targeted regions. We removed
211 polymorphic sites with sequencing depth below seven and minor allele frequencies below 3% to
212 filter out potentially erroneous segregating sites stemming from PCR stutter and sequencing
213 errors (Fumagalli 2013).

214 We also accessed the eDNA concentration of each sample. Among the 11 target
215 segments, T2, the second segment, was chosen as the target sequence for qPCR due to its length
216 (140 bp) falling within the recommended range for qPCR amplicons (Van Holm et al. 2021). To
217 create standard samples, we amplified the T2 target region by TakaRa Ex Taq® (RR001 Takara)
218 using the genomic DNA of CL1176_43 as the template. The amplified product was purified by
219 gel extraction using DNA gel extraction kit (Qiagen, Hilden). The purified product was further
220 diluted to create six standard samples for qPCR, with concentrations ranging from 1.75 ng/L to
221 1.75×10^{6} ng/L. The qPCR procedure followed the protocol outlined in KAPA SYBR FAST
222 Universal qPCR Kit (KAPABIOSYSTEMS, U.S.A). We assessed the specificity of the eDNA

223 amplified products through a melting curve analysis using the online BIO-RAD melting curve
224 generator. The copy number was estimated using the online BIO-RAD copy number calculator.
225 The number of DNA copies was estimated as the mean value of three repeats for the same water
226 sample (Fig. S2). We averaged the number of DNA copies across replicates and converted this
227 value to eDNA concentration (copies/L).

228 Statistical analysis was performed using Stata (https://www.stata.com/). Fixed effect
229 models were used to determine if fish number, body size, and feeding behavior affected the
230 number of segregating sites or eDNA concentration. We also constructed another simpler fix
231 effect model, which only considered the fish number as the variable. We used the F-test, Akaike
232 Information Criterion (AIC), and Bayesian information criterion (BIC) to discern whether
233 species abundance is the only factor attributed to the impact on the number of segregating sites
234 or eDNA concentration.

235 To take a step further from correlation analysis, we conducted an exhaustive leave-*p*-out
236 cross-validation test to compare the performance of estimating species abundance using
237 segregating sites and eDNA. Specifically, *p* samples were randomly selected from the mesocosm
238 dataset as testing samples, while the rest served as training samples. *p* ranges from 1 to 10.
239 Training samples were used to establish regression models between the number of segregating
240 sites/eDNA concentration and species abundance. Then, those models were applied to estimate
241 the species abundance in test samples. Notably, for every value of *p*, we exhaustively accessed
242 the performance for every possible combination of samples. If *p* was greater than one, the
243 species abundance of the same sample could be estimated multiple times by models built based
244 on different training samples. We evaluated estimation performance using absolute bias, i.e., the
245 absolute difference between ground truth and estimated species abundance. We further averaged

246 absolute bias for the same sample when assessing performance. The precision of the estimation
247 of each sample was calculated using the standard deviation of predicted species abundance.

248 **3 RESULTS**

249 **3.1 In Silico Analysis**

250 Simulation analysis revealed a strong positive correlation between the number of segregating
251 sites and species abundance across all mutation rates. All regression models had R^2 above 0.9.
252 The regression coefficient for the number of segregating sites increased with mutation rate,
253 reflecting the increased occurrence of segregating sites in the mtDNA of species with higher
254 mutation rates (Fig. 2). Regression models fit better with log-transformed species abundance
255 rather than its absolute value, since the number of segregating sites reached saturation when
256 sampling enough of the number of individuals.

257 **3.2 In Vitro Analysis**

258 Before the mesocosm experiment, we used baits designed from 11 target segments to enrich
259 DNA from tissues of *A. hasta* and *T. bifasciatus*, including target segments and their flanking
260 regions (Table S1). The result revealed the successful enrichment of the 11 targets in all samples.
261 After quality check and PCR duplication removal, an average of 5,304,806 bp reads for *A. hasta*
262 and 46,355 bp for *T. bifasciatus* remained. Among those, an average of 16,354 bp reads (9%)
263 were mapped to the assembly of enriched sequences, with an average sequencing depth of 119
264 for *A. hasta* and 35 for *T. bifasciatus* (Table S2). The length of aligned sequences of *A. hasta* was
265 9,183 bp on average. For *T. bifasciatus*, the length of aligned sequences was 6,469 bp on
266 average. We observed 507 segregating sites in *A. hasta*, and 163 segregating sites in *T.*

267 *bifasciatus* respectively. The raw reads and the assembled sequences are available at NCBI
268 Sequence Read Archive (SRA) with accession numbers PRJNA1100443.

269 Using *M. abei* as an outgroup (NCBI accession: NC_023353.1), phylogenetic analyses
270 showed that samples were grouped into two clades. The sequences of *A. hasta* were clustered
271 with the sequence of *A. hasta* retrieved from GenBank (NCBI accession: AY486321). The
272 sequences of *T. bifasciatus* were grouped with *T. trigonocephalus* (NCBI accession:
273 NC_029738) as their close sister species (Fig. S1). The topology of the phylogenetic tree and the
274 grouping pattern of samples ensure the correct identification of the species.

275 We examined the relationship between the number of segregating sites and species
276 abundance using enriched sequences from tissues. Regression analysis revealed that the number
277 of segregating sites was positively correlated with the species abundance in both *A. hasta* and *T.*
278 *bifasciatus* (Fig. 3). R^2 values for regression models of both species exceeded 0.9. In contrast to
279 the in silico study, the absolute value of species abundance fit better with models than log
280 transformed form since the number of segregating sites observed in collected individuals was far
281 from reaching saturation. A strong correlation was observed between the number of segregating
282 sites and species abundance in the enrichment data from both *A. hasta* and *T. bifasciatus*.
283 However, *A. hasta* exhibited higher genetic variability, resulting in stronger correlations.
284 Consequently, only *A. hasta* was selected for the subsequent mesocosm experiments.

285 **3.3 In Situ (Mesocosm) Experiments**

286 We enriched target sequences from eDNA samples using baits designed based on 11 targeted
287 segments. After quality filtering and deduplication, 164,145,279 reads remained. Among those,
288 20,571 reads (0.59 %) were successfully mapped to the majority consensus assembly of *A. hasta*
289 from in vitro analysis, with an average sequencing depth of 197 (Table S3). Subsequently, a

290 depth filter was applied to remove potential false positive polymorphic sites with sequencing
291 depth below seven and minor allele frequency below 3% (Fumagalli 2013). The raw reads are
292 available at SRA with accession numbers PRJNA1100453.

293 Mesocosm Experiments revealed a strong and significant positive correlation between the
294 number of segregating sites and species abundance ($r = 4.09$, $p < 0.01$). Although we also
295 observed a positive relationship ($r = 3.08$, $p < 0.01$) between eDNA concentration and species
296 abundance, species abundance accounted for a smaller proportion of variation in the eDNA
297 concentration ($R^2 = 0.46$) compared to the number of segregating sites ($R^2 = 0.70$) (Fig. 4).
298 According to the F-test, AIC and BIC value, species abundance dominated the impact on the
299 number of segregating sites. At the same time, mean weight and feeding behavior showed no
300 significant influence (F-test $p = 0.38$). On the contrary, mean weight or feeding behavior
301 demonstrated a nearly significant effect on eDNA concentration (F-test $p = 0.07$) (Table 2).

302 The cross-validation results indicated that the number of segregating sites is a more
303 accurate and precise proxy for estimating species abundance. The median absolute bias of
304 species abundance predicted based on the number of segregating sites was slightly lower than the
305 one derived from eDNA concentration. This difference in median absolute bias became more
306 pronounced when fewer samples were used as training data. Additionally, the upper limit of bias
307 significantly decreased when predicting species abundance using the number of segregating sites
308 compared to eDNA concentration (Fig. 5A). We evaluated the precision of species abundance
309 estimation by examining the standard deviation of predictions for the same sample across
310 different training datasets. As anticipated, we observed increasing standard deviation when fewer
311 samples were served as training data. The standard deviation of species abundance estimates

312 based on segregating sites was consistently lower than those predicted using eDNA
313 concentration (Fig. 5B).

