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: Segregating site, environmental DNA (eDNA), quantification, species abundance

Abstract: The advance of environmental DNA (eDNA) has enabled rapid and non-invasive species detection in aquatic environments. Although most studies focus on assessing the absence or presence of species, some recent studies explored the potential of using eDNA concentration of targeted genes to quantify species abundance. However, eDNA concentration could be easily perturbed by various biotic or abiotic factors, obscuring the concentration-species abundance relationship. Here, we propose using the number of segregating sites as a proxy for estimating species abundance. We examined the relationship between the number of segregating sites and species abundance in silico, in vitro, and in situ (mesocosm experiments) 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 mesocosm experiment further validated the correlation ($R^2 = 0.70$, $P < 0.01$). Such correlation was not affected by biotic factors, including body size and feeding behavior of the fish ($P > 0.05$). Results of the cross-validation test also showed that the number of segregating sites predicted species abundance more accurately and robustly than the eDNA concentration did. Overall, the number of segregating sites is a more accurate and robust indicator of species abundance 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 & Gleeson, 2016; Hosler, 2017; Sigsgaard et al., 2015). By sampling environment samples from various ecosystems, such as tropical freshwater (Robson et

al., 2016), intermittent river (Bylemans et al., 2016), brackish water (Forsström & Vasemägi, 2016), kelp forest (Port et al., 2016), and even terrestrial habitat (Ushio et al., 2017), the eDNA technique can capture ambient DNA released by organisms, which is subsequently analyzed to detect species. The eDNA approach has been widely applied in biomonitoring, including endangered (Fukumoto et al., 2015) or invasive species (Uchii et al., 2016) and allowing for monitoring single or multiple species simultaneously (DiBattista et al., 2017). Additionally, the eDNA technique has also been expanded into population studies, which provide a practical and nondestructive approach to determine species composition (Madduppa et al., 2021), symbiotic relationships (Shinzato et al., 2018), complex species structures (DiBattista et al., 2017) and breeding habits (Sakata et al., 2017).

Another frontier of eDNA research is to estimate species abundance based on eDNA data. Several studies have explored the potential of using eDNA concentration as a proxy by relating it to species abundance (Chambert et al., 2018; Dunn et al., 2017; Hänfling et al., 2016; Yusishen et al., 2020). Some studies reported strong positive relationships between eDNA concentration and species abundance (Karlsson et al., 2022), 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 concentration, such as species, life stage (Maruyama et al., 2014) and weight (Mizumoto et al., 2018). Some studies tried to model the effect of some of the factors (Jo et al., 2019; Sassoubre et al., 2016; Takeuchi et al., 2019), however, comprehensively considering all potential uncertainties that might affect eDNA concentration is a challenging task for researchers. Consequently, a metric resilient to uncertainties induced by biotic and abiotic factors is imperative for achieving an accurate and robust estimate of species abundance.

To establish a resilient metric against biotic and abiotic influences, 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 & Tajima, 2017). The inherent redundancy nature of segregating sites ensures that the estimates of species abundance based on the number of segregating sites should not be distorted by those factors influencing eDNA concentrations. We examined the relationship between the number of segregating sites and species abundance in silico (Fig. 1a), in vitro (Fig. 1b), and in situ (mesocosm) experiments (Fig. 1c) using two brackish water goby species, *Acanthogobius hasta* and *Tridentiger bifasciatus*. Results revealed that the number of segregating sites extracted from enriched sequences showed a stronger correlation and a better estimate of species abundance than eDNA concentration.

2 MATERIAL AND METHODS

2.1 Target selection

We chose 11 segments of the mtDNA genome as target markers for in vitro and in situ experiments. The segments were selected from alignment of mitochondrial genome sequences of *Danio rerio*, *Cyprinus carpio*, *Oryzias latipes*, *Tetraodon nigroviridis*, *Gasterosteus aculeatus*, *Lepisosteus oculatus*, *Anguilla japonica*, *Gadus morhua*, *Oreochromis niloticus*, *Homo sapiens*, *Xenopus tropicalis*, *Anolis carolinensis*, *Callorhinchus milii*, *Gallus gallus* (Table S1). We targeted conserved segments for better enrichment and pulled out the flanking fragments for sequencing.

2.2 In Silico Experiment

We first assessed the relationship between the number of segregating sites and species abundance by entirely simulated sequences. The length of simulated sequences was set at 17,000 bp, close to the total size of 11 target segments. The number of simulated sequences/individuals was 1000, and sequences were generated at the mutation rate of 10^{-6} /bp/gen. To account for mutation rate variation among different species, we also generated another two datasets at the mutation rate of 10^{-7} /bp/gen and 10^{-8} /bp/gen. All data were generated using the software Fastsimcoal2 (Excoffier & Foll, 2011). A subset of sequences were randomly chosen from the simulated data, ranging from 20 to 980 sequences with intervals of 20. Selected sequences were aligned using MUSCLE v1.0 (Edgar, 2004), then the number of segregating sites was counted from alignments. The simulation process was repeated three times at each specified number of sequences. The correlation between the number of segregating sites and the number of individuals/sequences was estimated by regression analysis using Microsoft Excel, in which the number of segregating sites was the dependent variable (y), the number of individuals/sequences was the independent variable (x) and the significance of the correlation was estimated using R^2 and p-value. The simulated sequences are available at https://doi.org/10.5061/dryad.w3r2280zz.

2.3 In Vitro Experiment

Before using 11 markers to enrich targets from eDNA samples, we need to test the performance of using those markers to enrich DNA extracted from tissue samples. Additionally, we examined the correlation between the number of segregating sites and species abundance using enriched sequences to get a sense of correlation from actual populations. We selected two kinds of fish for the in vitro experiment: *A. hasta* (92 samples) were bought from fishermen who caught fish from the East China Sea in August 2016, and *T. bifasciatus* (78 samples) caught from Dishui Lake in Shanghai. The DNA of two species was extracted from fin clips using an Ezup Column Animal

Genomic DNA Purification Kit (Sangon, Shanghai, China). The purified DNA was quantified with a NanoDrop 3300 Fluorospectrometer (Thermo Fisher Scientific, Wilmington, DE, U.S.A.) and visualized by agarose gel electrophoresis. The primers were designed according to the mitochondrial genome sequence of *A. hasta* (NCBI accession: AY486321) and *T. trigonocephalus* (NCBI accession: NC_029738) (Table S1). We evaluated the specificity of the primers using the software Primer3.

