Testing the applicability of environmental DNA metabarcoding to landscape genetics

Souta Nakajima[1] and Kenji Tsuri[1]

[1]Public Works Research Institute

April 22, 2024

Abstract

Landscape genetics is a field dealing with local genetic differences and contributes to strategic conservation planning. Recently, environmental DNA (eDNA) metabarcoding has proven useful not only for detecting species but also for assessing genetic diversity and genetic structure on a large scale such as in phylogeography. However, it remains unclear whether eDNA analysis also has sufficient power to perform the landscape genetics, which focuses on a local scale. To reveal the applicability of eDNA to landscape genetics, we conducted an eDNA metabarcoding analysis of the mitochondrial DNA D-loop region of the fluvial sculpin *Cottus nozawae* in the upper Sorachi River in Japan and compared the results with inferences based on traditional tissue-based approaches. As a result, the spatial distribution of haplotypes was generally consistent between the eDNA- and tissue-based approaches. In addition, the genetic differentiation statistics calculated using eDNA and tissue samples were highly correlated when compared in the same genetic region. Furthermore, if an inference based on genome-wide SNP data from tissue samples was taken as known truth, the inference by eDNA was not necessarily inferior to the inference by the same region from tissue samples. Finally, we confirmed that analyses using eDNA data can reveal patterns such as isolation-by-distance shown in previous studies on this species, indicating the applicability of eDNA to basic landscape genetics. Even though some limitations remain, eDNA may have great potential for conducting basic landscape genetics.

Introduction

Landscape genetics is an integrated field aiming to understand the relationship between landscape features and microevolutionary processes that generate local genetic differences (Manel et al. 2003). This field has developed not only as a basic discipline but also as an applied science, because knowledge of local genetic structure is useful for strategic conservation planning through the identification of dispersal barriers or corridors (Sommer et al. 2013; Bowman et al. 2016). Landscape genetics is becoming increasingly important under ongoing climate change and habitat loss (Manel and Holderegger 2013; Nakajima et al. 2023).

In freshwater ecosystems, environmental DNA (eDNA) metabarcoding is rapidly developing and becoming widespread as a cost-efficient and non-invasive tool for acquiring species information (Ruppert et al. 2019; Doi and Nakamura 2023). In recent years, eDNA metabarcoding has gradually gained attention not only for species detection but also for population-based analysis using intraspecific variation (Reviewed in Adams et al. 2019; Sigsgaard et al. 2020; Andres et al. 2023b; Couton et al. 2023). Previous studies drawing population-based inferences, such as by examining genetic diversity and differentiation, reported that despite some limitations such as the lack of individual information or the difficulty of distinguishing false positives and negatives from correct data, eDNA-based analysis has high applicability, given the sampling effort of traditional population studies (Tsuji et al. 2020a; Adams et al. 2023). However, most population-based studies using eDNA to date have been experimental or, when conducted in the field, have focused on intrapopulation genetic diversity, phylogenetic relationships, or fragmentation among populations at large scales (e.g., between watersheds or dams) (Turon et al. 2020; Snyder and Stepien 2020; Weitemier et al.

2021; Tsuji et al. 2023). Notably, there is a lack of research specifically focused on landscape genetics and its metrics dealing with local genetic variation. As landscape genetics typically requires a great effort to sample a large number of individuals, the application of eDNA has the potential to significantly simplify landscape genetics research.

The analysis of landscape genetics is essentially different from that of phylogeography (Wang 2010). In inferring recent gene flow and its limiting factors, the key data is the proportion of a given genotype in the gene pool within each population (gene frequency), not the ancestry or phylogenetic relationships between genotypes/individuals (Hudson et al. 1992; Bohonak and Roderick 2001; Bohonak and Vandergast 2011). Although some analytical methods require information on individuals (e.g., sibship analysis or assignment tests), the frequency-based statistics of genetic differentiation that are most often the focus when evaluating gene flow do not require information on individuals. Furthermore, since false positives and negatives usually display low abundance, they should only have a minor effect on the calculation of these frequency-based estimates (Couton et al. 2023). Consequently, the question of whether landscape genetic statistics can be calculated would come down to whether gene frequencies within each population reflect actual frequencies. In eDNA metabarcoding, gene frequencies can be obtained as relative read counts instead of numbers of individuals (or numbers of genomes for non-haploids). Despite differences in the nature of the obtained data, previous studies performed in tanks or even in the field have shown a good congruence in gene frequencies (typically as haplotype frequencies) between eDNA- and tissue-based approaches (Sigsgaard et al. 2016; Marshall and Stepien 2019; Andres et al. 2021, 2023a; Couton et al. 2023; Wakimura et al. 2023). Therefore, landscape genetics analyses using eDNA are qualitatively considered feasible. However, it remains unclear whether the statistics of genetic differentiation calculated from eDNA samples in the field are sufficient to withstand landscape genetic analysis. At the same time, analytical treatments that have been claimed to be effective in previous population-based eDNA studies, such as the removal of low-frequency reads (Tsuji et al. 2023) or the conversion of data to semi-quantitative rankings (Turon et al. 2020), would also need to be investigated; it is unclear whether they are also effective in landscape genetics.

As a model case, we targeted the fluvial sculpin *Cottus nozawae*in the upper watershed of the Sorachi River in Japan, for which detailed landscape genetics studies had been previously conducted. In this study, we performed an eDNA metabarcoding analysis of the mitochondrial DNA (mtDNA) D-loop region of this species and compared the results with inferences based on traditional tissue-based approaches. The aims of this study are (i) to clarify whether eDNA-based local genetic structure is consistent with tissue-based inferences, (ii) to reveal the extent to which the statistics of genetic differentiation calculated by eDNA are consistent with those obtained from tissue samples, as well as to what extent the analytical treatments of eDNA datasets change the results, and (iii) to discuss the applicability of eDNA to landscape genetics.