314 **4 DISCUSSION**

315 Estimating the species abundance in natural ecosystems could provide vital insights into a wide
316 array of evolutionary and ecological questions, and also hold implications for effective
317 conservation strategies. Emerging eDNA technologies, as a non-invasive and rapid approach,
318 represent a promising complement to traditional methods for estimating species abundance. This
319 study proposes using the number of segregating sites in eDNA as a proxy for estimating species
320 abundance. Segregating sites represent inherent genetic variability within a population, making
321 this metric less susceptible to differential individual DNA contributions in eDNA samples than
322 concentration-based methods. Thus, the segregating sites-based approach ensures a more precise
323 and less biased estimate of species abundance. Notably, our study has yet to estimate species
324 abundance in field tests, so the results from the in silico, in vitro, and mesocosm experiments
325 must be interpreted with caution. But still, our study is an essential step toward accurate and
326 precise estimates of species abundance using eDNA.

327 Previous studies have already leveraged genetic variability as an indicator of species
328 abundance in eDNA samples. Recent methods count distinct microsatellite alleles (Carreon-
329 Martinez et al. 2014) or mtDNA haplotypes (Yoshitake et al. 2019, 2021) as a representation of
330 species abundance. Another approach, the DNA mixture model, uses a probabilistic framework
331 to estimate the number of contributors to eDNA samples based on observed microsatellite alleles
332 or mtDNA haplotypes (Andres et al. 2021, 2023; Sethi et al. 2019). Both methods utilize distinct
333 alleles or haplotypes to infer species abundance but differ in complexity. Allele counting
334 provides a simple, conservative estimate of minimal possible abundance. The DNA mixture

335 model takes this concept further by considering the relative frequency of distinct
336 alleles/haplotypes and accounting for potentially unobserved alleles. In contrast, our method
337 employs a direct correlation between the number of segregating sites, and abundance. This
338 approach simplifies the analytical process, especially compared to the DNA mixture model,
339 while still capturing the underlying relationship between genetic variability and species
340 abundance. Another advantage is that our method does not require sequence assembly when
341 using haplotypes, allowing us to capture variation even when small pieces of sequencing data are
342 missing. To apply our method in the field, a pilot study is required to establish the relationship
343 between the number of segregating sites and species abundance, while theoretically, the allele
344 counting method or DNA mixture model does not need such a pilot study. On the other hand,
345 this initial investigation gives us a good sense of when genetic variability is exhausted, thus
346 helping us avoid scenarios where results might be underestimated without our knowledge.

347 In the in silico study, we found a strong correlation between genetic variation in
348 sequences and abundance estimates across all tested mutation rates. Notably, data generated
349 under higher mutation rates exhibited stronger correlations between segregating sites and
350 individual counts, with R^2 values of 0.99, 0.98, and 0.91 for mutation rates of 10^{-6}/bp/gen,
351 10^{-7}/bp/gen, and 10^{-8}/bp/gen. We also notice that the relationship between segregating sites and
352 species abundance can reach a plateau (Fig. 2), indicating that when genetic variability is
353 exhausted, the maximum detectable species abundance significantly depends on the population's
354 mutation rate. Our in silico study suggests that even at the lowest mutation rate (10^{-8}/bp/gen), we
355 can reliably estimate populations of up to 100 individuals, which suffices for most applications
356 (Fig. 2c). However, preliminary surveys are advisable for investigating larger populations.

357 Complementing the in silico study, the in vitro experiments also demonstrated a strong
358 relationship between the number of segregating sites and species abundance, particularly evident
359 in *A. hasta* ($R^2 = 0.98$). However, differing from in silico analysis, neither *A. hasta* nor *T.*
360 *bifasciatus* showed signs of reaching a plateau in the relationship between segregating sites and
361 abundance. This finding suggests a broader potential for using segregating sites to estimate
362 species abundance in large populations. We noticed lower sequence depth and coverage for
363 enriched data from *T. bifasciatus* in some samples, likely due to excessive PCR cycles during
364 DNA library preparation. This resulted in fewer on-target sequences for *T. bifasciatus* compared
365 to *A. hasta* (Table S2). Despite the relatively poor quality of the *T. bifasciatus* enrichment data,
366 we still identified a strong positive linear relationship between the number of segregating sites
367 and species abundance. This suggests that low sequencing depth did not significantly impact our
368 ability to extract segregating sites.

369 Mesocosm experiment results showed that the number of individuals predicted the
370 number of segregating sites ($R^2 = 0.70$) better than predicting the eDNA concentration ($R^2 =$
371 0.46) as expected. We observed that the overall R^2 was lower in the mesocosm experiments
372 compared to the in vitro and in silico studies. This difference can be attributed to variables
373 introduced from eDNA samples, such as DNA degradation and the influence of ambient water
374 chemistry on target enrichment. Counter-intuitively, both the number of segregating sites and the
375 number of DNA copies were not heavily affected by body size or feeding/non-feeding treatment
376 in this study, which was inconsistent with previous results (Klymus et al. 2015; Mizumoto et al.
377 2018). Moreover, eDNA concentrations measured in our study did not show overdispersed
378 distribution as described in the previous study (Chambert et al. 2018). One possible explanation
379 is that the experiment duration was too short, and the behavior of the fish was disrupted, so the

380 size of the fish and their feeding behavior had erratic effects on releasing DNA. Alternatively,
381 unforeseen factors, such as placing aquaria in proximity to corridors or disturbances, may have
382 led to varying stress responses and DNA release from the fish.

383 Our cross-validation analysis revealed that species abundance estimates using segregating
384 sites outperformed those based on eDNA concentration, showing lower absolute bias and
385 reduced standard deviation. The performance difference became more evident as the testing set
386 size increased. This is because multiple estimates per sample, derived from models built on
387 different training sample sets, provided a more comprehensive representation of the estimate bias
388 and variation distribution. While the segregating sites method showed substantially lower upper
389 quantile absolute mean bias, the difference in median absolute mean bias was less pronounced
390 but still favorable for segregating site methods. We anticipate that longer experiments, which
391 might introduce more confounding factors for the eDNA concentration-based method, could
392 further distinguish the performance of these two approaches.

393 Nevertheless, a couple of limitations need to be addressed when applying the
394 segregating-site method. Firstly, there are no global criteria for thresholding read depth and
395 minor allele frequency to distinguish polymorphism from sequencing errors. We set an ad hoc
396 criterion of more than seven reads and at least 3% minor-allele reads to call a segregating site.
397 Alternatively, duplex Sequencing (Schmitt et al. 2012) or adding unique molecular identifiers
398 (UMIs) (Yoshitake et al. 2021) may be applied to discriminate genuine polymorphic sites from
399 sequencing errors. Secondly, a pilot survey of the genetic diversity of the population has to be
400 done before extrapolating the correlation between the number of segregating sites and the
401 number of individuals in actual samples. There is often a limited number of individuals sampled
402 in the pilot study. Therefore, future research should explore the optimal number of individuals

403 for the pilot study and how much we can extrapolate the correlation between the number of
404 individuals and segregating sites from those pilot study samples. A recommended approach for
405 estimating the number of individuals needed for a pilot study is to conduct an in silico analysis
406 similar to the one in our study. Researchers can gather relevant parameters for the population of
407 target species, such as mutation rates, from existing literature or by estimating them based on
408 closely related species in the phylogeny. By simulating various scenarios, researchers can obtain
409 a reliable estimate of the sample size required for an effective pilot study.

410 In this study, we employed target enrichment instead of the conventional metabarcoding
411 approach to collect sequences from eDNA samples. Target enrichment has already been applied
412 in species detection, proving effective in capturing mitochondrial DNA (Jensen et al. 2021; Li et
413 al. 2023; Wilcox et al. 2018) and even nuclear DNA (Jensen et al. 2021) from water samples.
414 Our application extends the utility of target enrichment to species abundance estimates.
415 Compared to metabarcoding, target enrichment is particularly useful when eDNA is partially
416 degraded or when primer binding sites are inaccessible or partially accessible. This advantage is
417 crucial, as non-amplifiable sequences in eDNA could still contain novel alleles or haplotypes not
418 observed in amplifiable sequences, potentially leading to a loss of variability that could bias final
419 species abundance estimates. Moreover, unlike metabarcoding, target enrichment can extract
420 sequences from flanking genomic regions (Li et al. 2013), not just short amplifiable regions
421 suitable for PCR. In our study, this advantage enabled us to obtain segregating sites beyond
422 baits-targeted regions. While previous studies typically designed species-specific probes (Jensen
423 et al. 2021; Li et al. 2023; Wilcox et al. 2018), we designed RNA baits based on conserved
424 mitochondrial regions across vertebrates, allowing our probes to simultaneously capture eDNA
425 from multiple species, thereby reducing costs associated with species-specific baits design.

