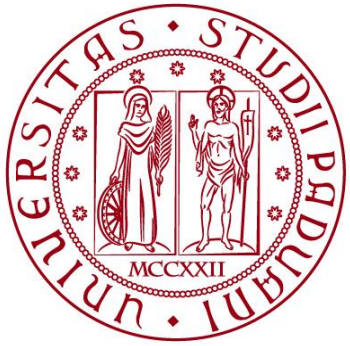
UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI BIOLOGIA

Corso di Laurea in Biologia Molecolare



Elaborato di Laurea

Il DNA ambientale (eDNA) nucleare stima le frequenze alleliche e l'abbondanza di una popolazione in mesocosmi sperimentali ed esperimenti sul campo

Tutor: Prof. Leonardo Congiu
Dipartimento di Biologia

Laureanda: Anna Maria Placentino

Anno accademico 2021/2022

Indice

1. Abstract

Gli approcci basati sul DNA ambientale (eDNA) costituiscono una frontiera in rapida e progressiva espansione nell'ambito della biologia della conservazione. Nello studio in esame, si dimostra come gli approcci basati sull'eDNA, associati a metodiche di multiplex-PCR, siano in grado di rilevare la variabilità genetica intraspecifica non solo a livello di siti variabili all'interno del genoma mitocondriale, ma anche nel genoma nucleare, consentendo così di valutare le frequenze degli alleli a livello di popolazione e di stimare l'abbondanza assoluta di una specie in un campione di eDNA. Allestendo un pannello di loci microsatelliti per la specie *Neogobius melanostomus*, ed effettuando campionamenti sia in mesocosmi sperimentali, sia in campo, sono state confrontate le frequenze degli alleli ottenute a partire da eDNA e campioni di tessuto della specie target. In entrambi i tipi di campionamento è emersa una significativa correlazione tra le stime delle frequenze alleliche ottenute da eDNA e quelle ottenute da campioni di tessuto, ad indicare come anche i marcatori nucleari possano venire amplificati in modo affidabile a partire da campioni ambientali. Pur necessitando di opportune validazioni, questo approccio pone le basi per la possibilità di condurre studi di genetica di popolazione basati su DNA ambientale, e, per questo, non invasivi.

2. Stato dell'arte

Gli organismi diffondono costantemente DNA nell'ambiente circostante attraverso le cellule e i prodotti di scarto che perdono ed espellono, includendo, tra questi, feci, muco, gameti, peli e pelle [2]. Tale DNA va sotto il nome di "DNA ambientale", abbreviato in "eDNA". L'impiego del DNA ambientale isolato da campioni come sedimenti antichi e terrestri, carote di ghiaccio ed ecosistemi acquatici, sta rivoluzionando lo studio della diversità e della distribuzione degli organismi. L'eDNA, infatti, garantisce agli studiosi la possibilità di monitorare la biodiversità in maniera non invasiva e sempre più sensibile e specifica, anche grazie alla capacità di rilevare organismi in bassa numerosità [3]. Gli approcci basati sul DNA ambientale si sono dimostrati efficienti nel rilevare la presenza/assenza delle specie, come pure nel restituire dati relativi ai loro habitat e alla loro varietà, distribuzione geografica e abbondanza relativa, portando all'affermazione dell'eDNA come valido, robusto ed economico strumento per gli studi sulla biodiversità, sull'ecologia di comunità e sulla biologia della conservazione. Inoltre, grazie alla maggiore sensibilità e ai costi ridotti, gli approcci basati sull'eDNA sono da preferire per il monitoraggio di specie acquatiche invasive, la cui individuazione tempestiva è funzionale alla loro gestione o eradicazione.
La maggior parte degli studi sul DNA ambientale finora condotti ha analizzato la diversità biologica a livello di specie o di ranghi tassonomici più elevati. Solo recentemente si sono cominciate ad indagare le prestazioni dell'eDNA come metodo per ottenere informazioni relativamente alla genetica di popolazione. A tal proposito, alcuni studi hanno guidato la messa a punto di approcci per la rilevazione della variabilità genetica intraspecifica nel genoma mitocondriale partendo da campioni ambientali; una futura espansione di tali approcci all'impiego di marcatori nucleari, come i microsatelliti o i polimorfismi a singolo nucleotide (SNP), promette di migliorare la capacità di condurre inferenze genetiche a livello di popolazione basandosi su eDNA.

2.1 Tecniche e marcatori molecolari applicati al DNA ambientale

2.1.1 Next- Generation Sequencing (NGS)

Il significativo sviluppo delle tecniche di next- generation sequencing (NGS) nel corso degli ultimi decenni ha reso possibile la combinazione dell'eDNA con i più avanzati metodi molecolari: ciò ha portato allo sviluppo dell'eDNA metabarcoding, che unifica il sequenziamento rapido e simultaneo di milioni di sequenze alla capacità di associare le sequenze ottenute dall'eDNA ad uno specifico taxa [4], e grazie a cui si è in grado di catturare contemporaneamente, mediante l'utilizzo di primer specifici per un certo gruppo tassonomico, o di un singolo primer disegnato su una porzione genica sufficientemente variabile, la presenza di taxa diversi in un campione [5]. In aggiunta, la massiva quantità di dati derivanti da approcci di NGS permette agli studiosi di osservare piccoli cambiamenti incorsi nelle comunità a seguito di fluttuazioni antropiche o naturali delle condizioni ambientali, impossibili

da riscontrare utilizzando metodi molecolari meno sensibili e tradizionali, come il sequenziamento Sanger [6].

2.1.2 DNA mitocondriale (mtDNA)

Il DNA mitocondriale rappresenta il marcatore genetico più utilizzato negli studi di diversità genetica in generale. Il fatto che abbia eredità matrilineare fa sì che possa essere considerato non ricombinante: di conseguenza, il mtDNA rimane invariato da una generazione all'altra a meno di non accumulare mutazioni, e ciò implica che il differenziamento avvenga solo sulla base di accumulo di mutazioni. L'aploidia ne permette una facile amplificazione, e consente di analizzare i dati di sequenziamento con risoluzione ottimale. Le regioni conservate del genoma mitocondriale possono venire utilizzate per confronti tra specie diverse oppure tra categorie sistematiche superiori; le due regioni ipervariabili, a livello della sequenza di controllo della replicazione, sono invece utilizzate per confronti tra individui della stessa specie. Essendo i mitocondri presenti in copia multipla per cellula, e possedendo ciascuno un proprio genoma in molte copie, il mtDNA rende possibile analizzare campioni molto critici sia per quantità che per conservazione. Tale aspetto fa inoltre sì che in ambienti acquatici possa rinvenirsi in concentrazioni superiori rispetto al DNA nucleare, e avere per questo una maggiore probabilità di rilevamento.

2.1.3 Marcatori microsatelliti

Essendo il DNA mitocondriale informativo del solo contributo materno, il suo utilizzo esclusivo potrebbe precludere una visione completa della distribuzione della variabilità genetica in una specie [7]. I marcatori nucleari consentono di considerare anche i caratteri ereditati in linea maschile, e, quindi, di ricostruire i processi a carico dell'intero genoma [7]. Tra i più utilizzati negli studi genetici vi sono i microsatelliti, piccole regioni di DNA ripetute un numero di volte compreso tra 3 e 80, molto frequenti nei genomi eucarioti, altresì noti come SSRs, STRs o SSLPs. Sono altamente polimorfi, e la loro variabilità deriva dal numero diverso di ripetizioni. La variabilità dei microsatelliti viene analizzata per PCR, poiché le sequenze fiancheggianti sono conservate all'interno delle specie. I microsatelliti sono marcatori codominanti, permettendo di distinguere omozigoti ed eterozigoti. L'elevata variabilità li rende molto utili in studi su sistemi a bassa distanza genetica.

2.2 Il DNA ambientale come strumento per la quantificazione delle specie

La biomassa delle specie costituisce un parametro fondamentale per la stima della produttività primaria e del ciclo di materiali di un ecosistema [8]. La conoscenza della biomassa risulta importante per la conservazione delle specie a rischio e per il controllo delle dimensioni delle popolazioni, soprattutto in presenza di specie aliene invasive. A dispetto della sua centralità negli studi di carattere ecologico, la biomassa di una specie, soprattutto se acquatica, risulta difficile da determinare con

precisione [8]. Lo sviluppo di metodi basati sull'eDNA per la quantificazione dell'abbondanza delle specie potrebbe portare, in tal senso, ad un miglioramento nella gestione delle specie invasive, e, più in generale, ad un monitoraggio rapido, semplice e non invasivo dello stato delle popolazioni.

2.2.1 Stima dell'abbondanza delle specie attraverso la concentrazione del DNA ambientale

Ipotizzando una relazione di proporzionalità diretta tra la quantità di eDNA rilasciato in acqua da vertebrati acquatici e la loro biomassa, quest'ultima potrebbe venire stimata semplicemente misurando il numero di copie di eDNA in un campione di acqua [8]. Tuttavia, se la correlazione tra concentrazione di eDNA e biomassa di una specie consente di ottenere un indice di abbondanza relativa affidabile, l'accuratezza e la precisione di questo metodo nel determinare l'abbondanza assoluta delle specie si sono rivelate difficili da indagare. Infatti, su tali correlazioni potrebbero significativamente influire bias nell'amplificazione dei templati taxa- specifiche, o fattori locali biotici e abiotici che influenzano la quantità di eDNA secreto da un organismo. Bisogna inoltre considerare come il tasso di produzione di eDNA possa variare in base alle dimensioni, al comportamento e al metabolismo degli individui, e come tali proprietà dipendano a loro volta dalle caratteristiche dell'ambiente.

2.2.2 Stima dell'abbondanza delle specie attraverso il metodo "DNA mixture estimators"

Un metodo innovativo per la conta degli individui in studi di carattere ecologico, originariamente sviluppato per le analisi forensi, si basa sull'impiego di miscele di DNA proveniente da più individui. Le miscele di DNA, infatti, contengono al loro interno una firma diretta dei contributori genetici originari, e rappresentano una recente interessante frontiera per la quantificazione dell'abbondanza assoluta di una specie [9]. Il numero totale di alleli che si ricava dai genotipi presenti in una miscela di DNA riflette quello degli individui che partecipano alla data miscela [9]. Possedendo informazioni in merito alle frequenze degli alleli di una popolazione, è altresì possibile applicare un modello statistico basato sul metodo della massima verosimiglianza che, dati i genotipi osservati, consente di inferire in maniera diretta il numero dei contributori genetici unici in una miscela [9]. Considerando organismi diploidi, il modello statistico restituisce, per un singolo locus, dati un insieme di *n* alleli distinti osservati e un insieme di frequenze alleliche associate, la probabilità che un certo numero di contributori genetici unici abbia prodotto il set di alleli osservati per quel locus; moltiplicando tra loro i valori di probabilità ottenuti per tutti i loci in esame si ottiene il numero stimato di individui che contribuiscono alla miscela. Poiché un campione ambientale può contenere DNA appartenente a più individui, il metodo degli stimatori è estendibile anche ad analisi condotte su eDNA.

2.3 Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples

Nel presente studio, si vuole esaminare in che misura sia possibile rilevare la variabilità genetica intraspecifica mediante DNA ambientale, e come questa possa venire utilizzata per la stima del numero dei contributori genetici presenti in un campione di eDNA. Partendo da campioni estratti da mesocosmi sperimentali e da esperimenti condotti sul campo, la caratterizzazione delle frequenze alleliche nella popolazione e la stima dell'abbondanza assoluta di individui della specie *Neogobius melanostomus* sono state effettuate combinando l'utilizzo di loci microsatelliti come marcatori nucleari a tecniche di sequenziamento Next- Generation (NGS). In primo luogo, è stata valutata la correlazione delle frequenze alleliche ottenute a partire da eDNA con quelle ottenute dalla genotipizzazione di singoli individui nei mesocosmi sperimentali, al fine di determinare la misura in cui i dati di sequenza recuperati dall'eDNA fossero rappresentativi degli alleli derivanti dai tessuti di *Neogobius melanostomus*. Successivamente, è stato statisticamente stimato il numero di individui geneticamente unici in ogni campione di eDNA. Infine, è stata testata la capacità del modello statistico di stimare accuratamente il numero di contributori genetici unici in combinazioni simulate di numeri variabili di individui, da un minimo di 2 fino ad un massimo di 58.

3. Approccio sperimentale

3.1 Caratterizzazione dei microsatelliti e allestimento della multiplex PCR

3.1.1 Isolamento del DNA e ligazione degli adattatori Y Stubby Illumina per la costruzione di una libreria di microsatelliti

Neogobius melanostomus, target dell'esperimento, è una specie di pesce nativa delle Steppe pontico- caspiche introdotta dall'uomo nel Nord America attraverso le acque di zavorra nel corso degli anni '90 del secolo scorso; da allora, ha cominciato a diffondere lungo i fiumi e i laghi dell'entroterra, portando al declino delle popolazioni native. Per la costruzione di una libreria di microsatelliti, 50-100 ng di DNA genomico sono stati estratti e purificati dal tessuto di 3 individui di *Neogobius melanostomus* catturati presso il lago di Cayuga, nello stato di New York, USA, e digeriti mediante gli enzimi di restrizione *Alu*I, *Rsa*I, e *Hpy166*II, che effettuano un taglio simmetrico producendo estremità piatte. I frammenti così ottenuti sono stati ligati adattatori Y "Stubby" Illumina mediante T4 DNA ligasi in presenza di 1 mM di ATP. Gli adattatori Illumina contengono, in generale, due sequenze, P5 e P7, che permettono alla libreria di legarsi alla flow cell di sequenziamento e generare cluster di cloni, due sequenze di legame per i primer di sequenziamento, Rd1 SP ed Rd2 SP, e, opzionalmente, una o due sequenze "index", identificative di un campione. Gli adattatori Y "Stubby" presentano solo le sequenze Rd1 SP ed Rd2 SP; le sequenze P5 e P7, insieme ad eventuali sequenze "index", vengono incorporate attraverso un successivo step di amplificazione. Poiché gli adattatori sono caratterizzati dalla presenza di una singola base T sporgente all'estremità 3'-terminale di ciascun filamento, i frammenti di DNA genomico sono stati

precedentemente sottoposti ad una reazione di adenilazione sfruttando l'attività polimerasica 5'-3' del frammento di Klenow exo- in presenza di dATP, al fine di ottenere estremità 3'- terminali debolmente sporgenti per una singola base A.