Materials and Methods

1. Study sites and sampling

This study was conducted in tributaries in the upper section of the Sorachi River, Hokkaido, Japan (Figure 1; Table 1). In this area, Nakajima et al. (2021, 2023) conducted population sampling of *C. nozawae* individuals and investigated the population structure of this species. From the downstream of riffles or runs of 21 sites where*C. nozawae* had been sampled before, water samples of 0.8 L were collected from the surface and filtered through Millipore 0.45 μ m Sterivex-HV filters (Merck KGaA, Darmstadt, Germany) with 100 mL syringes (JMS, Tokyo, Japan) in June 2023. Note that the site Pop 19 (in Figure 1) was approximately 500 m downstream from the population sampling site. Each 0.8 L of water at every site could be filtered with one Sterivex filter, except for one site (two filters were needed in Pop 18). The Sterivex filters were immersed in 1.3 mL of RNAlater (Thermo Fisher Scientific, Massachusetts, USA) and transported to the laboratory under refrigeration. After the RNAlater was removed using QIAvac Vacuum Systems (QIAGEN, Hilden, Germany) and a vacuum pump, total DNA was extracted using the DNeasy Blood & Tissue Kit (QIAGEN). The protocol of DNA extraction was basically followed by the "ATL-S procedure" in Fukuzawa et al. (2023), except that the volumes of the first addition of buffer ATL and proteinase K were changed to 400 μ L and 40μ L, respectively, and the final elution with buffer AE was performed in a volume of 120 μ

L. The DNA sample from Pop 18 was purified using the DNeasy PowerClean Pro Cleanup Kit (QIAGEN) to remove PCR inhibitors, as amplification in subsequent PCR was difficult.

2. Tissue-based sequencing and primer design

For reference, the partial control region sequences of approximately 16 individuals per population were amplified from tissue-based genomic DNA used in previous studies. Although Nakajima et al. (2021, 2023) did not include the distant populations Pop 20 and 21, tissue samples of some individuals were actually obtained in 2020, and these samples were also used as outgroups in the present study (Table 1). First, the entire control region was amplified with the primers L-Thr (5'-AGC TCA GCG YCA GAG CGC CGG TCT TGT AA-3') and H12Sr5 (5'-TGA TAA TAA AGT CAG GAC CAA G-3') (Yokoyama and Goto 2002) using TaKaRa Ex Taq Hot Start Version (Takara Bio, Shiga, Japan) with each 10 μ L reaction containing 1.0 μ L of 10×Ex Taq Buffer, 0.8 μ L of dNTPs (each 2.5 mM), each 0.5 μ L of 10 μ M primers, 0.05 μ L of TaKaRa Ex Taq HS, and 1.0 μ L of genomic DNA. The PCR conditions were initial denaturation at 94°C for 2 min and 30 cycles of denaturation at 98°C for 10 sec, annealing at 58°C for 30 sec, and extension at 72°C for 1 min. After the PCR products were purified using the ExoSAP-IT Express PCR Product Cleanup Reagent (Thermo Fisher Scientific), the 5'-end of the control region, which is the most used region in studies of Japanese sculpin (Yamamoto 2019), was sequenced by Eurofins Genomics (Tokyo, Japan) using the reverse internal primer H16498m (5'-CCT GAA RTA GGA ACC AAA TG-3') (Yokoyama and Goto 2002). Sequence data were aligned using Clustal W (Thompson et al. 1994) implemented in BioEdit (Hall 1999), and unique haplotypes were identified with the aid of DnaSP 6 (Rozas et al. 2017). Primers for eDNA were designed to amplify both the haplotypes from this study and those reported in Yokoyama and Goto (2002) (Table S2), as well as to have an amplification product length of around 400 bp, which can be sequenced by Illumina MiSeq:

CNdloopS1_F: 5'-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN NNN GCT CAA AGA AAG GAG ATT YTA ACT C-3'

CNdloopS1_R: 5'-GTG ACT GGA GTT CAG ACG TGT GCT CTT CCG ATC TNN NNN NCC GTT GGC ATT AAG AAA TCA ACT G-3'

Furthermore, not only tissue-based data from the same D-loop region, but genome-wide SNP data, which are considered more informative, were used as a partial reference. Data from multiplexed ISSR genotyping by sequencing (MIG-seq; Suyama and Matsuki 2015) at the studied 21 sites were downloaded from the DDBJ Sequence Read Archive (DRA) under accession number DRA017315, and SNPs were identified from the dataset at the 21 sites for examining the population structure and genetic differentiations (details found in Text S1).

3. Molecular protocol for the eDNA samples

A first-round PCR (1st PCR) of eDNA was performed using KOD Plus Neo (Toyobo, Tokyo, Japan), with each 20 μ L reaction containing 2.0μ L of 10×PCR Buffer, 2.0 μ L of dNTPs (each 2 mM), 1.2μ L of $MgSO_4$ (25 mM), each 0.5 μ L of 10μ M primers, 0.4 μ L of KOD Plus Neo polymerase, and 2.0μ L of template DNA. The PCR conditions were initial denaturation at 94 °C for 2 min and 35–40 cycles of denaturation at 94°C for 15 sec, annealing at 63°C for 30 sec, and extension at 68°C for 30 sec. The number of cycles was initially 35, and samples for which no target peak was identified by electrophoresis after the second-round PCR (2nd PCR) were analyzed again at 38 cycles and then 40 cycles (Table S1). Although the number of PCR cycles was kept modest to minimize errors (Wakimura et al. 2023), the haplotype accumulation curve suggested that the sequencing coverage was sufficient to detect haplotypes in the samples (Figure S1). The 1st PCR was conducted with eight replicates per sample (20 μ L × 8), and individual replicates were pooled and purified using 160 μ L of VAHTS DNA Clean Beads (Vazyme Biotech, Nanjing, China) as templates for the 2nd PCR. The 2nd PCR was conducted using KOD FX Neo (Toyobo) and primers with appropriate unique index sequences. For this PCR, each 10 μ L reaction contained 5.0 μ L of 2×PCR Buffer, 2.0 μ L of dNTPs (each 2 mM), 0.5 μ M of each forward and reverse primer, 0.2 μ L of KOD FX Neo polymerase, and 1.0 μ L of the template. The thermal conditions were as follows: initial denaturation at 94 °C for 2 min, 12