426 There are several challenges when applying target enrichment in the field study. eDNA
427 samples typically contain genetic material from a diverse range of species, resulting in the
428 capture of numerous non-target sequences when using the target enrichment method. While
429 mapping to the reference genome can filter out most non-target reads, this approach may still fail
430 to exclude some similar sequences from closely related species that differ by only a few base
431 pairs. This limitation could potentially lead to false positive segregating sites, resulting in
432 inflated species abundance estimates. Another challenge was the variability in capture efficiency.
433 Although most eDNA samples achieved satisfactory sequencing depth (over 50x) (Table S3),
434 some samples exhibited poor capture efficiency. For instance, one sample, FW1N5, only attained
435 an average sequencing depth of 18x, which might impede the detection of rare segregating sites.
436 These limitations underscore the need for careful consideration and potential refinement of target
437 enrichment protocols when applied to eDNA samples.

438 eDNA techniques provide a non-invasive and effective solution for biomonitoring,
439 providing insight into ecological and evolutionary processes in the wild. Qualitative studies, such
440 as species detection, are well-established using eDNA, while quantitative studies that require
441 accurate species abundance estimates still pose challenges. Our study proposes the use of
442 segregating site counts as a more precise and less biased proxy than eDNA concentration,
443 marking a significant advance toward accurate species abundance estimates from eDNA. The
444 following steps in developing this method will include (1) accurately identifying rare alleles with
445 low frequencies from eDNA data, (2) extended testing of different species in mesocosms and the
446 fields for longer durations, and (3) investigating effective sampling strategies for pilot studies.

447 **AUTHOR CONTRIBUTIONS**

448 Q.A. and C.L. designed the study. Q.A., H.Y., and Y.W. conducted the study and collected
449 specimens. Q.A. and H.Y. completed laboratory work, analyzed the data, and interpreted the
450 results. Q.A. and H.Y. wrote the manuscript. C.L. provided feedback on the manuscript. All
451 authors approved the manuscript for publication.

452 **ACKNOWLEDGEMENTS**

453 This work was supported by the National Key Research and Development Program of China
454 (2022YFC2601301). Dr Z. Zhang helped with model selection and correlation analyses.

455 **CONFLICT OF INTEREST STATEMENT**

456 The authors declare no conflict of interest.

457 **DATA ACCESSIBILITY STATEMENT**

458 Simulated sequences for in-silico study are available at https://doi.org/10.5061/dryad.w3r2280zz.
459 Raw sequences of targeted enrichment from tissues of *A. hasta* and *T. bifasciatus* were deposited
460 in NCBI's Sequence Read Archive (BioProject ID: PRJNA1100443). Raw sequences of enriched
461 sequences from eDNA samples were also available at NCBI's Sequence Read Archive
462 (BioProject ID: PRJNA1100453).

463 **BENEFIT-SHARING STATEMENT**

464 The study developed an accurate and precise method for estimating species abundance using
465 eDNA data. This advancement will guide future efforts in non-invasive biomonitoring
466 techniques for conservation ecology and population genetics. All data from the study are freely
467 accessible through the NCBI's Sequence Read Archive and DRYAD digital repository.

REFERENCE

[dataset] Ai, Q., Yuan, H., Wang, Y., Li, C.; 2024; Estimation of Species Abundance Based on the Number of Segregating Sites using Environmental DNA (eDNA); DRYAD; 10.5061/dryad.w3r2280zz

[dataset] Ai, Q., Yuan, H., Wang, Y., Li, C.; 2024; Estimation of Species Abundance Based on the Number of Segregating Sites using Environmental DNA (eDNA); Sequence Read Archive; PRJNA1100443

[dataset] Ai, Q., Yuan, H., Wang, Y., Li, C.; 2024; Estimation of Species Abundance Based on the Number of Segregating Sites using Environmental DNA (eDNA); Sequence Read Archive; PRJNA1100453

Allio, Remi, Stefano Donega, Nicolas Galtier, and Benoit Nabholz. 2017. "Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker." *Molecular Biology and Evolution* 34(11):2762–72. doi: 10.1093/molbev/msx197.

Andres, Kara J., David M. Lodge, Suresh A. Sethi, and Jose Andrés. 2023. "Detecting and Analysing Intraspecific Genetic Variation with eDNA: From Population Genetics to Species Abundance." *Molecular Ecology* 32(15):4118–32. doi: 10.1111/mec.17031.

Andres, Kara J., Suresh A. Sethi, David M. Lodge, and Jose Andrés. 2021. "Nuclear eDNA Estimates Population Allele Frequencies and Abundance in Experimental Mesocosms and Field Samples." *Molecular Ecology* 30(3):685–97. doi: 10.1111/mec.15765.

Barnes, Matthew A., and Cameron R. Turner. 2016. "The Ecology of Environmental DNA and Implications for Conservation Genetics." *Conservation Genetics* 17(1):1–17. doi: 10.1007/s10592-015-0775-4.

Bohmann, Kristine, and Christina Lynggaard. 2023. "Transforming Terrestrial Biodiversity Surveys Using Airborne eDNA." *Trends in Ecology & Evolution* 38(2):119–21. doi: 10.1016/j.tree.2022.11.006.

Brys, Rein, David Halfmaerten, Sabrina Neyrinck, Quentin Mauvisseau, Johan Auwerx, Michael Sweet, and Joachim Mergeay. 2021. "Reliable eDNA Detection and Quantification of the European Weather Loach (Misgurnus Fossilis)." *Journal of Fish Biology* 98(2):399–414. doi: 10.1111/jfb.14315.

Bylemans, Jonas, Elise M. Furlan, Luke Pearce, Trevor Daly, and Dianne M. Gleeson. 2016. "Improving the Containment of a Freshwater Invader Using Environmental DNA (eDNA) Based Monitoring." *Biological Invasions* 18(10):3081–89. doi: 10.1007/s10530-016-1203-5.

Carreon-Martinez, Lucia B., Kyle W. Wellband, Timothy B. Johnson, Stuart A. Ludsin, and Daniel D. Heath. 2014. "Novel Molecular Approach Demonstrates That Turbid River Plumes Reduce Predation Mortality on Larval Fish." *Molecular Ecology* 23(21):5366–77. doi: 10.1111/mec.12927.

Chambert, Thierry, David S. Pilliod, Caren S. Goldberg, Hideyuki Doi, and Teruhiko Takahara. 2018. "An Analytical Framework for Estimating Aquatic Species Density from Environmental DNA." *Ecology and Evolution* 8(6):3468–77. doi: 10.1002/ece3.3764.

DiBattista, Joseph D., Darren J. Coker, Tane H. Sinclair-Taylor, Michael Stat, Michael L.

Berumen, and Michael Bunce. 2017. "Assessing the Utility of eDNA as a Tool to Survey Reef-Fish Communities in the Red Sea." *Coral Reefs* 36(4):1245–52. doi: 10.1007/s00338-017-1618-1.

Dunn, Nicholas, Victoria Priestley, Alba Herraiz, Richard Arnold, and Vincent Savolainen. 2017. "Behavior and Season Affect Crayfish Detection and Density Inference Using Environmental DNA." *Ecology and Evolution* 7(19):7777–85. doi: 10.1002/ece3.3316.

Edgar, Robert C. 2004. "MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput." *Nucleic Acids Research* 32(5):1792–97. doi: 10.1093/nar/gkh340.

Excoffier, Laurent, and Matthieu Foll. 2011. "Fastsimcoal: A Continuous-Time Coalescent Simulator of Genomic Diversity under Arbitrarily Complex Evolutionary Scenarios." *Bioinformatics* 27(9):1332–34. doi: 10.1093/bioinformatics/btr124.

Ficetola, Gentile Francesco, Claude Miaud, François Pompanon, and Pierre Taberlet. 2008. "Species Detection Using Environmental DNA from Water Samples." *Biology Letters* 4(4):423–25. doi: 10.1098/rsbl.2008.0118.

Foppen, Jan Willem, Caesar Orup, Renaud Adell, Vincent Poulalion, and Stefan Uhlenbrook. 2011. "Using Multiple Artificial DNA Tracers in Hydrology." *Hydrological Processes* 25(19):3101–6. doi: 10.1002/hyp.8159.

Forsström, Tiia, and Anti Vasemägi. 2016. "Can Environmental DNA (eDNA) Be Used for Detection and Monitoring of Introduced Crab Species in the Baltic Sea?" *Marine Pollution Bulletin* 109(1):350–55. doi: 10.1016/j.marpolbul.2016.05.054.