3.1.2 Arricchimento del bersaglio mediante ibridazione e completamento della libreria di microsatelliti

In seguito, è stata effettuato l'arricchimento del bersaglio, ovvero la cattura dei frammenti genomici contenenti microsatelliti. A tale scopo si è utilizzato un approccio di ibridazione selettiva [10]: il DNA genomico è stato ibridato a sonde 5'-biotinilate contenenti diverse ripetizioni in tandem del motivo da arricchire e legate a biglie magnetiche streptavidinate; l'applicazione di un magnete ha poi reso possibile il recupero delle biglie. I frammenti genomici contenenti microsatelliti così selezionati sono stati sottoposti ad una reazione di PCR attraverso l'enzima DNA Polimerasi *Taq*: tale reazione, mediante l'impiego di un primer universale Illumina, recante una delle sequenze P5/P7, e di un primer "index" Illumina, recante l'altra delle due sequenze e un barcode molecolare, ha completato la sequenza degli adattatori. I risultati dell'esperimento di PCR sono stati verificati mediante elettroforesi su gel di agarosio all'1.0%, e la quantificazione dei prodotti è stata effettuata utilizzando il fluorimetro Qubit 2.0. Infine, prima del sequenziamento NGS, gli ampliconi ottenuti sono stati purificati e selezionati per un range di dimensioni compreso fra 300 e 600 pb utilizzando il kit di estrazione e purificazione Agencourt AMPure XP. La libreria è stata successivamente sottoposta a sequenziamento Illumina.

3.1.3 Identificazione informatica dei microsatelliti e disegno dei primer per la multiplex-PCR

L'identificazione informatica dei microsatelliti e il disegno dei primer per la successiva amplificazione mediante multiplex- PCR sono stati effettuati utilizzando il programma MSATCOMMANDER 1.0.3. Lo sperimentatore può selezionare sia il tipo di sequenza ripetuta da ricercare, sia il numero di ripetizioni. Nel presente lavoro, il numero di ripetizioni selezionato è compreso tra 10 e 24. Viene implementato il programma PRIMER3 per il disegno ottimale dei primer. I primer sono stati qui disegnati per amplificare regioni dalle dimensioni comprese tra 410-440 pb; la loro lunghezza è compresa fra 20 e 24 pb.

3.1.4 Test dei primer

La specificità dei primer per *Neogobius melanostomus* è stata testata attraverso il software di allineamento Primer Blast di NCBI. Quindi, coppie di primer forward e reverse per 43 loci microsatelliti sono stati ordinati dall'azienda Integrated DNA Technologies. La loro funzionalità è stata saggiata allestendo singole reazioni di PCR utilizzando come templato il DNA genomico estratto dal tessuto dei tre esemplari di *Neogobius Melanostomus* sopracitati. In seguito all'esclusione dei

primer aventi regioni di complementarietà o responsabili di una resa di amplificazione sub-ottimale, sono rimasti 35 dei 43 loci microsatelliti iniziali.

3.1.5 Multiplex PCR

I 35 loci microsatelliti sono stati suddivsi tra sette array di multiplex PCR, con ogni array contenente da 4 a 6 coppie di primer. Ogni multiplex PCR è stata così programmata: un'incubazione iniziale a 95°C per 15 min, per l'attivazione della DNA polimerasi *HotStarTaq*, seguita da 35 cicli composti di uno step di denaturazione a 94°C per 30 s, di uno step di ibridazione dei primer al DNA stampo a 59°C per 90 s, e di uno step di allungamento a 72°C per 90 s. I prodotti di PCR sono stati studiati attraverso una reazione di elettroforesi su gel di agarosio all'1%, che ha confermato la presenza di ampliconi delle dimensioni attese.

3.2 Mesocosmo sperimentale

58 esemplari di *Neogobius melanostomus* sono stati catturati in un sito del lago Cayuga e posti in mesocosmi sperimentali contenenti 12 L di acqua di rubinetto a temperatura ambiente. Sono stati allestiti un mesocosmo sperimentale contenente 1 individuo, uno contenente 3 individui, uno contenente 5 individui, e uno contenente 10 individui, ciascuno riprodotto in triplicato, per un totale di 12. 2 mesocosmi sperimentali contenenti sola acqua di rubinetto a temperatura ambiente sono stati preparati come controlli negativi. Dopo un'ora, i 58 esemplari sono stati sacrificati con tricaina metansolfonato, MS-222, seguendo il protocollo ACUP 306.02 approvato dalla Cornell Institutional Animal Care and Use Committee (IACUC). Da ogni individuo sono poi stati prelevati campioni di tessuto delle pinne caudali, e il DNA è stato estratto mediante il kit di estrazione "DNeasy Blood and Tissue extraction kit" della Qiagen Inc. Al fine di campionare il DNA ambientale rilasciato, da ogni mesocosmo sono stati raccolti in duplicato 2 L di acqua utilizzando contenitori in plastica Nalgene sterilizzati. Tali campioni sono stati conservati in ghiaccio fino alla filtrazione, eseguita mediante un filtro a membrana in nitrato di cellulosa. Per la loro conservazione, i filtri sono stati immersi in 700 µl di soluzione di Longmire e stoccati a -20 °C. L'eDNA è stato poi estratto dai filtri mediante il "DNeasy Blood and Tissue extraction kit" della Qiagen Inc., seguendo un protocollo modificato rispetto all'originale.

3.3 Esperimento sul campo

15 individui di *Neogobius melanostomus* sono stati prelevati a circa 20 miglia dal sito precedente. La panmissia tra gli individui dei due siti di campionamento ha permesso di considerare, nella stima delle frequenze alleliche della popolazione basata sui tessuti, tutti i 73 pesci complessivamente catturati. A circa 50 m di distanza da tale sito sono stati prelevati, in contenitori in plastica Nalgene sterilizzati, 3 campioni di 2 L di acqua. Il controllo negativo è qui rappresentato da un campione di 2 L di acqua distillata. Per la raccolta dei campioni di tessuto, la

filtrazione dell'acqua e l'estrazione del DNA, sono stati applicati i medesimi protocolli di cui sopra.

3.4 Preparazione della libreria e sequenziamento Illumina MiSeq

Sia per il mesocosmo sperimentale che per l'esperimento su campo, i loci microsatelliti sono stati amplificati mediante multiplex PCR a partire dal DNA ambientale e dal DNA estratto dai tessuti in reazioni separate. Il numero di cicli di PCR è stato incrementato a 45 per i campioni di eDNA a causa della bassa concentrazione del templato. Per ciascuno dei tre campioni di eDNA raccolti sul campo sono state allestite tre repliche. I prodotti delle sette multiplex PCR di ogni campione sono stati raccolti in un volume di 5 microlitri e sottoposti a indicizzazione in un secondo step di PCR utilizzando il kit di preparazione di librerie NEXTERA XT di Illumina. Il protocollo prevede l'aggiunta di un tag ad entrambe le estremità del frammento di PCR, cui segue una "index" PCR mediante coppie di "index" primer contenenti ciascuno una sequenza Rd1 SP/Rd2 SP complementare al tag, una sequenza P5/P7, e una sequenza di indicizzazione univoca. Gli "index" primer 1 e 2 del kit possiedono rispettivamente 27 (N701-N728) e 19 (N502- N521) sequenze alternative di indicizzazione, cosicché la loro combinazione possa conferire a ciascun campione un'etichetta molecolare specifica. Complessivamente sono state allestite due librerie: una contenente gli ampliconi provenienti dai campioni di tessuto e dall'eDNA del mesocosmo sperimentale, e una contenente gli ampliconi provenienti dagli analoghi campioni dell'esperimento sul campo. Le librerie sono state purificate con il kit Agencourt AMPure XP, e la loro concentrazione è stata stimata utilizzando il kit Qubit "dsDNA High-Sensitivity" e il fluorimetro Qubit 2.0. Dopo essere state diluite a 4 nM, sono state sottoposte a sequenziamento paired- end Illumina utilizzando il kit "MiSeq v2 500 bp" (PE 2 × 250 bp) presso il Cornell University's Institute of Biotechnology Genomics Facility.

3.5 Analisi bioinformatica e processamento dei dati di sequenziamento

Le sequenze tecniche degli adattatori sono state rimosse utilizzando il software TRIMMOMATIC 0.39. Per l'estrazione delle reads demultiplexate forward e reverse e la loro assegnazione al locus appropriato sono stati utilizzati degli script compilati in Perl. Per filtrare gli artefatti di PCR e i paraloghi è stato applicato un comando di corrispondenza che richiedeva che il 90% delle prime 40 pb di una read, corrispondenti alle regioni fiancheggianti i microsatelliti, si allineasse al reference. Per ognuno dei 73 individui di *Neogobius melanostomus* catturati è stato determinato il genotipo a ciascun locus microsatellite considerando l'allele con il più alto numero di reads mappate, partendo da una soglia minima di dieci; qualora almeno il 20% delle reads mappate corrispondesse ad un secondo allele, l'individuo è stato dichiarato eterozigote per quel locus. In seguito alla genotipizzazione sono stati esclusi due loci debolmente amplificati e cinque loci potenzialmente paraloghi. La frequenza degli alleli è stata quindi stimata dai genotipi. Nei campioni di eDNA,

un ulteriore filtraggio è avvenuto rimuovendo gli alleli aventi frequenza inferiore all'1% in ogni campione. Per i campioni di eDNA dei mesocosmi sperimentali state accorpate tra loro le reads di ogni duplicato. Anche per i campioni di eDNA raccolti sul campo sono state accorpate le reads delle tre repliche. Le frequenze alleliche sono state quindi stimate come la frequenza delle reads per ciascun locus nelle due condizioni sperimentali.

3.6 Confronto tra le frequenze alleliche ottenute dagli individui genotipizzati e le frequenze alleliche ottenute dal DNA ambientale

Per determinare la somiglianza tra le frequenze alleliche ricavate dall'eDNA e quelle ricavate dai tessuti nel mesocosmo sperimentale, è stata valutata per ogni locus, in R 3.5, la loro correlazione. La somiglianza è stata ulteriormente analizzata, tra mesocosmi corrispondenti, conducendo un'analisi delle componenti principali (PC) e costruendo una matrice di distanza euclidea utilizzando come input i valori PC sui due assi. Per l'esperimento sul campo, è stata valutata la correlazione tra le frequenze alleliche determinate dai campioni di eDNA e quelle determinate dai 73 pesci genotipizzati in tutto.

3.7 Stima dei contributori genetici unici

Per determinare quale set di genotipi diploidi avesse prodotto con maggiore probabilità il set di alleli osservato, dati un insieme di alleli osservati a ciascun locus, $A = \{a_1, ..., a_n\}$, e di frequenze associate, $p = \{p_1, ..., p_n\}$, è stata applicata l'equazione di probabilità per il numero di contributori genetici unici [9].
Nella condizione sperimentale controllata, l'equazione è stata applicata indipendentemente ai campioni genotipizzati e a quelli di eDNA di ogni mesocosmo, determinando da ciascuno l'insieme di alleli A osservato. La sensibilità del metodo è stata valutata filtrando le reads delle sequenze dell'eDNA per soglie sempre più severe, al di sotto delle quali venivano rimosse (0.001, 0.01, 0.1). A causa del numero variabile di alleli per locus, le reads sono state anche filtrate per soglie variabili da 0.001 a 0.1, tanto minori quanto maggiore era la ricchezza allelica di un locus. Il numero di contributori genetici unici è stato stimato dapprima calcolando le frequenze alleliche p della popolazione dai genotipi degli individui; successivamente, la stessa operazione è stata ripetuta considerando come p le frequenze alleliche combinate dei tre campioni di eDNA. Essendo noti gli individui presenti nei mesocosmi (58), è stata calcolata poi la differenza tra i contributori genetici stimati e reali per ciascuna miscela. Per valutare le prestazioni del metodo rispetto a campioni di eDNA corrispondenti ad un maggior numero di individui, il modello è stato applicato a 1,000 miscele simulate contenenti da 2 fino a tutti i 58 individui utilizzati nel mesocosmo, costruite combinando tra loro i conteggi delle reads dell'eDNA provenienti da un minimo di 2 mesocosmi fino ad un massimo di 12. La stima del numero dei contributori genetici è stata effettuata anche per ogni campione di eDNA proveniente dal campo, determinando l'insieme degli alleli osservati A per ogni

campione. *P* è stato stimato prima dal genotipo dei 73 individui utilizzati nell'esperimento, poi delle reads combinate dei 3 campioni di eDNA.

4. Risultati

4.1 Sequenziamento e genotipizzazione

Dopo la demultiplazione delle reads e la rimozione degli adattatori, sono rimaste 35,583,440 delle 47,920,390 reads iniziali. Nessuno dei loci microsatelliti è stato identificato nei controlli negativi. Ciascun campione tissutale ha esibito un'elevata profondità di sequenziamento totale (media = 45.534 reads, SD = 19.958) e per singolo locus (media = 1.626 reads, SD = 1.714). Tutti i 73 individui sono stati genotipizzati per 26 o più dei 28 loci, che sono tutti risultati multiallelici, con una media di 9.4 alleli per locus. I loci microsatelliti sono stati amplificati con successo per tutti i campioni di eDNA provenienti dai mesocosmi sperimentali, con una profondità media di sequenziamento totale di 37,151 reads per campione (SD = 9.161) e di 1,327 reads per singolo locus (SD = 1.393), senza variazioni tra mesocosmi con diverse densità di *Neogobius melanostomus*. I campioni di eDNA provenienti dal campo hanno mostrato invece profondità di sequenziamento inferiori, con una profondità media di sequenziamento totale di 4,305 reads per campione (SD =3,796) e di 154 reads per locus (SD =283).