cycles of denaturation at 98°C for 10 sec, annealing at 60°C for 30 sec, and extension at 68°C for 30 sec, and a final extension at 68°C for 2 min. The 2nd PCR products were purified again using VAHTS DNA Clean Beads, and positive bands of targeted amplicons were confirmed using the Fragment Analyzer System and the dsDNA 915 Reagent Kit (Agilent Technologies, California, USA). The prepared libraries were sequenced on the Illumina MiSeq platform using the MiSeq Reagent Kit v3 (2 × 300 cycles) (Illumina, California, USA). The procedures from the purification of 1st PCR products to this MiSeq run were performed by Bioengineering Lab. (Kanagawa, Japan). Fastq files containing raw reads were denoised using the DADA2 1.18.0 (Callahan et al. 2016), an algorithm known to have high accuracy in the context of population-level inference (Tsuji et al. 2020b; Macé et al. 2022). DADA2 was performed in Qiime2 2022.11 (Bolyen et al. 2019) with the default parameters except: *–p-trunc-len-f* 240,*–p-trunc-len-r* 180, and *–p-mim-overlap* 20. The obtained sequences were annotated by a local BLAST search against a reference database consisting of sequences of the studied species obtained from Yokoyama and Goto (2002), Ito et al. (2018), and the present study (Table S2). Sequences with an identity of 90.0% or more were picked up (but no sequences were matched at 90.0%–99.4%; see Results).

4. Population structure, genetic diversity, and genetic differentiation

Haplotype frequencies based on relative read counts (eDNA) or the number of individuals (tissue) within each site were plotted and compared. Additionally, the spatial patterns of haplotype frequencies were compared with the population structure inferred from genome-wide SNP data.

To identify overall trends in intrapopulation genetic diversity, haplotype diversity (h_S;$h_S = 1 - \sum_i {p_i}^2$, where p_i is the haplotype frequency of the i -th haplotype) and haplotype richness (hr ; number of haplotypes corrected for differences in the number of individuals or total read counts among populations; haplotype version of allelic richness (El Mousadik and Petit 1996)) of each population were calculated. For interpopulation genetic differentiation, we first calculated Nei's F_{ST} (G_{ST}; Nei 1973), as haplotype differentiation in Hudson et al. (1992) (i.e., it does not matter whether two haplotypes differ by one nucleotide or by tens). Other than G_{ST}, we also used the allele frequency distance (AFD; Berner 2019) with the modification of using haplotypes instead of alleles. AFD is considered less susceptible to differences in sample size between populations and more sensitive in the range of weak differentiation that is of interest in local scale studies (Berner 2019). In addition, this metric is identical toD_{PS} (1 - proportion of shared alleles; Bowcock et al. 1994), which is known to reflect recent gene flow (˜10 generations) particularly relevant to management and the purpose of landscape genetics (Leroy et al. 2018; Savary et al. 2021). Hereafter, we refer to this metric as D_{PS}, although this notation has not yet been used for haplotype-based analysis. The calculation process of D_{PS} for two arbitrary populations Pop A and Pop B is$D_{PS} = (\sum_i |p_{iPopA} - p_{iPopB}|)/2$, where p_{iPopA} and p_{iPopB} are the haplotype frequencies of the i -th haplotype in Pop A and Pop B, respectively. As this study focuses on local genetic differentiation reflecting recent gene flow, subsequent analyses and discussions are primarily based on D_{PS}, while analyses using more common and well-known G_{ST} statistics were also conducted.

To clarify the spatial scale at which gene flow is more dominant than genetic drift for genetic differentiation, Mantel correlograms displaying the spatial correlation of D_{PS} (as simple genetic distance available in eDNA) and geographic distance for each 7.5 km distance class until 60 km and then from 60 km to 128 km (maximum value) were generated. Within the distance classes where the correlation coefficient is significantly positive, gene flow among populations is considered to be particularly active (Diniz-Filho and De Campos Telles 2002). The correlogram in each distance class was assessed with 9999 permutations using the package ecodist 2.0.9 (Goslee and Urban 2007) in R 4.2.1 (R Core Team 2022).

Genetic differentiation statistics were also calculated by SNP data (Text S1). Generally, genetic diversity and genetic drift pressure differ between nuclear DNA and mtDNA (Toews and Brelsford 2012; Morin et al. 2018; Saitoh 2021). However, at the spatial scale at which gene flow is dominant, because the theoretical pattern of being inversely proportional to migration is common (Allendorf et al. 2022), genetic differentiation should be synchronized regardless of markers in the absence of sex differences. On the spatial scale explicitly stated, the more informative SNP data were used as the *known truth.*

5. Comparison of eDNA- and tissue-based genetic differentiation

Correlations between the metrics calculated by eDNA- and tissue-based approaches were examined for each statistic of genetic diversity and differentiation. For genetic differentiation, statistics were calculated in the three datasets: the entire dataset, the dataset upstream of the Kanayama dam (upstream dataset), and the dataset only between population pairs with a waterway distance of <15 km (15-km dataset). The upstream dataset was used to exclude the effects of a long spatial gap and a dam, and the 15-km dataset was used to reveal the pattern at the spatial scale where gene flow is especially dominant (Nakajima et al. 2021). Correlations of genetic differentiations for the entire dataset and upstream dataset were tested by Mantel tests (9999 permutations and 10000 times bootstrapping to estimate confidence intervals) using the package ecodist in R, and the statistics for datasets not in matrix style (genetic differentiation in the 15-km dataset and genetic diversity) were tested by simple bootstrapping (10000 times) using the package CarletonStats 2.2 (Chihara and Loy 2023).