Fukumoto, Sou, Atushi Ushimaru, and Toshifumi Minamoto. 2015. "A Basin-Scale Application of Environmental DNA Assessment for Rare Endemic Species and Closely Related Exotic Species in Rivers: A Case Study of Giant Salamanders in Japan." *Journal of Applied Ecology* 52(2):358–65. doi: 10.1111/1365-2664.12392.

Fumagalli, Matteo. 2013. "Assessing the Effect of Sequencing Depth and Sample Size in Population Genetics Inferences." *PLOS ONE* 8(11):e79667. doi: 10.1371/journal.pone.0079667.

Furlan, Elise M., and Dianne Gleeson. 2016. "Environmental DNA Detection of Redfin Perch, Perca Fluviatilis." *Conservation Genetics Resources* 8(2):115–18. doi: 10.1007/s12686-016-0523-1.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29(7):644–52. doi: 10.1038/nbt.1883.

Hänfling, Bernd, Lori Lawson Handley, Daniel S. Read, Christoph Hahn, Jianlong Li, Paul Nichols, Rosetta C. Blackman, Anna Oliver, and Ian J. Winfield. 2016. "Environmental DNA Metabarcoding of Lake Fish Communities Reflects Long-Term Data from Established Survey Methods." *Molecular Ecology* 25(13):3101–19. doi: 10.1111/mec.13660.

Hosler, Denise. 2017. "Where Is the Body? Dreissenid Mussels, Raw Water Testing, and the Real Value of Environmental DNA." *Management of Biological Invasions* 8(3):335–41. doi: 10.3391/mbi.2017.8.3.07.

Hughes, A. Randall, Brian D. Inouye, Marc T. J. Johnson, Nora Underwood, and Mark Vellend. 2008. "Ecological Consequences of Genetic Diversity." *Ecology Letters* 11(6):609–23.

556 doi: 10.1111/j.1461-0248.2008.01179.x.
557 Jensen, Mads Reinholdt, Eva Egelyng Sigsgaard, Shenglin Liu, Andrea Manica, Steffen Sanvig
558 Bach, Michael Møller Hansen, Peter Rask Møller, and Philip Francis Thomsen. 2021.
559 "Genome-Scale Target Capture of Mitochondrial and Nuclear Environmental DNA from
560 Water Samples." *Molecular Ecology Resources* 21(3):690–702. doi: 10.1111/1755-
561 0998.13293.
562 Jerde, Christopher L., Andrew R. Mahon, W. Lindsay Chadderton, and David M. Lodge. 2011.
563 "'Sight-Unseen' Detection of Rare Aquatic Species Using Environmental DNA."
564 *Conservation Letters* 4(2):150–57. doi: 10.1111/j.1755-263X.2010.00158.x.
565 Karlsson, Erik, Martin Ogonowski, Göran Sundblad, Josefin Sundin, Ofir Svensson, Ilkka
566 Nousiainen, and Anti Vasemägi. 2022. "Strong Positive Relationships between eDNA
567 Concentrations and Biomass in Juvenile and Adult Pike (Esox Lucius) under Controlled
568 Conditions: Implications for Monitoring." *Environmental DNA* 4(4):881–93. doi:
569 10.1002/edn3.298.
570 Klymus, Katy E., Catherine A. Richter, Duane C. Chapman, and Craig Paukert. 2015.
571 "Quantification of eDNA Shedding Rates from Invasive Bighead Carp
572 Hypophthalmichthys Nobilis and Silver Carp Hypophthalmichthys Molitrix." *Biological*
573 *Conservation* 183:77–84. doi: 10.1016/j.biocon.2014.11.020.
574 Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. "MEGA7: Molecular Evolutionary
575 Genetics Analysis Version 7.0 for Bigger Datasets." *Molecular Biology and Evolution*
576 33(7):1870–74. doi: 10.1093/molbev/msw054.
577 Lankau, Richard A., and Sharon Y. Strauss. 2011. "Newly Rare or Newly Common:
578 Evolutionary Feedbacks through Changes in Population Density and Relative Species
579 Abundance, and Their Management Implications." *Evolutionary Applications* 4(2):338–
580 53. doi: 10.1111/j.1752-4571.2010.00173.x.
581 Li, Chenhong, Michael Hofreiter, Nicolas Straube, Shannon Corrigan, and Gavin J. P. Naylor.
582 2013. "Capturing Protein-Coding Genes across Highly Divergent Species."
583 *BioTechniques* 54(6):321–26. doi: 10.2144/000114039.
584 Li, Heng, and Richard Durbin. 2010. "Fast and Accurate Long-Read Alignment with Burrows–
585 Wheeler Transform." *Bioinformatics* 26(5):589–95. doi: 10.1093/bioinformatics/btp698.
586 Li, Juan, Peter Seeber, Jan Axtner, Rachel Crouthers, Milou Groenenberg, Arnulf Koehncke,
587 Alexandre Courtiol, Pin Chanratana, and Alex D. Greenwood. 2023. "Monitoring
588 Terrestrial Wildlife by Combining Hybridization Capture and Metabarcoding Data from
589 Waterhole Environmental DNA." *Biological Conservation* 284:110168. doi:
590 10.1016/j.biocon.2023.110168.
591 Madduppa, Hawis, Ni Kadek Dita Cahyani, Aji Wahyu Anggoro, Beginer Subhan, Edwin Jefri,
592 Lalu M. Iqbal Sani, Dondy Arafat, Nebuchadnezzar Akbar, and Dietriech Geoffrey
593 Bengen. 2021. "eDNA Metabarcoding Illuminates Species Diversity and Composition of
594 Three Phyla (Chordata, Mollusca and Echinodermata) across Indonesian Coral Reefs."
595 *Biodiversity and Conservation* 30(11):3087–3114. doi: 10.1007/s10531-021-02237-0.
596 Maruyama, Atsushi, Keisuke Nakamura, Hiroki Yamanaka, Michio Kondoh, and Toshifumi
597 Minamoto. 2014. "The Release Rate of Environmental DNA from Juvenile and Adult
598 Fish." *PLOS ONE* 9(12):e114639. doi: 10.1371/journal.pone.0114639.
599 Matthews, Thomas J., and Robert J. Whittaker. 2015. "REVIEW: On the Species Abundance
600 Distribution in Applied Ecology and Biodiversity Management." *Journal of Applied*
601 *Ecology* 52(2):443–54. doi: 10.1111/1365-2664.12380.

Merkes, Christopher M., S. Grace McCalla, Nathan R. Jensen, Mark P. Gaikowski, and Jon J. Amberg. 2014. "Persistence of DNA in Carcasses, Slime and Avian Feces May Affect Interpretation of Environmental DNA Data." *PLOS ONE* 9(11):e113346. doi: 10.1371/journal.pone.0113346.

Mizumoto, Hiroki, Hirokazu Urabe, Takashi Kanbe, Michio Fukushima, and Hitoshi Araki. 2018. "Establishing an Environmental DNA Method to Detect and Estimate the Biomass of Sakhalin Taimen, a Critically Endangered Asian Salmonid." *Limnology* 19(2):219–27. doi: 10.1007/s10201-017-0535-x.

Port, Jesse A., James L. O'Donnell, Ofelia C. Romero-Maraccini, Paul R. Leary, Steven Y. Litvin, Kerry J. Nickols, Kevan M. Yamahara, and Ryan P. Kelly. 2016. "Assessing Vertebrate Biodiversity in a Kelp Forest Ecosystem Using Environmental DNA." *Molecular Ecology* 25(2):527–41. doi: 10.1111/mec.13481.

Poté, John, Rafaël Ackermann, and Walter Wildi. 2009. "Plant Leaf Mass Loss and DNA Release in Freshwater Sediments." *Ecotoxicology and Environmental Safety* 72(5):1378–83. doi: 10.1016/j.ecoenv.2009.04.010.

Robson, Heather L. A., Tansyn H. Noble, Richard J. Saunders, Simon K. A. Robson, Damien W. Burrows, and Dean R. Jerry. 2016. "Fine-Tuning for the Tropics: Application of eDNA Technology for Invasive Fish Detection in Tropical Freshwater Ecosystems." *Molecular Ecology Resources* 16(4):922–32. doi: 10.1111/1755-0998.12505.

Sakata, Masayuki K., Nobutaka Maki, Hideki Sugiyama, and Toshifumi Minamoto. 2017. "Identifying a Breeding Habitat of a Critically Endangered Fish, Acheilognathus Typus, in a Natural River in Japan." *The Science of Nature* 104(11):100. doi: 10.1007/s00114-017-1521-1.

