Population genomics of New Zealand pouched lamprey (kanakana; piharau; *Geotria australis*)

Allison K. Miller[1], Nataliya Timoshevskaya[2], Jeramiah J. Smith[2], Joanne Gillum[1], Saeed Sharif[1], Shannon Clarke[3], Cindy Baker[4], Jane Kitson[5], Neil J. Gemmell[1*], and Alana Alexander[1*]

1. *Anatomy Department, School of Biomedical Sciences, University of Otago, 270 Great King Street, Dunedin, 9016, New Zealand*
2. *Department of Biology, University of Kentucky, 101 Morgan Building, Lexington, Kentucky, 40506-0225 USA*
3. *AgResearch, Invermay Agricultural Centre, Mosgiel 9053, New Zealand*
4. *National Institute of Water and Atmospheric Research Limited, PO Box 11 115, Hamilton 3251 New Zealand*
5. *Ngāi Tahu, Kitson Consulting Ltd, Invercargill/Waihopai, 9879, New Zealand*

Allison K. Miller, allison.miller@postgrad.otago.ac.nz*, ORCID 0000-0002-5334-2771*

Nataliya Timoshevskaya, nti225@uky.edu*, ORCID 0000-0002-9775-5877*

Jeramiah J. Smith*, jjsmit3@uky.edu, ORCID 0000-0001-5333-5531*

Joanne Gillum*, joanne.gillum@otago.ac.nz*

Saeed Sharif, saeed.sharif@gribbles.co.nz

Shannon Clarke, Shannon.clarke@agresearch.co.nz*, ORCID 0000-0002-4615-8917*

Cindy F Baker*, Cindy.Baker@niwa.co.nz*

Jane Kitson, jane@kitsonconsulting.co.nz*, ORCID 0000-0003-3022-5123*

Neil J. Gemmell*, neil.gemmell@otago.ac.nz*, ORCID* 0000-0003-0671-3637

Alana Alexander*, alana.alexander@otago.ac.nz*, ORCID 0000-0002-6456-7757*

* These authors contributed equally to this work

Corresponding author: Allison K. Miller, active e-mail: akmiller333@gmail.com or milal712@student.otago.ac.nz

Abstract

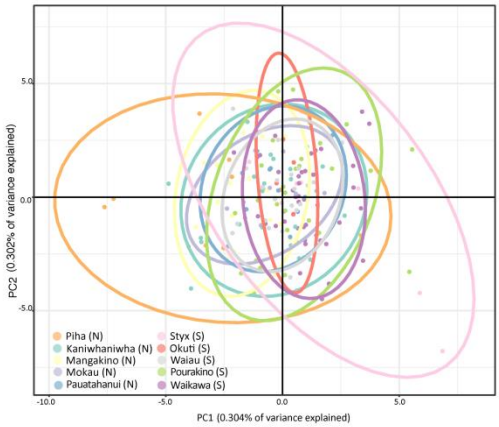
Pouched lamprey (*Geotria australis*) or kanakana/piharau is a culturally and ecologically significant jawless fish that is distributed throughout Aotearoa New Zealand. Despite its importance, much remains unknown about historical relationships and gene flow between populations of this enigmatic species within New Zealand. To help inform management, we assembled a draft *Geotria australis* genome and completed the first comprehensive population genomics analysis of pouched lamprey within New Zealand using targeted gene sequencing (Cyt-b and COI) and restriction site-associated DNA sequencing (RADSeq) methods. Employing 16,000 genome-wide single nucleotide polymorphisms (SNPs) derived from RADSeq (n=186) and sequence data from Cyt-b (766 bp, n=94) and COI (589 bp, n=20), we reveal low levels of structure across 10 sampling locations spanning the species range within New Zealand. F-statistics, outlier analyses, and STRUCTURE suggest a single panmictic population, and Mantel and EEMS tests reveal no significant isolation by distance. This implies either ongoing gene flow among populations or recent shared ancestry among New Zealand pouched lamprey. We can now use the information gained from these genetic tools to assist managers with monitoring effective population size, managing potential diseases, and conservation measures such as artificial propagation programs. We further demonstrate the general utility of these genetic tools for acquiring information about elusive species.

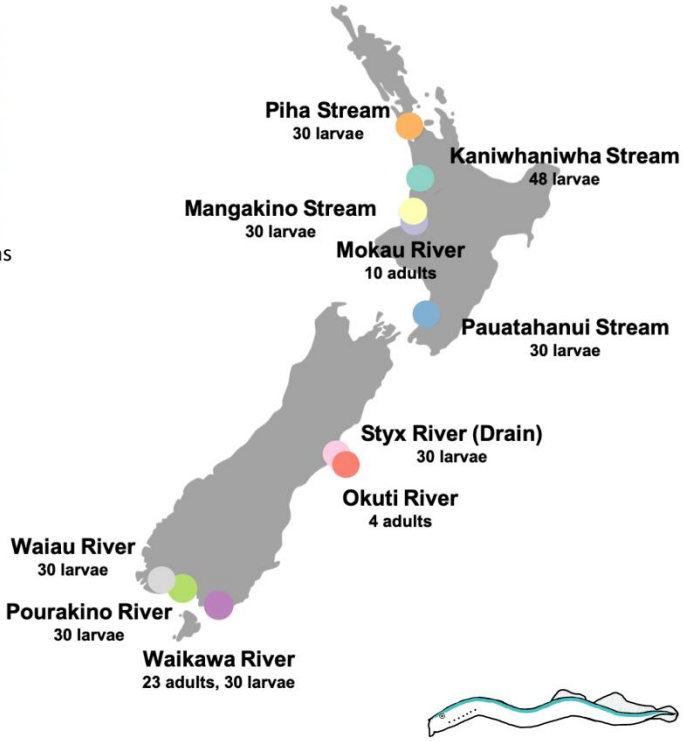
Keywords

conservation genetics, gene flow, genome, historical demography, Southern Hemisphere lamprey, effective population size

Graphical abstract

Aotearoa New Zealand lamprey sampling locations





Abstract te reo Māori

E mihi ana mātou ki a Papatūānuku rāua ko Ranginui, me ngā awa kāinga o te kanakana/piharau hoki. Ka mihi hoki mātou ki ngā whānau, ngā hapū, me ngā iwi e tiaki nei i tēnei taonga, ā, ki te kanakana/piharau hoki nāna i whakatakoto te ara o tēnei rangahau. Nā ngā ingoa maha mō tēnei kararehe (tae ana ki ko pihapiharau, rātou ko kanakana, ko korokoro, ko kanakana wairaki, ko nganangana, ko pipiharau, ko puhikorokoro, ko tuna korokoro, ko pia, ko piharau) ko kanakana/piharau te ingoa ka whakamahi mō tēnei tuhinga. He taonga tēnei kararehe kauae kore, te kanakana/piharau (*Geotria australis*), ā, he hirahira hauropi hoki ki a Aotearoa whānui. Ahakoa tōna whakahirahira me te hōhonu o te mātauranga Māori, he nui ngā pātai e toe ana e pā ana ki te whakapapa tawhito me te whakawhiti o ngā ira kei waenganui ngā taupori o tēnei momo manganga kei Aotearoa, te kanakana/piharau. Hei tautoko ngā kaitiaki, i hangaia e mātou tētahi huinga ira hukihuki o te *Geotria australis*. Waihoki, nā te pānui o te pītau ira o ētahi ira (Cyt-b me COI) me ngā ira karihi pīwawa (RADseq), i whakaputa te tātari matai iranga tuatahi o ngā taupori o te piharau/kanakana ki Aotearoa ki ngā ira karihi. Nā te whakamahi o ngā pītau ira (SNPs) 16,000 mai te RADSeq (tīpako = 186), te Cyt-b (766 te roanga, tīpako = 94), me te COI (589 te roanga, tīpako = 20) i whāwhāki mātou he iti rawa te rerekētanga o ngā rohe pītau ira i waenganui ngā wahi tīpako e horahia ngā nohonga whānui o ngā piharau/kanakana. E mea ana ētahi tatauranga (arā, F-statistics, outlier analyses, me STRUCTURE) kōtahi anake te taupori kanakana/piharau o Aotearoa. I kite hoki i roto i ngā tatauranga Mantel me EEMS kahore he wehewehenga o ngā pītau ira nā te pāmamao o ngā awa. Nā reira, e pūheke tonu ana te ai o ngā kanakana/piharau i waenganui ngā taupori o te motu, he whānaunga tata ngā tupuna o ngā piharau/kanakana nō Aotearoa rānei. E ai ki te whakataukī, he kaha te manawa o te piharau. Ko te tūmanako he orite te oranga o te piharau/kanakana ki te manawa piharau, ā, ka āwhina tēnei rangahau i ngā kaitiaki kia āta mātakitaki te rahi o te taupori, kia whakamauru ngā tahumaero, ā, kia tautoko te whakamakurutanga o te piharau/kanakana hoki. He pai hoki ēnei tātari ira mō te rangahau i ngā kararehe matawhiwhiu.

Introduction

Diadromous fish species, fishes which spend portions of their life cycles in both fresh and marine environments, comprise ~1.5% of all described fishes yet they amount to ~3% of all endangered species (McDowall, 1992). These species are likely at enhanced risk since they must find passages through freshwater barriers and combat the threats of both freshwater and marine systems. These threats include habitat degradation, overexploitation, invasive species, poor forestry practices, pollution, climate change, disease, marine overexploitation, and bycatch (McDowall, 1992; Dudgeon *et al*, 2006; He *et al*, 2017; WWF, 2018).

Geotria australis Gray, 1851 (pouched lamprey) is known as kanakana by southern iwi (Indigenous New Zealand tribes) and piharau by northern iwi in Aotearoa New Zealand. It is a diadromous, jawless vertebrate in Geotriidae (Petromyzontiformes) (Figure 1) and is one of the most widely distributed lampreys; *G. australis* is found in Chile, Australia, and throughout Aotearoa New Zealand, including Te Ika-a-Māui (the North Island), Te Waipounamu (the South Island), Rakiura (Stewart Island), and Rēkohu/Wharekauri (the Chatham Islands) (James, 2008; McDowall, 1990; McDowall, 2002; Tedd and Kelso, 1993). Similar to salmon, *Geotria* are anadromous; eggs and larvae develop in freshwater for approximately three to four years, juveniles migrate downstream to the ocean where they feed and mature, and mature adults migrate back into freshwater where they spawn and die (Baker *et al*, 2017; Miller *et al*, 2021). At all life stages *Geotria* are prey for multiple species, including endangered species, and likely act as ecosystem engineers during their larval stages (Docker *et al*, 2015; McDowall, 1990; Moore and Wakelin, 1997; Tickell, 1964).

Despite their cultural importance, biological importance, and moderate size (migrating mature adults ~500 mm length), pouched lamprey are often physically overlooked in both freshwater and marine environments. As a nocturnally-moving species that hides during the day, pouched lamprey are often missed or visually misidentified as eels (Jellyman and Glova, 2002; Kelso and Glova, 1993). This has led them to be referred to as an “elusive” or “secretive” species. Adults and juveniles are not large enough to fit satellite tags, hampering attempts to both monitor their marine life stage and determine if adults are philopatric, i.e. if they return to natal or “home” streams.