6. Differences by the additional data treatment

Since data obtained in eDNA metabarcoding are different in nature from conventional tissue-based approaches and contain erroneous sequences, previous studies have indicated that data filtering or the conversion of relative read counts to semi-quantitative rankings are effective in population-based analysis (although their scope is mainly phylogeography). To investigate their effectiveness in landscape genetics analysis, D_{PS} and G_{ST} were calculated from the eDNA dataset in which the following treatments were performed and compared to tissue-based statistics: treatment 1, haplotypes with a low frequency ($p_i < 0.01$) in each population were removed from the data (replaced to 0 reads) (Tsuji et al. 2023); treatment 2, haplotypes with less than half of the proportion of the most predominant haplotype in each population ($p_i < max\ p_i/2$) were removed from the data (Tsuji et al. 2023); treatment 3, haplotype frequency was converted into the following semi-quantitative rankings: rank 1 if $p_i \leq 0.5$; rank 2 if $0.5 < p_i \leq 0.75$; rank 3 if$0.75 < p_i \leq 0.9$; rank 4 if $0.9 < p_i$ (Turon et al. 2020).

7. Application to basic landscape genetics

Finally, to verify that eDNA can address basic landscape genetics questions, we examined whether two patterns previously shown in the targeted system (Nakajima et al. 2021) were also detected by analysis using eDNA. These patterns were: (i) genetic differentiation is significantly correlated with waterway geographic distance (isolation-by-distance) and (ii) genetic differentiation is not correlated with differences in summer water temperature between populations. From the 19 populations except Pop 20 and Pop 21 where water temperature data are unavailable and 13 populations in the upstream dataset (assuming local scale studies), correlations ofD_{PS} with waterway geographic distance and water temperature differences were calculated by Mantel tests with 9999 permutations.

Results

An average of 2848 reads per sample were assigned to the reference of*C. nozawae* (Table 1), all with sequence identity of >99.4%. A total of 66 haplotypes were detected from eDNA, whereas 58 haplotypes were detected from tissue samples. Among the haplotypes detected from eDNA, 35 haplotypes matched with 100% identity to haplotypes detected from tissue samples, and the other haplotypes matched with 99.4%–99.7% identity (representing differences of 1–2 bases). The number of haplotypes detected in each population was 2–17 (mean: 8.62) in eDNA and 1–11 (mean: 6.43) in tissue samples, and the haplotype diversity h_{S} ranged from 0.01–0.83 (mean: 0.57) in eDNA and 0.00–0.89 (mean: 0.64) in tissue samples. Global G_{ST} was 0.33 in eDNA and 0.29 in tissue samples, and the average D_{PS} was 0.86 in both eDNA and tissue samples. Pairwise G_{ST} ranged from 0.01–0.90 (eDNA) or 0.01–0.89 (tissue), and D_{PS}ranged from 0.17–1.00 (eDNA) or 0.25–1.00 (tissue) (Table S3).

The spatial distribution of haplotypes was generally consistent between eDNA and tissue samples (Figure 2). Some major patterns of the spatial structure could be commonly found, such as the dominance of a unique haplotype in Pop 3, the presence of haplotypes commonly found at many sites in Pop 1–11, and Pop 12

being composed almost entirely of one haplotype also found in the downstream sites (Pop 13–15). Compared with the results of the STRUCTURE analysis using SNP data, some patterns such as the presence of spatial structure or the uniqueness of Pop 3 were common, but there were some differences from D-loop region results, such as the absence of a clear differentiation between Pop 1–11 and Pop 12–13 in SNP data.

The statistics of genetic diversity and differentiation calculated from eDNA were positively correlated with those calculated from tissue samples (Figures 3 and S2). For genetic differentiation, the correlation coefficient was r = 0.73 ($p < 0.001$) for G_{ST} and r = 0.83 ($p < 0.001$) for D_{PS}. From the spatial autocorrelation analysis by the Mantel correlogram, r values were consistently positive up to approximately 30 km and significant up to at least 15 km in all genetic markers, indicating that gene flow can be considered particularly strong in this range. Correlations of genetic differentiation calculated between eDNA and SNP data were low when calculated for all 21 populations, but there was some correlation when calculated only for population pairs less than 15 km apart (Figure 3, Table 2).

Additional data treatment of the eDNA dataset did not increase the correlation of its statistics with tissue-based statistics (Tables 2 and S4). Treatment 2, which involved bold data removal, significantly worsened the correlation in most datasets (95% CI not overlapping). In terms of the correlation coefficient alone, treatment 3 (semi-quantitative approach) sometimes outperformed the no treatment dataset, although the differences were not significant.

From the basic analysis including spatial and water temperature data, the detection patterns from our eDNA data were congruent with those of previous studies, with a significant correlation between genetic differentiation and waterway distance (r = 0.50, $p < 0.001$ for 19 populations; r = 0.52, $p < 0.01$ for upstream populations; Figures 4 and S3) and no significant correlation between genetic differentiation and water temperature differences (r = 0.20, $p = 0.16$ for 19 populations; r = 0.32, $p = 0.06$ for upstream populations).

Discussion

In this study, we investigated the local genetic structure of *C. nozawae* from eDNA analysis for the D-loop region and compared the results with those obtained from tissue samples, to reveal the applicability of eDNA to landscape genetics.

The haplotype distributions obtained from eDNA and tissue samples showed similar patterns, although there were some differences in each haplotype frequency. Major spatial patterns, including the presence of genetically isolated sites and differences corresponding to spatial structure, were also detected from both approaches. Note that the population structure inferred from the D-loop region is only based on a single locus; thus, the strength of the population structure was not as apparent as that inferred from genome-wide SNP data, which is commonly used in present-day tissue studies (Figure 2). Nevertheless, showing the haplotype distribution obtained from eDNA should be useful for understanding a rough spatial pattern within an entire watershed.