Schmitt, Michael W., Scott R. Kennedy, Jesse J. Salk, Edward J. Fox, Joseph B. Hiatt, and Lawrence A. Loeb. 2012. "Detection of Ultra-Rare Mutations by next-Generation Sequencing." *Proceedings of the National Academy of Sciences* 109(36):14508–13. doi: 10.1073/pnas.1208715109.

Sethi, Suresh A., Wesley Larson, Keith Turnquist, and Dan Isermann. 2019. "Estimating the Number of Contributors to DNA Mixtures Provides a Novel Tool for Ecology." *Methods in Ecology and Evolution* 10(1):109–19. doi: 10.1111/2041-210X.13079.

Shinzato, Chuya, Yuna Zayasu, Miyuki Kanda, Mayumi Kawamitsu, Noriyuki Satoh, Hiroshi Yamashita, and Go Suzuki. 2018. "Using Seawater to Document Coral-Zoothanthella Diversity: A New Approach to Coral Reef Monitoring Using Environmental DNA." *Frontiers in Marine Science* 5.

Sigsgaard, Eva Egelyng, Henrik Carl, Peter Rask Møller, and Philip Francis Thomsen. 2015. "Monitoring the Near-Extinct European Weather Loach in Denmark Based on Environmental DNA from Water Samples." *Biological Conservation* 183:46–52. doi: 10.1016/j.biocon.2014.11.023.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30(9):1312–13. doi: 10.1093/bioinformatics/btu033.

Strickler, Katherine M., Alexander K. Fremier, and Caren S. Goldberg. 2015. "Quantifying Effects of UV-B, Temperature, and pH on eDNA Degradation in Aquatic Microcosms." *Biological Conservation* 183:85–92. doi: 10.1016/j.biocon.2014.11.038.

Tajima, F. 1989. "The Effect of Change in Population Size on DNA Polymorphism." *Genetics* 123(3):597–601. doi: 10.1093/genetics/123.3.597.

Takahashi, Takuya, and Fumio Tajima. 2017. “The Amount of DNA Polymorphism When Population Size Changes Linearly.” *Genes & Genetic Systems* 92(1):55–57. doi: 10.1266/ggs.16-00055.

Turner, Cameron R., Karen L. Uy, and Robert C. Everhart. 2015. “Fish Environmental DNA Is More Concentrated in Aquatic Sediments than Surface Water.” *Biological Conservation* 183:93–102. doi: 10.1016/j.biocon.2014.11.017.

Uchii, Kimiko, Hideyuki Doi, and Toshifumi Minamoto. 2016. “A Novel Environmental DNA Approach to Quantify the Cryptic Invasion of Non-Native Genotypes.” *Molecular Ecology Resources* 16(2):415–22. doi: 10.1111/1755-0998.12460.

Van Holm, Wannes, Justien Ghesquière, Nico Boon, Tim Verspecht, Kristel Bernaerts, Naiera Zayed, Ioanna Chatzigiannidou, and Wim Teughels. 2021. “A Viability Quantitative PCR Dilemma: Are Longer Amplicons Better?” *Applied and Environmental Microbiology* 87(5):e02653-20. doi: 10.1128/AEM.02653-20.

Wilcox, Taylor M., Katherine E. Zarn, Maxine P. Piggott, Michael K. Young, Kevin S. McKelvey, and Michael K. Schwartz. 2018. “Capture Enrichment of Aquatic Environmental DNA: A First Proof of Concept.” *Molecular Ecology Resources* 18(6):1392–1401. doi: 10.1111/1755-0998.12928.

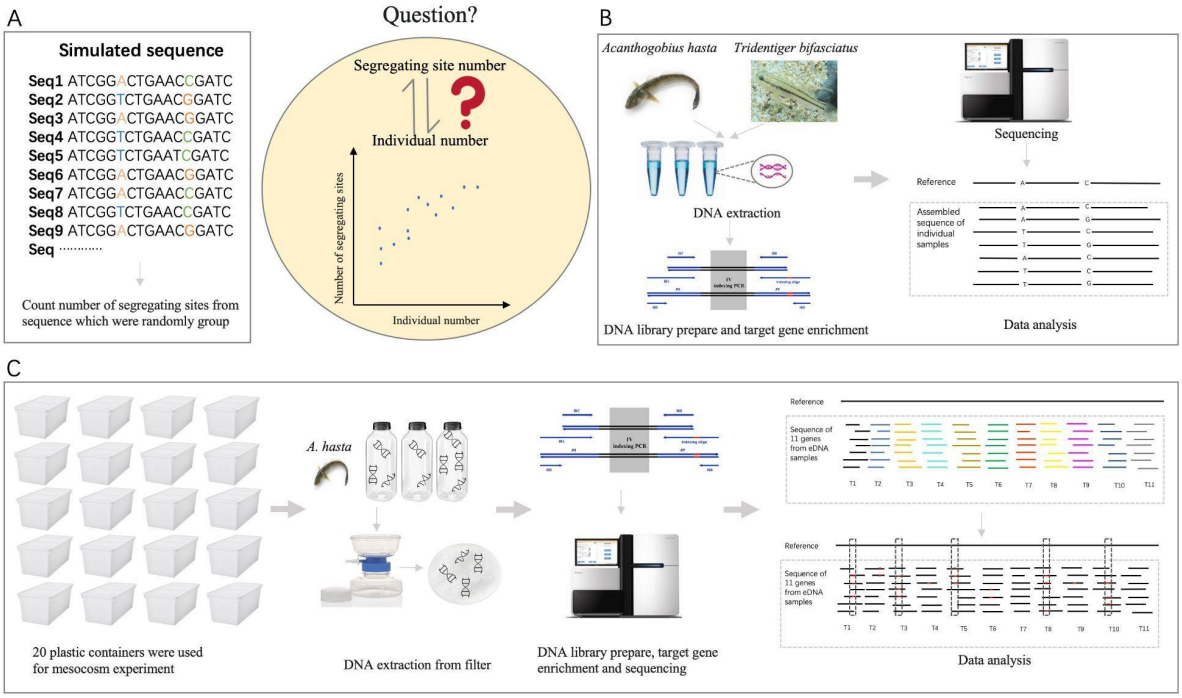
Yates, Matthew C., Dylan J. Fraser, and Alison M. Derry. 2019. “Meta-Analysis Supports Further Refinement of eDNA for Monitoring Aquatic Species-Specific Abundance in Nature.” *Environmental DNA* 1(1):5–13. doi: 10.1002/edn3.7.

Yoshitake, Kazutoshi, Atushi Fujiwara, Aiko Matsuura, Masashi Sekino, Motoshige Yasuike, Yoji Nakamura, Reiichiro Nakamichi, Masaaki Kodama, Yumiko Takahama, Akinori Takasuka, Shuichi Asakawa, Kazuomi Nishikiori, Takanori Kobayashi, and Shugo Watabe. 2021. “Estimation of Tuna Population by the Improved Analytical Pipeline of Unique Molecular Identifier-Assisted HaCeD-Seq (Haplotype Count from eDNA).” *Scientific Reports* 11(1):7031. doi: 10.1038/s41598-021-86190-6.

Yoshitake, Kazutoshi, Tatsuki Yoshinaga, Chikaya Tanaka, Nanami Mizusawa, Md. Shaheed Reza, Atsumi Tsujimoto, Takanori Kobayashi, and Shugo Watabe. 2019. “HaCeD-Seq: A Novel Method for Reliable and Easy Estimation About the Fish Population Using Haplotype Count from eDNA.” *Marine Biotechnology* 21(6):813–20. doi: 10.1007/s10126-019-09926-6.

Yusishen, Michael E., Frances-Claire Eichorn, W. Gary Anderson, and Margaret F. Docker. 2020. “Development of Quantitative PCR Assays for the Detection and Quantification of Lake Sturgeon (Acipenser Fulvescens) Environmental DNA.” *Conservation Genetics Resources* 12(1):17–19. doi: 10.1007/s12686-018-1054-8.

684 **TABLES AND FIGURES**



685

686 **Fig. 1.** Three main components of this study. (A) In silico study. We tested the correlation
687 between the number of segregating sites and the species abundance using simulated data. (B) In
688 vitro study. We designed markers and synthesized baits to capture target genes from tissues of *A.*
689 *hasta* and *T. bifasciatus* individuals. Datasets mixed from captured sequences were used to
690 assess the correlation between the number of segregating sites and the species abundance. (C) In
691 situ (mesocosm) study. We conducted controlled mesocosm experiments using *A. hasta*.
692 Segregating sites were extracted from captured sequences in eDNA samples, and used to
693 evaluate their correlation to species abundance and their performance in estimating species
694 abundance.