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

The process of gene enrichment followed Li *et al.*'s protocol (C. Li et al., 2013). We sheared the 1000 ng purified DNA into around 250 bp using a Covaris M220 Focused-ultrasonicator (Woburn, Massachusetts, U.S.A.). Sheared DNA was used to prepare libraries with two different indexes incorporated in both ends. All indexed libraries were mixed into one tube equimolarity for gene enrichment and Illumina sequencing on a HiSeq 2500 sequencer (Illumina, Inc, San Diego, CA, U.S.A.).

The enriched sequencing data were assembled using the software Trinity2.06 (Grabherr et al., 2011). We evaluated the enrichment quality based on raw read numbers, mapped read numbers, enriched loci numbers and sequencing depth. All assembled sequences of 92 *A. hasta*

and 78 *T. bifasciatus* were used to build maximum likelihood trees using GTRGAMMA model in RAxML v8.0.0 (Stamatakis, 2014) with *Mugilogobius abei* (NCBI accession: NC_023353.1) as outgroup. We used 92 sequences enriched from *A. hasta* and 78 sequences enriched from *T. bifasciatus* to evaluate the correlation between the number of segregating sites and the number of individuals. The sequences were randomly selected across diverse group sizes, ranging from 5 to 90 for *A. hasta*, and 5 to 70 for *T. bifasciatus*. We generated three repeats for each group size. The sequences in each group were aligned using MUSCLE v1.0 (Edgar, 2004), then the number of segregating sites was counted using MEGA7 (Kumar et al., 2016). The correlation between the number of segregating sites and the number of individuals was evaluated by regression analysis using Microsoft Excel. The number of segregating sites was set as the dependent variable (y), while the number of individuals was the independent variable (x) and the significance of the correlation was estimated by R^2, p-value.

2.3 In Situ Experiment (Mesocosm experiment)

A. hasta used for the mesocosm experiment were captured from the East China Sea in September, 2017. The fish were kept in dechlorinated water with 5‰ salinity and fed with frozen shrimp before the experiment. The water in the aquaria was changed at least four times, and the dead fish were removed in a timely manner until all the fish were accustomed to the aquaria. All equipment, including containers, air stones and pipes were sterilized one week before the experiment. To reduce the mortality of fish, we did not weigh the individuals for size grouping, but divided them into four groups according to the body length (< 10 cm, 10-12.5 cm, 12.5-15 cm, > 15 cm). The average weight of fish was measured after the experiment (Table 1). Twenty 65 L plastic containers were prepared to keep *A. hasta* (Table 1), and one aquarium without fish was set as control. We set five fish number gradients, including 7, 14, 21, 28, 35. To account for

the effect of feeding behavior in the estimation of species abundance, *Penaeus Vannamei* was cut into pieces to feed the fish, at approximately 5% of the estimated total weight of fish in each aquarium. One liter of water in every aquarium was collected after three days. The collected water was filtered using 47-mm GF/F glass filters (pore size c. o.7 μm; Whatman International, Ltd), then the filters were soaked in pure ethanol and stored at -80 °C until DNA extraction. DNA samples were extracted using an Ezup Column Animal Genomic DNA Purification Kit (Sangon, Shanghai, China). The purified eDNA was checked using agarose gel electrophoresis and stored at -20 °C until the target gene enrichment and qPCR.

For better target capture efficiency, we redesigned primers for the in situ experiment according to sequences we obtained from the in vitro experiment (Table S4). PCR products were further transcribed and prepared as baits following instructions in section 2.2. By checking agarose gel electrophoresis, we found that the length of purified eDNA was too large for next-generation sequencing. Thus, we sheared DNA to around 250 bp using a Covaris M220 Focused-ultrasonicator (Woburn, Massachusetts, U.S.A.). Libraries preparation, target gene enrichment and sequencing followed the above procedure. After preprocessing raw reads including trimming adaptors, removing low quantity reads and PCR duplicates, the reads of each sample were mapped to the reference of *A. hasta* using BWA-MEM (H. Li & Durbin, 2010). The number of segregating sites was counted from the targeted regions as well as flanking sequences, spanning 50 bp upstream and downstream of the targeted regions. We removed polymorphic sites with sequencing depth below seven and minor allele frequencies below 3% to filter out potentially erroneous segregating sites stemming from PCR stutter and sequencing errors.

We also accessed the eDNA concentration of each sample. Among the 11 target segments, T2, the second segment, was chosen as the target sequence for qPCR due to its length

(140 bp) falling within the recommended range for qPCR amplicons (Van Holm et al., 2021). To create standard samples, we amplified the T2 target region by TakaRa Ex Taq® (RR001 Takara) using the genomic DNA of CL1176_43 as the template. The amplified product was purified by gel extraction using DNA gel extraction kit (Qiagen, Hilden). The purified product was further diluted to create six standard samples for qPCR, with concentrations ranging from 1.75 ng/L to 1.75×10^{6} ng/L. The qPCR procedure followed the protocol outlined in KAPA SYBR FAST Universal qPCR Kit (KAPABIOSYSTEMS, U.S.A). We assessed the specificity of the eDNA amplified products through a melting curve analysis using the online BIO-RAD melting curve generator. The copy number was estimated using the online BIO-RAD copy number calculator. The number of DNA copies was estimated as the mean value of three repeats for the same water sample (Fig. S2). We averaged the number of DNA copies across replicates and converted this value to eDNA concentration (copies/L).

Statistical analysis was performed using Stata (*Stata Statistical Software: Release 18.*, 2023). Fixed effect models were used to determine if fish number, body size and feeding behavior affected the number of segregating sites or eDNA concentration. We also constructed another simpler fix effect model, which only considered the fish number as the variable. We used the F-test, Akaike Information Criterion (AIC) and Bayesian information criterion (BIC) to discern whether species abundance is the only factor attributed to the impact on the number of segregating sites or eDNA concentration.