4.2 Confronto tra le frequenze alleliche ottenute dagli individui genotipizzati e le frequenze alleliche ottenute dal DNA ambientale

Nei mesocosmi, le frequenze alleliche ottenute dall'eDNA sono risultate strettamente associate a quelle ottenute dalla genotipizzazione dei 58 individui, secondo un coefficiente di correlazione di Pearson *r* di 0.95 per tutti i loci e di 0.88-1.00 per singolo locus. Dall'analisi delle componenti principali (PCA) è emersa una significativa somiglianza tra i campioni di eDNA e i corrispondenti individui genotipizzati (*Figura 1c*). La matrice di distanza euclidea ha restituito quanto ottenuto dalla PCA, con una distanza tra coppie di campioni di eDNA e di tessuto corrispondenti minore all'interno di ciascun mesocosmo; la distanza tra coppie è maggiore tra mesocosmi aventi densità di *Neogobius melanostomus* pari ad 1, mentre decresce all'aumentare della densità (*Figura 1d*).

Anche nell'esperimento sul campo le frequenze alleliche ottenute dall'eDNA sono risultate strettamente associate a quelle ottenute dalla genotipizzazione dei 73 individui, secondo un coefficiente di correlazione di Pearson *r* di 0.84 per tutti i loci e di 0.41-1.00 per singolo locus. Gli alleli con frequenza > 0.24 nei 73 individui genotipizzati sono stati recuperati in almeno uno dei tre campioni di eDNA, mentre 132 dei 253 alleli totali non sono stati rilevati dall'eDNA in nessuna delle tre repliche, e si sono verificati a basse frequenze nella popolazione (media =0.03, SD = 0.04). D'altra parte, i campioni di eDNA hanno anche identificato diversi alleli non documentati dagli individui genotipizzati, anche se a basse frequenze (media =

0.02, SD = 0.02): questi possono rappresentare veri alleli a bassa frequenza, o sequenze errate.

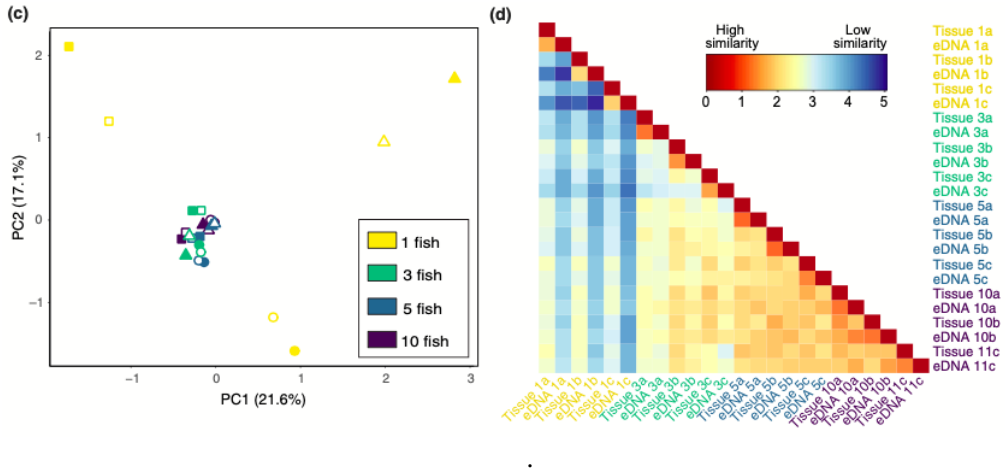


Figura 1c e 1d. (c) PCA delle frequenze alleliche dei 28 loci per i campioni di tessuto (simboli pieni) e di eDNA (simboli vuoti) dei 12 mesocosmi. I colori rappresentano le diverse densità del mesocosmo (1, 3, 5, o 10 pesci) e i simboli rappresentano i replicati. (d) Matrice di distanza euclidea delle frequenze alleliche tra coppie campioni di eDNA e di tessuto di ogni mesocosmo. I colori rappresentano le diverse densità dei mesocosmi, e le lettere (a, b, o c) i replicati.

4.3 Stima dei contributori genetici unici

Nella *figura 2* sono riportati i risultati ottenuti dal calcolo dei contributori genetici unici nella condizione sperimentale controllata, condotto sia per i campioni genotipizzati che per quelli di eDNA di ogni mesocosmo. La colonna a sinistra raccoglie i risultati ottenuti utilizzando come input *p* le frequenze alleliche dedotte dai genotipi dei 58 individui, quella a destra quelli ottenuti utilizzando come *p* le frequenze combinate delle reads dei campioni di eDNA. Considerando l'insieme *A* di alleli osservati dai campioni di tessuto, le stime ottenute specificando *p* dai genotipi (primo pannello in alto a sinistra) hanno mostrato uno scarto di ± 2 individui rispetto al valore reale, e di fino a -5 specificandolo dalle reads dell'eDNA (primo pannello in alto a destra). Considerando l'insieme *A* di alleli osservati dai campioni di eDNA, sono emersi bias positivi e negativi per le soglie di frequenza selezionate (0.001, 0.01, 0.1), indipendentemente dal fatto che *p* fosse stato stimato dai tessuti o dai campioni di eDNA (dall'alto, pannelli 2, 3 e 4 a destra e sinistra), anche se bias positivi sono stati più accentuati specificando *p* dalle reads dell'eDNA. Utilizzando le frequenze alleliche *p* stimate dai genotipi, il bias è risultato positivo in tutti i mesocosmi applicando la soglia più bassa: esso indica la presenza di alleli falsi positivi, probabilmente frutto di artefatti introdotti durante la PCR o il sequenziamento, essendo stata dimostrata l'assenza di contaminazione tra campioni. Alla soglia di 0.01 il bias è risultato positivo in tutti i mesocosmi ad eccezione di quelli di 10 individui, dove lo scarto è stato in media di ±1 (dall'alto, pannelli 2 e 3 a sinistra). Alla soglia di 0.1 il bias è risultato negativo nei mesocosmi

con cinque o dieci individui: lo scarto delle reads al di sotto di tale frequenza ha evidentemente significato la perdita di numerosi alleli aventi frequenza minore o uguale a 0.1, risultando in una perdita di genotipi e, quindi, di individui. In tutti i mesocosmi, la soglia variabile basata sulla ricchezza allelica ha dimostrato di fornire la stima più accurata dell'abbondanza assoluta. Nonostante bias positivi siano stati più accentuati specificando *p* dalle reads dell'eDNA, in tutti i mesocosmi i risultati ottenuti si sono dimostrati simili: ciò suggerisce che, in condizioni controllate, potrebbe non essere necessaria la genotipizzazione degli individui per ottenere stime affidabili dell'abbondanza assoluta di una specie in campioni di eDNA. Negli esperimenti sul campo, ricavando *p* dai 73 individui genotipizzati sono stati stimati rispettivamente cinque, tre e tre individui in ciascuno dei tre campioni di eDNA. Tuttavia, poiché 132 dei 253 alleli totali non sono stati rilevati dall'eDNA in nessuna delle tre repliche, si tratta probabilmente di una sottostima. Quando *p* è stato specificato dalle frequenze combinate delle reads dei tre campioni di eDNA, 13, 7 e 6 individui geneticamente distinti sono stati stimati per ciascun campione. La figura 3 rappresenta i risultati ottenuti dalla stima dei contributori nelle miscele simulate. La soglia più alta (0.1), risultando superiore alle frequenze di tutti gli alleli che non fossero i più comuni, ha dato bias negativi (da destra, secondo pannello): per questo, è stato stimato un massimo di circa 15 individui, anche per densità simulate > 50 individui. Le soglie di filtraggio più basse (0.01, 0.001) hanno sovrastimato il numero di contributori per tutte le densità (da sinistra, primo e secondo pannello). La soglia variabile basata sulla ricchezza allelica (da destra, primo pannello) ha mostrato una minore distorsione in tutte le simulazioni eccetto in quella contenente il maggior numero di contributori (58), dove si è verificato un bias negativo.

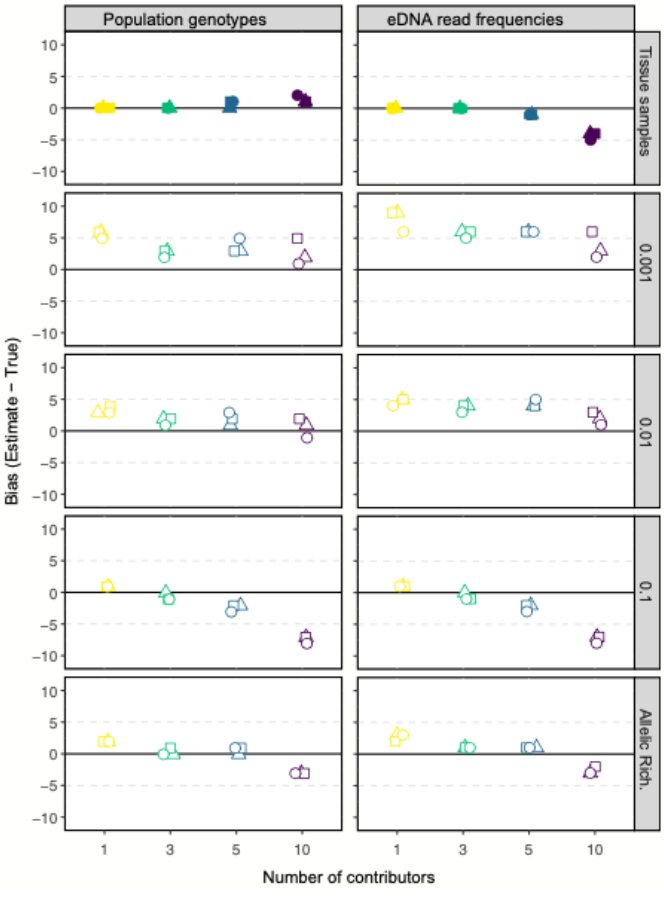
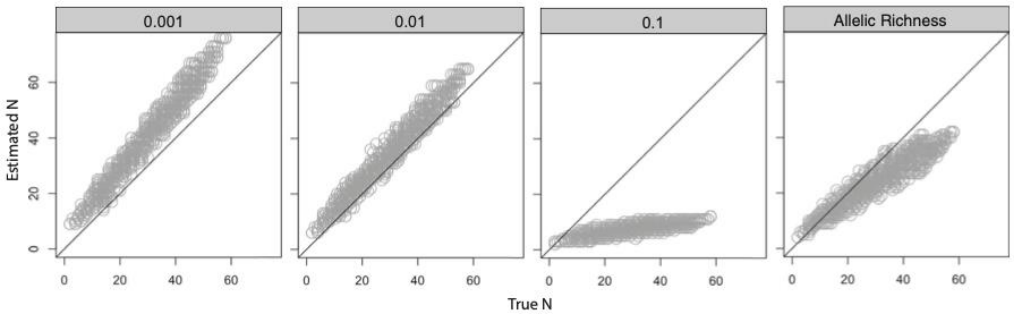


Figura 2 (a sinistra). Bias nella stima del numero dei contributori genetici unici utilizzando campioni di tessuto (simboli pieni) e di eDNA (simboli cavi) per mesocosmi sperimentali a differenti densità di Neogobius melanostomus (1, 3, 5, o 10 pesci). Le frequenze alleliche p della popolazione sono state derivate dai genotipi dei 58 individui (a sinistra) o dalle frequenze delle reads

dell'eDNA combinate tra tutti i mesocosmi. I simboli rappresentano i replicati, e i pannelli indicano le soglie al di sotto delle quali sono state rimosse le letture di sequenza (0.1, 0.01, 0.001), o una soglia variabile basata sulla ricchezza allelica per locus.

Figura 3 (a capo). Stima del numero di contributori genetici unici in miscele di eDNA simulate (range 2-58 individui) utilizzando alleli provenienti da 1,000 miscele di eDNA generate combinando numeri variabili di mesocosmi, fino ad un massimo di 12. I pannelli indicano le soglie al di sotto delle quali sono state rimosse le letture di sequenza (0.1, 0.01, 0.001), o una soglia variabile basata sulla ricchezza allelica per locus.



5. Discussione

I dati genetici delle popolazioni sono alla base di molti studi sui processi comportamentali, ecologici ed evolutivi nelle popolazioni selvatiche e contribuiscono a una gestione efficace della conservazione. Utilizzando loci microsatelliti come marcatori nucleari, molto polimorfi e dunque utilizzabili per la genotipizzazione, nel presente lavoro l'eDNA è stato testato come strumento per la rilevazione della variabilità genetica di una popolazione recuperando le frequenze alleliche ottenute dal DNA estratto da campioni acquatici di *Neogobius melanostomus* e confrontandole con quelle ottenute dalla genotipizzazione di campioni di tessuto. Ciò è stato svolto sia in condizioni sperimentali controllate, in cui erano noti gli individui presenti, e che quindi hanno funto da modello per la capacità di caratterizzazione degli alleli, sia in natura. Applicando poi il metodo "DNA mixture estimators", è stato stimato il numero di contributori genetici unici relativo ai campioni ambientali prelevati nelle due condizioni sperimentali, dimostrando come sia possibile utilizzare le informazioni sulla diversità genetica intraspecifica per determinare il numero di individui catturati in un campione di eDNA.

Nei mesocosmi, le frequenze alleliche per i loci microsatelliti ottenute dall'eDNA hanno mostrato significativa somiglianza con quelle ottenute dalla genotipizzazione dei tessuti, nonostante l'approccio abbia esibito una ridotta sensibilità nel distinguere mesocosmi aventi densità simili di individui; tuttavia, avendo impiegato per l'esperimento il pool genico di una singola popolazione, ciò è comprensibile. Anche nell'esperimento sul campo le frequenze alleliche ottenute dall'eDNA si sono dimostrate sufficientemente accurate, nonostante non siano state rilevate

quelle degli alleli più rari nella popolazione. Saranno quindi necessari interventi per risolvere il problema della bassa rilevabilità degli alleli più rari.

Rilevare la variabilità genetica intraspecifica da campioni di eDNA è anche utile per stimare il numero di individui geneticamente distinti che hanno contribuito al campione. Nella condizione sperimentale controllata, filtrando le frequenze alleliche alla soglia più alta (0.1) i risultati sono stati i più accurati per i mesocosmi con 1 e 3 pesci, mentre i bias negativi riscontrati per i mesocosmi con 5 e 10 pesci riflettono probabilmente la rimozione di veri alleli a frequenza minore di 0.1. Alle soglie di 0.01 e 0.001, il numero di contributori per mesocosmi ad alte densità è stato risolto sufficientemente, mentre i bias positivi riscontrati per i mesocosmi con 1 e 3 pesci riflettono l'introduzione di falsi alleli. Gli autori, a tal proposito, raccomandano, per future applicazioni, di eseguire un filtraggio bioinformatico basato su soglie moderate o basate sulla ricchezza allelica dei loci. Avvertono anche del fatto che l'impatto dei falsi alleli e della perdita di veri alleli sulla stima dei contributori genetici unici rimane da indagare, con particolare attenzione agli esperimenti sul campo, dove i falsi alleli sono più difficili da distinguere da quelli veri a bassa frequenza e la rilevazione di alleli rari può essere insufficiente, e suggeriscono l'implementazione di algoritmi di correzione degli errori e di procedure per la rimozione del rumore di fondo che aiutino ad eliminare le sequenze errate e a mantenere i veri alleli a bassa frequenza.