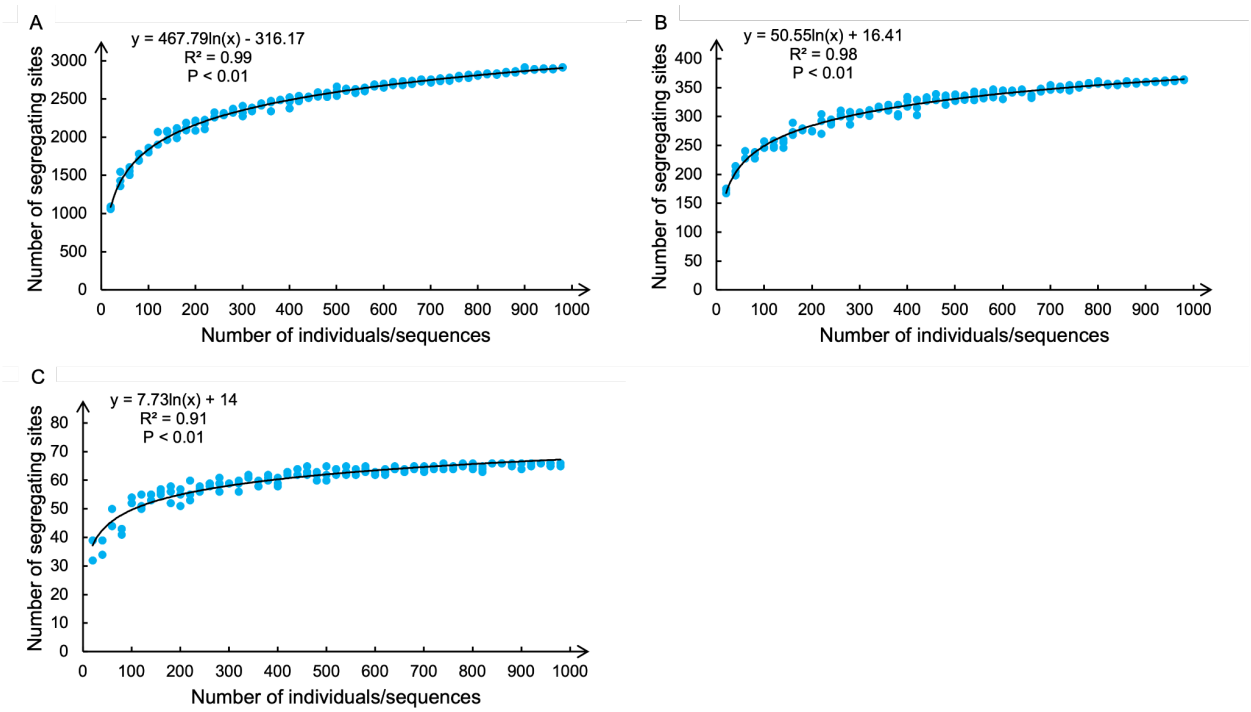


Fig. 2. Relationship between the number of segregating sites and the number of simulated individuals/sequences from in-silico analysis. Sequences of 1,000 individuals, each with a length of 17,000 bp, were simulated with a mutation rate of 10^{-6} /bp/gen (A), 10^{-7} /bp/gen (B) and 10^{-8} /bp/gen (C) respectively. For each mutation rate, 20 to 980 samples were randomly drawn from the population, and the number of segregating sites was counted. The sampling process was repeated three times for each number of individuals. Each dot in the plots represents one random sampling. The fitted line represents the relationship between the number of segregating sites and the species abundance.

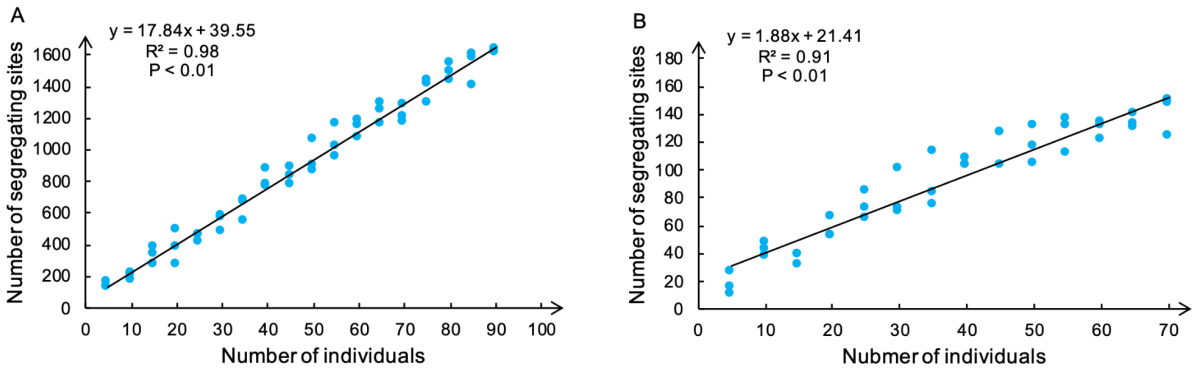
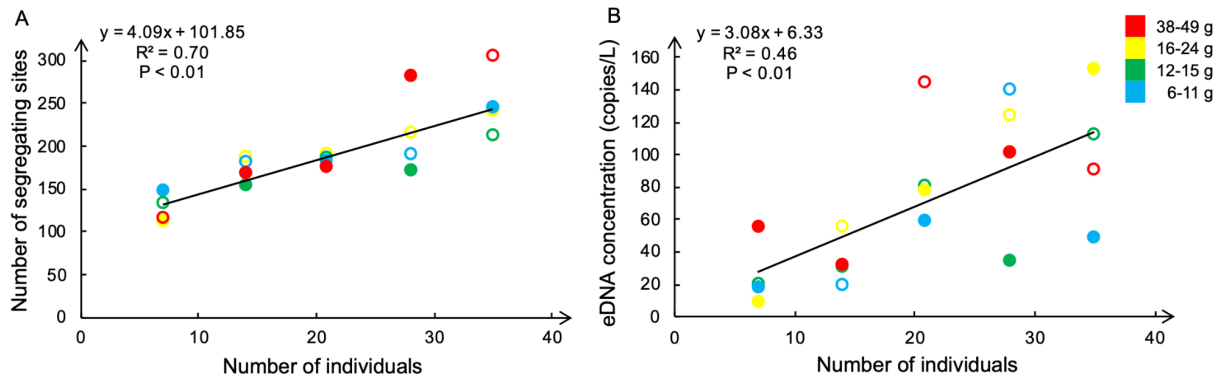


Fig. 3. Relationship between the number of individuals and the number of segregating sites observed in the assembled sequences of *A. hasta* and *T. bifasciatus*. For *A. hasta*, 5 to 90 individuals were randomly sampled from the population, and the number of segregating sites was counted (A). Similarly, for *T. bifasciatus*, 5 to 70 were randomly sampled from the population, then the number of segregating sites was counted (B). The sampling process was repeated three times for each number of individuals. Each dot in the plots represents one random sampling. Fitted lines represent the relationship between the number of individuals and the species abundance.



713

714 **Fig. 4.** Relationship between the number of individuals and the number of segregating sites in
715 eDNA samples (A), or between the number of individuals and the eDNA concentration (B). The
716 eDNA concentration was estimated based on the number of DNA copies at locus T2 using
717 qPCR. The colors of the dots depict the weight range of average weight for individuals in each
718 aquarium: blue (6-11 g), green (12-15 g), yellow (16-24 g), and red (38-49 g). The solid dots
719 indicate that the individuals in these aquariums were fed, and the circles indicate that the
720 individuals were not fed. The number of individuals stocked in each tank ranged from 7,14, 21,
721 28 to 35. Four replicates were set up at each stocking density level, with different average
722 weights and feeding treatments.