To take a step further from correlation analysis, we conducted an exhaustive leave-*p*-out cross validation test to compare the performance of estimating species abundance using segregating sites and eDNA. Specifically, *p* samples were randomly selected from the mesocosm dataset as testing samples, while the rest served as training samples. *p* ranges from 1 to 10.

Training samples were used to establish regression models between the number of segregating sites/eDNA concentration and species abundance. Then, those models were applied to estimate the species abundance in test samples. Notably, for every value of *p*, we exhaustively accessed the performance for every possible combination of samples. If *p* was greater than one, the species abundance of the same sample could be estimated multiple times by models built based on different training samples. We evaluated estimation performance using absolute bias, i.e., the difference between ground truth and estimated species abundance. We further averaged absolute bias for the same sample when assessing performance. The robustness of the estimation of each sample was calculated using the standard deviation of predicted species abundance.

3 RESULTS

3.1 In Silico Analysis

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

3.2 In Vitro Analysis

Prior to the mesocosm experiment, we used baits designed from 11 target segments to enrich DNA from tissues of *A. hasta* and *T. bifasciatus*, including target segments and their flanking regions (Table S1). The result revealed the successful enrichment of the 11 targets in all samples.

After quality check and PCR duplication removal, an average of 5,304,806 bp reads for *A. hasta* and 46,355 bp for *T. bifasciatus* remained. Among those, an average of 16,354 bp reads (9%) were mapped to the assembly of enriched sequences, with an average sequencing depth of 119 for A. hasta and 35 for *T. bifasciatus* (Table S2). The length of aligned sequences of *A. hasta* was 9,183 bp on average. For *T. bifasciatus*, the length of aligned sequences was 6,469 bp on average. We observed 507 segregating sites in *A. hasta*, and 163 segregating sites in *T. bifasciatus* respectively. The raw reads and the assembled sequences are available at NCBI Sequence Read Archive (SRA) with accession numbers PRJNA1100443.

Using *Mugilogobius abei* as an outgroup (NCBI accession: NC_023353.1), phylogenetic analyses showed that samples were grouped into two clades. The sequences of *A. hasta* were clustered with the sequence of *A. hasta* retrieved from GenBank (NCBI accession: AY486321). However, the samples of *T. bifasciatus* were grouped with *T. trigonocephalus* (NCBI accession: NC_029738) as their close sister species (Fig. S1).

We examined the relationship between the number of segregating sites and species abundance using enriched sequences from tissues. Regression analysis revealed that the number of segregating sites was positively correlated with the species abundance in both *A. hasta* and *T. bifasciatus* (Fig. 3). R^2 values for regression models of both species exceeded 0.9. In contrast to the in silico study, the absolute value of species abundance fit better with models than log transformed form since the number of segregating sites observed in collected individuals was far from reaching saturation.

3.3 In Situ (Mesocosm) Experiments

In the in vitro study, the regression results for *A. hasta* were better than those for *T. bifasciatus*, so we only used *A. hasta* in mesocosm experiments. We enriched target sequences from eDNA

samples using baits designed based on 11 targeted segments. After quality filtering and deduplication, 164,145,279 reads remained. Among those, 20,571 reads (0.59 %) were successfully mapped to the majority consensus assembly of *A. hasta* from in vitro analysis, with an average sequencing depth of 197 (Table S3). Subsequently, a depth filter was applied to remove potential false positive polymorphic sites with sequencing depth below seven and minor allele frequency below 3%. The raw reads are available at NCBI Sequence Read Archive (SRA) with accession numbers PRJNA1100453.

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

The cross-validation results indicated that the number of segregating sites is a more accurate and robust proxy for estimating species abundance. The median absolute bias of species abundance predicted based on the number of segregating sites was slightly lower than the one derived from eDNA concentration. This difference in median absolute bias became more pronounced when fewer samples were used as training data. Additionally, the upper limit of bias significantly decreased when predicting species abundance using the number of segregating sites compared to eDNA concentration (Fig. 5A). We evaluated the robustness of species abundance

estimation by examining the standard deviation of predictions for the same sample across different training datasets. As anticipated, we observed increasing standard deviation when fewer samples were served as training data. The standard deviation of species abundance estimates based on segregating sites was consistently lower than those predicted using eDNA concentration (Fig. 5B).

4 DISCUSSION

Estimating the species abundance in natural ecosystems could provide vital insights into a wide array of evolutionary and ecological questions, and also hold implications for effective conservation strategies. Emerging eDNA technologies, as a non-invasive and rapid approach, represent a promising complement to traditional methods for estimating species abundance. In this study, we proposed using the number of segregating sites from eDNA as a robust proxy to estimate species abundance. In silico, in vitro and in situ (mesocosm) experiments revealed a strong positive relationship between the number of segregating sites and species abundance. The cross-validation test results based on mesocosm data showed that the number of segregating sites is a more accurate and robust proxy than eDNA concentration for estimating species.

Our in silico and in vitro experiments showed that the number of segregating sites was highly correlated with the number of individuals ($P < 0.01$, $R^2 > 0.9$). The results also indicated that more genetic variation in the sequences might provide better correlation results. For instance, in the in silico results, data simulated under a higher mutation rate exhibited a stronger correlation between segregating sites and individuals, with R^2 values of 0.99, 0.98, and 0.91 for mutation rates of 10-6 /bp/gen, 10-7 /bp/gen, and 10-8 /bp/gen, respectively. Similarly, in vitro experiment, *A. hasta* had more segregating sites than *T. bifasciatus* had (507 vs. 163), and the correlation between the number of individuals and the number of segregating sites was higher in