Against this dearth of knowledge around population connectivity, pouched lamprey abundance appears to be declining in Aotearoa New Zealand (Best, 1941; Dunn *et al*, 2018; Kitson *et al*, 2012; McDowall, 1990; McDowall, 2011). Reasons for these declines are unknown, although, anthropogenic stressors and disease are possible causes. Lamprey reddening syndrome (LRS), a syndrome associated with external haemorrhaging around the gill-pores and along the body, has been recorded on upstream migrating pouched lamprey from the South Island since 2011 (Brosnahan *et al*, 2018). The lack of rudimentary knowledge concerning the geographical relationships (population structure), degree of philopatry, and migration patterns of pouched lamprey in New Zealand has greatly hindered the ability of managers to implement robust conservation policy for this species. The potential risk of LRS transmission between the North and South Islands is a particular concern. In addition, understanding why certain pouched lamprey population patterns exist (e.g., recent origin/limited gene flow, and [or] more ancient origin/elevated gene flow) will allow managers to better understand the vulnerability and stability of the current pouched lamprey population(s).

Here we investigate population connectivity throughout Aotearoa New Zealand based on 186 pouched lamprey sampled from 5 North Island and 5 South Island locations. We use a newly assembled draft genome, restriction site associated DNA sequencing (RADSeq; Dodds *et al*, 2015; Elshire *et al*, 2011), and targeted sequencing of mitochondrial DNA (mtDNA) to investigate genetic structure, gene flow, and historical demography of pouched lamprey in Aotearoa New Zealand. Using these inferences, we then demonstrate how genomic methods, such as RADSeq, can be used to enhance information about the current (and historical) population structure, size, and migration of pouched lamprey and other elusive species. We further illustrate how this information can be used to assist managers with translocation, propagation, disease management, and population monitoring through time.

Materials and Methods

Sample collections

New Zealand National Institute of Water and Atmospheric Research (NIWA) scientists and iwi (Māori tribal) representatives collected pouched lamprey adults and larvae from 2016 to 2019 using electrofishing, hand-catching, and trapping methods. Specifically, 258 larvae and 37 adults were collected (Table S1). Engagement with local iwi was undertaken before all collection events. Following recommendations by Hale *et al* (2012), approximately 30 individuals were targeted during each sampling event; however, tissue quality and relatedness scores following initial data analysis determined the final number of individuals incorporated in our analyses. Tissues were collected from the 295 total individuals during 12 sampling events at 10 sites (North Island: Mokau, Piha, Kaniwhaniwha, Pauatahanui, and Mangakino; South Island: Styx, Waiau, Waikawa, Okuti, and Pourakino) (Figure 2). The Kaniwhaniwha and Waikawa sites were sampled twice. Sample identification (ID), location, year sampled, life stage, library number, raw read counts, number of loci, Aotearoa Genomic Data Repository accession links, targeted markers used, and NCBI GenBank accession numbers are provided in Table S1. For adult pouched lamprey, a fin clip from the caudal fin or from the second dorsal fin was placed into RNAlater® [Thermo Fisher Scientific], and for larvae, the entire individual was placed in 15ml vials of RNAlater®. Additional muscle and gonadal tissue was collected from one mature male sampled at the Okuti site, preserved in RNAlater®, and used for the draft genome assembly. All tissues were stored at -20 °C until they were used for DNA extractions.

DNA extraction and sequencing

DNA was extracted from subsampled tissues following Gemmell and Akiyama (1996), quantified via Qubit Fluorometer (Thermo Fisher Scientific), and DNA integrity assessed using gel electrophoresis.

Targeted mitochondrial DNA markers

Mitochondrial DNA cytochrome b (Cyt-b) sequences were amplified from 94 New Zealand pouched lamprey using Geotria496L and Phe1612H primers (Lang *et al*, 2009). The 20 µL polymerase chain reaction (PCR) was carried out using Bioline Biotaq DNA polymerase and comprised of the following:

14.1 µL ddH_2O, 2 µL of 10 × reaction buffer, 1.2 µL of $MgCl_2$ (50 mM), 0.5 µL of each dNTPs (10 mM), 0.2 µL *Taq* polymerase (5 µ/µL), 0.5 µL of forward primer (10 µM), 0.5 µL reverse primer (10 µM), and 1 µL template DNA. The thermoprofile consisted of an initial denaturation of 94 °C for 3 min, followed by 30 cycles of 94 °C for 1 min, annealing at 60 °C for 1 min, and extension at 72 °C for 2 min, with a final 10 min extension. Cytochrome c oxidase I (COI) sequences were amplified using the primers (FishF1 and FishR1) following Ward *et al* (2005). PCR products were purified using either PALL AcroPrep96 filter plates (http://www.pall.com) or exonuclease 1 and shrimp alkaline phosphatase (https://www.nucleics.com/DNA_sequencing_support/exonucleaseI-SAP-PCR-protocol.html), and sequenced (reverse and forward directions) using BigDye v3.1 dye terminator chemistry, on an Applied Biosystems 3730xl DNA Analyzer (Applied Biosystems, USA), following dye-terminator removal using columns prepared with Sephadex G-50 Fine grade (Cytiva Life Sciences). All available *Mordacia* (lamprey outgroup taxa, Mordaciidae) and *Geotria australis* Cyt-b and COI sequences were downloaded from NCBI GenBank and included with the newly sequenced Cyt-b and COI regions in downstream analyses (sequences and accession numbers in Table S1). All new sequences from Aotearoa New Zealand were deposited into the Aotearoa Genomic Data Repository (https://repo.data.nesi.org.nz/TAONGA-LAMPREY/collection/TLCMC00005-00002).

Geotria australis genome sequencing and assembly

Since reference-based RADSeq analyses were found to outperform *de novo* approaches (Rochette *et al*, 2019) and position-relative-to-scaffold information was needed for downstream outlier analyses, a genome was sequenced and assembled from a male *Geotria australis*. DNA was extracted from RNAlater preserved gonadal and muscle tissues from one male adult pouched lamprey (Table S1 NZKIN19_M1_T1) using a Qiagen DNeasy Blood and Tissue Kit. The extracted DNA was submitted to the Otago Genomics Facility where it was quality checked, prepared into two separate libraries (Rubicon Thruplex DNA-seq), and HiSeq 2500 V4 sequenced (paired end, 2x125 bp) on one lane. Initial assessment of genome size and read sampling was performed by examining the distribution of *k*-mers for *k*=21 that was generated using Jellyfish v.2.2.4 (Marçais and Kingsford, 2011) and analysed using GenomeScope2.0 (Ranallo-Benavidez *et al*, 2020). Evaluation of the sequenced reads with FastQC (v0.11.8) showed a high level of duplication primarily due to the presence of telomere-like sequences. Duplicates were removed with reference-free duplicate remover hts_SuperDeduper from HTStream_1.0.0 with default parameters. Retained reads were assembled using Abyss-pe v.2.2.3 (Simpson *et al*, 2009) based on the construction of a de Bruijn graph for *k*-mers with an experimentally chosen *k*=100. To assess the completeness/quality of the assembly, the assembly was searched for expected low copy orthologs using BUSCO v. 5.1.3 (Seppey *et al*, 2019) and *k*-mer based metrics were generated using Merqury v1.1 (Rhie *et al*, 2020). BUSCO was run with parameters: --limit=20 (number of candidate contigs considered per ortholog) and the parameter -tcov of MetaEuk (Levy Karin *et al*, 2020) subroutine was set to 0.2 to accommodate the fragmentary nature of the assembly. Searches for conserved single copy orthologs were performed using databases for metazoan conserved orthologs (metazoa_odb10: 2021-02-23) and a previously reported set of 233 Core Vertebrate Genes (Hara *et al*, 2015) that was reformatted to be compatible with BUSCO v.5. To further resolve the presence/absence of missing conserved orthologs, missing vertebrate BUSCOs were aligned to the genome using tblastn from blastall suite of the Basic Local Alignment Search Tool (BLAST; blast-2.2.17) with disabled dust filter (parameter -F) to preserve

search at low complexity regions (e.g., GC rich) that are known to be enriched in lamprey genomes (Smith *et al*, 2013). To investigate the presence of muscle-specific *k*-mers, we extracted the corresponding subregions (lengths 21 to 3088 bases) from 9104 contigs and aligned them to NCBI NR with blastn (Altschul *et al*, 1990) with word size 11.

Restriction-site associated DNA sequencing (RADSeq)

A specific RADSeq method, termed genotyping-by-sequencing (GBS; Elshire *et al*, 2011), was used following modifications proposed by Dodds *et al* (2015). The *Pst*I restriction enzyme was chosen for its ability to interrogate vertebrate genomes at a high depth of coverage (De Donato *et al*, 2013). A total of 5 libraries of a 193–318 bp size range (selected via a Labgene Scientific Pippin Prep) were sequenced on an Illumina HiSeq 2500 platform (single-end 1× 101 bp). Of these 5 libraries, one was created in 2016–2017 with 39 samples and sequenced over 2 lanes; three were created in 2018 with 252 samples and sequenced over 3 lanes; and one was created in 2019 with 4 samples and sequenced in 1 lane (Table S1). Negative controls and repeated individuals for each sequencing run, i.e. the same individual was included in the 2017 and 2018 sequencing rounds (not included in the totals above), were included to detect contamination and ensure genotyping consistency. The raw RADSeq reads were deposited into the Aotearoa Genomic Data Repository (https://repo.data.nesi.org.nz/TAONGA-LAMPREY/collection/TLCMC00005-00002).

Bioinformatics and analyses

Targeted mitochondrial DNA markers (Cyt-b and COI)

Sanger-sequenced Cyt-b and COI sequences were manually inspected and edited (low-quality bases and primer sites removed) using Geneious Prime 2020.1.1. The resulting forward and reverse consensus sequences for each individual sequenced were aligned using MAFFT (Katoh and Standley, 2016) with the E-INS-i strategy. Each alignment was then manually inspected with Geneious. Sequence similarity and uncorrected distances were evaluated using BLAST and Geneious Prime, respectively. Haplotype networks (median joining network; Bandelt *et al*, 1999) were generated with PopART (Leigh and Bryant, 2015).

Restriction-site associated DNA sequencing (RADSeq)

Raw sequencing reads were quality assessed with FastQC (0.11.9) and assembled to reference using Stacks2 v2.52 (Rochette *et al*, 2019). First, raw reads were demultiplexed, filtered for adapters, and quality filtered (--clean, --quality, --rescue) with Stacks *process_radtags* following default parameters for each of the six lanes individually. The resulting reads (~67% retained) were then mapped (~61% retained) to the indexed *G. australis* genome scaffolds with BWA-MEM v0.7.17 (Li, 2013). Samtools v1.10 (Li *et al*, 2009) was used to convert output BWA-MEM files to sorted BAM files. The Stacks pipeline was then run manually using *gstacks* with default parameters to identify and genotype SNPs. Using these preliminary data, duplicate and closely related individuals were removed using the custom R script *SNP_mismatch* (https://github.com/laninsky/SNP_comparisons/blob/master/SNP_mismatch.R, created by AA), and

the VCFtools v0.1.16 (Danecek *et al*, 2011) relatedness statistical analyser *--relatedness2* was used to confirm that 1st degree relatives were identified and removed. VCFtools was used to remove individuals with more than 50% missing data, and Stacks2 *populations* was used to phase haplotypes at each locus for the remaining individuals, retaining only loci genotyped in a minimum of 80% of individuals in a population (-r 80). The resulting Stacks-assembled dataset was exported and subjected to further filtering. VCFtools was used to retain only biallelic sites with minor allele count > 2, a minimum read depth of 10, and loci found in 80% or more individuals. The final filtered dataset was exported for downstream analyses, excluding the *CubSFS* (R package, Waltoft and Hobolth, 2018) downstream analyses of which no minor allele count filter was applied.