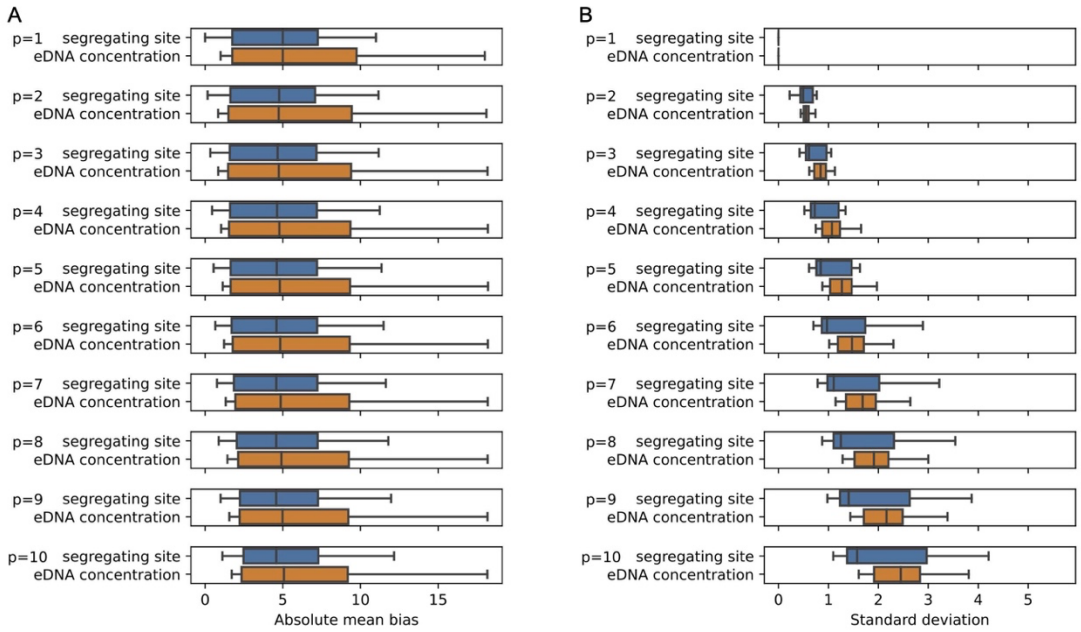


Fig. 5. Cross-validation test results for predicting species abundance. Boxplots show (A) absolute mean bias and (B) standard deviation of predicted species abundance based on segregating sites (blue) and eDNA concentration (orange) with increasing size of testing sets (top to bottom, 1 to 10 samples). Lower bias and standard deviation indicate better model accuracy and precision.

729 **Table 1.** Design of the mesocomic experiments.

Feeding\No. fish	7[a]	14	21	28	35
F[b]	W_1 (6)[c]	W_2 (12)	W_3 (18)	W_4 (39)	W_1 (11)
F_0	W_2 (13)	W_1 (7)	W_4 (38)	W_3 (24)	W_2 (15)
F	W_3 (19)	W_4 (44)	W_1 (7)	W_2 (13)	W_3 (20)
F_0	W_4 (49)	W_3 (16)	W_2 (14)	W_1 (8)	W_4 (42)

730 [a]The number (7, 14, 21, 28 and 35) indicates the number of fish kept in each aquarium.

731 [b]F means the fish were fed during the experiment. F_0 means the fish were not fed.

732 [c]The number in parentheses is the average weight of fish in this aquarium. The number in

733 subscript denotes the rank of the average weight of fish in the treatment group. Smaller rank in

734 subscript means lower average fish weight in this aquarium.

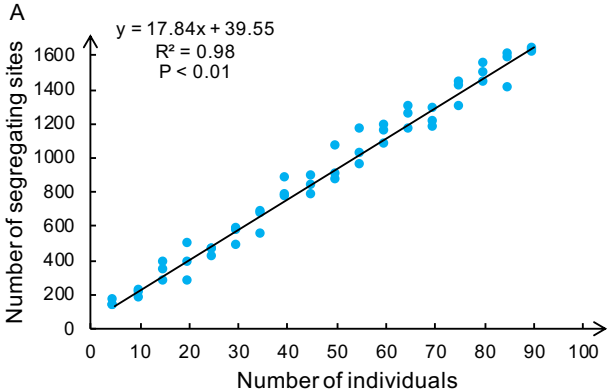
735 **Table 2.** Comparison of linear models for predicting the number of segregating sites or eDNA

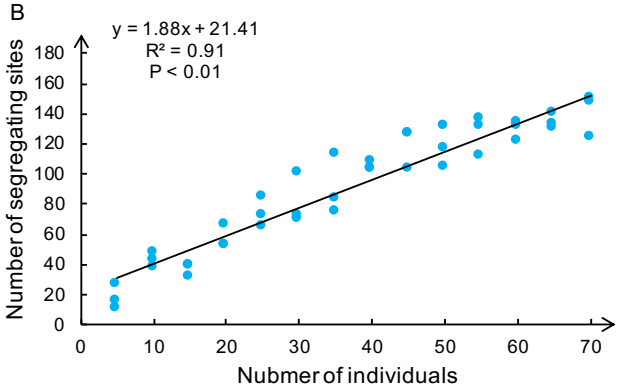
736 concentration using fish number, mean weight and feeding behavior.

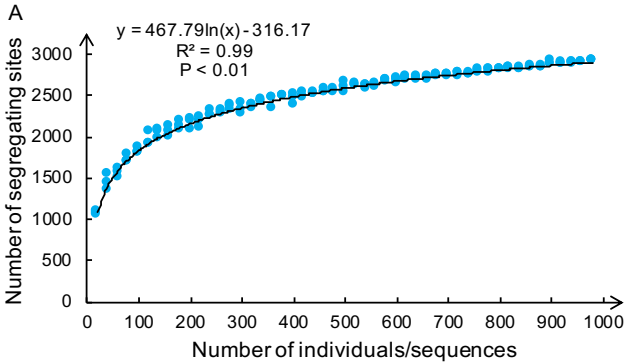
Variable	Segregating sites		eDNA concentration	
	Model 1	Model 2	Model 1	Model 2
Fish number (Coef.)	4.09**	4.08**	3.08**	3.06**
Mean weight (Coef.)		2.54		-25.95
Feeding (Coef.)		0.63		0.66
R^2	0.70	0.73	0.46	0.60
AIC	191.6	193.5	200.7	198.9
BIC	193.6	197.5	202.7	202.9
F-test (p-value)[a]		0.38		0.07

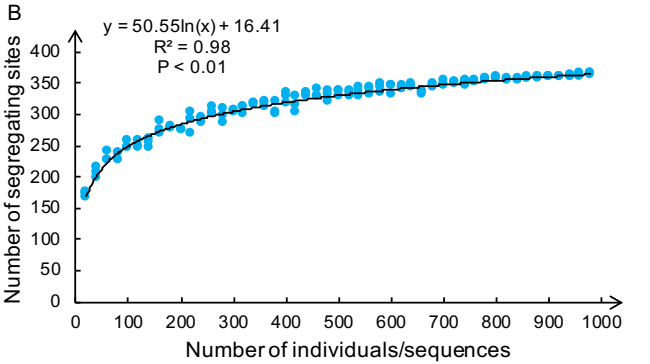
737 * $P < 0.05$; ** $P < 0.01$.

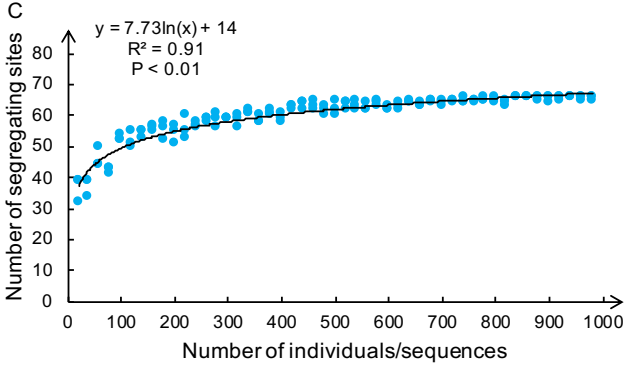
738 [a] Comparing model 1 vs. model 2 using F-test statistics.