All statistics of genetic diversity and differentiation calculated from eDNA were significantly correlated with those obtained from tissue samples. Regarding genetic diversity, the correlation with tissue-based approaches was lower in hr than h_S (Figure S2). This is likely because statistics based on the number of haplotypes tend to be more affected by potentially erroneous sequences generated in next-generation sequencing than those based on gene frequencies. As for genetic differentiation, the main subject of this study, both statistics used were based on gene frequencies and were highly correlated between eDNA- and tissue-based calculations using the same marker. The correlation coefficients were similar to those of a previous study that calculated a statistic identical to D_{PS} from eDNA and found a correlation of r = 0.76 with tissue samples (Andres et al. 2023a). What is novel and particularly important in this study is the fact that these high correlations are observed not only for the entire dataset but also for datasets featuring nearby sites (upstream dataset and 15-km dataset) (Table 2). This indicates that genetic differentiation can be calculated with nearly the same accuracy as for tissue samples even at spatial scales where gene flow is the dominant factor in shaping the strength of genetic differentiation. On the other hand, some distant sites did not share any haplotypes

(i.e.,$D_{\mathrm{PS}} = 1$; Table S3), and comparisons of the strength of genetic differentiation between such pairs of sites are deemed difficult in the used marker.

Another important aspect in local scale inference is that eDNA-based differentiation was not necessarily inferior to tissue-based differentiation using the same region when inferences based on SNP data are regarded as more reliable (see D_{PS} in upstream dataset and 15-km dataset in Table 2). In population genetics, a higher number of individuals x loci will lead to better accuracy in inference, and the inference obtained from small samples introduces biases associated with the individuals sampled. In eDNA, although there are usually fewer available loci, it is possible that the samples contain information from many individuals (Tsuji et al. 2020a), which may be more reflective of the "gene pool" in each site. How many individuals are generally reflected in eDNA samples is still under study (Couton et al. 2023), but because the studied area displays high population densities in *C. nozawae* (about 50 individuals per 100 m^2; Suzuki et al. 2021), information on a large number of individuals might be obtained from eDNA. Nevertheless, a clear difference in the accuracy between the eDNA- and tissue-based approaches could not be identified. The take-home message at this time is that eDNA analysis is not necessarily inferior to the tissue-based analysis from approximately 16 individuals per population when the same markers are used.

When exploring regional haplotypes or their new distributions from the perspective of phylogeography, it is paramount to minimize false positives and negatives (Turon et al. 2020; Tsuji et al. 2023). However, genetic differentiation at the local scale, such as that addressed in this study, is fundamentally dependent on gene frequencies within each population, and their overall spatial patterns or relationship to the environment are the subject of analysis. In addition, although the effect of haplotypes detected at low frequencies on statistics is relatively low, the sharing of the same haplotype at low frequencies at multiple sites sometimes provides important information for the comparison of the strength of gene flow (Slatkin and Barton 1989). On the other hand, since denoising primarily removes erroneous sequences that occur in sync with the correct sequences (Callahan et al. 2016), the remaining false positives are likely to be randomly distributed in small amounts across all samples and may not have a large impact on inferences. For these reasons, bulk removal of low-frequency sequences may not be a beneficial option in landscape genetics. Conversely, it may be worth considering the results that a semi-quantitative approach (treatment 3) did not decrease the correlation with tissue samples but rather increased it in some datasets (Table 2). The fact that the accuracy did not change much when detailed values were converted to rough values suggests that the calculated statistics should be used to capture overall trends rather than to find meaning in slight differences in values.

Future challenges and perspectives

This study revealed the feasibility and usefulness of eDNA in landscape genetics. Even if the degree of accuracy demonstrated in this study is accepted by users as a usable level, there are certain limitations that cannot be immediately resolved. The first is the low information content of short DNA fragments available in eDNA. This study used a single region of approximately 400 bp of mtDNA, but longer sequences would be difficult to amplify due to the degradation of DNA in water (Jo et al. 2017). In addition, mtDNA, which is generally used in eDNA studies, has limited information that is equivalent to that of a single locus (Andres et al. 2023b). Landscape genetics studies after the 2010s are commonly conducted by analyzing multiple independent loci (Wang 2011). For more accurate analysis, it is essential to use loci in nuclear DNA, which is not linked to mtDNA and is regarded as independent loci (Couton et al. 2023). The use of nuclear DNA in eDNA is still challenging, however, this issue could be solved through the implementation of microsatellite and SNP analysis currently under development (Andres et al. 2021; Liu et al. 2024). Another limitation is that eDNA cannot provide individual information. The statistics calculated in this study do not require individual information because they deal with gene frequency in gene pools, and eDNA sampling may even be possible to reduce bias due to the selection of some individuals in the population. However, in order to address more complex biological questions in the field, analyses requiring individual information will sometimes be required (Couton et al. 2023). Actually, in the studied species in the Sorachi River, source-sink structures (asymmetric gene flow) from low-temperature tributaries to high-temperature tributaries have been detected through an analysis using the theory of assignment test (Paetkau et al. 1995),

which assigns individuals to groups (Nakajima et al. 2021), but this analysis cannot be applied without individual information.

Nevertheless, this study shows that eDNA has the potential to perform basic analyses via summary statistics with the same level of accuracy as that seen in tissue samples. Given that the "basic landscape genetics," as practiced in the 2000s when only short sequence data could be handled, could be rapidly performed through cost-efficient and non-invasive sampling, the usefulness of eDNA analyses is evident. With the increasing number of available markers, landscape genetics using eDNA is expected to be more useful and be utilized for various purposes such as the evaluation and conservation of habitat connectivity under environmental changes.

Acknowledgments

We are grateful to Noriko Amagai for her support with the laboratory work. Some of the tissue-based D-loop sequences were obtained with the help of Kaiji Suzuki, Masanao Sueyoshi, and Futoshi Nakamura. We thank Kazutaka Sakiya and Keiko Muraoka for their continuous support. This study was supported by JSPS KAKENHI Grant Number JP23K11515.