Le condizioni controllate del mesocosmo sperimentale potrebbero non riflettere la complessità esistente in natura. Gli autori sottolineano come, nonostante popolazioni naturali di *Neogobius melanostomus* possano esibire hotspot di densità elevata, la densità media delle popolazioni nel lago Cayuga rimaga inferiore a quella nei mesocosmi sperimentali, e come questo possa rappresentare un limite per la profondità di sequenziamento ottenibile da esperimenti sul campo utilizzando marcatori nucleari, specie se confrontata con quella ottenibile utilizzando marcatori mitocondriali, da cui la necessità di ottimizzare gli attuali limiti di rilevamento. Il mesocosmo non rende conto nemmeno della complessità biofisica dei sistemi naturali, dove particelle e organismi presenti contribuiscono ai campioni di eDNA e potrebbero inibire la reazione di PCR compromettendo l'identificazione degli alleli, specie se il DNA target si trovasse a basse concentrazioni. In presenza di specie non bersaglio strettamente correlate a quella target, la specificità del primer deve essere accuratamente testata per evitare aspecifici.

L'approccio qui sviluppato è più probabile che integri le metodologie esistenti per lo studio della variabilità genetica intraspecifica negli studi di popolazione, anziché sostituirle. Lo sviluppo di pannelli specie-specifici di marcatori microsatelliti, la PCR multipla, e la stima del numero di contributori, richiedono numerosi dati di sequenza, verosimilmente provenienti dalla genotipizzazione di tessuti. Tuttavia, avendo dimostrato che in condizioni sperimentali controllate le frequenze alleliche

di una popolazione ottenute dalle frequenze delle reads dell'eDNA sono altamente correlate con quelle degli individui genotipizzati, e che anche le stime dei contributori genetici unici sono simili, indipendentemente da come le frequenze alleliche siano derivate, questo studio convalida sperimentalmente l'impiego di microsatelliti nucleari come marcatori per la stima delle frequenze alleliche di popolazione e dell'abbondanza assoluta delle specie utilizzando partendo da DNA ambientale.

Con un'adeguata convalida e miglioramenti tecnici appropriati, la genetica delle popolazioni basata sull'eDNA mostra il potenziale per migliorare la conservazione e la gestione delle specie. Per esempio, nell'ambito della prevenzione della diffusione e della minimizzazione degli impatti di specie invasive, potrebbe contribuire in modo sostanziale al loro monitoraggio, informando sulla variabilità genetica e sulla dimensione della popolazione aliena nei siti di colonizzazione. Data l'elevata sensibilità e la facilità di raccolta dei campioni, questo approccio può anche essere vantaggioso per il monitoraggio di specie in cui piccole dimensioni delle popolazioni, habitat vasti o complessi, comportamento elusivo, o il desiderio di ridurre al minimo il campionamento invasivo, possano impedire valutazioni efficaci. Non meno importante, le tempistiche e i costi sempre più ridotti delle tecniche NGS potrebbero contribuire alla diffusione di tali metodi.

Per concludere, questo studio dimostra l'avanzamento degli approcci basati sull'eDNA nell'includere anche marcatori genetici diretti al genoma nucleare, provando la capacità di ottenere da questo informazioni sulla variabilità genetica intraspecifica, e getta le basi per una potenziale alternativa alla stima dell'abbondanza delle specie mediante la concentrazione di eDNA. Pur necessitando di ulteriori convalide e ottimizzazioni, pone i presupposti per la possibilità di svolgere studi di genetica di popolazioni a partire da campioni ambientali.

Bibliografia

1. Andres KJ, Sethi SA, Lodge DM, Andrés J. Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. *Mol Ecol*. 2021; 30:685–697.
2. Thomsen, P. F., Kielgast, J., Iversen, L. L., Wiuf, C., Rasmussen, M., Gilbert, M. T. P., ... Willerslev, E. (2012). Monitoring endangered freshwater biodiversity using environmental DNA. *Molecular Ecology*, *21*(11), 2565–2573.
3. Williams, M.A, O'Grady, J, Ball, B, Carlsson, J, de Eyto, E, McGinnity, P, Jennings, E, Regan, F, McDermott, A.P. (2019). The application of CRISPR-Cas for single species identification from environmental DNA. *Molecular Ecology Recources,19*(5),1106-1114.
4. Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., & De Vere, N. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895.
5. Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.-M., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929–942.
6. Shokralla, S., Spall, J. L., Gibson, J. F., Mehrdad, H. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology, 21*(8), 1794-1805.
7. Marino, I. A. M. (2009). Applicazioni di marcatori microsatellite per lo studio della fileogeografia di organismi lagunari dell'Adriatico. Tesi di dottorato in Biologia Evoluzionistica. Università degli studi di Padova.
8. Takahara, T., Minamoto, T., Yamanaka, H., Doi, H., & Kawabata, Z. I. (2012). Estimation of fish biomass using environmental DNA. *PLoS One*, *7*(4), e35868.
9. Sethi, S. A., Larson, W., Turnquist, K., & Isermann, D. (2019). Estimating the number of contributors to DNA mixtures provides a novel tool for ecology. *Methods in Ecology and Evolution*, *10*(1), 109–119.
10. Zane, L., Bargelloni, L., Patarnello, T. (2002). Strategies for microsatellite isolation: a review. *Molecular Ecology*, *11*, 1-16.

Received: 14 November 2019 | Revised: 23 November 2020 | Accepted: 1 December 2020

DOI: 10.1111/mec.15765

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples

Kara J. Andres[1] | Suresh A. Sethi[2] | David M. Lodge[1,3] | Jose Andrés[1]

[1]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA

[2]U.S. Geological Survey, New York Cooperative Fish and Wildlife Unit, Cornell University, Ithaca, NY, USA

[3]Cornell Atkinson Center for Sustainability, Cornell University, Ithaca, NY, USA

Correspondence
Kara J. Andres, Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, NY, USA.
Email: kja68@cornell.edu

Funding information
National Science Foundation (NSF), Grant/Award Number: 1748389; Department of Defence (DoD), Grant/Award Number: RC19-1004

Abstract

Advances in environmental DNA (eDNA) methodologies have led to improvements in the ability to detect species and communities in aquatic environments, yet the majority of studies emphasize biological diversity at the species level by targeting variable sites within the mitochondrial genome. Here, we demonstrate that eDNA approaches also have the capacity to detect intraspecific diversity in the nuclear genome, allowing for assessments of population-level allele frequencies and estimates of the number of genetic contributors in an eDNA sample. Using a panel of microsatellite loci developed for the round goby (*Neogobius melanostomus*), we tested the similarity between eDNA-based and individual tissue-based estimates of allele frequencies from experimental mesocosms and in a field-based trial. Subsequently, we used a likelihood-based DNA mixture framework to estimate the number of unique genetic contributors in eDNA samples and in simulated mixtures of alleles. In both mesocosm and field samples, allele frequencies from eDNA were highly correlated with allele frequencies from genotyped round goby tissue samples, indicating nuclear markers can be reliably amplified from water samples. DNA mixture analyses were able to estimate the number of genetic contributors from mesocosm eDNA samples and simulated mixtures of DNA from up to 58 individuals, with the degree of positive or negative bias dependent on the filtering scheme of low-frequency alleles. With this study we document the application of eDNA and multiple amplicon-based methods to obtain intraspecific nuclear genetic information and estimate the absolute abundance of a species in eDNA samples. With proper validation, this approach has the potential to advance noninvasive survey methods to characterize populations and detect population-level genetic diversity.

KEYWORDS
DNA mixtures, environmental DNA, intraspecific diversity, invasive species, microsatellites, round goby

1 | INTRODUCTION

Environmental DNA (eDNA) approaches are transforming how scientists and resource managers assess the diversity and distributions of organisms (Deiner, Bik, et al., 2017; Taberlet et al., 2012). Using DNA isolated from environmental samples such as ancient and terrestrial sediments, ice cores, and aquatic ecosystems, eDNA methodologies capture the genetic material organisms release into the environment through cells, hair, skin, and faeces (Thomsen et al., 2012; Willerslev et al., 2003, 2007). Such approaches can provide an efficient way to detect species presence/absence (Ficetola et al., 2008; Pilliod et al., 2013), habitat use (Stewart et al., 2017), and relative abundance (Hänfling et al., 2016; Jerde et al., 2011). With greater detection probability and reduced cost over traditional sampling methods, eDNA methods are particularly well-suited for surveillance of aquatic invasive species, where early detection may be vital for their management or eradication (Dejean et al., 2012; Jerde et al., 2011; Lodge et al., 2016; Vander Zanden et al., 2010). Furthermore, technical advancements in next-generation sequencing (NGS) methods have led to the development of eDNA metabarcoding, or the simultaneous detection of multiple species with a single molecular marker (e.g., Deiner, Bik, et al., 2017; Kelly et al., 2014; Margulies et al., 2005; Taberlet et al., 2012; Valentini et al., 2016). Environmental DNA can therefore provide information about species distributions, relative abundance, or composition that can be broadly applied in studies of biodiversity, community ecology, and conservation biology (Bohmann et al., 2014; Lodge et al., 2012; Thomsen & Willerslev, 2015).

The majority of eDNA studies to date have assessed biological diversity at or above the species level, with relatively little attention given to intraspecific genetic diversity (Adams et al., 2019; Sigsgaard et al., 2020). However, some recent studies have developed approaches to detect intraspecific genetic variation in the mitochondrial genome from environmental samples (Deiner, Renshaw, et al., 2017; Elbrecht et al., 2018; Parsons et al., 2018; Sigsgaard et al., 2017; Tsuji et al., 2019; Turon et al., 2020). Due to its high copy number per cell, mitochondrial DNA (mtDNA) may occur at higher concentrations in water than nuclear DNA (but see Bylemans et al., 2017; Minamoto et al., 2017; Piggott, 2016), potentially leading to higher detection probability in environmental samples. On the other hand, mtDNA is haploid and nonrecombining so, as a single locus, may be limited in providing the high resolution of genetic variation required for detailed population genetic analyses (Ballard & Whitlock, 2004; Hurst & Jiggins, 2005; Rubinoff et al., 2006; Teske et al., 2018). Expanding eDNA approaches to detect intraspecific variation in nuclear DNA markers such as microsatellites or single nucleotide polymorphisms (SNPs) can therefore enhance our ability to make genetic inferences at the population level.

In this study, we explore the extent to which intraspecific genetic diversity can be detected in eDNA and used to estimate the number of unique genetic contributors to an eDNA sample. As a proof of concept, we use nuclear microsatellite markers and NGS methods to characterize population allele frequencies and estimate the absolute abundance of round gobies (*Neogobius melanostomus*) using eDNA samples from experimental mesocosms and in a field-based trial. The round goby, a fish species native to the Ponto-Caspian region, was initially introduced to North America via ballast water in 1990 and has since spread throughout the Laurentian Great Lakes (Charlebois et al., 1997; Jude et al., 1992; Schaeffer et al., 2005). More recently, round gobies have spread to inland lakes and rivers, where they can cause native species declines through competition, predation, and contaminant cycling (Janssen & Jude, 2001; Kornis et al., 2012; Krakowiak & Pennuto, 2008). Due to the short time interval between arrival and establishment, round gobies present a high invasion risk even at low densities (Vélez-Espino et al., 2010), and control strategies may require information on species abundance due to the rapid decline in the success of eradication efforts as invasive populations grow and spread (Vander Zanden et al., 2010). Thus, the development of eDNA methods to quantify species abundance at the invasion front could lead to improved management strategies for this invader.

Several previous efforts to assess species abundance with eDNA have used correlative relationships between eDNA concentration and indices of species abundance or biomass (e.g., Kelly et al., 2014; Pilliod et al., 2013; Takahara et al., 2012). While these methods can provide an index of relative abundance, their accuracy and precision with respect to absolute species abundance has been difficult to establish, and such correlative relationships can be heavily impacted by taxon-specific amplification biases (Kelly et al., 2019) or local biotic and abiotic factors influencing the amount of DNA shed by an organism (Barnes & Turner, 2015). For instance, the production rate of eDNA can vary with an organism's size, behaviour, or metabolism, all of which may vary across a range of abiotic conditions (Klymus et al., 2015; Lacoursière-Roussel et al., 2016; Maruyama et al., 2014; Takahara et al., 2012). The difficulty in obtaining robust quantitative measurements of eDNA production among individuals and its relationship to the amount of DNA in an eDNA sample currently limits our ability to reliably link measurements of eDNA concentration to species abundance, density, or biomass (Iversen et al., 2015).

In contrast to correlative relationships between eDNA concentrations and relative species abundance, DNA mixture estimators take a radically different approach to estimating absolute abundance in a sample (Sethi et al., 2019). Originally developed in criminal forensics, DNA mixture estimators provide an inferential framework that uses the genetic signature of mixtures to estimate the number of unique genetic contributors in a mixture of DNA based on population allele frequencies and the number of unique alleles identified (Curran et al., 1999; Weir et al., 1997). While these estimators have previously been applied to tissue-based mixtures of DNA for diet analysis (Sethi et al., 2019), environmental samples can also contain DNA from multiple individuals. If intraspecific genetic diversity can be detected eDNA, mixture estimators may therefore provide a means of estimating the number of contributors to environmental samples that relies on the detected presence of haplotypes or alleles rather than eDNA concentrations.

Here, we applied DNA mixture estimators to eDNA using species-specific nuclear genetic markers we developed for round gobies. We first assessed the similarity of allele frequencies from eDNA and individually genotyped individuals in experimental mesocosms to evaluate the extent to which alleles derived from round goby tissues are represented in sequence data recovered from eDNA. We then used a likelihood-based DNA mixture model to estimate the number of genetically unique individuals contributing genetic material to each eDNA sample. Finally, we tested the ability of the DNA mixture estimator to accurately estimate the number of unique genetic contributors in simulated combinations of up to 58 individuals.