To ensure our results were robust to the choice of assembly program, we also generated a SNP dataset with Ipyrad v0.7.28 (Eaton and Overcast, 2016) *de novo* assembling sequence reads using default parameters with a few customisations ("gbs" datatype, "2" stricter filter adapters, "5" max uncalled bases in consensus, "8" max heterozygotes in consensus, and "20" max number of snps per locus). The resulting Ipyrad- *de novo*-assembled dataset was further pruned for duplicate individuals/close-kin and filtered (minor allele count > 2, min depth 10, loci in 80% or more individuals, biallelic sites only).

Population structure

STRUCTURE v2.3.4 (Pritchard *et al*, 2000) was used with an admixture model (Pritchard *et al*, 2000) and Bayesian iterative algorithm to determine the underlying number of genetic clusters (K) for pouched lamprey. The final filtered dataset (see above) followed the STRUCTURE-specific filtering recommendations of Linck and Battey (2019). Lambda was first inferred for a K of one and was fixed at the obtained value for all subsequent runs. Five replicates were run for each K between one and seven, each replicate run starting from a different random seed. Replicates were run for 100,000 MCMC replicates after a burn-in of 50,000. Trials with longer chains did not produce different results and likelihood values were consistent across replicates for each given K, suggesting burn-in was long enough to reach stationarity within runs, and convergence across runs. To determine the optimal K values, the Evanno method (Evanno *et al*, 2005) was then attempted with both Clumpak v1.1.2 (Jakobsson and Rosenberg, 2007) "best K" and Structure Harvester v0.6.94 (Earl, 2012). Individual assignments to the best-fitting K genetic clusters and mean likelihood probabilities were then determined from the STRUCTURE Q matrices output using Clumpak and plotted using the Clumpak "main pipeline."

Model-free principal component analysis (PCA) and discriminant analysis of principal components (DAPC) were performed using the R package *adegenet* v2.1.3 (Jombart, 2008; Jombart and Ahmed, 2011). DAPC analyses were performed de novo using the *find.clusters* to determine the number of groups, *optim.a.score* to determine the optimal number of PCs, and the lowest BIC value to select the optimal K. PCA and DAPC methods were additionally used to evaluate whether genetic structure was concordant with geographical gradients (latitude and longitude), islands, life stages, and collection years (2015–2016 adults, 2016–2018 larvae).

Patterns of diversity

Expected heterozygosity, observed heterozygosity (H_O), Wright's inbreeding coefficient (F_{IS}), standardized multilocus heterozygosity (sMLH), and genetic diversity (locus and population) were estimated for the ten Aotearoa New Zealand sampling sites (Figure 2) with *HIERFSTAT* (R package, Goudet, 2005). Nei (1987) F_{ST} was estimated with *HEIRFSTAT*, and Weir and Cockerham (1984) F_{ST} using *StAMPP* (R package; Pembleton *et al*, 2013). Pairwise F_{st} comparisons and global F-statistics were tested for significance using 1,000 bootstrap replicates (95% confidence interval). Effective population size (N_e) was estimated with NeEstimator v2.1 (Do *et al*, 2014) for the entire Aotearoa New Zealand region (both with the 10 sampling locations considered as separate populations and as one metapopulation) and the North and South Islands separately utilising a random mating model and the linkage disequilibrium method. A < 0.05 allele frequency cut off value (P_{crit} =0.05) and a 0 cutoff value (P_{crit} =0+, all alleles used) were used, and jackknife 95% confidence intervals constructed.

Isolation by distance (IBD) analyses were performed using a pairwise per individual genetic dissimilarity matrix calculated from the Stacks2 final filtered dataset using two R packages: *Adegenet* v2.1.3 to create a Genlight-class object, and *dartR* v1.8.3 to convert this to a similarity matrix (Gruber *et al*, 2018; Jombart, 2008; Jombart and Ahmed, 2011). Distance between sampling locations was measured from the corresponding river mouth (coastal entrance) as individuals exit the river mouth upon maturity (e.g., they do not traverse land). The 10 sampling locations (datum WGS84) were then linked to individuals and a distance matrix between the locations was calculated using a custom R script (https://github.com/wpearman1996/PopGenScripts/blob/main/LC_distance_calculation.R; Pearman *et al*, 2020; Van Etten and Hijmans, 2010; van Etten and van Etten, 2020). The resulting geographic pairwise distance matrix and genetic pairwise distance matrix were then vectorised in R and the *ecodist* v2.0.7 R package was used to estimate a Mantel coefficient and corresponding confidence limits (p-values; Goslee and Urban, 2007). In addition, a Mantel test was performed using a genetic distance matrix of mean F_{ST} per population (Weir and Cockerham, 1984; F_{ST} values calculated with *StAMPP*).

Gene flow and genetic diversity among sampling regions were estimated with EEMS (estimated effective migration surfaces; Petkova *et al*, 2016). EEMS, like PCA, can summarise population structure visually; however, unlike PCA, EEMS is not influenced by uneven collection of data among sampling sites (see Petkova *et al*, 2016 Figure 2). EEMS is useful for datasets where population structure is not entirely consistent with isolation by distance as it identifies areas where genetic similarity deviates from the genetic similarity expected under an isolation by distance model. We used the pairwise genetic dissimilarity matrix and sample coordinates previously calculated for the Mantel test and created a habitat file using the Google Maps API v3 Tool (http:www.birdtheme.org/useful/v3tool.html), including only marine habitat as the environment that pouched lamprey migrates within. A custom R script (created by D. Petkova) was used to run EEMS (File S2) as the habitat map included holes (the land extent of the two main New Zealand Islands) and setup was complicated by New Zealand's close proximity to the international date line (longitude 180°). Three independent runs for multiple deme sizes (291, 561, 749, 996; EEMS manual recommended; Petkova *et al*, 2016) were completed using 6 million Markov chain Monte Carlo (MCMC) iterations and 2 million burn-in iterations for a total of 12 runs. The *rEEMSplots* R package

was then used to produce plots (run conversions, observed and fitted dissimilarities, posterior probabilities) and surfaces (migration [m] and diversity [q]). Migration was considered statistically significant if the posterior probability [Pr] of effective migration rates [m] was greater than the overall mean migration rate [0], given the average squared genetic difference [D], by over 95% for all runs except one run [deme 561].

Outlier analysis

Outlier analyses were conducted to identify loci showing higher than background differentiation (F_{ST}) to detect loci that were potentially under selection. The R package *OutFLANK* (Whitlock and Lotterhos, 2015) was used following an adjusted *OutFLANK* script (T. Oosting, https://github.com/tomoosting/). A binary biallelic genotype table (.bim file) and sample information file (.fam file) were created using PLINK v1.9 (S. Purcell and C. Chang www.cog-genomics.org/plink/1.9/, Chang *et al*, 2015) and used with chromosome information created from scaffolds of the assembled *G. australis* genome to detect outlier SNPs. The analysis was performed using a false discovery threshold of 0.05%, minimum heterozygosity of 0.1, and left and right trim fractions of 0.1. Outliers with the highest F_{ST} estimate were further thinned to every 500,000 base pairs ("sliding_window") to ensure putative outliers were unlikely to be in physical linkage. A Discontiguous Megablast search (max E-value 0.05, max target sequences 20) of the NCBI nucleotide collection (nr/nt) database was performed with Geneious Prime v2022.0.1 on the sequences potentially under divergent selection to attempt to identify function.

Historical demography

Demographic history was estimated with *CubSFS* v1.0 considering all samples as representing a single population. Methods followed those listed in (Cole *et al*, 2019). The final filtered dataset, a VCF file containing only variant sites, was first modified to include monomorphic sites following the GitHub fastsimcoal2_inputs protocol (Alexander 2021; https://github.com/laninsky/bats_and_rats/tree/master/fastsimcoal2_inputs). This VCF file (monomorphic sites included) was then used to create a SFS file using easySFS (Overcast 2021; https://github.com/isaacovercast/easySFS), and *CubSFS* (used with R version 4.0.2) was used to estimate demographic history using 29 knots and a t_m of 0.35 coalescent units.

Results

Sequencing, assembly, and filtering

Targeted mitochondrial DNA markers

The 766 bp Cyt-b sequence region was identical in all of the 63 sequenced New Zealand pouched lamprey individuals and showed pairwise similarity of 98.04% (15 fixed substitutions) and 99.86% (1 fixed substitution) to 2 previously sequenced (> 390 bp) *G. australis* sequences from GQ206164 (Chile) and GQ206165 (Western Australia), respectively (Figure 3). A single transition was observed between the GenBank mitogenome-trimmed Cyt-b sequence from New Zealand (NC_029404; Ren *et*

al, 2016) and our samples, which may be the result of a sequencing error. The New Zealand pouched lamprey haplotype shared ~85.025% sequence identity with *G. macrostoma* haplotypes (GenBank accessions MT478645–53). The 589 bp trimmed targeted COI sequence region was identical in all 20 New Zealand individuals sequenced during this study and with the 14 Australian (Victoria, South Australia, and Tasmania) sequences from GenBank. The New Zealand pouched lamprey COI region additionally showed pairwise similarity of 98.30% (10 fixed substitutions) and 99.32–99.83% (3–4 fixed substitutions) to the 2 Chile and 8 Western Australia sequences listed in GenBank.

Geotria australis genome sequencing and assembly

Nearly 200 million paired end reads were sequenced from DNA of gonadal tissue and a similar number from muscle DNA of a male adult pouched lamprey (Table S3). Based on the distribution of *k*-mers within these reads the haploid genome length was estimated to be 1.07 Gb with 38.5% of the genome existing as single-copy sequence with an estimated heterozygosity of 0.67% (on average one polymorphism per 149 bp). Inclusion of both gonad and muscle sequences in the same assembly yielded a total assembly length of 0.8 Gb that was distributed across 1,899,277 scaffolds with N50 scaffold of 966 bases and a maximal scaffold length of 46,391 nucleotides. Searches for single-copy orthologs detected 75% of the expected core vertebrate orthologs (of 233 total) and 79% of the expected metazoan orthologs (of 954 total; Hara *et al*, 2015). For comparison, the chromosome-scale assembly of the sea lamprey genome contains 90% of the expected core vertebrate orthologs and 75% of metazoan orthologs (Smith *et al*, 2018). It is also worth noting that we were able to recover partial alignments of 58/59 missing core vertebrate orthologs by aligning them directly to the assembly with tblastn. Partial alignments with at least 60% identity covered from 5.5 to 89% of the orthologs' length.

The high rate of BUSCO detection is consistent with *k*-mer based estimates of genome completeness and quality, which indicate a consensus error rate of <1/million bases and 91% assembly completeness when testes and muscle reads are considered together, though it is important to recognize that the assembly was generated from these same reads. Of the total assembly *k*-mers, 0.17 million *k*-mers were found to be muscle-specific, i.e. were missing from the testes read set, compared to 1.66 million *k*-mers that were found to be testes-specific, i.e., were missing from the muscle read set. This is consistent embryonically programmed DNA elimination events that result in the removal of specific sequences from somatic cell lineages, as has been observed in other lamprey species (Smith *et al*, 2009; Smith *et al*, 2018; Smith *et al*, 2020). However, a more complete assessment of programmed DNA eliminations in pouched lamprey will await assembly improvement and the coordinated collection of additional samples from reproductively mature individuals. Investigation of muscle-specific *k*-mers, or those undersampled in testes, revealed that 42 corresponding contigs aligned to high copy rDNA sequences from flatworms (several from *Stegodexamene anguillae*), which might reflect the presence of this parasite within the muscle tissue of the sequenced animal. These sequences were pruned from the assembly prior to submission to Aotearoa Genomic Data Repository. Taken together our analyses indicate that while the assembly is fragmented due to the presence of repeats and reliance on Illumina short fragment libraries, the vast majority of sequence content is present in the assembly and the assembly can act as a valuable reference for reduced representation sequencing datasets.