References

Adams CIM, Hepburn C, Jeunen G, et al (2023) Environmental DNA reflects common haplotypic variation. Environ DNA 5:906–919. https://doi.org/10.1002/edn3.352

Adams CIM, Knapp M, Gemmell NJ, et al (2019) Beyond Biodiversity: Can environmental DNA (eDNA) cut it as a population genetics tool? Genes 10:192. https://doi.org/10.3390/genes10030192

Allendorf FW, Funk WC, Aitken SN, et al (2022) Population Subdivision. In: Allendorf FW, Funk WC, Aitken SN, et al. (eds) Conservation and the Genomics of Populations. Oxford University Press, Oxford, pp 172–203

Andres KJ, Lodge DM, Andres J (2023a) Environmental DNA reveals the genetic diversity and population structure of an invasive species in the Laurentian Great Lakes. Proc Natl Acad Sci 120:e2307345120. https://doi.org/10.1073/pnas.2307345120

Andres KJ, Lodge DM, Sethi SA, Andres J (2023b) Detecting and analysing intraspecific genetic variation with eDNA: From population genetics to species abundance. Mol Ecol 32:4118–4132. https://doi.org/10.1111/mec.17031

Andres KJ, Sethi SA, Lodge DM, Andres J (2021) Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. Mol Ecol 30:685–697. https://doi.org/10.1111/mec.15765

Berner (2019) Allele Frequency Difference *AFD* –An intuitive alternative to F_{ST} for quantifying genetic population differentiation. Genes 10:308. https://doi.org/10.3390/genes10040308

Bohonak AJ, Roderick GK (2001) Dispersal of invertebrates among temporary ponds: Are genetic estimates accurate? Isr J Zool 47:367–386

Bohonak AJ, Vandergast AG (2011) The value of DNA sequence data for studying landscape genetics. Mol Ecol 20:2477–2479. https://doi.org/10.1111/j.1365-294X.2011.05122.x

Bolyen E, Rideout JR, Dillon MR, et al (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9

Bowcock AM, Ruiz-Linares A, Tomfohrde J, et al (1994) High resolution of human evolutionary trees with polymorphic microsatellites. Nature 368:455–457. https://doi.org/10.1038/368455a0

Bowman J, Greenhorn JE, Marrotte RR, et al (2016) On applications of landscape genetics. Conserv Genet 17:753–760. https://doi.org/10.1007/s10592-016-0834-5

Callahan BJ, McMurdie PJ, Rosen MJ, et al (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583. https://doi.org/10.1038/nmeth.3869

Chihara L, Loy A (2023) CarletonStats: Functions for Statistics Classes at Carleton College

Couton M, Viard F, Altermatt F (2023) Opportunities and inherent limits of using environmental DNA for population genetics. Environ DNA 5:1048–1064. https://doi.org/10.1002/edn3.448

Diniz-Filho JAF, De Campos Telles MP (2002) Spatial autocorrelation analysis and the identification of operational units for conservation in continuous populations. Conserv Biol 16:924–935. https://doi.org/10.1046/j.1523-1739.2002.00295.x

Doi H, Nakamura K (2023) Special issue: Environmental DNA as a practical tool for aquatic conservation and restoration. Landsc Ecol Eng 19:1–2. https://doi.org/10.1007/s11355-022-00534-6

El Mousadik A, Petit RJ (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. Theor Appl Genet 92:832–839. https://doi.org/10.1007/BF00221895

Fukuzawa T, Shirakura H, Nishizawa N, et al (2023) Environmental DNA extraction method from water for a high and consistent DNA yield. Environ DNA 5:627–633. https://doi.org/10.1002/edn3.406

Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. J Stat Softw 22:1–19. https://doi.org/10.18637/jss.v022.i07

Hall T (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser 41:95–98.

Hudson RR, Boos DD, Norman L. Kaplan (1992) A statistical test for detecting geographic subdivision. Mol Biol Evol 9:138–151. https://doi.org/10.1093/oxfordjournals.molbev.a040703

Ito N, Gotoh RO, Shirakuma T, et al (2018) Genetic structure of glacial-relict populations of a freshwater sculpin, *Cottus nozawae* , in Yamagata area of the Tohoku district. Biogeography 20:96–102. https://doi.org/10.11358/biogeo.20.96

Jo T, Murakami H, Masuda R, et al (2017) Rapid degradation of longer DNA fragments enables the improved estimation of distribution and biomass using environmental DNA. Mol Ecol Resour 17:e25–e33. https://doi.org/10.1111/1755-0998.12685

Leroy G, Carroll EL, Bruford MW, et al (2018) Next-generation metrics for monitoring genetic erosion within populations of conservation concern. Evol Appl 11:1066–1083. https://doi.org/10.1111/eva.12564

Liu Z, Kishe MA, Gabagambi NP, et al (2024) Nuclear environmental DNA resolves fine-scale population genetic structure in an aquatic habitat. iScience 27:108669. https://doi.org/10.1016/j.isci.2023.108669

Mace B, Hocde R, Marques V, et al (2022) Evaluating bioinformatics pipelines for population-level inference using environmental DNA. Environ DNA 4:674–686. https://doi.org/10.1002/edn3.269

Manel S, Holderegger R (2013) Ten years of landscape genetics. Trends Ecol Evol 28:614–621. https://doi.org/10.1016/j.tree.2013.05.012

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol 18:189–197. https://doi.org/10.1016/S0169-5347(03)00008-9

Marshall NT, Stepien CA (2019) Invasion genetics from eDNA and thousands of larvae: A targeted metabarcoding assay that distinguishes species and population variation of zebra and quagga mussels. Ecol Evol 9:3515–3538. https://doi.org/10.1002/ece3.4985

Morin PA, Foote AD, Baker CS, et al (2018) Demography or selection on linked cultural traits or genes? Investigating the driver of low mtDNA diversity in the sperm whale using complementary mitochondrial and

nuclear genome analyses. Mol Ecol 27:2604–2619. https://doi.org/10.1111/mec.14698

Nakajima S, Sueyoshi M, Hirota SK, et al (2021) A strategic sampling design revealed the local genetic structure of cold-water fluvial sculpin: a focus on groundwater-dependent water temperature heterogeneity. Heredity 127:413–422. https://doi.org/10.1038/s41437-021-00468-z