2 | MATERIALS AND METHODS

2.1 | Microsatellite characterization and multiplex assay development

Genomic DNA (50–100 ng) from a pool of three round goby (*Neogobius melanostomus*) individuals collected from Cayuga Lake, New York, USA was endonuclease-digested with *AluI*, *RsaI*, and *Hpy166II*. The digestions were pooled for subsequent adenylation with Klenow (exo-) and dATP, and the resulting products were ligated to an Illumina Y-adaptor sequence using T4 DNA ligase in the presence of 1 mM ATP. Genomic fragments containing repeats were captured by hybridization to biotinylated repeats and streptavidin-coated magnetic beads followed by amplification with Platinum Taq DNA polymerase and indexing with Illumina primers (one universal primer and one index primer). PCR products were quantified with a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA), verified by electrophoresis on a 1.0% agarose gel, and size-selected (300–600 base pairs [bp]) with Agencourt AMPure XP beads (Beckman Coulter, Indianapolis, IN). The "design primers" function of MSATCOMMANDER 1.0.3 (Faircloth, 2008; Rozen & Skaletsky, 2000) was then used to create a library of microsatellite tetramer repeats based on the number of motif repeats (10–24) and PCR product length (410–440 bp).

Primer specificity was inspected using NCBI Primer Blast, where no other species were detected as matches to the designed primer pairs. Forward and reverse primers (range 20–24 bp) for 43 loci were ordered from Integrated DNA Technologies (http://www.idtdna.com) and tested for functionality in single reactions using genomic DNA extracted from the tissue of three round gobies. Following exclusion of primers with complementary sequences or suboptimal PCR amplification, 35 microsatellite loci remained (Table S1).

Microsatellite loci were grouped into seven multiplex PCR assays, with each multiplex containing 4–6 primer pairs (Table S1). We tested the performance of each multiplex in a PCR containing 1 µl (20–30 ng) of round goby genomic DNA, 1 µl of primer pairs in equimolar concentrations (2 µM), and 5 µl of Qiagen Multiplex PCR Master Mix (Qiagen Inc.). The program for multiplex PCR is as follows: initial denaturation at 95°C for 15 min followed by 35 cycles of 94°C for 30 s, 59°C for 90 s, and 72°C for 90 s. Gel electrophoresis in 1% agarose stained with ethidium bromide confirmed the presence of PCR products within the expected size range for all multiplexes.

2.2 | Mesocosm experiment

We collected live round gobies (n = 58) from a site on Cayuga Lake via beach seining and placed them in one of 12 experimental mesocosms containing 12 L of aged room temperature tap water. Each mesocosm treatment was conducted in triplicate and contained round gobies (approximately 7–12 cm length) at densities of one, three, five, or 10 individuals. An additional round goby was erroneously added to a single replicate of the n = 10 treatment to total 11 individuals (labelled mesocosm 10c), but is hereafter grouped into the density treatment of 10 individuals. Two additional mesocosms served as negative controls (mesocosms with aged room temperature tap water only). After 1 h, round gobies were removed from the mesocosms and euthanized with MS-222 according to the Cornell IACUC Animal Care and Use Procedure (ACUP 306.02). Tissues were sampled from caudal fins of each individual and DNA was extracted with a DNeasy Blood and Tissue extraction kit (Qiagen Inc.) following the manufacturer's protocols. Following the removal of all fish from the mesocosms, duplicate 2 L water samples were collected from each mesocosm in sterilized wide-mouth Nalgene plastic bottles and stored on ice until vacuum-filtration through a cellulose nitrate membrane filter (47 mm diameter, 1 µm pore size). Filters were immersed in 700 µl Longmire's solution (100 mM Tris, 100 mM EDTA, 10 mM NaCl, 0.5% SDS, 0.2% sodium azide) and stored at −20°C until DNA extraction. Environmental DNA was extracted from filters following a modified protocol from the DNeasy Blood and Tissue extraction kit (Qiagen Inc.) as in Spens et al. (2017). To minimize contamination, eDNA sample filtration and pre-PCR laboratory protocols were carried out in separate rooms within dedicated pre-PCR facilities, and stringent precautions were followed according to Goldberg et al. (2016). Round goby tissues were handled and processed in a separate facility. All reusable equipment including collection bottles, forceps, and the vacuum filtration apparatus was cleaned between samples by soaking in a 50% commercial bleach solution, rinsing in DI water, and treating under UV bulbs for 30 min each. In addition to the two field controls described above, one filtration blank and one PCR blank served as negative controls.

2.3 | Field trial

To determine the feasibility of estimating population allele frequencies from eDNA samples in a field-based setting, we collected eDNA samples and additional round gobies (n = 15) from another site on Cayuga Lake (c. 20 miles away from the site of round goby collection for the mesocosm experiment; Figure S1A). Sampling was conducted during the summer months when round goby densities peak in nearshore waters, and density estimates from a previous study using benthic videography and direct

observation report round goby densities of 0.34 fish/m^2 in this section of the lake (Andres et al., 2020). We confirmed the round gobies collected from the two sites in Cayuga Lake are panmictic using genotyped tissue samples from both sites and the "find.clusters" function of the ADEGENET package in R version 3.5 (Jombart, 2008; R Core Team, 2016), where a single cluster (k) exhibited the lowest Bayesian Information Criterion (BIC; Figure S1B). Thus, we consider all 73 round gobies (58 in the mesocosm experiment and 15 in the field trial) when estimating tissue-based population allele frequencies in the field trial. To sample eDNA, three 2 L water samples were collected from shoreline locations approximately 50 m apart in sterilized wide-mouth Nalgene plastic bottles. A negative field control of 2 L of distilled water was also collected at the site. Water filtration, tissue sampling, and DNA extraction protocols were identical to those described for the mesocosm experiment above.

2.4 | Library preparation and MiSeq sequencing

Microsatellite loci were amplified from eDNA and tissue samples in separate reactions using multiplex PCR methods described above, with the number of PCR cycles increased to 45 for eDNA samples due to low template DNA concentrations. Three PCR replicates were performed for each of the three eDNA samples from the field trial. Products from all seven multiplexes were pooled from each sample in equal volumes (5 µl each) and uniquely barcoded in a second-stage PCR using Illumina Nextera XT tags. Each 20 µl second-stage PCR included 2 µl pooled PCR product diluted 1:1 with molecular H_2O, 4 µl 5× HF buffer, 0.4 µl 10 mM dNTPs, 0.1 µl OneTaq DNA polymerase, 0.4 µl each of 10 µM Nextera Index Primer 1 (N701–N728) and Nextera Index Primer 2 (N502–N521). One library was constructed from the pooled PCR products for all tissue and eDNA samples in the mesocosm experiment, while another library was constructed from the tissue and eDNA samples from the field trial. DNA libraries were purified with Agencourt AMPure XP beads and the concentration of each library was estimated using the Qubit dsDNA High-Sensitivity Kit and Qubit 2.0 fluorometer. The libraries were diluted to 4 nM with PCR-grade water and paired-end sequenced on an Illumina MiSeq sequencing platform (Illumina, San Diego, CA) with the MiSeq v2 500 bp kit (PE 2 × 250 bp) by Cornell University's Institute of Biotechnology Genomics Facility.

2.5 | Bioinformatic analysis

Demultiplexed reads from each Miseq run were processed with TRIMMOMATIC v0.39 (Bolger et al., 2014) to remove adapter sequences. We then ran a custom Perl script to extract forward and reverse reads and assign them to each locus as described in D'Aloia et al. (2017). The script includes the following steps: (i) trim low-quality reads with Phred scores less than 20; (ii) create contigs from overlapping paired-end reads with a minimum overlap of at least 20 bp and mismatch rate of less than 0.05; (iii) identify and sort reads corresponding to each locus using the forward primer; (iv) collapse identical reads (100% identity) for each sample; and (v) collapse reads across all samples. To filter out most PCR artifacts and paralogues while retaining true microsatellite repeats and SNPs, we required 90% of the first 40 bp of a read to align with and match the reference contig constructed from the most common allele at each locus across all of the samples. We determined the multilocus diploid genotype for each round goby tissue sample based on the allele with the highest read count at each locus. Individuals were considered heterozygous at a locus if at least 20% of the reads corresponded to a second allele, and only alleles with a read depth of at least 10 reads per individual were considered (as in D'Aloia et al., 2017). Following individual genotyping, we excluded two poorly amplified loci and five potentially paralogous loci exhibiting significant deviations from Hardy-Weinberg equilibrium (Paradis, 2010) and heterozygote excess. The remaining 28 loci were used in further analyses (Table S1).

For eDNA samples, we excluded alleles with fewer than 10 total reads in each sample and scaled read counts to 100 reads per sample to account for differences in read depth. To further filter out potentially erroneous sequence data arising from PCR stutter and sequencing error, we removed alleles below 1% frequency in each eDNA sample from analysis. Due to low variation in read depth and allele frequencies between duplicate mesocosm eDNA samples (Figures S2–S3), we pooled the scaled reads from the two eDNA samples for each mesocosm. We also pooled the scaled reads from the three replicate eDNA samples from the field trial. eDNA allele frequencies were then estimated as the read frequencies of alleles in each mesocosm and in the field eDNA sample. Thus, while allele frequency estimations in tissue samples are derived from genotyped individuals, allele frequency estimations in eDNA samples are taken directly from sequence read frequencies.

2.6 | Comparison of genotyped individuals and eDNA samples

All further analyses were performed in R version 3.5 (R Core Team, 2016). To determine the similarity between allele frequencies derived from eDNA reads and genotyped tissues in the mesocosm experiment, we combined allele frequencies across all 12 mesocosm eDNA samples and evaluated the correlation between eDNA allele frequencies and tissue allele frequencies for all alleles across all loci, as well as on a per-locus basis. We further examined the similarity between eDNA-based and tissue-based allele frequencies in corresponding mesocosms by conducting a principal components (PC) analysis on the scaled and centred allele frequencies from eDNA reads and genotyped individuals. Subsequently, we constructed a Euclidean distance matrix for all samples using principal components values along all PC axes described above as inputs. For the field trial, we evaluated the correlation between allele frequencies determined from the eDNA samples collected from Cayuga Lake and from the 73 genotyped round gobies.

2.7 | DNA mixture contributor estimation

To estimate the number of unique genetic contributors to a DNA mixture (e.g., the number of individuals captured in each eDNA sample), we implemented a likelihood-based model described in Sethi et al. (2019). At each locus *j*, the model estimates the likelihood that a proposed number of diploid contributors, *x*, produces the observed set of *n* alleles, $A = \{a_1, \ldots, a_n\}$, given a set of associated population allele frequencies, $p = \{p_1, \ldots, p_n\}$, using the following equation:

$$L_j(x|A,p) = \sum_{d_1=0}^{d} \sum_{d_2=0}^{d-d_1} \cdots \sum_{d_{n-1}}^{d-d_1-\ldots-d_{n-2}} \left[\left(\frac{(2x)!}{\prod_{i=1}^{n} g_i!} \right) \prod_{i=1}^{n} p_i^{g_i} \right] \quad (1)$$

This equation accounts for all of the combinations of alleles that may arise in a mixture due to redundancy within or among individuals, where $d = 2x - n$ is the total number of "masked" alleles calculated as the difference between the total number of alleles present for *x* diploid organisms and the total number of unique alleles observed in the mixture genotype, and g_i is the total number of copies of allele a_i truly present in the mixture plus any masked copies of the allele d_i, with $\sum_{i=1}^{n} g_i = 2x$. As in Sethi et al. (2019), we calculated this likelihood with custom R scripts using a numerically equivalent but more computationally efficient form of Equation 1 derived by Weir et al. (1997).

The estimated number of individuals contributing to the DNA mixture is therefore identified as the maximum likelihood estimate of the number of contributors given the product of this likelihood across all loci:

$$\max_x \prod_j L_j(x|A,p) \quad (2)$$

For the mesocosm samples, we applied this equation to pooled individual genotypes and eDNA mixtures from each mesocosm with a proposed number of contributors (1–100), where the set of observed alleles A was determined per mesocosm and the population allele frequencies *p* were estimated directly from the 58 genotyped individuals used in the experiment. To evaluate the sensitivity of the contributor estimation to false alleles and allelic dropout, we filtered eDNA sequence reads according to a succession of increasingly strict thresholds, or frequencies below which reads were removed (0.001, 0.01, 0.1). Due to variation in the number of alleles present at each locus (Table S1), we also filtered reads using variable thresholds according to per-locus allelic richness, where the threshold decreased from 0.1 to 0.001 as the number of alleles at a locus increased. We repeated the contributor estimations using the allele frequencies combined across all eDNA samples to represent population allele frequencies *p*. Bias in the contributor estimation (estimated # contributors - true # contributors) was calculated for each eDNA-based and tissue-based DNA mixture.

To assess the performance of the contributor estimation on eDNA samples representing a greater number of individuals, we applied the maximum likelihood estimator to simulated mixtures of up to our total sample of 58 round gobies in the mesocosm experiment. Using a bootstrapping procedure, we combined eDNA read counts from mesocosms in simulated mixtures ranging from 2–12 mesocosms per draw. We estimated the number of genetic contributors to mixtures with 1,000 bootstrap replicates at fixed thresholds and a variable threshold based on allelic richness as described above.

We also applied the contributor estimation to each eDNA sample from the field trial, where the set of observed alleles A was determined from each eDNA sample and population allele frequencies *p* were estimated from the 73 genotyped individuals used in the experiment. We repeated the contributor estimations with allele frequencies combined across the three replicate eDNA samples taken from the field used to represent population-level allele frequencies *p*.