RADSeq sequencing

RADSeq sequencing yielded ~900 million single-end 101 bp Illumina sequences (178,524–4,830,235 reads per individual). Since reference-based RADSeq analyses were found to outperform *de novo* approaches (Rochette *et al*, 2019), we used the Stacks reference genome pipeline that resulted in 372,081 genotyped loci across the 295 samples (258 larvae and 37 adults), with 31.42% of variants scored as missing. The Stacks dataset used for downstream analyses consisted of 16,672 SNPs characterized across 186 samples with 17 samples removed due to missing data and 92 removed through relatedness filters. Each of the 92 identified sibling-pairs were collected from the same sampling site, of which 6.28% proportion of variants were scored as missing, and ranged from 4 to 45 individuals per location (Table S1).

The Ipyrad *de novo* assembly pipeline resulted in an initial dataset consisting of 290 pouched lamprey (258 larvae and 32 adults), with 5 individuals removed due to missing data, 115,407 variants, and 62.1% of variants scored as missing. Following additional filtering and pruning of sibling-pairs, the final Ipyrad dataset consisted of 8,884 SNPs (with 6.65% of variants scored as missing) across 202 individuals. When datasets were subject to the same filtering methods (e.g., retained reads with minor allele count > 2; data not shown), downstream analyses were concordant between the Stacks and Ipyrad datasets. Given this concordance, we present the results based on our Stacks dataset for the remainder of this paper.

Population structure

The data used for Structure consisted of 186 individuals and 16,629 SNPs. Preliminary Structure runs identified 0.3544 as the optimal lambda value, which was then fixed for subsequent runs. As K=1 had the highest likelihood, the Evanno method could not be implemented, and unsurprisingly, structure plots for higher levels of K revealed admixture over all sampling locations (Figure 4).

A PCA also demonstrated little separation of sampling site locations when plotted against PC1 and PC2, explaining only 0.304% and 0.302% of the genomic variance, respectively (Figure 5). However, PCA plots of individuals coloured by gradients in latitude and longitude suggested slight separation along PC1 (Figure 5). This trend appears to be driven by subtle population structure between the islands, and non-overlapping locations, as no association was found with latitudinal and longitudinal coordinates when the dataset was broken into each island and analysed separately (Figure S4). No structure was seen by collection years and life stages, however because collection years and life stages were confounded with sampling location we cannot state unequivocally that such structure does not exist (Figure S5).

De novo DAPC analyses of population structure using an optimal a-score (55 principal components retained, 9 discriminant functions) identified the lowest Bayesian Information Criterion (BIC) value (970.0136) at k=1, thus suggesting the data fits best in a single cluster. Plots of discriminant function one also demonstrated a single contiguous cluster (no gaps between clusters of individuals). However, subtle structure was observed in DAPC scatter plots when individuals were assigned to sample sites; Piha Stream – the most northerly of our sample locations – showed slight

separation except from the North Island locations of Mangakino Stream and Mokau River (Figure 6). The 12 most admixed individuals (defined as those having no more than 0.5 probability of membership to any group) were from Waiau, Pauatahanui, Kaniwhaniwha, Waikawa, Pourakino, and Okuti (Figure S6). Of all the sampled locations, the least amount of admixture appeared in Piha Stream and the Styx River (Figure S7).

We did not detect isolation by distance (IBD) using a Mantel test between coastal river entrances corresponding to the 10 sampling sites (allelic data per individual: Mantel r=0.0005 and two-tailed p-value=0.9730, and pairwise F_{ST} between populations: Mantel r=0.1606 and two-tailed p-value=0.2220). In agreement with the Mantel test results, estimated effective migration surfaces (EEMS) plots of distances between demes and observed dissimilarity between demes suggested a poor linear relationship (R^2~0.20–0.48) i.e. that spatial data patterns were poorly explained by IBD.

Results from the EEMS analyses demonstrated higher than average historical gene flow between some sampling sites. It is ideal for the MCMC chains of the EEMS individual runs to converge upon one posterior value, and all of our runs converged around a log posterior value of ~88580. Higher than average migration rates were consistently seen between the southernmost sites (Waiau, Pourakino, and Waikawa; Figure 7 with mean migration rates log[m] > 1), and were statistically significant. Higher than average migration was also seen for the four most northerly sites (Piha, Kaniwhaniwha, Mangakino, and Mokau; Figure 7 with mean migration rates log[m] > 0.5), albeit with less support: posterior probability Pr{m > 0 | D} exceeded 90% for 7 out of the 12 runs. The "central" sites (Pauatahanui Stream, Styx River, and Okuti River) did not have consistent migratory corridors or migration barriers across all runs. Scatter plots of between-deme and within-deme genetic dissimilarity components, however, suggested a poor fit of the EEMS model to our data. A strong linear relationship was expected between the observed and fitted values yet was only occasionally produced (R^2~0.86–0.94 within demes, R^2~0.08–0.73 between demes).

Patterns of diversity

Locus-specific F_{ST} estimates ranged from -0.0480 to 0.51940 (mean=0.0010) among the 10 sampling locations, with only a small number of SNPs yielding higher values (i.e., > 0.05) (Figure S8). F_{ST} values averaged over all loci equalled 0.0022 and 0.0018 for Nei (1987) and Weir and Cockerham (1984), respectively. Observed heterozygosity (H_O), observed gene diversity, and inbreeding coefficient (F_{IS}), averaged over all loci equalled 0.1106, 0.1156, and 0.0434, respectively. Mean pairwise F_{ST} values among populations ranged between -0.0001 and 0.0067 according to Nei (1987) and -0.00008 and 0.00660 according to Weir and Cockerham (1984) (Table 1). Although we detected statistically significant genetic differentiation among our sampling locations (at alpha = 0.05), the values of F_{ST} for these comparisons were low indicating very little restriction in gene flow between locations and/or extremely recent shared ancestry between locations. Similarly, no statistically significant genomic differentiation was found among collection years (2015–2017; pairwise F_{ST}=-0.0117–0.0287) or life stages (migrating adults and larvae; pairwise F_{ST}=0.0081 adults, -0.0066 larvae).

Observed heterozygosity, expected heterozygosity (H_S), inbreeding coefficient, and standardized multilocus heterozygosity (sMLH; takes into consideration the variation of the analysed loci) ranged from 0.1038–0.1156, 0.1078–0.1181, 0.0218– 0.0619, 0.96108–1.03938 respectively,

across populations (Table 2). All metrics were very similar among populations, consistent with the recent shared ancestry/high gene flow suggested by the population structure analyses.

When individuals were grouped by sampling location (10 sites) N_e was estimated for all populations as negative or "infinite". NeEstimator sampling error is calculated from unbiased estimators using a known sample size, and a larger than expected sampling error can occur by chance resulting in a negative estimate of N_e. When a negative estimate of N_e occurs, it is often interpreted as "infinite", or that "there is no evidence for variation in the genetic characteristic caused by a finite number of parents — it can all be explained by sampling error" (see Do et al, 2014). When considering all sampled pouched lamprey as a single population, N_e was estimated at 3,326.7 for the < 0.05 allele frequency (95% confidence interval: 3,186.2–3,480.0) and 8,066.6 when all alleles were used (95% confidence interval: 7791.1–8,362.1). When each island was assessed as a single population (North and South), N_e was estimated at 2,847.2 for the < 0.05 allele frequency (95% confidence interval: 2,615.8–3,123.1) and 8,917.6 when all alleles were used (95% confidence interval: 8,101.9–9,915.7) for the North Island, while N_e was estimated at 4,111.4 for the < 0.05 allele frequency (95% confidence interval: 3,758.2–4,537.3) and 11,535.3 when all alleles were used (95% confidence interval: 10,569.9–12,694.4) for the South Island.

Outlier loci

OutFLANK identified 12 potential adaptive SNPs from all sampling locations (Figure S9). Principal component (PC) analyses of the SNPs revealed a lack of structure as was observed for neutral loci. The NCBI blast analysis of the 12 outlier sequences recovered 5 unique hits (Table S10). These included a beta-1,3-galactosyl-O-glycosyl-glycoprotein beta-1,6-N-acetylglucosaminyltransferase (GCNT1)-like miscRNA that is possibly involved in protein modification, a sorting nexin-6 (SNX6)-like mRNA possibly involved in intracellular trafficking, a zinc finger MYM-type protein 3 (ZMYM3)-like miscRNA possibly involved in the regulation of cell morphology, a solute carrier organic anion transporter family member 1C1 (SLCO1C1)-like mRNA possibly involved in the transport of organic anions, and a ruthenus protein promyelocytic leukemia (PML)-like mRNA possibly involved in apoptosis/senescence/DNA damage response/viral defence (UniProt Consortium, 2021).

Historical demography

Demographic reconstructions of the single pouched lamprey population using *CubSFS* suggested that there was a recent decline in the effective population size (N_e) since the last glacial maximum (LGM, ~20,000 years ago). *CubSFS* runs testing variations in the knots and coalescent units all produced plots with peak N_e around the LGM. The 0–30,000 year plot created using 29 knot with a t_m of 0.35 coalescent units had knots at 0.00, 20605.49, and 41433.01 years before present and thus appeared to provide adequate resolution for the assessment of the N_e around the LGM (Figure S11).

Discussion

Determinants for the lack of differentiation: Gene flow and/or recent colonisation

We assessed population connectivity, historical migration, and population size of pouched lamprey from Aotearoa New Zealand using single-gene and RADSeq data in order to guide management of pouched lamprey and similar "elusive" species. Both our single-gene and RADSeq datasets demonstrated a lack of strong population structure within Aotearoa New Zealand, which is not surprising given that several anadromous Northern Hemisphere lamprey species also show limited population structure (e.g., Bracken *et al*, 2015; Bryan *et al*, 2005; Goodman *et al*, 2008; Yamazaki *et al*, 2014). Similar to Northern Hemisphere lampreys, the lack of structure in New Zealand pouched lamprey is likely a consequence of gene flow and/or recent colonisation.

Northern Hemisphere anadromous lampreys are thought to be non-philopatric (do not home to natal streams; Bjerselius *et al*, 2000; Mateus *et al*, 2021; Moser *et al*, 2015; Spice *et al*, 2012; Waldman *et al*, 2008), which means they generally show increased gene flow and weak population structure relative to philopatric species (e.g., salmonids; Bracken *et al*, 2015; Bryan *et al*, 2005; Goodman *et al*, 2008; Yamazaki *et al*, 2014). Given the lack of population structure detected in our dataset, natal philopatry is also unlikely for pouched lamprey in Aotearoa New Zealand. Reduced population structure amongst pouched lamprey and other lampreys may also arise from behavioural, morphological, and reproductive traits that increase the chances of gene flow across a wider geographic range. For instance, reduced population structure may occur through non-specific highly-mobile host/prey choice, or through continuous selection for individuals with larger body sizes (possibly associated with longer marine migrations).