A. hasta than in *T. bifasciatus* ($R^2 = 0.98$ vs. $R^2 = 0.91$). Mesocosm experiment results showed that the number of individuals predicted the number of segregating sites ($R^2 = 0.70$) better than predicting the eDNA concentration ($R^2 = 0.46$) as expected. Cross-validation tests showed more accurate and robust estimation when using the number of segregating sites as a proxy compared to eDNA concentration. But counter-intuitively, both the number of segregating sites and the number of DNA copies were not heavily affected by body size or feeding/non-feeding treatment in this study, which was inconsistent with previous results (Klymus et al., 2015; Mizumoto et al., 2018). Moreover, eDNA concentrations measured in our study did not show overdispersed distribution as described in the previous study (Chambert et al., 2018). In cross-validation analysis, our estimate of species abundance using segregating sites did not significantly outperform the one based on eDNA concentration. One possible explanation is that the experiment duration was too short, and the behavior of the fish was disrupted, so the size of the fish and their feeding behavior had erratic effects on releasing DNA. Alternatively, unforeseen factors, such as placing aquarium proximity to corridors or disturbances, may have led to varying stress responses and DNA release from the fish. Notably, our study has yet to estimate species abundance in field tests, so the results from the in silico, in vitro, and mesocosm experiments must be interpreted with caution. Undoubtedly, our study is an essential step toward accurate and robust estimates of species abundance using eDNA.

The problem of using eDNA concentration to quantify species abundance has been challenged many times. Several methods, such as modeling overdispersed data (Chambert et al., 2018), seeking correlations between rank biomass and rank read abundance (Kelly et al., 2014), or using the number of sequence reads per species and the proportion of sampling sites in which a species was detected with eDNA (i.e., site occupancy) as proxies for abundance (Hänfling et

al., 2016) have been proposed. All those methods are based on eDNA concentration, which are likely to fail if the factors affecting eDNA concentration vary among samples.

Nevertheless, the predominant studies still use eDNA concentration as a proxy to estimate species abundance. These studies attempted to establish a correlation between eDNA concentration and species abundance, and only some tried to estimate species abundance from the correlation. However, uncertainty in the distribution of eDNA concentration arising from various biotic and abiotic factors may introduce bias into estimates. For example, the size of animals (Doi et al., 2017), temperature (Lacoursière-Roussel et al., 2016), season (Laramie et al., 2015; Ostberg et al., 2018), and presence of eggs may affect the release of eDNA (Dunn et al., 2017), whereas the flow of water may change the spatial distribution of eDNA (Jane et al., 2015). Besides factors related to eDNA releasing, other variables may cause variation in the estimation of copy numbers of eDNA, such as volume of water samples, methods of filtering, DNA extraction, PCR protocols, and number of PCR replicates (Deiner et al., 2015; Schultz & Lance, 2015; Takahara et al., 2015). Nevertheless, those factors should not change the number of segregating sites, because the number of segregating sites only increases along with genetic variation but not the eDNA concentration per se. As long as there is enough sampling and sequencing depth, whether there are more or fewer DNA copies will not change the number of segregating sites detected.

Another challenge in DNA quantification methods is how big the area the estimates could be applied. When one liter of water sample was taken, the estimates certainly should not be implied for the number of animals that live in that one liter of water, but how far the estimates can be projected is a difficult question. Furthermore, if water samples were taken from two adjacent sites, the estimated numbers might not be additive. Therefore, we cannot add up the

numbers because the animals may travel back and forth, and their DNA copies may be sampled at both sites. Using the number of segregating sites to estimate species abundance may mitigate this problem. The number of segregating sites would reach a plateau that should reflect the species abundance of the area with increasing numbers of samples.

A couple of limitations need to be addressed when applying the segregating-site method. Firstly, there are no global criteria for thresholding read depth and minor allele frequency to distinguish polymorphism from sequencing errors. We set an ad hoc criterion of more than seven reads and at least 3% minor-allele reads to call a segregating site. Alternatively, "Duplex Sequencing" may be applied to discriminate genuine polymorphic sites from sequencing errors (Schmitt et al., 2012). Secondly, a pilot survey of the genetic diversity of the population has to be done before extrapolating the correlation between the number of segregating sites and the number of individuals in actual samples. There is often a limited number of individuals sampled in the pilot study. Therefore, future research should explore the optimal number of individuals for the pilot study and how much we can extrapolate the correlation between the number of individuals and segregation sites from those pilot study samples.

eDNA techniques provide a non-invasive and effective solution for biomonitoring, providing insight into ecological and evolutionary processes in the wild. Qualitative studies, such as species detection, are well-established using eDNA, while quantitative studies that require accurate species abundance estimates still pose challenges. Our study suggested using the number of segregating sites as a less biased proxy compared to eDNA concentration, representing a crucial step toward achieving accurate and robust species abundance estimates from eDNA. The following steps in developing this method will include (1) accurately identifying rare alleles with low frequencies from eDNA data, (2) extended testing of different

species in mesocosms and the fields for longer durations, and (3) investigating effective sampling strategies for pilot studies.

ACKNOWLEDGEMENTS

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

DATA ACCESSIBILITY AND BENEFIT SHARING

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

AUTHOR CONTRIBUTIONS

All authors conceived and designed the study, interpreted the results, and contributed to writing the manuscript. Q.A., H.Y. and Y.W. conducted the study and collected specimens. Q.A. and H.Y. completed laboratory work and analyzed the data. All authors approved the manuscript for publication.

REFERENCE

Brys, R., Halfmaerten, D., Neyrinck, S., Mauvisseau, Q., Auwerx, J., Sweet, M., & Mergeay, J. (2021). Reliable eDNA detection and quantification of the European weather loach (Misgurnus fossilis). *Journal of Fish Biology*, *98*(2), 399–414. https://doi.org/10.1111/jfb.14315

Bylemans, J., Furlan, E. M., Pearce, L., Daly, T., & Gleeson, D. M. (2016). Improving the

containment of a freshwater invader using environmental DNA (eDNA) based monitoring. *Biological Invasions*, *18*(10), 3081–3089. https://doi.org/10.1007/s10530-016-1203-5

Chambert, T., Pilliod, D. S., Goldberg, C. S., Doi, H., & Takahara, T. (2018). An analytical framework for estimating aquatic species density from environmental DNA. *Ecology and Evolution*, *8*(6), 3468–3477. https://doi.org/10.1002/ece3.3764