3 | RESULTS

3.1 | Sequencing and genotyping

The full data set contained 47,920,390 reads, of which 35,583,440 remained after demultiplexing and trimming adapters. Following exclusion of alleles below the minimum read depth of 10 reads, the target loci were not identified in any of the negative control blanks from the field, extraction, or amplification processes, indicating there was no detectable cross-contamination. Round goby tissue samples exhibited a high total read depth per sample (mean = 45,534 reads, SD = 19,958; Figure S4A) and total read depth per locus (mean = 1,626 reads, SD = 1,714, Figure S4B). All individuals were genotyped at ≥26 of the 28 loci in all individuals (i.e., fewer than two loci per sample were considered missing data in our pipeline). All microsatellites were multiallelic with an average of 9.4 alleles per locus (range: 2–21 alleles per locus) among the 73 round gobies comprising the sample population (Table S1). Microsatellite loci were successfully amplified in all eDNA samples from mesocosms containing fish with an average total read depth of 37,151 reads per sample (SD = 9,161) and average read depth of 1,327 reads per locus (SD =1,393). The average per-locus read depth and total read depth did not vary across mesocosm densities (Figure S5), indicating round goby density did not have an impact on sequence recovery in the mesocosm experiment. Read depths were lower in eDNA samples from the field trial, with an average total read depth of 4,305 reads per sample (SD =3,796;; Figure S4A) and 154 reads per locus (SD =283; Figure S4B).

3.2 | Comparison of genotyped individuals and eDNA samples

In the mesocosm experiment, allele frequencies from eDNA sequence reads across all mesocosms closely resembled allele frequencies from the 58 genotyped individuals (Pearson's correlation coefficient r = 0.95 across all loci, range r = 0.88–1.00 per locus;

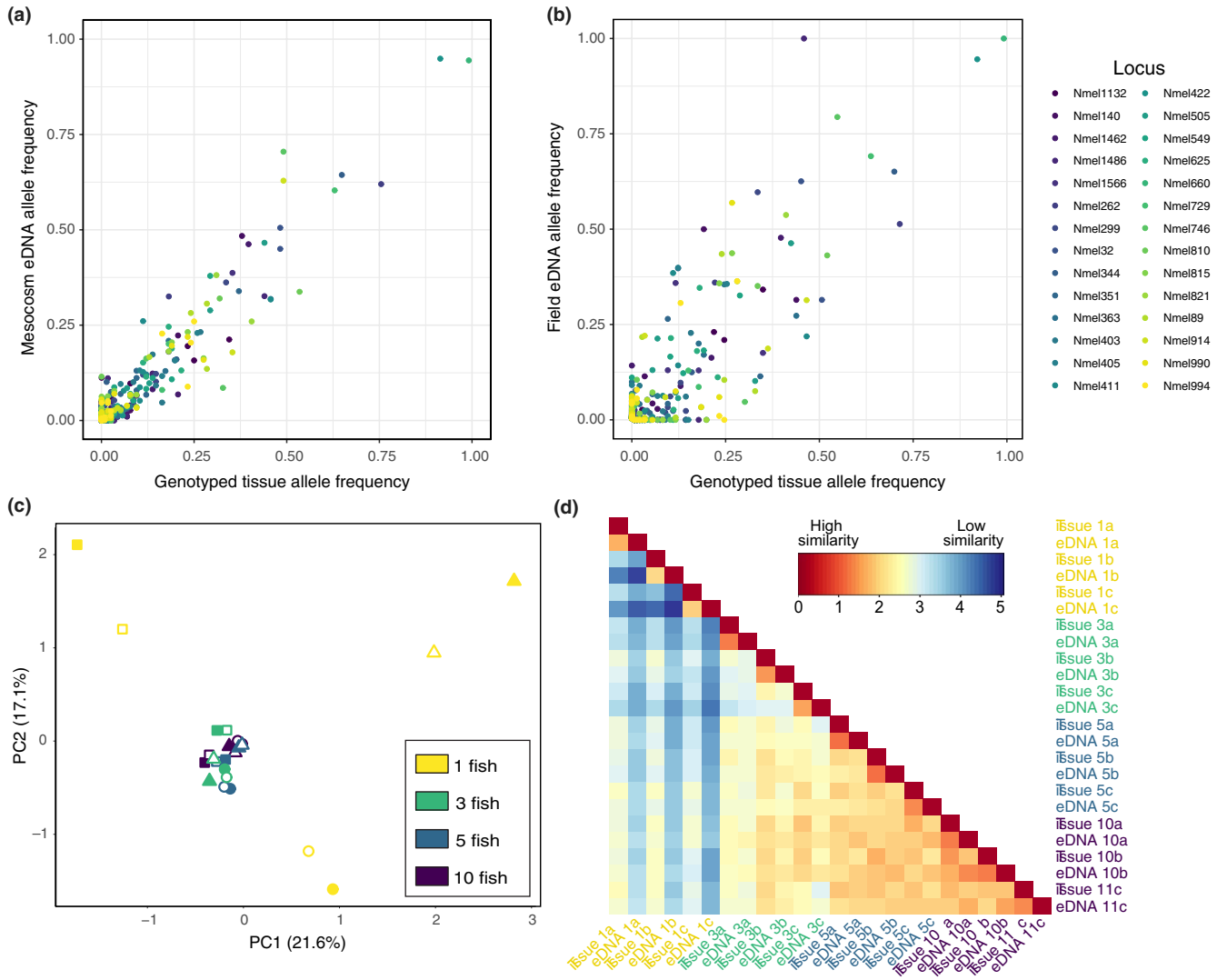


FIGURE 1 (a) Correlation between eDNA-derived and tissue-derived allele frequencies for all alleles across 28 loci in the mesocosm experiment. (b) Correlation between eDNA-derived and tissue-derived allele frequencies for all alleles across 28 loci in the field trial. (c) PCA of allele frequencies across 28 loci for round goby tissue samples (filled symbols) and eDNA samples (hollow symbols) from 12 mesocosms varying in round goby density. Colours represent mesocosm density treatments (1, 3, 5, or 10 fish) and symbols represent treatment replicates. (d) Heatmap of the pairwise Euclidean distances across all PC axes of allele frequencies from mesocosm eDNA and tissue samples, with blue colours indicating far distances (low similarity) and red colours indicating close distances (high similarity). Samples are arranged in pairs (eDNA/tissue samples) from each mesocosm, with colours representing mesocosm density treatments and letters (a, b, or c) representing treatment replicates [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 1a). Principal component analysis results showed high similarity between eDNA samples and genotyped individuals in each mesocosm, where eDNA samples clustered tightly with the individuals from the associated mesocosm (Figure 1c). Across all PC axes, the pairwise Euclidean distance was noticeably smaller within a mesocosm than between mesocosms (Figure 1d). eDNA and tissue sample pairs were most differentiated from other samples when derived from mesocosms with single round gobies, becoming more genetically similar to other mesocosms as the number of round gobies per mesocosm increased.

Allele frequencies from eDNA samples and genotyped individuals were also highly correlated in the field trial, with a Pearson's correlation coefficient of $r = 0.84$ across all loci (range $r = 0.41$–1.00 per locus; Figure 1b). Several alleles at low frequency in the population were not recovered by eDNA samples, with only 121 of 253 total alleles identified from genotyped individuals occurring in at least one of the three eDNA samples. However, all alleles with a frequency >0.24 in the 73 genotyped individuals were recovered by at least one of the three eDNA samples, and alleles not detected with eDNA occurred at low frequencies in the population (mean =0.03, SD = 0.04). On the other hand, eDNA samples also identified several alleles not documented in the genotyped individuals, albeit at low read frequencies (mean = 0.02, SD = 0.02). Such alleles may represent true low-frequency alleles not included in the genotyped individuals or may be the product of erroneous sequences.

3.3 | Contributor estimation

Estimates of the number of genetic contributors in mesocosms using observed alleles from genotyped tissue samples were within ±2 contributors at all round goby densities when population-level allele frequencies were specified using genotyped tissues and in mixtures of up to five individuals when allele frequencies were specified from eDNA read frequencies (top panel, Figure 2). When estimating the number of genetic contributors using observed alleles from eDNA samples, patterns of bias emerged across frequency thresholds below which reads were removed (0.001, 0.01, 0.1) regardless of how population-level allele frequencies were characterized. The contributor estimation was positively biased the lowest thresholds (0.001 and 0.01) across all mesocosm densities with the exception of the 10-individual mixtures using population allele frequencies from genotyped individuals, where estimates were within ±1 genetic contributor. Contributor estimations were also within ±1 contributors in mesocosms with one or three round gobies at the highest threshold (0.1), while negative bias was more apparent in mesocosms with five or 10 individuals at this threshold (Figure 2). Across all mesocosm densities, the variable threshold based on allelic richness outperformed all other thresholds (maximum bias = +5 associated with a 10-individual mixture using population allele frequencies from eDNA reads). Thus, adjusting the threshold according

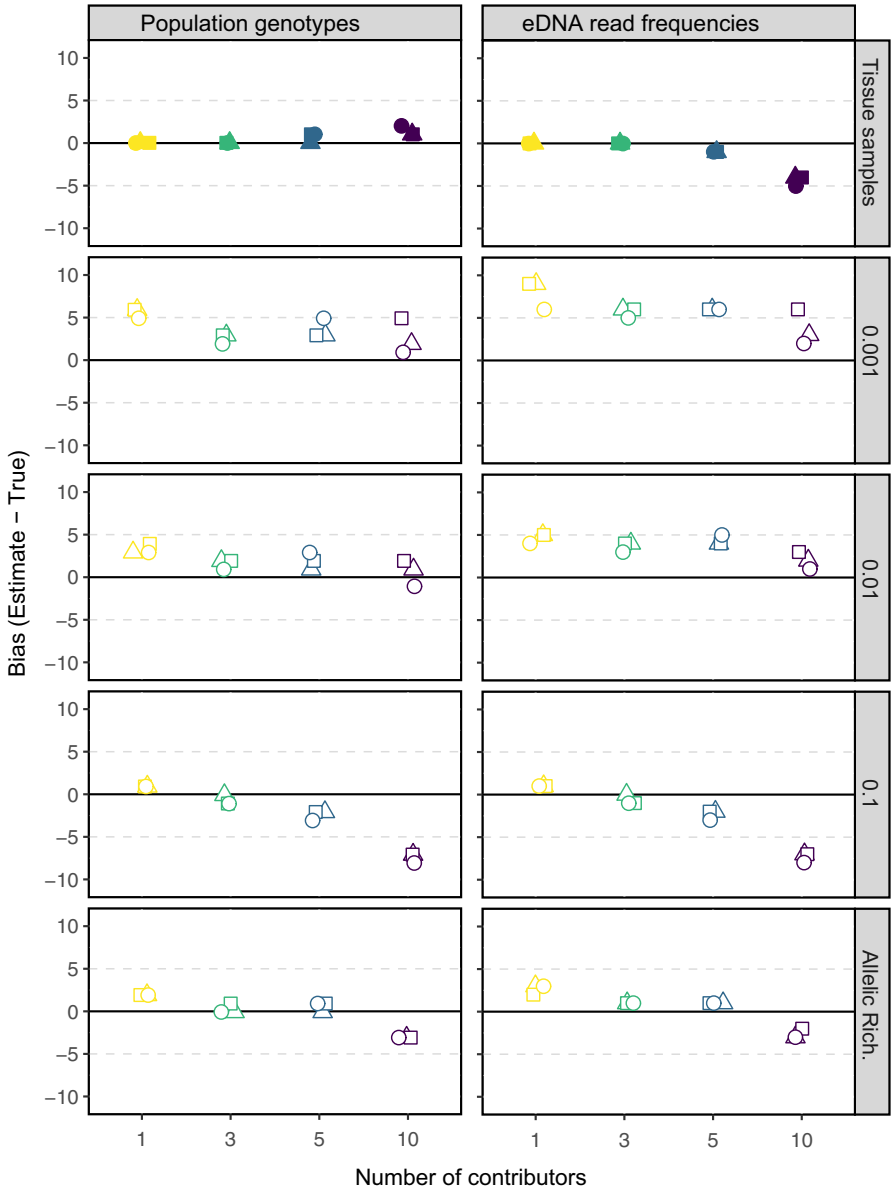


FIGURE 2 Bias of the contributor estimation using genotypes from round goby tissue samples (filled symbols) and eDNA samples (hollow symbols) across mesocosm treatments of round goby density (1, 3, 5, or 10 fish). The population allele frequencies for mixture estimation input were derived from 58 genotyped individual round gobies (left) or from eDNA read frequencies combined across all mesocosms. Symbols represent treatment replicates and panels indicate fixed threshold frequencies below which sequence reads were removed (0.1, 0.01, 0.001) or a variable threshold based on per-locus allelic richness (Allelic Rich.) [Colour figure can be viewed at wileyonlinelibrary.com]

to the per-locus allelic richness provides the most accurate estimate of absolute abundance in a mixed-DNA sample.

At the lowest threshold (0.001), contributor estimations of eDNA samples exhibited a positive bias across all densities, indicating the presence of false positive alleles in eDNA reads. However, the lack of cross-contamination in the negative control samples indicates the false positive alleles probably arose from artifacts introduced during PCR or sequencing, rather than cross-contamination between eDNA and tissue samples. The issue of positive bias in contributor estimations was more prevalent when population allele frequencies were specified using eDNA read frequencies. Nonetheless, patterns of bias and estimates of the number of genetic contributors in mesocosms were similar regardless of whether the input population allele frequencies were characterized using eDNA read frequencies or tissue-based allele frequencies (Figure 2). Thus, under controlled conditions, genotyped individuals may not be required to obtain reliable estimates of the absolute abundance of a species in eDNA samples.

In simulated mesocosm mixtures, the number of individuals could be reasonably estimated in mixtures up to 58 individuals, although bias associated with threshold values were apparent (Figure 3). The highest threshold (0.1) often exceeded the allele frequencies for all but the most common alleles, resulting in a negative bias in the contributor estimation. At this threshold, the maximum contributor estimation peaked at around 15 individuals, even at simulated densities >50 individuals. On the other hand, filtering the eDNA reads according to lower thresholds (0.01, 0.001) appeared to overestimate the number of contributors across all densities. The variable threshold based on allelic richness showed lower variation in the estimated number of contributors and performed well for all but the largest numbers of contributors, where a negative bias occurred. Thus, while our 28-locus panel was able to resolve mixtures of up to 58 individuals, filtering decisions can have a large effect on the estimated number of contributors in eDNA samples.

In the field trial, the contributor estimation resulted in an estimated five, three, and three genetically distinct individuals captured by the three replicate eDNA samples when population-level allele frequencies p were estimated from the 73 genotyped individuals. However, because the contributor estimation calculations only consider alleles from the specified population-level allele frequencies, this is probably an underestimate as we did not recover several low-frequency alleles from the genotyped tissues in the eDNA samples. When population-level allele frequencies were specified using the combined reads from the three replicate eDNA samples, an estimated 13, 7, and six genetically distinct individuals contributed to the mixture of DNA from each sample, respectively.