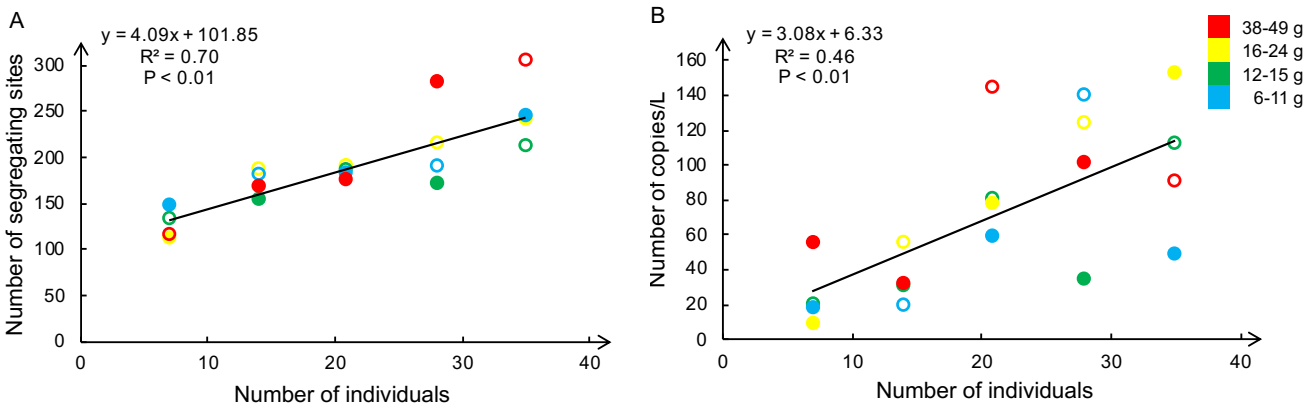

A
y = 17.84x + 39.55
R² = 0.98
P < 0.01
Number of segregating sites
Number of individuals
0
200
400
600
800
1000
1200
1400
1600
0
10
20
30
40
50
60
70
80
90
100

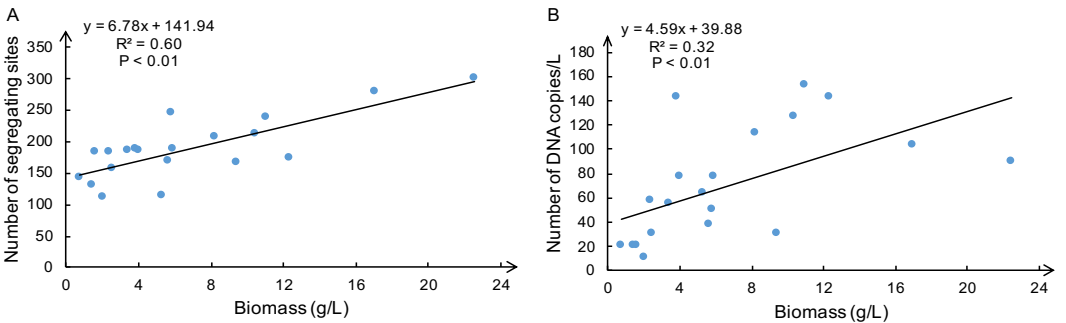

B
y = 1.88x + 21.41
R² = 0.91
P < 0.01
Number of segregating sites
Nubmer of individuals
0
20
40
60
80
100
120
140
160
180
0
10
20
30
40
50
60
70

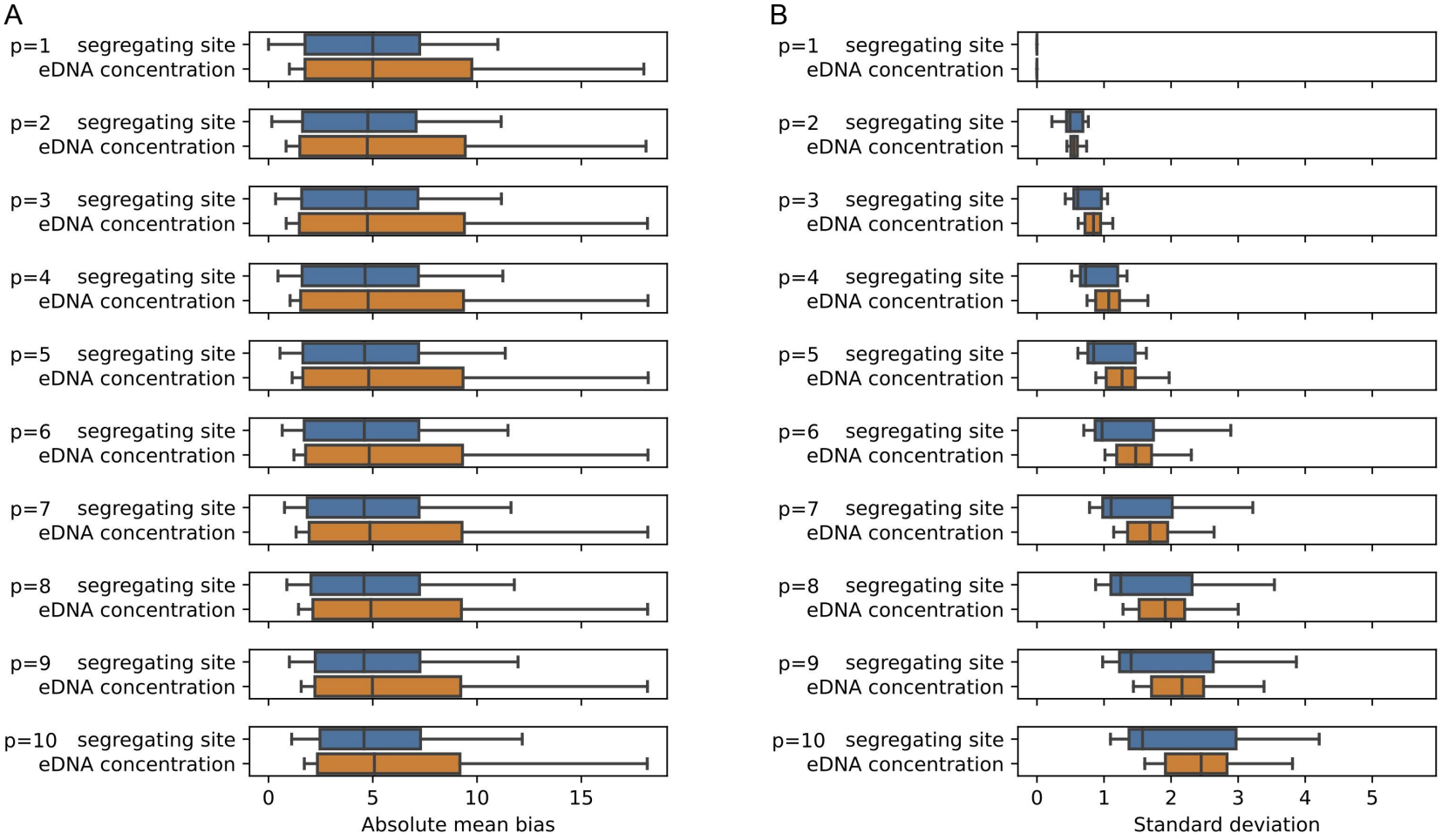

A
y = 467.79ln(x) - 316.17
R² = 0.99
P < 0.01
Number of segregating sites
Number of individuals/sequences


B
y = 50.55ln(x) + 16.41
R² = 0.98
P < 0.01
Number of segregating sites
Number of individuals/sequences


C
y = 7.73ln(x) + 14
R² = 0.91
P < 0.01
Number of segregating sites
Number of individuals/sequences


A
y = 4.09x + 101.85
R² = 0.70
P < 0.01
Number of segregating sites
Number of individuals
B
y = 3.08x + 6.33
R² = 0.46
P < 0.01
Number of copies/L
Number of individuals
38-49 g
16-24 g
12-15 g
6-11 g


A
y = 6.78x + 141.94
R² = 0.60
P < 0.01
Number of segregating sites
0
50
100
150
200
250
300
350
0
4
8
12
16
20
24
Biomass (g/L)
B
y = 4.59x + 39.88
R² = 0.32
P < 0.01
Number of DNA copies/L
0
20
40
60
80
100
120
140
160
180
0
4
8
12
16
20
24
Biomass (g/L)


A
p=1 segregating site
eDNA concentration
p=2 segregating site
eDNA concentration
p=3 segregating site
eDNA concentration
p=4 segregating site
eDNA concentration
p=5 segregating site
eDNA concentration
p=6 segregating site
eDNA concentration
p=7 segregating site
eDNA concentration
p=8 segregating site
eDNA concentration
p=9 segregating site
eDNA concentration
p=10 segregating site
eDNA concentration
0
5
10
15
Absolute mean bias
B
p=1 segregating site
eDNA concentration
p=2 segregating site
eDNA concentration
p=3 segregating site
eDNA concentration
p=4 segregating site
eDNA concentration
p=5 segregating site
eDNA concentration
p=6 segregating site
eDNA concentration
p=7 segregating site
eDNA concentration
p=8 segregating site
eDNA concentration
p=9 segregating site
eDNA concentration
p=10 segregating site
eDNA concentration
0
1
2
3
4
5
Standard deviation