Deiner, K., Walser, J.-C., Mächler, E., & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, *183*, 53–63. https://doi.org/10.1016/j.biocon.2014.11.018

DiBattista, J. D., Coker, D. J., Sinclair-Taylor, T. H., Stat, M., Berumen, M. L., & Bunce, M. (2017). Assessing the utility of eDNA as a tool to survey reef-fish communities in the Red Sea. *Coral Reefs*, *36*(4), 1245–1252. https://doi.org/10.1007/s00338-017-1618-1

Doi, H., Inui, R., Akamatsu, Y., Kanno, K., Yamanaka, H., Takahara, T., & Minamoto, T. (2017). Environmental DNA analysis for estimating the abundance and biomass of stream fish. *Freshwater Biology*, *62*(1), 30–39. https://doi.org/10.1111/fwb.12846

Dunn, N., Priestley, V., Herraiz, A., Arnold, R., & Savolainen, V. (2017). Behavior and season affect crayfish detection and density inference using environmental DNA. *Ecology and Evolution*, *7*(19), 7777–7785. https://doi.org/10.1002/ece3.3316

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Excoffier, L., & Foll, M. (2011). fastsimcoal: A continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*,

27(9), 1332–1334. https://doi.org/10.1093/bioinformatics/btr124

Forsström, T., & Vasemägi, A. (2016). Can environmental DNA (eDNA) be used for detection and monitoring of introduced crab species in the Baltic Sea? *Marine Pollution Bulletin*, *109*(1), 350–355. https://doi.org/10.1016/j.marpolbul.2016.05.054

Fukumoto, S., Ushimaru, A., & Minamoto, T. (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), 358–365. https://doi.org/10.1111/1365-2664.12392

Furlan, E. M., & Gleeson, D. (2016). Environmental DNA detection of redfin perch, Perca fluviatilis. *Conservation Genetics Resources*, *8*(2), 115–118. https://doi.org/10.1007/s12686-016-0523-1

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), Article 7. https://doi.org/10.1038/nbt.1883

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., Blackman, R. C., Oliver, A., & Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, *25*(13), 3101–3119. https://doi.org/10.1111/mec.13660

Hosler, D. (2017). Where is the body? Dreissenid mussels, raw water testing, and the real value of environmental DNA. *Management of Biological Invasions*, *8*(3), 335–341. https://doi.org/10.3391/mbi.2017.8.3.07

Jane, S. F., Wilcox, T. M., McKelvey, K. S., Young, M. K., Schwartz, M. K., Lowe, W. H., Letcher, B. H., & Whiteley, A. R. (2015). Distance, flow and PCR inhibition: eDNA dynamics in two headwater streams. *Molecular Ecology Resources*, *15*(1), 216–227. https://doi.org/10.1111/1755-0998.12285

Jo, T., Murakami, H., Yamamoto, S., Masuda, R., & Minamoto, T. (2019). Effect of water temperature and fish biomass on environmental DNA shedding, degradation, and size distribution. *Ecology and Evolution*, *9*(3), 1135–1146. https://doi.org/10.1002/ece3.4802

Karlsson, E., Ogonowski, M., Sundblad, G., Sundin, J., Svensson, O., Nousiainen, I., & Vasemägi, A. (2022). Strong positive relationships between eDNA concentrations and biomass in juvenile and adult pike (Esox lucius) under controlled conditions: Implications for monitoring. *Environmental DNA*, *4*(4), 881–893. https://doi.org/10.1002/edn3.298

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using Environmental DNA to Census Marine Fishes in a Large Mesocosm. *PLOS ONE*, *9*(1), e86175. https://doi.org/10.1371/journal.pone.0086175

Klymus, K. E., Richter, C. A., Chapman, D. C., & Paukert, C. (2015). Quantification of eDNA shedding rates from invasive bighead carp Hypophthalmichthys nobilis and silver carp Hypophthalmichthys molitrix. *Biological Conservation*, *183*, 77–84. https://doi.org/10.1016/j.biocon.2014.11.020

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874. https://doi.org/10.1093/molbev/msw054

Lacoursière-Roussel, A., Rosabal, M., & Bernatchez, L. (2016). Estimating fish abundance and

biomass from eDNA concentrations: Variability among capture methods and environmental conditions. *Molecular Ecology Resources*, *16*(6), 1401–1414. https://doi.org/10.1111/1755-0998.12522

Laramie, M. B., Pilliod, D. S., & Goldberg, C. S. (2015). Characterizing the distribution of an endangered salmonid using environmental DNA analysis. *Biological Conservation*, *183*, 29–37. https://doi.org/10.1016/j.biocon.2014.11.025

Li, C., Hofreiter, M., Straube, N., Corrigan, S., & Naylor, G. J. P. (2013). Capturing protein-coding genes across highly divergent species. *BioTechniques*, *54*(6), 321–326. https://doi.org/10.2144/000114039

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Madduppa, H., Cahyani, N. K. D., Anggoro, A. W., Subhan, B., Jefri, E., Sani, L. M. I., Arafat, D., Akbar, N., & Bengen, D. G. (2021). eDNA metabarcoding illuminates species diversity and composition of three phyla (chordata, mollusca and echinodermata) across Indonesian coral reefs. *Biodiversity and Conservation*, *30*(11), 3087–3114. https://doi.org/10.1007/s10531-021-02237-0

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M., & Minamoto, T. (2014). The Release Rate of Environmental DNA from Juvenile and Adult Fish. *PLOS ONE*, *9*(12), e114639. https://doi.org/10.1371/journal.pone.0114639

Mizumoto, H., Urabe, H., Kanbe, T., Fukushima, M., & Araki, H. (2018). Establishing an environmental DNA method to detect and estimate the biomass of Sakhalin taimen, a critically endangered Asian salmonid. *Limnology*, *19*(2), 219–227. https://doi.org/10.1007/s10201-017-0535-x

Ostberg, C. O., Chase, D. M., Hayes, M. C., & Duda, J. J. (2018). Distribution and seasonal differences in Pacific Lamprey and Lampetra spp eDNA across 18 Puget Sound watersheds. *PeerJ*, *6*, e4496. https://doi.org/10.7717/peerj.4496

Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., & Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, *25*(2), 527–541. https://doi.org/10.1111/mec.13481

Robson, H. L. A., Noble, T. H., Saunders, R. J., Robson, S. K. A., Burrows, D. W., & Jerry, D. R. (2016). Fine-tuning for the tropics: Application of eDNA technology for invasive fish detection in tropical freshwater ecosystems. *Molecular Ecology Resources*, *16*(4), 922–932. https://doi.org/10.1111/1755-0998.12505

Sakata, M. K., Maki, N., Sugiyama, H., & Minamoto, T. (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. https://doi.org/10.1007/s00114-017-1521-1

Sassoubre, L. M., Yamahara, K. M., Gardner, L. D., Block, B. A., & Boehm, A. B. (2016). Quantification of Environmental DNA (eDNA) Shedding and Decay Rates for Three Marine Fish. *Environmental Science & Technology*, *50*(19), 10456–10464. https://doi.org/10.1021/acs.est.6b03114

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., & Loeb, L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences*, *109*(36), 14508–14513. https://doi.org/10.1073/pnas.1208715109

Schultz, M. T., & Lance, R. F. (2015). Modeling the Sensitivity of Field Surveys for Detection of

Environmental DNA (eDNA). *PLOS ONE*, *10*(10), e0141503. https://doi.org/10.1371/journal.pone.0141503

Shinzato, C., Zayasu, Y., Kanda, M., Kawamitsu, M., Satoh, N., Yamashita, H., & Suzuki, G. (2018). Using Seawater to Document Coral-Zoothanthella Diversity: A New Approach to Coral Reef Monitoring Using Environmental DNA. *Frontiers in Marine Science*, *5*. https://www.frontiersin.org/articles/10.3389/fmars.2018.00028

Sigsgaard, E. E., Carl, H., Møller, P. R., & Thomsen, P. F. (2015). Monitoring the near-extinct European weather loach in Denmark based on environmental DNA from water samples. *Biological Conservation*, *183*, 46–52. https://doi.org/10.1016/j.biocon.2014.11.023

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stata Statistical Software: Release 18. (2023). [Computer software]. College Station, TX: StataCorp LLC.

Tajima, F. (1989). The effect of change in population size on DNA polymorphism. *Genetics*, *123*(3), 597–601. https://doi.org/10.1093/genetics/123.3.597

Takahara, T., Minamoto, T., & Doi, H. (2015). Effects of sample processing on the detection rate of environmental DNA from the Common Carp (Cyprinus carpio). *Biological Conservation*, *183*, 64–69. https://doi.org/10.1016/j.biocon.2014.11.014

Takahashi, T., & Tajima, F. (2017). The amount of DNA polymorphism when population size changes linearly. *Genes & Genetic Systems*, *92*(1), 55–57. https://doi.org/10.1266/ggs.16-00055

Takeuchi, A., Iijima, T., Kakuzen, W., Watanabe, S., Yamada, Y., Okamura, A., Horie, N.,

Mikawa, N., Miller, M. J., Kojima, T., & Tsukamoto, K. (2019). Release of eDNA by different life history stages and during spawning activities of laboratory-reared Japanese eels for interpretation of oceanic survey data. *Scientific Reports*, *9*(1), Article 1. https://doi.org/10.1038/s41598-019-42641-9

Uchii, K., Doi, H., & Minamoto, T. (2016). A novel environmental DNA approach to quantify the cryptic invasion of non-native genotypes. *Molecular Ecology Resources*, *16*(2), 415–422. https://doi.org/10.1111/1755-0998.12460

Ushio, M., Fukuda, H., Inoue, T., Makoto, K., Kishida, O., Sato, K., Murata, K., Nikaido, M., Sado, T., Sato, Y., Takeshita, M., Iwasaki, W., Yamanaka, H., Kondoh, M., & Miya, M. (2017). Environmental DNA enables detection of terrestrial mammals from forest pond water. *Molecular Ecology Resources*, *17*(6), e63–e75. https://doi.org/10.1111/1755-0998.12690

Van Holm, W., Ghesquière, J., Boon, N., Verspecht, T., Bernaerts, K., Zayed, N., Chatzigiannidou, I., & Teughels, W. (2021). A Viability Quantitative PCR Dilemma: Are Longer Amplicons Better? *Applied and Environmental Microbiology*, *87*(5), e02653-20. https://doi.org/10.1128/AEM.02653-20

Yusishen, M. E., Eichorn, F.-C., Anderson, W. G., & Docker, M. F. (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. https://doi.org/10.1007/s12686-018-1054-8