Nakajima S, Suzuki H, Nakatsugawa M, et al (2023) Inferring future changes in gene flow under climate change in riverscapes: a pilot case study in fluvial sculpin. Landsc Ecol 38:1351–1362. https://doi.org/10.1007/s10980-023-01633-x

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci 70:3321–3323. https://doi.org/10.1073/pnas.70.12.3321

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. Mol Ecol 4:347–354. https://doi.org/10.1111/j.1365-294X.1995.tb00227.x

R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, et al (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 34:3299–3302. https://doi.org/10.1093/molbev/msx248

Ruppert KM, Kline RJ, Rahman MS (2019) Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Glob Ecol Conserv 17:e00547. https://doi.org/10.1016/j.gecco.2019.e00547

Saitoh T (2021) High variation of mitochondrial DNA diversity as compared to nuclear microsatellites in mammalian populations. Ecol Res 36:206–220. https://doi.org/10.1111/1440-1703.12190

Savary P, Foltete J, Moal H, et al (2021) Analysing landscape effects on dispersal networks and gene flow with genetic graphs. Mol Ecol Resour 21:1167–1185. https://doi.org/10.1111/1755-0998.13333

Sigsgaard EE, Jensen MR, Winkelmann IE, et al (2020) Population-level inferences from environmental DNA—Current status and future perspectives. Evol Appl 13:245–262. https://doi.org/10.1111/eva.12882

Sigsgaard EE, Nielsen IB, Bach SS, et al (2016) Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. Nat Ecol Evol 1:0004. https://doi.org/10.1038/s41559-016-0004

Slatkin M, Barton NH (1989) A comparison of three indirect methods for estimating average levels of gene flow. Evolution 43:1349–1368. https://doi.org/10.1111/j.1558-5646.1989.tb02587.x

Snyder MR, Stepien CA (2020) Increasing confidence for discerning species and population compositions from metabarcoding assays of environmental samples: case studies of fishes in the Laurentian Great Lakes and Wabash River. Metabarcoding Metagenomics 4:e53455. https://doi.org/10.3897/mbmg.4.53455

Sommer S, McDevitt AD, Balkenhol N (2013) Landscape genetic approaches in conservation biology and management. Conserv Genet 14:249–251. https://doi.org/10.1007/s10592-013-0473-z

Suyama Y, Matsuki Y (2015) MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. Sci Rep 5:16963. https://doi.org/10.1038/srep16963

Suzuki K, Ishiyama N, Koizumi I, Nakamura F (2021) Combined effects of summer water temperature and current velocity on the distribution of a cold-water-adapted sculpin (*Cottus nozawae*). Water 13:975. https://doi.org/10.3390/w13070975

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680. https://doi.org/10.1093/nar/22.22.4673

Toews DPL, Brelsford A (2012) The biogeography of mitochondrial and nuclear discordance in animals. Mol Ecol 21:3907–3930. https://doi.org/10.1111/j.1365-294X.2012.05664.x

Tsuji S, Maruyama A, Miya M, et al (2020a) Environmental DNA analysis shows high potential as a tool for estimating intraspecific genetic diversity in a wild fish population. Mol Ecol Resour 20:1248–1258. https://doi.org/10.1111/1755-0998.13165

Tsuji S, Miya M, Ushio M, et al (2020b) Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. Environ DNA 2:42–52. https://doi.org/10.1002/edn3.44

Tsuji S, Shibata N, Inui R, et al (2023) Environmental DNA phylogeography: Successful reconstruction of phylogeographic patterns of multiple fish species from cups of water. Mol Ecol Resour 23:1050–1065. https://doi.org/10.1111/1755-0998.13772

Turon X, Antich A, Palacin C, et al (2020) From metabarcoding to metaphylogeography: separating the wheat from the chaff. Ecol Appl 30:e02036. https://doi.org/10.1002/eap.2036

Wakimura K, Uchii K, Kikko T (2023) Evaluation of genetic diversity in an endangered fish *Gnathopogon caerulescens* using environmental DNA and its potential use in fish conservation. Environ DNA 5:973–986. https://doi.org/10.1002/edn3.408

Wang IJ (2010) Recognizing the temporal distinctions between landscape genetics and phylogeography. Mol Ecol 19:2605–2608. https://doi.org/10.1111/j.1365-294X.2010.04715.x

Wang IJ (2011) Choosing appropriate genetic markers and analytical methods for testing landscape genetic hypotheses. Mol Ecol 20:2480–2482. https://doi.org/10.1111/j.1365-294X.2011.05123.x

Weitemier K, Penaluna BE, Hauck LL, et al (2021) Estimating the genetic diversity of Pacific salmon and trout using multigene eDNA metabarcoding. Mol Ecol 30:4970–4990. https://doi.org/10.1111/mec.15811

Yamamoto S (2019) Genetic population structure of Japanese river sculpin *Cottus pollux* (Cottidae) large-egg type, inferred from mitochondrial DNA sequences. J Fish Biol 94:325–329. https://doi.org/10.1111/jfb.13890

Yokoyama R, Goto A (2002) Phylogeography of a freshwater sculpin, *Cottus nozawae* , from the northeastern part of Honshu Island, Japan. Ichthyol Res 49:147–155. https://doi.org/10.1007/s102280200019

Data Accessibility and Benefit-Sharing

Data Accessibility Statement: Raw MiSeq reads were deposited in the DRA under the accession number DRA017646. D-loop sequences from tissue samples are available in the DDBJ/EMBL/GenBank databases under accession numbers LC591956–LC591978, LC591981, and LC790392–LC790425. Other data are available on Figshare (doi: 10.6084/m9.figshare.25305250).

Benefit-Sharing Statement: This study was conducted in Japan by Japanese researchers affiliated with a Japanese institute; the Nagoya Protocol is not applicable. Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Author Contributions

S.N. conceived and designed the study. S.N. and K.T. performed the field sampling. S.N. led the laboratory work. S.N. analyzed the data and wrote the manuscript, with contributions from K.T.