Colonisation of pouched lamprey from a recent ancestral population may also have led to the lack of overall variation seen across Aotearoa New Zealand. At the Last Glacial Maximum (LGM; 18–26 kyr; Shulmeister, 2017), if pouched lamprey were excluded from most of New Zealand due to environmental barriers such as glaciers (permanent ice) or steepened sea surface temperature gradients (see Lorrey and Bostock, 2017), and then recolonized New Zealand when conditions became more favourable, we might expect to see the high population genetic homogeneity we observe. This scenario, however, seems unlikely for two reasons. First, pollen, macrofossil, insect, and geomorphic evidence suggests that permanent ice did not cover all of New Zealand during the LGM; a large majority of the vegetation was shrubland-grassland, especially on the coastlines which extended further than the present day (Wood *et al*, 2017). Our historical demographic reconstructions suggest that effective population size of pouched lamprey peaked around the LGM, thus it may be that pouched lamprey in New Zealand were benefited by the expanded coastlines, widespread bogs, and more prevalent shrublands. Second, population genetic studies of lamprey have demonstrated that genetic differentiation can occur in less than 200 years. For example, studies using microsatellite loci demonstrated that significant genetic differentiation occurred between sea lamprey populations in the lower and upper Great Lakes after the opening of the Welland Canal in 1892 (Bryan *et al*, 2005). If pouched lamprey had roughly similar migration rates between populations, it would be expected that at least weak population structure would have developed over the 20,000 years since the LGM. Unfortunately, pouched lamprey fossil data are lacking from New Zealand, and many other areas, since lampreys lack ossified anatomical structures

and do not fossilise well. Future studies that evaluate the population genomics and divergence times of *Geotria* across the Southern Hemisphere may help determine if pouched lamprey are a recently introduced population to Aotearoa New Zealand and whether recent colonisation has contributed to the observed lack of differentiation in pouched lamprey.

Migration and dispersal in the marine environment

The lack of genetic structure between the North and South Islands of Aotearoa New Zealand is indicative of ongoing gene flow or recent colonisation (see previous section). This suggests pouched lamprey are able to migrate between islands and overcome hurdles in the marine environment such as strong currents, steep temperature gradients, and upwelling (Chiswell *et al*, 2017; Stevens *et al*, 2012; Vincent *et al*, 1991) either through prey attachment or free swimming behaviours. This dispersal capability is further supported by observer and commercial at-sea data, as well as museum vouchers that demonstrated that New Zealand pouched lamprey may travel at least 400 km from the New Zealand shoreline (Fisheries New Zealand Official Information Act request dated 4 March 2019 and A. Miller 2019, Pers. Commun. with Auckland War Memorial Museum curators, 27 May 2020). Suspected prey species of lampreys (e.g., cetaceans; McDowall, 1990; Miočić-Stošić *et al*, 2020; Pike, 1951) are thought to make long-range migrations across New Zealand and pouched lamprey may make long-range migrations with them by attaching to them as "hitchhikers".

However, despite the overall suggestion of a largely panmictic metapopulation, we detected some subtle neutral-loci population structure. Marine currents around New Zealand (Figure 2 inset) may also influence the dispersal ability of pouched lamprey — either when free-swimming, or due to the impacts on prey that they attach to, or "hitchhike", with. Piha Stream and the Styx River appear to have lower admixture which may be a result of converging currents, thus possibly reducing the gene flow to and from these locations. The number of migrants, not the proportion of migrants (demographic connectivity), determines the genetic connectivity of a population, regardless of population size (Lowe and Allendorf, 2010). Thus, populations can sometimes be demographically uncoupled even if they have small F_{ST} values. Consequently, although philopatry has not been shown to occur in other lampreys it cannot be ruled out entirely as a factor for the subtle structure seen in pouched lamprey, potentially resulting from environmental preferences, or limited breeding habitats in the streams available to them. For example, possible anthropogenic stressors on streams located on the North Island may explain the slight structure observed between islands. However, a limited amount of suitable breeding habitat does not appear to be supported in the Auckland region (North Island) as there are streams containing suitable habitat in the region, yet semi-extensive surveys have determined that pouched lamprey are only found in select streams within these regions. Furthermore, philopatric anadromous species such as sturgeons, American shad, and salmonids generally appear to have higher pairwise Fst values across neutral loci (e.g., Hasselman *et al*, 2013; Whitaker *et al*, 2020; White *et al*, 2021), and our neutral loci F_{ST} values between pouched lamprey sampling sites were low (-.0001–0.0067) even for anadromous lampreys with comparable geographic ranges (Table 2 in Mateus *et al*, 2021). Thus, we believe philopatry is unlikely in pouched lamprey.

Adaptive variation could also be an explanation for the subtle neutral-loci population structure through selection for phenotypes (e.g., total length and weight) that are more fit for

environments with certain temperature, flow, and lunar conditions, as distributional differences of Northern Hemisphere species and *G. australis* have been observed associated with variation in these environmental parameters (Arakawa and Yanai, 2020; Golovanov *et al*, 2019; Macey and Potter, 1978; Miller *et al*, 2021; Moser *et al*, 2015; Potter *et al*, 2015; Sloane, 1984). Our outlier analyses recovered 12 potential adaptive SNPs and suggested that 5 genes, involved in multiple cell functions, may be contributing to the subtle structure of the neutral loci. Previous RADSeq studies have found SNPs that localised to genes of multiple functions in *Lampetra* and *Entosphenus* species (Hess *et al*, 2013; 2014; Rougemont *et al*, 2017; Mateus *et al*, 2013). Genes predicted to be involved in development and body morphology in particular, appear limportant to the genomic diversity of lampreys (Hess *et al*, 2020). Here we find possible commonality for several genes and gene functions. The genes BAX/CASP7 and SLC35F5 (solute carrier) that were associated with apoptosis and transmembrane functions, respectively, were discovered as outliers in *E. tridentatus* RNASeq data (Hess *et al*, 2013), and may hold similar functions to the PML and SLCO1C1 (solute carrier) genes also associated with apoptosis and membrane transport from our putative adaptive loci. Additional genes associated with adaptive loci in *E. tridentatus*, such as ZNF385D and PCDH15, were identified later in *E. tridentatus* (Hess *et al*, 2014) and may be similar to our zinc finger protein (ZMYM3) and cell adhesion (GCNT1) genes. In addition, immune genes were associated with RADSeq sequence differences between *L. fluviatilis* and *L. planeri* (Mateus *et al*, 2013; Rougemont *et al*, 2017) and our PML gene may be playing a similar immunological role. Immunity genes such as PML, GCNT1, as well as others, may prove important for future studies that aim to better understand lamprey syndromes and diseases such as LRS. Overall, since only 12 SNPs were recovered it difficult to determine how informative the SNPs are to the structure of the population in Aotearoa New Zealand. Future studies that include joint genetic sampling and morphological traits, such as those seen in association analyses by Hess *et al* (2020) on Pacific lamprey (*Entosphenus tridentatus*), will help determine if selection could be impacting population structure in pouched lamprey.

More broadly, colonisation history and rates of intercontinental dispersal of *Geotria* in the Southern Hemisphere are uncertain. Presently our understanding of the systematics of the Southern Hemisphere lampreys is based on morphological characters and mitochondrial DNA markers.

Mitochondrial DNA markers and morphological characters suggest a deep split between Atlantic *Geotria macrostoma* sampled from Argentina and Pacific *G. australis* (Nardi *et al*, 2020; Riva-Rossi *et al*, 2020). Within *G. australis,* Chilean populations show mitochondrial sequence distinctions from Australasian populations, and within Australasia, Western Australian populations appear genetically differentiated from Aotearoa New Zealand, Tasmania, and South Australian populations (this study [Figure 3] and published works; Nardi *et al*, 2020; Riva-Rossi *et al*, 2020). Although limited to three Chilean samples, the genetic differences of the Chilean samples from the Australian and New Zealand samples suggest, at least from mtDNA, that little to no migration occurs between the Pacific and Atlantic coasts of South America, that trans-Pacific migration is rare, and that migration between Western Australia and other Australasian locations may also be restricted.

In this study we aimed to better understand the population structure of pouched lamprey within Aotearoa New Zealand and to better understand how they are related to other Southern Hemisphere pouched lamprey. RADSeq data and mtDNA data were used, respectively. RADSeq data has been shown to be more suitable for examining historical relationships and contemporary fine-scale gene flow (e.g., Longo et al 2020), while mtDNA is suitable for examining larger-scale

phylogenetic relationships. An investigation into the finer-scale population structure of Southern Hemisphere is warranted and will require nuclear gene regions to better understand these relationships and resolve if dispersal and/or vicariance explains the current distribution of *Geotria* in New Zealand and in the Southern Hemisphere.

Management and conservation of pouched lamprey and other elusive species

Estimating population size, habitat preferences, behaviours, population structure, and migration corridors is challenging for elusive species. This study, however, illustrates the benefits of genetics for acquiring conservation-relevant information for these species.

Elusive species are those that have a low probability of detection and include both abundant and rare species. The reasons for their low probability of detection vary by species, however, the impacts this has on the management and conservation of rare species is often universally detrimental. Since elusive species are under-detected by surveying and monitoring equipment they often get counted incorrectly, or not counted at all (undefined/"NA") in population censuses. This lack of detection also creates large knowledge gaps about their preferred environments, inter- and intra-species behaviours, and migrations/movements, among others. For example, many threatened tropical mammals are poorly researched due to the difficulties detecting them (Jambari et al, 2019). This even includes large species such as the tiger (*Panthera tigris*) in India where monitoring and surveying difficulties have led to decades of misleading information and poor conservation practices (Karanth *et al*, 2003). Large elusive fishes, such as eels are also considered difficult to sample quantitatively since species such as the European eel (*Anguilla anguilla*) are thought to avoid detection during electrofishing and trapping surveys, leading to inadequate abundance estimates (Degerman *et al*, 2019). As one of these elusive species, pouched lamprey are routinely under-detected in general stream surveys using standard methods (e.g., electrofishing, spotlighting, and trapping) due to their secretive habitats and tendency to migrate on receding flood waters. For instance, standard electrofishing pulse settings are often inadequate for pulsing large adults or sedentary larvae (Joy *et al*, 2013). This has made it difficult for pouched lamprey managers, stakeholders, and kaitiaki to accurately infer important population demographics and implement robust conservation policies. Here we show that molecular methods such as RADSeq can increase our understanding of the elusive pouched lamprey and can be used on other elusive species, such as eels, to detect similar important population structure and migration information. Molecular methods such as RADSeq often require only a limited amount of physical sampling — sometimes only a single sampling event is needed. Thus, unlike most standard monitoring methods that require the routine collection of individuals, these molecular methods are much more feasible for elusive species. In addition, the samples/tissues collected for molecular analyses can often be stored and used for multiple studies across multiple disciplines, such as genetic and isotope analyses. The information gained from these molecular methods can then be used to improve conservation initiatives and assist future management decisions (see below).

The lack of genetic structure in our data suggests the presence of ongoing gene flow between regions within Aotearoa New Zealand. This has ramifications for potential future translocation efforts between neighbouring streams as well as sourcing individuals for potential propagation programs. Multiple case studies, led by tribal and government agencies, of Northern

Hemisphere lampreys have demonstrated the success of artificial propagation programs, and have shown that successful programs can be created by reusing existing hatchery facilities with some adjustments (Lampman *et al*, 2020). These methods are transferrable to pouched lamprey; however, both translocations and propagation programs would warrant considerable research, careful consultation, and be investigated in collaboration with iwi partners, to ensure cultural values are adhered to and that any genetic variation, even if subtle, is maintained (such as those protocols created for Pacific lamprey, see Hess *et al*, 2014; Lampman *et al*, 2016).