TABLES AND FIGURES

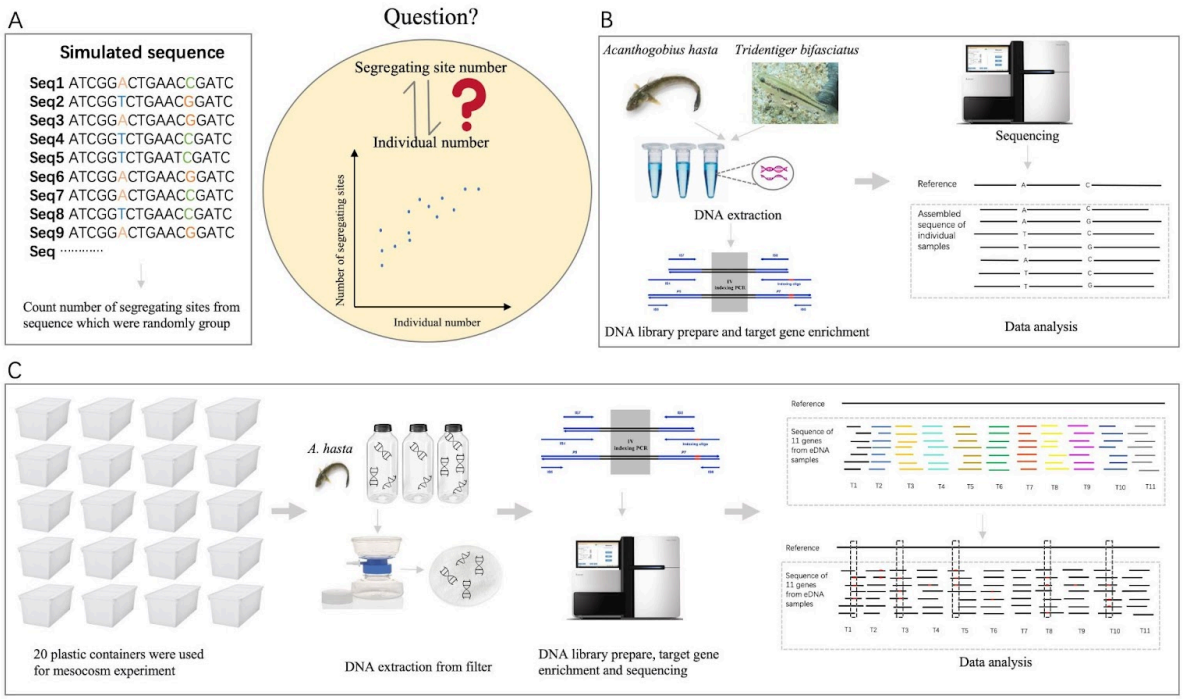

Fig. 1. Three main components of this study. (A) In silico study. We tested the correlation between the number of segregation sites and the species abundance using simulated data. (B) In vitro study. We designed markers and synthesized baits to capture target genes from tissues of *A. hasta* and *T. bifasciatus* individuals. Dataset mixed from captured sequences were used to assess the correlation between the number of segregation sites and the species abundance. (C) In situ (mescosm) study. We conducted controlled mescosm experiments using *A. hasta*. Segregation sites were extracted from captured sequences in eDNA samples, and used to evaluate its correlation to species abundance and its performance of estimating species abundance.

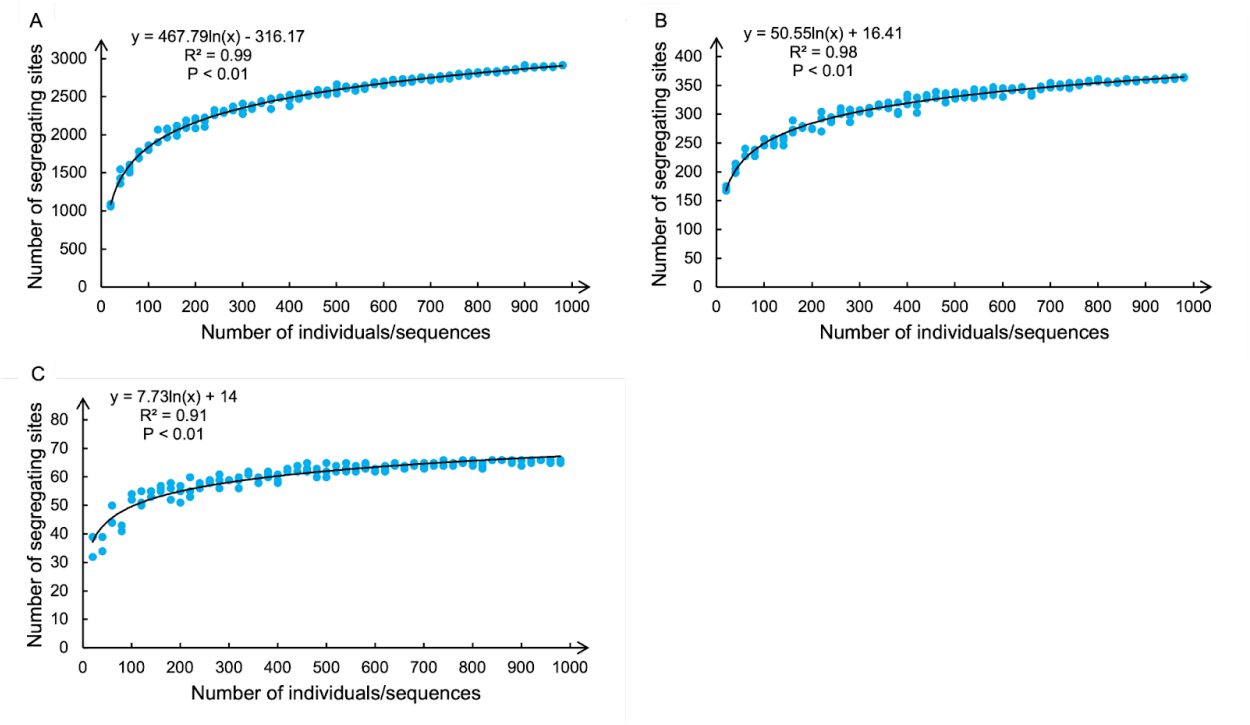

Fig. 2. Relationship between the number of segregating sites and the number of simulated individuals/sequences from in-silico analysis. Sequences of 1,000 individuals, each with the length of 17,000 bp, were simulated with a mutation rate of 10^{-6} /bp/gen (A), 10^{-7} /bp/gen (B) and 10^{-8} /bp/gen (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. Fitted line represents the relationship the number of segregating sites and the species abundance.

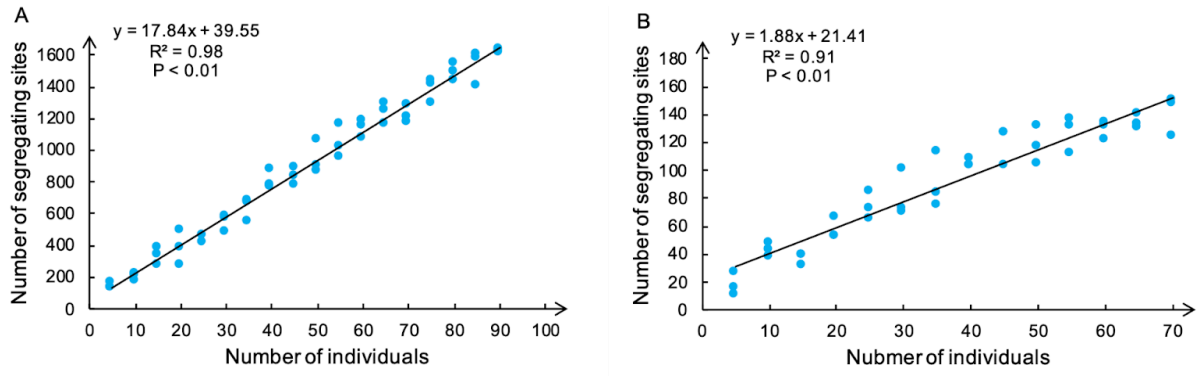

Fig. 3. Relationship between the number of individuals and the number of segregating sites observed in the assembled sequences of *A. hasta* and *T. bifasciatus*. For *A. hasta*, 5 to 90 individuals were randomly sampled from the population, then counted the number of segregation sites (A). Similarly, for *T. bifasciatus*, 5 to 70 were randomly sampled from the population, then the number of segregation 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.