4 | DISCUSSION

Estimating the genetic diversity and abundance of a species provides insights into a wide range of ecological and evolutionary processes and may have important implications for conservation management opportunities. While analysis of eDNA is a well-established approach for detecting species, it also holds potential to detect genetic diversity within species (Adams et al., 2019; Sigsgaard et al., 2020). With this study, we use eDNA and NGS methods to detect intraspecific genetic diversity of an aquatic invasive species by recovering microsatellite allele frequencies that are similar to those derived from genotyped tissue samples in experimental mesocosms and in field-based eDNA collections. Using DNA mixture analyses, we estimated the number of genetic contributors of the target species within environmental samples, demonstrating the ability to use intraspecific genetic information to estimate the number of individuals captured in an eDNA sample. Although technical challenges regarding the parsing out of sequencing noise from low-abundance alleles in eDNA samples remain, this study experimentally validates the use of nuclear microsatellites to estimate population-level allele frequencies and absolute abundance of aquatic species using eDNA methods, a requisite step toward population-level inferences using nuclear eDNA.

To date, studies using eDNA approaches to characterize intraspecific genetic variation in aquatic species have been limited to

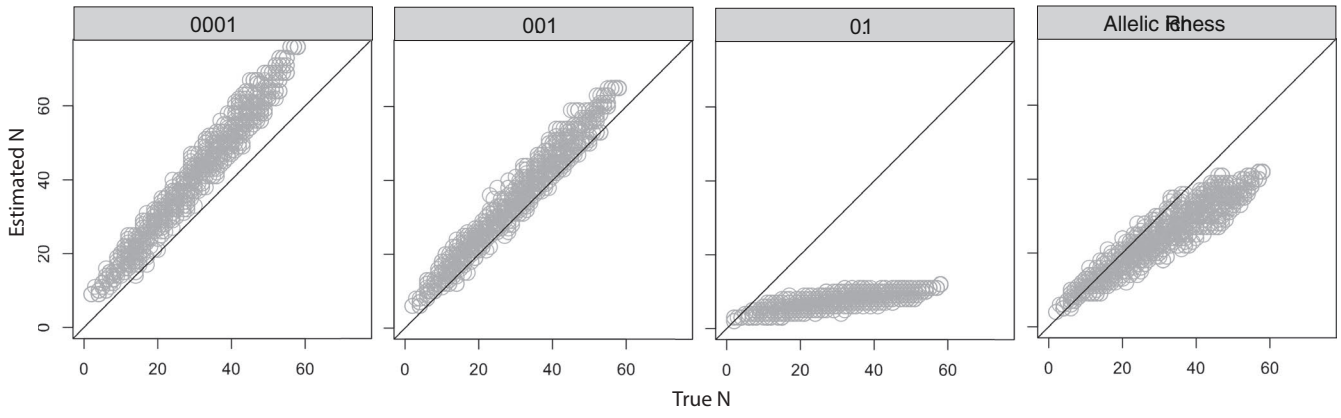


FIGURE 3 Estimated number of individuals contributing to simulated eDNA mixtures (range 2–58 individuals) using alleles from 1,000 simulated eDNA mixtures generated by constructing combinations of up to 12 mesocosms. Panels correspond to fixed threshold frequencies below which sequence reads were removed (0.001, 0.01, 0.1) or a variable threshold based on per-locus allelic richness. Diagonal lines represent a 1:1 relationship (i.e., zero bias for mixture contributor estimates)

a single locus in the mitochondrial genome (Elbrecht et al., 2018; Parsons et al., 2018; Sigsgaard et al., 2017; Tsuji et al., 2019; Turon et al., 2020). The expansion of eDNA approaches to target multiallelic nuclear DNA markers could allow for the detection of robust higher-resolution population-level genetic information from water samples, as is common practice in contemporary tissue-based population genetics studies. In controlled mesocosms, we document microsatellite allele frequencies from eDNA closely resembled tissue-based allele frequencies across all mesocosms and on a per-mesocosm basis, although our approach exhibited decreased sensitivity in genetically distinguishing mesocosms from one another at high round goby densities (Figure 1c–d). Because we used round gobies derived from a single population source, this is to be expected. We also demonstrate reasonably accurate allele frequency estimates from eDNA samples collected in natural conditions in a field trial, albeit with reduced detection of rare alleles in the population. Such eDNA-based estimates of population-level allele frequencies could potentially be used in population genetic inferences and demographic analyses using eDNA sampling methods. However, because eDNA samples contain a pool of DNA from many individuals, this approach is unable to determine multilocus genotypes or assign genotypes to individuals, and methods designed to analyse population genetics using individual genotypes will need to be adapted into an eDNA framework. Theoretical and analytical frameworks for estimating population genetic parameters from pooled tissue samples of many individuals (Pool-seq) have already been developed (Boitard et al., 2013; Gautier et al., 2013; Hivert et al., 2018), and similar frameworks may be useful for eDNA-based population genetics. As emphasized in Sigsgaard et al. (2020), however, such frameworks may need to account for additional potential sources of bias affecting the precision of population allele frequency estimates from eDNA, including variation in the number of individuals sequenced, unequal contributions of DNA from individuals, and variation from library preparation and sequencing.

Detecting intraspecific genetic variation in eDNA samples is also useful for estimating the number of genetically distinct individuals detected in a sample, which may be advantageous over approaches using DNA concentrations to predict species abundance or biomass. With the number of loci used in this study, the number of genetic contributors in simulated mixtures of up to 58 individuals could be resolved. While contributor estimations at the highest allele frequency threshold provided the most accurate estimates at low round goby densities in the mesocosm experiment, they were insufficient in resolving high numbers of round gobies, probably due to the removal of true low-frequency alleles below the threshold limits. In contrast, low thresholds sufficiently resolved the number of contributors at high round goby densities but erroneously inflated the number of contributors at low densities due to the introduction of false alleles. We therefore recommend bioinformatic filtering based upon moderate thresholds or variable thresholds associated with locus-specific allelic richness in future applications of DNA mixture analysis. However, we also caution future studies to further investigate the possible impacts of false alleles and allelic dropout on contributor estimations, particularly in field-based settings where false alleles are more difficult to distinguish from true low-abundance alleles and detection of rare alleles may be low. Because low-frequency alleles provide strong information on the number of individuals present in a sample (Sethi et al., 2019), efforts to maximize the recovery of low-frequency alleles through optimization of field and laboratory protocols may be required to obtain accurate estimates of the number of individuals captured in eDNA samples. Additionally, applications of error-correction algorithms and denoising procedures may be required to aid in the detection and removal of erroneous sequences while retaining true low-frequency alleles (Elbrecht et al., 2018; Tsuji et al., 2019; Turon et al., 2020).

Future eDNA studies may consider the use of single nucleotide polymorphisms (SNPs) as a target nuclear marker, as they are an abundant and widespread form of variation throughout the genome of most species (Morin et al., 2004). However, because the inferential power of the DNA mixture model is limited by the number of recovered alleles, much larger marker panels of biallelic SNPs will be needed to resolve eDNA mixtures into the number of genetic contributors, particularly as the number of contributors grows (Sethi et al., 2019). Rather than targeting single SNPs, a potential solution may be to target several SNPs occurring in the same genomic region that can be jointly genotyped (Kidd et al., 2013). Such multiallelic "microhaplotype" markers have high per-locus information content in a small length of DNA and may reduce the potential for analysis errors that arise when targeting microsatellites including PCR stutter and allelic dropout.

Although our approach demonstrates promise for future applications of noninvasive population genetic sampling using nuclear eDNA, the controlled settings of our mesocosm experiments and limited spatial and temporal scale of the field trial may not reflect the complexity of in situ conditions. Thus, several obstacles may need to be addressed before this approach can be broadly applied in field-based settings. For instance, although round gobies may exhibit localized hotspots of high density, the average density of round gobies in occupied habitats of Cayuga Lake (1.82 fish/m^2) is lower than in our mesocosm experiments (Andres et al., 2020), and read depths we observed in mesocosm eDNA samples may not be achievable in field settings. Indeed, even with targeted eDNA sampling in areas of high expected round goby densities, read depths in eDNA samples from the field trial averaged 4,305 reads per sample, which is much lower than reported in other eDNA studies using targeted field sampling and markers in the mitochondrial genome (e.g., average 263,111 reads per sample at sites where whale sharks were reported, Sigsgaard et al., 2017; average 237,434.5 reads per sample taken from harbour porpoise fluke prints, Parsons et al., 2018). To ensure genetic data obtained from eDNA samples sufficiently reflects the genetic diversity of the population of interest when targeting loci in the nuclear genome, efforts to evaluate the limit of detection and optimize field and laboratory strategies to achieve sufficient eDNA copy numbers may be required (Adams et al., 2019; Sigsgaard et al., 2020).

Mesocosm conditions also lacked the biophysical complexity inherent in natural systems, where many other particles and organisms are present and contributing to eDNA samples (Barnes & Turner, 2015). PCR inhibition from nontarget particles may restrict accurate molecular identification of alleles, particularly when coupled with low eDNA concentrations of target species DNA (Hunter et al., 2019). Importantly, if closely related nontarget species are found in the sampled habitats, primer specificity must be thoroughly tested to ensure DNA from co-occurring nontarget species is not amplified. While no congeners of the round goby are found in North America, the freshwater tubenose goby (*Proterorhinus semilunaris*, formerly known as *P. marmoratus*; Stepien & Tumeo, 2006) is found throughout the Great Lakes. Although we tested primer specificity in silico using DNA databases, in vitro testing using tissue-derived DNA from nontarget species may also be required if reference sequence data is lacking for closely related co-occurring species.

With proper validation and appropriate analytical frameworks, eDNA-based population genetics has the potential to enhance the use of eDNA methods in conservation and management of species. For example, preventing the spread and minimizing the undesirable impacts of invasive species will require effective monitoring of non-native populations, including evaluating population-level genetic variation and population size at the sites of initial colonization (Le Roux & Wieczorek, 2009). With further development, this method might someday inform management strategies at early stages in the invasion process when eradication efforts are most likely to be successful in preventing proliferation and future spread (Leung et al., 2002; Lodge et al., 2016). This approach may also be beneficial for monitoring species where small population sizes, expansive or complex habitats, elusive behaviour, or a desire to minimize invasive sampling can prevent effective population assessments. For instance, Parsons et al. (2018) used eDNA approaches to generate mitochondrial sequence data in a highly elusive marine mammal where physical tissue-based sampling presents logistical challenges and limits the detection of population genetic structure. The high sensitivity of eDNA methods and relative ease of sample collection therefore present a noninvasive and potentially cost-effective opportunity to study the population genetics of aquatic organisms for which traditional sampling is difficult or impossible.

As with other eDNA methods such as DNA metabarcoding, the approach developed here is likely to complement, rather than replace, existing methods of evaluating intraspecific diversity in population genetics studies (Yoccoz, 2012). Indeed, developing species-specific panels of microsatellite DNA markers requires sufficient DNA sequence data for the species of interest, and optimization of multiplex PCR requires testing on tissue-derived DNA samples. Estimating the number of contributors to eDNA samples also requires an assessment of population allele frequencies, an estimate that may be derived from tissue-based genotyping of the population of interest. However, we demonstrate that under controlled experimental conditions, population allele frequencies from eDNA read frequencies are highly correlated with allele frequencies from genotyped individuals and contributor estimations are similar regardless of where the population allele frequencies are derived. Estimating the number of contributors in eDNA samples may therefore be feasible even in the absence of population-level sequence information from tissue samples.

As the time and costs associated with obtaining and analysing molecular data continue to decline, eDNA methodologies may become an increasingly effective approach for detecting and quantifying the presence of invasive, rare, or threatened species. Moreover, with the recent expansion of eDNA approaches into studies of intraspecific diversity, the scope of eDNA applications has broadened to population-level inferences. With this study, we demonstrate the advancement of eDNA approaches to encompass genetic markers in the nuclear genome, with implications for future studies of population genetics using next-generation sequencing of environmental samples. By incorporating DNA mixture analyses into a nuclear genome-based eDNA framework, we estimate the number of unique contributors to eDNA samples, providing the first steps to a potential alternative to correlation-based estimates of species abundance using DNA concentration. Furthermore, we demonstrate the ability to obtain population-level genetic information from nuclear eDNA, supporting the potential for future assessments of population genetics from environmental samples. Provided further validation and optimization in field-based settings, such an advancement could transform the ways in which we obtain population-level genetic information on species of conservation or management concern.

ACKNOWLEDGEMENTS

This study was supported with funding provided by the National Science Foundation (NSF) Coastal SEES program (grant number 1748389) and the Department of Defence (DoD) Strategic Environmental Research and Development Program (SERDP) (grant number RC19-1004). We thank Amanda Wong and Amelia Weiss for assistance with collecting specimens. We also thank Wes Larson and Timothy Lambert for discussions regarding the application and analysis of contributor estimations using eDNA samples. We are grateful to three anonymous reviewers for their valuable comments on the manuscript. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the US Government.

AUTHOR CONTRIBUTIONS

All authors conceived and designed the study, interpreted results, and contributed to writing the manuscript. K.J.A. conducted the study and collected specimens. K.J.A., and J.A completed laboratory work and analysed the data. All authors approved the manuscript for publication.

DATA AVAILABILITY STATEMENT

Illumina MiSeq raw sequence data are uploaded to NCBI's Sequence Read Archive (BioProject ID: PRJNA680257). Microsatellite primers are available in Table S1. All scripts used in the data processing and analysis are available on GitHub (https://github.com/karaandres/eDNA_goby_mesocosms).

ORCID

Kara J. Andres https://orcid.org/0000-0003-4822-7047

REFERENCES

Adams, C. I., Knapp, M., Gemmell, N. J., Jeunen, G. J., Bunce, M., Lamare, M. D., & Taylor, H. R. (2019). Beyond biodiversity: Can environmental DNA (eDNA) cut it as a population genetics tool? *Genes*, *10*(3), 192.