Tables

Table 1 Details of sampling sites and data used in this study.

Pop ID	Longitude (E°)	Latitude (N°)	eDNA (D-loop)	eDNA (D-loop)	eDNA (D-loop)	Tissue (D-loop)	Tissue
			Reads	*hr*	h_S	n	*hr*

Pop ID	Longitude (E°)	Latitude (N°)	eDNA (D-loop)	eDNA (D-loop)	eDNA (D-loop)	Tissue (D-loop)	Tissue
1	142.670	43.057	2015	10.000	0.830	16	5.762
2	142.672	43.085	1044	8.998	0.706	16	6.681
3	142.699	43.107	1379	1.984	0.012	17	1.529
4	142.703	43.218	1997	6.000	0.683	16	5.412
5	142.699	43.218	1631	4.996	0.338	16	3.683
6	142.678	43.163	1278	6.000	0.756	16	5.274
7	142.699	43.155	3535	6.490	0.206	16	2.688
8	142.680	43.133	2290	9.983	0.406	16	5.874
9	142.739	43.136	3570	10.833	0.743	16	5.083
10	142.680	43.129	4024	16.968	0.747	16	5.724
11	142.624	43.137	2394	8.987	0.533	18	5.529
12	142.665	43.196	1404	2.000	0.091	14	1.000
13	142.603	43.173	2207	9.000	0.485	16	3.625
14	142.346	43.135	7422	16.679	0.782	16	5.913
15	142.393	43.149	3013	15.990	0.753	16	6.324
16	142.558	43.207	2730	6.000	0.747	16	4.305
17	142.501	43.216	3569	6.994	0.632	15	5.468
18	142.401	43.216	556	2.000	0.498	15	4.653
19	142.403	43.218	2201	15.990	0.778	16	7.350
20	142.543	43.323	7166	7.999	0.628	9	2.000
21	142.241	43.625	4382	5.995	0.552	9	3.000

Reads, number of sequence reads assigned to each site and studied species after BLAST; n, number of individuals; hr , haplotype richness; h_{S}, haplotype diversity;H_{E}, expected heterozygosity

Table 2 Summary of the correlation between eDNA- and tissue-based genetic differentiation statistics.

			Correlat
	Stat.	Treatment	Entire d
(A) eDNA vs tissue (both D-loop)	(A) eDNA vs tissue (both D-loop)	(A) eDNA vs tissue (both D-loop)	(A) eDN
	G_{ST}	No treatment	0.734 [0
		Treatment 1	0.728 [0
		Treatment 2	0.496 [0
		Treatment 3	0.757 [0
	D_{PS}	No treatment	0.825 [0
		Treatment 1	0.823 [0
		Treatment 2	0.673 [0
		Treatment 3	0.784 [0
(B) eDNA (D-loop) vs tissue (SNP)	(B) eDNA (D-loop) vs tissue (SNP)	(B) eDNA (D-loop) vs tissue (SNP)	(B) eDN
	G_{ST}	No treatment	0.091 [0
	D_{PS}	No treatment	0.456 [0
(C) tissue (D-loop) vs tissue (SNP)	(C) tissue (D-loop) vs tissue (SNP)	(C) tissue (D-loop) vs tissue (SNP)	(C) tissu
	G_{ST}	No treatment	0.299 [0
	D_{PS}	No treatment	0.487 [0

Upstream dataset, only 13 populations upstream of the Kanayama dam; 15-km dataset, only between population pairs with a waterway distance of <15 km. The confidence interval (95% CI) in the entire dataset and upstream dataset was obtained by bootstrapping testing in Mantel tests, while the 95% CI for the

15-km dataset, which is not matrix style, was obtained by simple bootstrapping. In the comparison of the same region, the results calculated using eDNA datasets obtained after the three types of treatment are also displayed: Treatment 1, removal of low read frequencies; Treatment 2, bold removal of low read frequencies; Treatment 3, conversion to semi-quantitative rankings. See Table S4 for the full combinations.

Figure Legends

Figure 1 Sampling localities. The number on the map indicates the population ID. Pop 1–13 are located upstream of the Kanayama dam and can be particularly assumed to be within the local scale. The map was created using the National Land Numerical Information from the MLIT of Japan (nlftp.mlit.go.jp/)

Figure 2 Spatial population structure among 21 sites estimated by three types of markers. (A, B) The haplotype compositions in each site estimated by the D-loop region using eDNA-based (A) and tissue-based (B) approaches. Haplotype names correspond to Table S2. Within each site, boxes separated by black lines indicate different haplotypes, and the box size reflects haplotype frequency. Only the haplotypes with high frequency (the sum of p_i [?] 0.5 in either eDNA or tissue) are shown in color (except white). Different colors except white indicate different haplotypes. (C) Population structure inferred by the STRUCTURE analysis using SNP data obtained by MIG-seq. Barplots display the proportion of the membership coefficients in the inferred subpopulations (clusters) at K = 6 for all individuals. The number indicates the population ID.

Figure 3 Congruence of eDNA- and tissue-based genetic differentiation. (A) Mantel correlograms showing the spatial autocorrelation of genetic data. The r values obtained from the three types of markers are displayed. Filled markers indicate significant correlations from a null model of spatial structure ($p< 0.05$). (B, C) Correlation of genetic differentiation statistics calculated by haplotype frequencies estimated from eDNA- and tissue-based approaches for statistics D_{PS} (B) and G_{ST}(C). (D, E) Correlation between genetic differentiation statistics calculated by eDNA samples and genome-wide SNP data for statisticsD_{PS} (D) and G_{ST} (E). Filled circles indicate the population pairs less than 15 km apart, and correlations in this 15-km dataset are also shown.

Figure 4 Relationship of eDNA-based genetic differentiation with geographic distance and water temperature in the 13 sites upstream of the Kanayama dam. Scatter plots of D_{PS} versus waterway distance (A) and summer water temperature differences (B) between populations are displayed.