Geotria australis is known to have lamprey reddening syndrome (LRS) and haemorrhagic septicaemia in Aotearoa New Zealand and Australia, respectively. In New Zealand pouched lamprey were first seen with LRS, a syndrome causing external haemorrhaging and body reddening, in Southland rivers in 2011 and it has since been annually documented in rivers of the South Island. The cause of LRS is unknown and most documented cases were lethal or thought to be associated with mass mortalities (Brosnahan *et al*, 2018). A lack of organised reporting has made monitoring LRS difficult, and there is concern that LRS may spread across cohorts (Kitson *et al*, 2012) and to the North Island. Our results suggest that if LRS is not associated with particular environmental conditions of Southland, i.e. if it is transmittable to other individuals and or has a genetic basis for susceptibility, that it will likely be seen in North Island individuals since gene flow appears to occur between the North and South Islands. Thus, LRS should be the subject of ongoing monitoring and North Island managers should be especially observant for it.

In addition to genetic structure, we also investigated the estimated effective population size (N_e) of New Zealand pouched lamprey, finding our N_e estimate of a single population to be ~3,327-8,067. This was lower than the lowest N_e estimates, 50,000, for European sea lamprey (Almada *et al*, 2008) that were sampled on an approximately similar geographic scale. However, the European sea lamprey N_e estimates were inferred using a different method than ours: the control region (d-loop) with the estimate of diversity index ($\theta = 2N_e\mu$). N_e estimates of other lamprey species have been inferred using similar methods to ours (e.g., RADSeq data and NeEstimator), but these estimates were focused on much smaller geographic scales such as creeks and rivers (e.g., Rougemont *et al*, 2017; Sard *et al*, 2020). N_e between islands in Aotearoa New Zealand, however, did meet expectations. Greater N_e estimates were found to occur in the South Island than in the North Island, which is similar to anecdotal observations that pouched lamprey are less abundant in the North Island compared to the South Island. New Zealand pouched lamprey are considered "Threatened – Nationally Vulnerable" since they were predicted to occupy ≤ 100 ha (1 km^2) and to decline by 10–50% over three generations (qualifiers: "data poor" and "secure overseas"; Dunn *et al*, 2018; Goodman *et al*, 2014). Direct comparisons of N_e to adult census size (N_c) are problematic because most natural populations have some degree of population structure, lack discrete generations, and do not maintain a constant population size. For example, effective population size estimates of Chinook salmon were commonly underrepresented due to their fluctuating population size between years and since only a few individuals will produce many offspring; Shrimpton and Heath, 2003). However, N_e analyses, such as the one performed in this study, will be useful tools for monitoring pouched lamprey N_e through time. This monitoring may act as a rough proxy for monitoring the expected declines in N_c and the successes of specific conservation measures such as translocation and propagation programs.

Conclusions

Here we used RADSeq and targeted markers to improve our knowledge of the population structure and genetic variation of pouched lamprey (*G. australis*) in Aotearoa New Zealand: the first genome-wide population study of any Southern Hemisphere lamprey species. Our mitochondrial DNA (Cyt-b and COI) analyses and RADSeq data revealed that gene flow occurs throughout Aotearoa New Zealand and that this lack of population structure may be influenced by recolonisation following the last glacial maximum. However, our historical demographic reconstructions differ from those expected under a recolonisation scenario and thus support the hypothesis that long-distance migration explains the lack of structure in New Zealand pouched lamprey. We recovered fine scale structure between sites and along geographical gradients (latitudinal and longitudinal); however, this was not reflected in our putative outlier analyses. Our isolation by distance tests were insignificant, suggesting other potential factors such as philopatry and ocean currents may be subtly influencing pouched lamprey gene flow. The techniques used in this study can be applied broadly to other elusive species (e.g., eels) to better understand their distribution, effective population size, and population structure. This structure and abundance information will be highly useful for ongoing pouched lamprey monitoring, conservation, and disease-management strategies.

Funding

This work was supported by a Ministry of Business, Innovation, and Employment New Zealand contract C01X1615 to CB and administered by NIWA; the University of Otago; a University of Otago Doctoral Scholarship to AKM; a National Institutes of Health Minority Health International Research Training Programme run through the University of California Santa Cruz to NJG and AKM; and a Ministry of Business, Innovation & Employment's Research Infrastructure programme grant that supported the New Zealand eScience Infrastructure (NeSI) high performance computing facilities utilised in this research.

Acknowledgements

The authors would like to extend their sincere thanks to the mana whenua who are kaitiaki for kanakana/piharau. In particular, we would like to thank Murihiku tāngata whenua and Waikawa mana whenua, including the tāngata tiaki/kaitiaki and Waikawa whānau research advisory group. We would like to extend a special thank you to Peter Stockman and also to Phoenix Hale for his help with the pouched lamprey collections. Thank you to members of the AgResearch team for their sequencing assistance; to Desislava Petkova for the EEMS support and scripts; to William Pearman for the use of his R scripts; to Tom Oosting for his help with OutFLANK; and to Matthew Dale and Jonah Yick for the use of their photographs. Additional thanks to the editors and reviewers for their valuable comments. We acknowledge the use of New Zealand eScience Infrastructure (NeSI) high performance computing facilities, consulting support (especially by Dinindu Senanayake), and training services as part of this research. We lastly thank Peter Williamson and Hanareia Ehau-Taumaunu for their te reo Māori translation guidance. Pouched lamprey were collected in accordance with the NIWA Animal Ethics Committee approval AEC189. Fish capture and handling procedures were carried out under a permit from the Ministry for Primary Industries, Fisheries New Zealand (Special Permit 666/2).

Data Availability

We have deposited the primary data underlying these analyses as follows:

- The RADSeq DNA sequences and targeted DNA sequences are available in Aotearoa Genomic Data Repository at https://repo.data.nesi.org.nz/TAONGA-LAMPREY/collection/TLCMC00005-00002. The draft genome assembly is available in Aotearoa Genomic Data Repository at https://repo.data.nesi.org.nz/TAONGA-LAMPREY/collection/TLCMC00005-00003. All data will be shared upon reasonable request.

References

Almada VC, Pereira AM, Robalo JI, Fonseca JP, Levy A, Maia C *et al* (2008). Mitochondrial DNA fails to reveal genetic structure in sea-lampreys along European shores. *Molecular Phylogenetics and Evolution***:** 391-396.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *Journal of molecular biology* **215**(3)**:** 403-410.

Arakawa H, Yanai S (2020). Upper thermal tolerance of larval Arctic lamprey (*Lethenteron camtschaticum*). *Ichthyological Research***:** 1-6.

Baker CF, Jellyman DJ, Reeve K, Crow S, Stewart M, Buchinger T *et al* (2017). First observations of spawning nests in the pouched lamprey (*Geotria australis*). *Canadian J of Fisheries* **74**(10)**:** 1603-1611.

Bandelt H-J, Forster P, Röhl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution* **16**(1)**:** 37-48.

Best E (1941). *The Maori. Volume 2.* Wellington: Polynesian Society. *Memoirs of the Polynesian Society*, 5, 1924.

Bjerselius R, Li W, Teeter JH, Seelye JG, Johnsen PB, Maniak PJ *et al* (2000). Direct behavioral evidence that unique bile acids released by larval sea lamprey (*Petromyzon marinus*) function as a migratory pheromone. *Can J Fish Aquat Sci* **57**(3)**:** 557-569.

Bracken FS, Hoelzel AR, Hume JB, Lucas MC (2015). Contrasting population genetic structure among freshwater-resident and anadromous lampreys: the role of demographic history, differential dispersal and anthropogenic barriers to movement. *Molecular Ecology* **24**(6)**:** 1188-1204.

Brosnahan CL, Pande A, Keeling SE, van Andel M, Jones JB (2018). Lamprey (*Geotria australis*; Agnatha) reddening syndrome in Southland rivers, New Zealand 2011–2013: laboratory findings and epidemiology, including the incidental detection of an atypical *Aeromonas salmonicida*. *New Zealand Journal of Marine and Freshwater Research* (Available Online).

Bryan M, Zalinski D, Filcek K, Libants S, Li W, Scribner K (2005). Patterns of invasion and colonization of the sea lamprey (*Petromyzon marinus*) in North America as revealed by microsatellite genotypes. *Molecular Ecology* **14**(12)**:** 3757-3773.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**(1)**:** s13742-13015-10047-13748.

Chiswell SM, Zeldis JR, Hadfield MG, Pinkerton MH (2017). Wind-driven upwelling and surface chlorophyll blooms in Greater Cook Strait. *New Zealand Journal of Marine and Freshwater Research* **51**(4)**:** 465-489.

Cole TL, Dutoit L, Dussex N, Hart T, Alexander A, Younger JL *et al* (2019). Receding ice drove parallel expansions in Southern Ocean penguins. *Proceedings of the National Academy of Sciences* **116**(52)**:** 26690-26696.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA *et al* (2011). The variant call format and VCFtools. *Bioinformatics* **27**(15)**:** 2156-2158.

De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG (2013). Genotyping-by-sequencing (GBS): a novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. *PLoS One* **8**(5)**:** e62137.

Degerman E, Tamario C, Watz J, Nilsson PA, Calles O (2019). Occurrence and habitat use of European eel (*Anguilla anguilla*) in running waters: lessons for improved monitoring, habitat restoration and stocking. *Aquatic ecology* **53**(4)**:** 639-650.

Do C, Waples RS, Peel D, Macbeth G, Tillett BJ, Ovenden JR (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular ecology resources* **14**(1)**:** 209-214.

Docker MF, Hume JB, Clemens BJ (2015). Introduction: a surfeit of lampreys. In: Docker MF (ed) *Lampreys: Biology, Conservation and Control Volume 1*, Vol 37. Springer Netherlands: Dordrecht, pp. 1-74.

Dodds KG, McEwan JC, Brauning R, Anderson RM, Stijn TC, Kristjánsson T *et al* (2015). Construction of relatedness matrices using genotyping-by-sequencing data. *BMC genomics* **16**(1)**:** 1047.

Dudgeon D, Arthington AH, Gessner MO, Kawabata Z-I, Knowler DJ, Lévêque C *et al* (2006). Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological reviews* **81**(2)**:** 163-182.

Dunn NR, Allibone RM, Closs G, Crow S, David BO, Goodman J *et al* (2018). *Conservation status of New Zealand freshwater fishes, 2017*. Publishing Team, Department of Conversation.

Earl DA (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**(2)**:** 359-361.

Eaton D, Overcast I. (2016). URL http://ipyrad. [accessed 1 August 2016].

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES *et al* (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**(5)**:** e19379.

Evanno G, Regnaut S, Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology* **14**(8)**:** 2611-2620.

Gemmell NJ, Akiyama S (1996). An efficient method for the extraction of DNA from vertebrate tissues. *Trends in Genetics* **12**(9)**:** 338-339.

Golovanov V, Nekrutov N, Zvezdin A, Smirnov A, Tsimbalov I (2019). Thermoadaptation characteristics of European river lamprey *Lampetra fluviatilis* smolts. *Journal of Ichthyology* **59**(5)**:** 805-809.

Goodman D, Reid S, Docker M, Haas G, Kinziger A (2008). Mitochondrial DNA evidence for high levels of gene flow among populations of a widely distributed anadromous lamprey *Entosphenus tridentatus* (Petromyzontidae). *Journal of Fish Biology* **72**(2)**:** 400-417.