Andres, K. J., Sethi, S. A., Duskey, E., Lepak, J. M., Rice, A. N., Estabrook, B. J., Fitzpatrick, K. B., George, E., Marcy-Quay, B., Paufve, M. R., Perkins, K., & Scofield, A. E. (2020). Seasonal habitat use indicates that depth may mediate the potential for invasive round goby impacts in inland lakes. *Freshwater Biology*, *65*(8), 1337–1347.

Ballard, J. W. O., & Whitlock, M. C. (2004). The incomplete natural history of mitochondria. *Molecular Ecology*, *13*, 729–744.

Barnes, M. A., & Turner, C. R. (2015). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*(1), 1–17. https://doi.org/10.1007/s10592-015-0775-4.

Bohmann, K., Evans, A., Gilbert, M. T. P., Carvalho, G. R., Creer, S., Knapp, M., Yu, D. W., & de Bruyn, M. (2014). Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, *29*(6), 358–367. https://doi.org/10.1016/j.tree.2014.04.003

Boitard, S., Kofler, R., Françoise, P., Robelin, D., Schlötterer, C., & Futschik, A. (2013). Pool- hmm: A Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. *Molecular Ecology Resources*, *13*, 337–340.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120.

Bylemans, J., Furlan, E. M., Hardy, C. M., McGuffie, P., Lintermans, M., & Gleeson, D. M. (2017). An environmental DNA-based method for monitoring spawning activity: a case study, using the endangered Macquarie perch (*Macquaria australasica*). *Methods in Ecology and Evolution*, *8*(5), 646–655.

Charlebois, P. M., Marsden, J. E., Goettel, R. G., Wolfe, R. K., Jude, D. J., & Rudnika, S. (1997). The round goby, *Neogobius melanostomus* (Pallas), a review of European and North American Literature (Vol. INHS Special Publication No. 20): Illinois-Indiana Sea Grant Program and Illinois Natural History Survey.

Curran, J. M., Triggs, C. M., Buckleton, J., & Weir, B. (1999). Interpreting DNA mixtures in structured populations. *Journal of Forensic Science*, *44*(5), 987–995.

D'Aloia, C. C., Bogdanowicz, S. M., Harrison, R. G., & Buston, P. M. (2017). Cryptic genetic diversity and spatial patterns of admixture within Belizean marine reserves. *Conservation Genetics*, *18*(1), 211–223.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., & De Vere, N. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895.

Deiner, K., Renshaw, M. A., Li, Y., Olds, B. P., Lodge, D. M., & Pfrender, M. E. (2017). Long-range PCR allows sequencing of mitochondrial genomes from environmental DNA. *Methods in Ecology and Evolution*, *8*(12), 1888–1898.

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E., & Miaud, C. (2012). Improved detection of an alien invasive species through environmental DNA barcoding: the example of the American bullfrog *Lithobates catesbeianus*. *Journal of Applied Ecology*, *49*(4), 953–959.

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, *6*, e4644.

Faircloth, B. C. (2008). MSATCOMMANDER: Detection of microsatellite repeat arrays and automated, locus-specific primer design. *Molecular Ecology Resources*, *8*(1), 92–94.

Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, *4*(4), 423–425. https://doi.org/10.1098/rsbl.2008.0118.

Gautier, M., Foucaud, J., Gharbi, K., Cézard, T., Galan, M., Loiseau, A., Thomson, M., Pudlo, P., Kerdelhué, C., & Estoup, A. (2013). Estimation of population allele frequencies from next-generation sequencing data: Pool- versus individual-based genotyping. *Molecular Ecology*, *22*, 3766–3779.

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., & Cornman, R. S. (2016). Critical considerations for the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307.

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.

Hivert, V., Leblois, R., Petit, E. J., Gautier, M., & Vitalis, R. (2018). Measuring genetic differentiation from pool-seq data. *Genetics*, *210*, 315–330.

Hunter, M. E., Ferrante, J. A., Meigs-Friend, G., & Ulmer, A. (2019). Improving eDNA yield and inhibitor reduction through increased water volumes and multi-filter isolation techniques. *Scientific Reports*, *9*(1), 1–9.

Hurst, G. D., & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1572), 1525–1534.

Iversen, L. L., Kielgast, J., & Sand-Jensen, K. (2015). Monitoring of animal abundance by environmental DNA — An increasingly obscure perspective: A reply to Klymus et al,Monitoring of animal abundance by environmental DNA — An increasingly obscure perspective: A reply to Klymus. *Biological Conservation*, *192*, 479–480. https://doi.org/10.1016/j.biocon.2015.09.024

Janssen, J., & Jude, D. J. (2001). Recruitment failure of mottled sculpin *Cottus bairdi* in Calumet Harbor, southern Lake Michigan, induced by the newly introduced round goby *Neogobius melanostomus*. *Journal of Great Lakes Research*, *27*(3), 319–328. https://doi.org/10.1016/s0380-1330(01)70647-8

Jerde, C. L., Mahon, A. R., Chadderton, W. L., & Lodge, D. M. (2011). "Sight-unseen" detection of rare aquatic species using environmental DNA. *Conservation Letters*, *4*(2), 150–157.

Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*, 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jude, D. J., Reider, R. H., & Smith, G. R. (1992). Establishment of Gobiidae in the Great Lakes Basin. *Canadian Journal of Fisheries and Aquatic Sciences*, *49*(2), 416–421. https://doi.org/10.1139/f92-047

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.

Kelly, R. P., Shelton, A. O., & Gallego, R. (2019). Understanding PCR processes to draw meaningful conclusions from environmental DNA studies. *Scientific Reports*, *9*(1), 1–14.

Kidd, K. K., Pakstis, A. J., Speed, W. C., Lagace, R., Chang, J., Wootton, S., & Ihuegbu, N. (2013). Microhaplotype loci are a powerful new type of forensic marker. *Forensic Science International: Genetics Supplement Series*, *4*(1), e123–e124.

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.

Kornis, M. S., Mercado-Silva, N., & Vander Zanden, M. J. (2012). Twenty years of invasion: a review of round goby *Neogobius melanostomus* biology, spread and ecological implications. *Journal of Fish Biology*, *80*(2), 235–285. https://doi.org/10.1111/j.1095-8649.2011.03157.x

Krakowiak, P. J., & Pennuto, C. M. (2008). Fish and macroinvertebrate communities in tributary streams of eastern Lake Erie with and without round gobies (*Neogobius melanostomus*, Pallas 1814). *Journal of Great Lakes Research*, *34*(4), 675–690. https://doi.org/10.3394/0380-1330-34.4.675

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.

Le Roux, J., & Wieczorek, A. (2009). Molecular systematics and population genetics of biological invasions: towards a better understanding of invasive species management. *Annals of Applied Biology*, *154*(1), 1–17.

Leung, B., Lodge, D. M., Finnoff, D., Shogren, J. F., Lewis, M. A., & Lamberti, G. (2002). An ounce of prevention or a pound of cure: bioeconomic risk analysis of invasive species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *269*(1508), 2407–2413.

Lodge, D. M., Simonin, P. W., Burgiel, S. W., Keller, R. P., Bossenbroek, J. M., Jerde, C. L., Kramer, A. M., Rutherford, E. S., Barnes, M. A., Wittmann, M. E., Chadderton, W. L., Apriesnig, J. L., Beletsky, D., Cooke, R. M., Drake, J. M., Egan, S. P., Finnoff, D. C., Gantz, C. A., Grey, E. K., … Zhang, H. (2016). Risk analysis and bioeconomics of invasive species to inform policy and management. *Annual Review of Environment and Resources*, *41*, 453–488.

Lodge, D. M., Turner, C. R., Jerde, C. L., Barnes, M. A., Chadderton, L., Egan, S. P., Feder, J. L., Mahon, A. R., & Pfrender, M. E. (2012). Conservation in a cup of water: estimating biodiversity and population abundance from environmental DNA. *Molecular Ecology*, *21*(11), 2555–2558.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376.

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.

Minamoto, T., Uchii, K., Takahara, T., Kitayoshi, T., Tsuji, S., Yamanaka, H., & Doi, H. (2017). Nuclear internal transcribed spacer-1 as a sensitive genetic marker for environmental DNA studies in common carp Cyprinus carpio. *Molecular Ecology Resources*, *17*(2), 324–333.

Morin, P. A., Luikart, G., Wayne, R. K., & the SNP Workshop Group (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*, *19*(4), 208–216.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, *26*(3), 419–420.

Parsons, K. M., Everett, M., Dahlheim, M., & Park, L. (2018). Water, water everywhere: environmental DNA can unlock population structure in elusive marine species. *Royal Society Open Science*, *5*(8), 180537.

Piggott, M. P. (2016). Evaluating the effects of laboratory protocols on eDNA detection probability for an endangered freshwater fish. *Ecology and Evolution*, *6*(9), 2739–2750.

Pilliod, D. S., Goldberg, C. S., Arkle, R. S., & Waits, L. P. (2013). Estimating occupancy and abundance of stream amphibians using environmental DNA from filtered water samples. *Canadian Journal of Fisheries and Aquatic Sciences*, *70*(8), 1123–1130.

R Core Team (2016). *R: A language and environment for statistical computing*.

S. Rozen, & H. Skaletsky (Eds.) (2000). Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols* (pp. 365–386). Springer.

Rubinoff, D., Cameron, S., & Will, K. (2006). A genomic perspective on the shortcomings of mitochondrial DNA for "barcoding" identification. *Journal of Heredity*, *97*(6), 581–594.

Schaeffer, J. S., Bowen, A., Thomas, M., French, J. R. III, & Curtis, G. L. (2005). Invasion history, proliferation, and offshore diet of the round goby *Neogobius melanostomus* in western Lake Huron, USA. *Journal of Great Lakes Research*, *31*(4), 414–425.

Sethi, S. A., Larson, W., Turnquist, K., & Isermann, D. (2019). Estimating the number of contributors to DNA mixtures provides a novel tool for ecology. *Methods in Ecology and Evolution*, *10*(1), 109–119.

Sigsgaard, E. E., Jensen, M. R., Winkelmann, I. E., Møller, P. R., Hansen, M. M., & Thomsen, P. F. (2020). Population-level inferences from environmental DNA—Current status and future perspectives. *Evolutionary Applications*, *13*(2), 245–262.

Sigsgaard, E. E., Nielsen, I. B., Bach, S. S., Lorenzen, E. D., Robinson, D. P., Knudsen, S. W., Pedersen, M. W., Jaidah, M. A., Orlando, L., Willerslev, E., Møller, P. R., & Thomsen, P. F. (2017). Population characteristics of a large whale shark aggregation inferred from seawater environmental DNA. *Nature Ecology & Evolution*, *1*(1).

Spens, J., Evans, A. R., Halfmaerten, D., Knudsen, S. W., Sengupta, M. E., Mak, S. S., & Hellström, M. (2017). Comparison of capture and storage methods for aqueous macrobial eDNA using an optimized extraction protocol: advantage of enclosed filter. *Methods in Ecology and Evolution*, *8*(5), 635–645.

Stepien, C. A., & Tumeo, M. A. (2006). Invasion genetics of Ponto-Caspian gobies in the Great Lakes: a 'cryptic' species, absence of founder effects, and comparative risk analysis. *Biological Invasions*, *8*(1), 61–78.

Stewart, K., Ma, H., Zheng, J., & Zhao, J. (2017). Using environmental DNA to assess population-wide spatiotemporal reserve use. *Conservation Biology*, *31*(5), 1173–1182.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, *21*(8), 1789–1793.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050.

Takahara, T., Minamoto, T., Yamanaka, H., Doi, H., & Kawabata, Z. I. (2012). Estimation of fish biomass using environmental DNA. *PLoS One*, *7*(4), e35868.

Teske, P. R., Golla, T. R., Sandoval-Castillo, J., Emami-Khoyi, A., van der Lingen, C. D., von der Heyden, S., Chiazzari, B., Jansen van Vuuren, B., & Beheregaray, L. B. (2018). Mitochondrial DNA is unsuitable to test for isolation by distance. *Scientific Reports*, *8*(1), 8448.

Thomsen, P. F., Kielgast, J., Iversen, L. L., Møller, P. R., Rasmussen, M., & Willerslev, E. (2012). Detection of a diverse marine fish fauna using environmental DNA from seawater samples. *PLoS One*, *7*(8), e41732.

Thomsen, P. F., & Willerslev, E. (2015). Environmental DNA – An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, *183*, 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2019). Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA*.

Turon, X., Antich, A., Palacín, C., Præbel, K., & Wangensteen, O. S. (2020). From metabarcoding to metaphylogeography: separating the wheat from the chaff. *Ecological Applications*, *30*(2), e02036.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J.-M., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA

metabarcoding. *Molecular Ecology*, *25*(4), 929–942. https://doi.org/10.1111/mec.13428

Vander Zanden, M. J., Hansen, G. J., Higgins, S. N., & Kornis, M. S. (2010). A pound of prevention, plus a pound of cure: early detection and eradication of invasive species in the Laurentian Great Lakes. *Journal of Great Lakes Research*, *36*(1), 199–205.

Vélez-Espino, L. A., Koops, M. A., & Balshine, S. (2010). Invasion dynamics of round goby (*Neogobius melanostomus*) in Hamilton Harbour. *Lake Ontario. Biological Invasions*, *12*(11), 3861–3875. https://doi.org/10.1007/s10530-010-9777-9

Weir, B. S., Triggs, C., Starling, L., Stowell, L., Walsh, K., & Buckleton, J. (1997). Interpreting DNA mixtures. *Journal of Forensic Science*, *42*(2), 213–222.

Willerslev, E., Cappellini, E., Boomsma, W., Nielsen, R., Hebsgaard, M. B., Brand, T. B., & Dahl-Jensen, D. (2007). Ancient biomolecules from deep ice cores reveal a forested southern Greenland. *Science*, *317*(5834), 111–114.

Willerslev, E., Hansen, A. J., Binladen, J., Brand, T. B., Gilbert, M. T. P., Shapiro, B., & Cooper, A. (2003). Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, *300*(5620), 791–795.

Yoccoz, N. G. (2012). The future of environmental DNA in ecology. *Molecular Ecology*, *21*(8), 2031–2038.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Andres KJ, Sethi SA, Lodge DM, Andrés J. Nuclear eDNA estimates population allele frequencies and abundance in experimental mesocosms and field samples. *Mol Ecol*. 2021;30:685–697. https://doi.org/10.1111/mec.15765