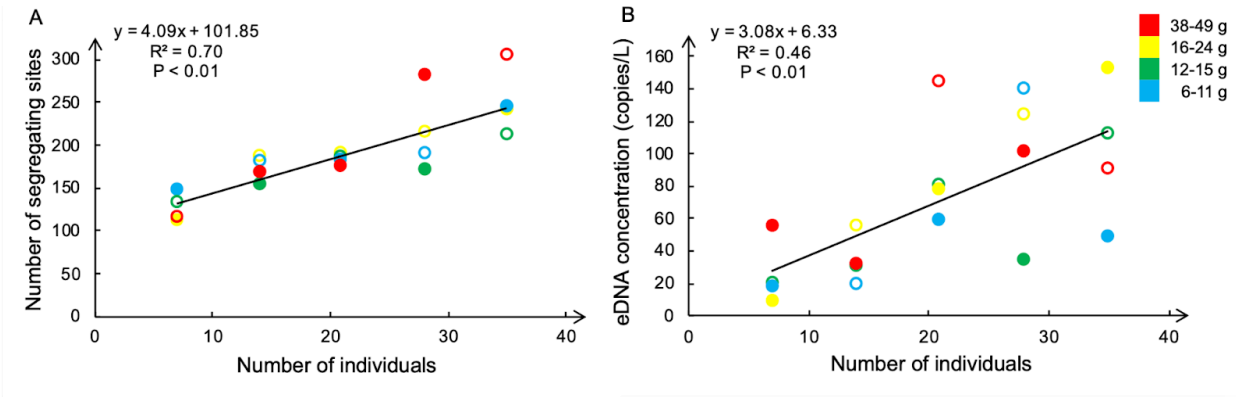

Fig. 4. Relationship between the number of individuals and the number of segregating sites in eDNA samples (A), or between the number of individuals and the eDNA concentration (B). The eDNA concentration was estimated based on the number of DNA copies at locus T2 using qPCR. Colors of the dots depict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

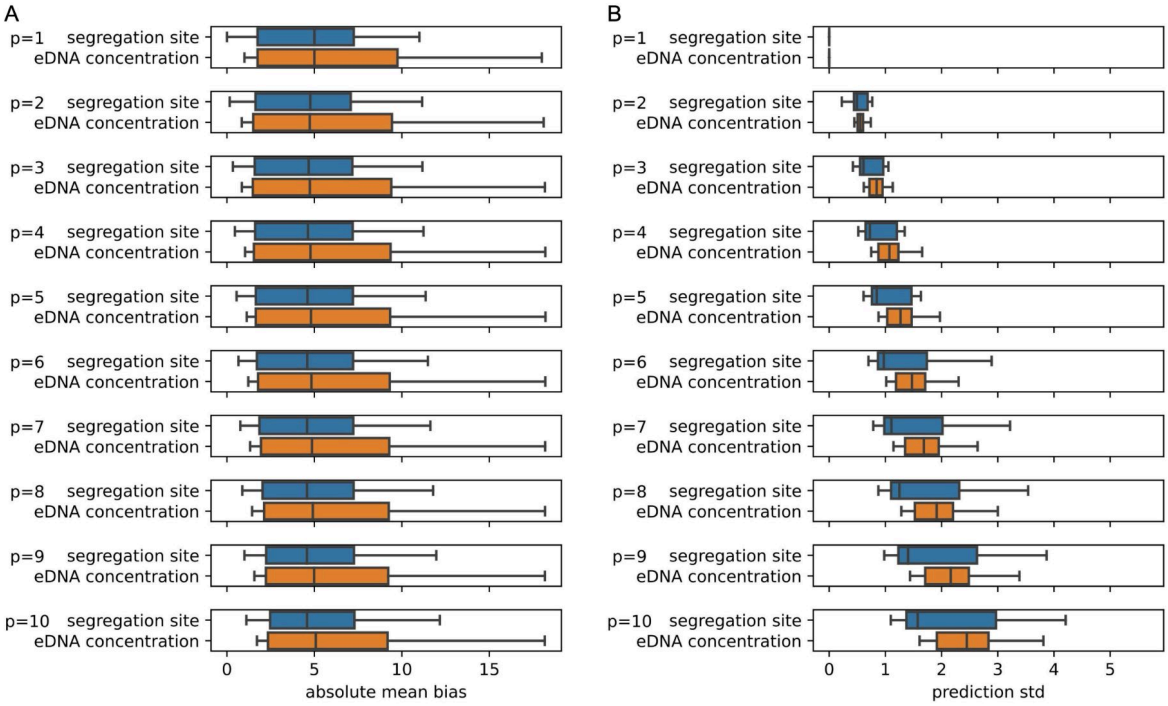

Fig. 5. Cross-validation test results for predicting species abundance. Boxplots show (A) absolute mean bias and (B) standard deviation of predicted species abundance based on segregation sites (blue) and eDNA concentration (orange) with increasing size of testing sets (top to bottom, 1 to 10 samples).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

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)

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

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

[c]The number in parentheses is the average weight of fish in this aquarium. The number in subscript denotes the rank of the average weight of fish in the treatment group. Smaller rank in subscript means lower average fish weight in this aquarium.

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

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

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