Goodman J, Dunn NR, Ravenscroft PJ, Allibone RM, Boubee JA, David BO *et al* (2014). *Conservation status of New Zealand freshwater fish, 2013*. Department of Conversation: New Zealand.

Goslee SC, Urban DL (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software* **22**(7)**:** 1-19.

Goudet J (2005). Hierfstat, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes* **5**(1)**:** 184-186.

Gruber B, Unmack PJ, Berry OF, Georges A (2018). dartr: An r package to facilitate analysis of SNP data generated from reduced representation genome sequencing. *Molecular Ecology Resources* **18**(3)**:** 691-699.

Hale ML, Burg TM, Steeves TE (2012). Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PloS one* **7**(9)**:** e45170.

Hara Y, Tatsumi K, Yoshida M, Kajikawa E, Kiyonari H, Kuraku S (2015). Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. *BMC genomics* **16**(1)**:** 1-12.

Hasselman DJ, Ricard D, Bentzen P (2013). Genetic diversity and differentiation in a wide ranging anadromous fish, A merican shad (*Alosa sapidissima*), is correlated with latitude. *Molecular ecology* **22**(6)**:** 1558-1573.

He F, Zarfl C, Bremerich V, Henshaw A, Darwall W, Tockner K *et al* (2017). Disappearing giants: a review of threats to freshwater megafauna. *Wiley Interdisciplinary Reviews: Water* **4**(3)**:** e1208.

Hess JE, Campbell NR, Close DA, Docker MF, Narum SR (2013). Population genomics of P acific lamprey: adaptive variation in a highly dispersive species. *Molecular Ecology* **22**(11)**:** 2898-2916.

Hess JE, Caudill CC, Keefer ML, McIlraith BJ, Moser ML, Narum SR (2014). Genes predict long distance migration and large body size in a migratory fish, Pacific lamprey. *Evolutionary Applications* **7**(10)**:** 1192-1208.

Hess JE, Smith JJ, Timoshevskaya N, Baker C, Caudill CC, Graves D *et al* (2020). Genomic islands of divergence infer a phenotypic landscape in Pacific lamprey. *Molecular Ecology* **29**(20)**:** 3841-3856.

Jakobsson M, Rosenberg N (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**(14)**:** 1801-1806.

Jambari A, Sasidhran S, Halim HRA, Mohamed KA, Ashton-Butt A, Lechner AM *et al* (2019). Quantifying species richness and composition of elusive rainforest mammals in Taman Negara National Park, Peninsular Malaysia. *Global Ecology and Conservation* **18:** e00607.

James AB (2008). *Ecology of the New Zealand lamprey (Geotria australis): a literature review*. Department of Conservation, Wanganui Conservancy: New Zealand.

Jellyman D, Glova G (2002). Habitat use by juvenile lampreys (*Geotria australis*) in a large New Zealand river. *New Zealand Journal of Marine and Freshwater Research* **36**(3)**:** 503-510.

Jombart T (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**(11)**:** 1403-1405.

Jombart T, Ahmed I (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**(21)**:** 3070-3071.

Karanth KU, Nichols JD, Seidenstricker J, Dinerstein E, Smith JLD, McDougal C *et al*. (2003). *Animal Conservation forum, Vol. 6*. Cambridge University Press, pp 141-146.

Katoh K, Standley DM (2016). A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* **32**(13)**:** 1933-1942.

Kelso J, Glova G (1993). Distribution, upstream migration and habitat selection of maturing lampreys, *Geotria australis*, in Pigeon Bay Stream, New Zealand. *Marine and Freshwater Research* **44**(5)**:** 749-759.

Kitson J, Leith V, Whaanga D, Hay J, Quarterman A, Ledington S *et al*. (2012). Final Technical Report for Ngā Pae o Te Māramatanga.

Lampman R, Moser ML, Jackson A, Rose R, Gannam A, Barron J (2016). Chapter twenty two Developing techniques for artificial propagation and early rearing of Pacific lamprey (*Entosphenus tridentatus*) for species recovery *Jawless Fishes of the World: Volume 2*. Vol. 2, p 160.

Lang NJ, Roe KJ, Renaud CB, Gill HS, Potter IC, Freyhof J *et al* (2009). Novel relationships among lampreys (Petromyzontiformes) revealed by a taxonomically comprehensive molecular data set. *American Fisheries Society Symposium* **7**(2)**:** 1-15.

Leigh JW, Bryant D (2015). popart: full-feature software for haplotype network construction. *Methods in Ecology Evolution* **6**(9)**:** 1110-1116.

Levy Karin E, Mirdita M, Söding J (2020). MetaEuk—Sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8:** 1-15.

Li H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* preprint arXiv:1303.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al* (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16)**:** 2078-2079.

Linck E, Battey C (2019). Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources* **19**(3)**:** 639-647.

Longo GC, Lam L, Basnett B, Samhouri J, Hamilton S, Andrews K *et al* (2020). Strong population differentiation in lingcod (*Ophiodon elongatus*) is driven by a small portion of the genome. *Evolutionary applications* **13**(10)**:** 2536-2554.

Lorrey A, Bostock H (2017). *The climate of New Zealand through the Quaternary Landscape and Quaternary Environmental Change in New Zealand*. Springer, pp 67-139.

Lowe WH, Allendorf FW (2010). What can genetics tell us about population connectivity? *Molecular ecology* **19**(15)**:** 3038-3051.

Macey DJ, Potter IC (1978). Lethal temperatures of ammocoetes of the Southern Hemisphere lamprey, *Geotria australis* Gray. *Environmental Biology of Fishes* **3**(2)**:** 241-243.

Marçais G, Kingsford C (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6)**:** 764-770.

Mateus CS, Docker MF, Evanno G, Hess JE, Hume JB, Oliveira IC *et al* (2021). Population structure in anadromous lampreys: Patterns and processes. *Journal of Great Lakes Research* **47**(Suppl. 1) S38-S58.

Mateus CS, Stange M, Berner D, Roesti M, Quintella BR, Alves MJ *et al* (2013). Strong genome-wide divergence between sympatric European river and brook lampreys. *Current Biology* **23**(15)**:** R649-R650.

McDowall R (1990). *New Zealand Freshwater Fishes: A Natural History and Guide* (Illustrated), revised from 1978 edn. Heinemann Reed: Auckland.

McDowall R (1992). Particular problems for the conservation of diadromous fish. *Aquatic Conservation: Marine and freshwater ecosystems* **2**(4)**:** 351-355.

McDowall RM (2002). Accumulating evidence for a dispersal biogeography of southern cool temperate freshwater fishes. *Journal of Biogeography* **29**(2)**:** 207-219.

McDowall RM (2011). *Ikawai: freshwater fishes in Māori culture and economy*. Canterbury University Press: Christchurch.

[dataset]* Miller AK, Baker C, Alexander A, Gemmell NJ, (2021), *Geotria australis* genome assembly data, Aotearoa Genomic Data Repository, https://repo.data.nesi.org.nz/TAONGA-LAMPREY/collection/TLCMC00005-00003

Miller AK, Baker C, Kitson JC, Yick JL, Manquel PEI, Alexander A *et al* (2021). The Southern Hemisphere lampreys (Geotriidae and Mordaciidae). *Reviews in Fish Biology and Fisheries* **31**(2)**:** 201-232.

[dataset]* Miller AK, Gillum J, Sharif S, Clarke S, Baker C, Kitson JC, Alexander A, Gemmell NJ, (2021), New Zealand kanakana genotype by sequencing (GBS) and targeted sequencing (COI and Cyt-b) data, Aotearoa Genomic Data Repository, https://repo.data.nesi.org.nz/TAONGA-LAMPREY/collection/TLCMC00005-00002

Miočić-Stošić J, Pleslić G, Holcer D (2020). Sea Lamprey (*Petromyzon marinus*) Attachment to the Common Bottlenose Dolphin (*Tursiops truncatus*). *Aquatic Mammals* **46**(2).

Moore P, Wakelin M (1997). Diet of the yellow-eyed penguin *Megadyptes antipodes*, South Island, New Zealand, 1991–1993. *Marine Ornithology* **25:** 17-29.

Moser ML, Almeida PR, Kemp PS, Sorensen PW (2015). Lamprey spawning migration. In: Docker MF (ed) *Lampreys: biology, conservation and control*. Springer, pp 215-263.

Nardi CF, Sánchez J, Fernández DA, Casalinuovo MÁ, Rojo JH, Chalde T (2020). Detection of lamprey in Southernmost South America by environmental DNA (eDNA) and molecular evidence for a new species. *Polar Biology***:** 1-15.

Nei M (1987). *Molecular evolutionary genetics*. Columbia university press.

Pearman WS, Wells SJ, Silander OK, Freed NE, Dale J (2020). Concordant geographic and genetic structure revealed by genotyping-by-sequencing in a New Zealand marine isopod. *Ecology and Evolution* **10**(24)**:** 13624-13639.

Pembleton LW, Cogan NO, Forster JW (2013). St AMPP: An R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular ecology resources* **13**(5)**:** 946-952.

Petkova D, Novembre J, Stephens M (2016). Visualizing spatial population structure with estimated effective migration surfaces. *Nature genetics* **48**(1)**:** 94-100.

Pike GC (1951). Lamprey marks on whales. *Journal of the Fisheries Board of Canada* **8**(4)**:** 275-280.

Potter IC, Gill HS, Renaud CB, Haoucher D (2015). The Taxonomy, Phylogeny, and Distribution of Lampreys. In: Docker MF (ed) *Lampreys: Biology, Conservation and Control: Volume 1*. Springer Netherlands: Dordrecht, pp 35-73.

Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**(2)**:** 945-959.

Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**(1)**:** 1-10.

Ren J, Pu J, Buchinger T, Zhu X, Baker C, Li W (2016). The mitogenomes of the pouched lamprey (*Geotria australis*) and least brook lamprey (*Lampetra aepyptera*) with phylogenetic considerations. *Mitochondrial DNA Part A* **27**(5)**:** 3560-3562.

Rhie A, Walenz BP, Koren S, Phillippy AM (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology* **21**(1)**:** 1-27.

Riva-Rossi C, Barrasso DA, Baker C, Quiroga AP, Baigún C, Basso NG (2020). Revalidation of the Argentinian pouched lamprey *Geotria macrostoma* (Burmeister, 1868) with molecular and morphological evidence. *Plos one* **15**(5)**:** e0233792.

Rochette NC, Rivera-Colón AG, Catchen JM (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular ecology* **28**(21)**:** 4737-4754.

Rougemont Q, Gagnaire PA, Perrier C, Genthon C, Besnard AL, Launey S, Evanno G (2017). Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. *Molecular ecology* **26**(1): 142-162.

Sard NM, Smith SR, Homola JJ, Kanefsky J, Bravener G, Adams JV, Holbrook CM, Hrodey PJ, Tallon K, Scribner KT (2020). RAPTURE (RAD capture) panel facilitates analyses characterizing sea lamprey reproductive ecology and movement dynamics. *Ecology and evolution* **10**(3): 1469-1488.

Seppey M, Manni M, Zdobnov EM (2019). BUSCO: assessing genome assembly and annotation completeness. *Methods in molecular biology (Clifton, NJ)* **1962:** 227-245.

Shrimpton JM, Heath DD (2003). Census vs. effective population size in chinook salmon: large-and small-scale environmental perturbation effects. *Molecular Ecology* **12**(10)**:** 2571-2583.

Shulmeister J (2017). Blowing on the West Wind. The Most Recent Quaternary Glaciation of New Zealand *Landscape and Quaternary Environmental Change in New Zealand*. Springer, pp 171-204.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009). ABySS: a parallel assembler for short read sequence data. *Genome research* **19**(6)**:** 1117-1123.

Sloane R. (1984). *Papers and proceedings of the Royal Society of Tasmania, Vol. 118*, pp 163-171.

Smith JJ, Antonacci F, Eichler EE, Amemiya CT (2009). Programmed loss of millions of base pairs from a vertebrate genome. *Proceedings of the National Academy of Sciences* **106**(27)**:** 11212-11217.

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS *et al* (2013). Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nature genetics* **45**(4)**:** 415-421.

Smith JJ, Timoshevskaya N, Ye C, Holt C, Keinath MC, Parker HJ *et al* (2018). The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nature genetics* **50**(2)**:** 270.

Smith JJ, Timoshevskiy VA, Saraceno C (2020). Programmed DNA Elimination in Vertebrates. *Annual Review of Animal Biosciences* **9**.

Spice EK, Goodman DH, Reid SB, Docker MF (2012). Neither philopatric nor panmictic: microsatellite and mtDNA evidence suggests lack of natal homing but limits to dispersal in Pacific lamprey. *Molecular Ecology* **21**(12)**:** 2916-2930.

Stevens C, Smith M, Grant B, Stewart C, Divett T (2012). Tidal energy resource complexity in a large strait: The Karori Rip, Cook Strait. *Continental Shelf Research* **33:** 100-109.

Tedd P, Kelso J (1993). Distribution, growth and transformation timing of larval *Geotria australis* in New Zealand. *Ecology of Freshwater Fish* **2**(3)**:** 99-107.

Tickell W (1964). Feeding preferences of the albatrosses *Diomedea melanophris* and *D. chrysostoma* at South Georgia. *Symp Biol Antarct* **1:** 383-387.

UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* **49**(D1)**:** D480-D489.

Van Etten J, Hijmans RJ (2010). A geospatial modelling approach integrating archaeobotany and genetics to trace the origin and dispersal of domesticated plants. *PLoS One* **5**(8)**:** e12060.

van Etten J, van Etten MJ (2020). Package 'gdistance'. URL https://cran.r-project.org/web/packages/gdistance/index.html. [accessed 9 August 2020].

Vincent W, Howard-Williams C, Tildesley P, Butler E (1991). Distribution and biological properties of oceanic water masses around the South Island, New Zealand. *New Zealand journal of marine and freshwater research* **25**(1)**:** 21-42.

Waldman J, Grunwald C, Wirgin I (2008). Sea lamprey *Petromyzon marinus*: an exception to the rule of homing in anadromous fishes. *Biology letters* **4**(6)**:** 659-662.

Waltoft BL, Hobolth A (2018). Non-parametric estimation of population size changes from the site frequency spectrum. *Statistical applications in genetics and molecular biology* **17**(3).

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**(1462)**:** 1847-1857.

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *evolution***:** 1358-1370.

Whitaker J, Price L, Boase J, Bernatchez L, Welsh A (2020). Detecting fine-scale population structure in the age of genomics: a case study of lake sturgeon in the Great Lakes. *Fisheries Research* **230:** 105646.

White SL, Kazyak DC, Darden TL, Farrae DJ, Lubinski BA, Johnson RL *et al* (2021). Establishment of a microsatellite genetic baseline for North American Atlantic sturgeon (*Acipenser o. oxyrhinchus*) and range-wide analysis of population genetics. *Conservation Genetics* **22**(6)**:** 977-992.

Whitlock MC, Lotterhos KE (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of F ST. *The American Naturalist* **186**(S1)**:** S24-S36.

Wood J, Wilmshurst J, Newnham R, McGlone M (2017). *Evolution and ecological change during the New Zealand Quaternary Landscape and Quaternary Environmental Change in New Zealand*. Springer, pp 235-291.

WWF (2018). Living planet report - 2018: Aiming Higher. Grooten, M. and Almond, R.E.A. (eds). WWF: Gland, Switzerland, pp 1-144.

Yamazaki Y, Yokoyama R, Nagai T, Goto A (2014). Population structure and gene flow among anadromous Arctic lamprey (*Lethenteron camtschaticum*) populations deduced from polymorphic microsatellite loci. *Environmental biology of fishes* **97**(1)**:** 43-52.

Figure Legends

Figure 1 *Geotria australis* (kanakana/piharau; pouched lamprey). A. migrating adult B. oral disc of migrating adult C. juvenile D. adult. Photo credit: A=AKM, B=Matthew Dale, C=Jonah Yick, D=Robert Holdaway. Photo D was reprinted from Population structure in anadromous lampreys: Patterns and processes, In Press, Copyright (2021), with permission from Elsevier.

Figure 2 Aotearoa New Zealand pouched lamprey sampling locations and marine currents (inset). Blue points represent Te Ika-a-Māui/North Island sites. Green points represent Te Waipounamu/South Island sites. Striped blue and striped green points, along with asterisks (*), denote a site sampled on more than one occasion. Numbers below the site names denote the total number of larvae and adults collected from the site.

Figure 3 Median joining haplotype network (created with PopArt) and pouched lamprey sampling locations. Haplotype network: colours correspond to locations, sizes of circles correspond to the frequency of the sequences belonging to a specific haplotype, numbers correspond to the number of fixed substitutions (mutations) from the large Aotearoa New Zealand haplotype, and the Aotearoa New Zealand haplotype acquired from mitochondrial data (NC_029404) is coloured in grey-blue. All of the Aotearoa New Zealand samples represented in the haplotype networks, except for the mitogenome haplotype, were collected during the current study. All of the Australian and Chilean samples were acquired from GenBank.

Figure 4 STRUCTURE v2.3.4 (Pritchard *et al*, 2000) plots of the final filtered Stacks2-assembled RADSeq dataset (see text) for pouched lamprey from 10 sites in Aotearoa New Zealand. Plots for each of the seven genetic clusters (K) are displayed as rows and the ten sampling sites are displayed as columns. Colours represent different genetic clusters. “N” and “S” denote North and South Island sites respectively. K=1 (not shown) had the highest likelihood and therefore the Evanno method could not be implemented.

Figure 5 Principal component analysis (PCA) plots of pouched lamprey from 10 sites in Aotearoa New Zealand coloured by A. sampling location B. latitudinal position C. longitudinal position. Symbols “(N)” and “(S)” represent South Island and North Island respectively. Ellipses = 95% confidence.

Figure 6 Discriminant analysis of principle components (DAPC) plot of pouched lamprey from 10 sites in Aotearoa New Zealand coloured by sampling location. A plot of the cumulated variance explained by the eigenvalues is positioned in the top left corner. A plot of the eigenvalues that were retained is

positioned in the bottom right corner. Sampling sites are listed in a latitudinal manner with the most northern site at the top left and the most southern site at the bottom right.

Figure 7 EEMS posterior mean migration rates m (on the log10 scale) for pouched lamprey in Aotearoa New Zealand. Green indicates higher than average migration and brown indicates lower than average migration. Black circles represent deems consisting of one or more sampling sites and are proportional in size to the number of isolates in the deme. Major marine currents are drawn in blue.

Table 1 Pairwise F_{ST} estimates (Wright [1949], updated by Weir and Cockerham [1984]) of Aotearoa New Zealand pouched lamprey from the ten collection sites. Asterisks denote pairwise F_{ST} p-values (1000 bootstraps): "*"=significant at < 0.05, "**"=significant at < 0.005. F_{ST} values are shaded by value from highest (darkest) to lowest (lightest). Symbols "(N)" and "(S)" represent South Island and North Island respectively.

Table 2 Estimates of genetic diversity for Aotearoa New Zealand pouched lamprey sampling sites. Abbreviations: average observed heterozygosity (H_O), average gene diversities within populations heterozygosity (H_S), average Wright's inbreeding coefficient (F_{IS}, following Nei [1987]), and standardized multilocus heterozygosity (sMLH) over all loci. Symbols "(N)" and "(S)" represent South Island and North Island respectively.

Table 1 Pairwise F_{ST} estimates (Wright [1949], updated by Weir and Cockerham [1984]) of Aotearoa New Zealand pouched lamprey from the ten collection sites. Asterisks denote pairwise F_{ST} p-values (1000 bootstraps): "*"=significant at < 0.05, "**"=significant at < 0.005. F_{ST} values are shaded by value from highest (darkest) to lowest (lightest). Symbols "(N)" and "(S)" represent South Island and North Island respectively.

	Piha (N)	**Kaniw (N)**	**Manga (N)**	**Mokau (N)**	**Paua (N)**	**Styx (S)**	**Okuti (S)**	**Waika (S)**	**Poura (S)**	**Waiau (S)**
Piha Stream (N)										
Kaniwhaniwha Stream (N)	0.00153**									
Mangakino Stream (N)	0.00000	0.00044								
Mokau River (N)	0.00341**	0.00147**	0.00174**							
Pauatahanui Stream (N)	0.00323**	0.00075	0.00153**	0.00331**						
Styx River (Drain) (S)	0.00505**	0.00320**	0.00314**	0.00598**	0.00336**					
Okuti River (S)	0.00660**	0.00422**	0.00247*	0.00721**	0.00124	0.00333*				
Waikawa River (S)	0.00346**	0.00217**	0.00291**	0.00338**	0.00153**	0.00152**	0.00269**			
Pourakino River (S)	0.00310**	0.00187**	0.00185**	0.00300**	0.00198**	0.00241**	0.00345**	0.00001		
Waiau River (S)	0.00187**	0.00081**	0.00028	0.00222**	0.00005	0.00231**	0.00231*	0.00132**	0.00141**	

Table 2 Estimates of genetic diversity for Aotearoa New Zealand pouched lamprey sampling sites. Abbreviations: average observed heterozygosity (H_O), average gene diversities within populations heterozygosity (H_S), average Wright's inbreeding coefficient (F_{IS}, following Nei [1987]), and standardized multilocus heterozygosity (sMLH) over all loci. Symbols "(N)" and "(S)" represent South Island and North Island respectively.

Sampling Location	n	H_O	H_S	F_{IS}	sMLH
Piha Stream (N)	12	0.1111	0.1164	0.0455	1.0016
Kaniwhaniwha Stream (N)	25	0.1095	0.1153	0.0508	0.9875
Mangakino Stream (N)	19	0.1106	0.1162	0.0481	0.9963
Mokau River (N)	9	0.1048	0.1078	0.0277	0.9792
Pauatahanui Stream (N)	11	0.1119	0.1162	0.0374	1.0092
Styx River (Drain) (S)	10	0.1087	0.1148	0.0533	0.9795
Okuti River (S)	4	0.1038	0.1106	0.0619	0.9611
Waikawa River (S)	45	0.1118	0.1163	0.039	1.0110
Pourakino River (S)	30	0.1102	0.1153	0.0449	0.9923
Waiau Stream (S)	21	0.1156	0.1181	0.0218	1.0394

Fig. 1 *Geotria australis* (kanakana/piharau; pouched lamprey). A. migrating adult B. oral disc of migrating adult C. juvenile D. adult. Photo credit: A=AKM, B=Matthew Dale, C=Jonah Yick, D=Robert Holdaway. Photo D was reprinted from Population structure in anadromous lampreys: Patterns and processes, In Press, Copyright (2021), with permission from Elsevier.

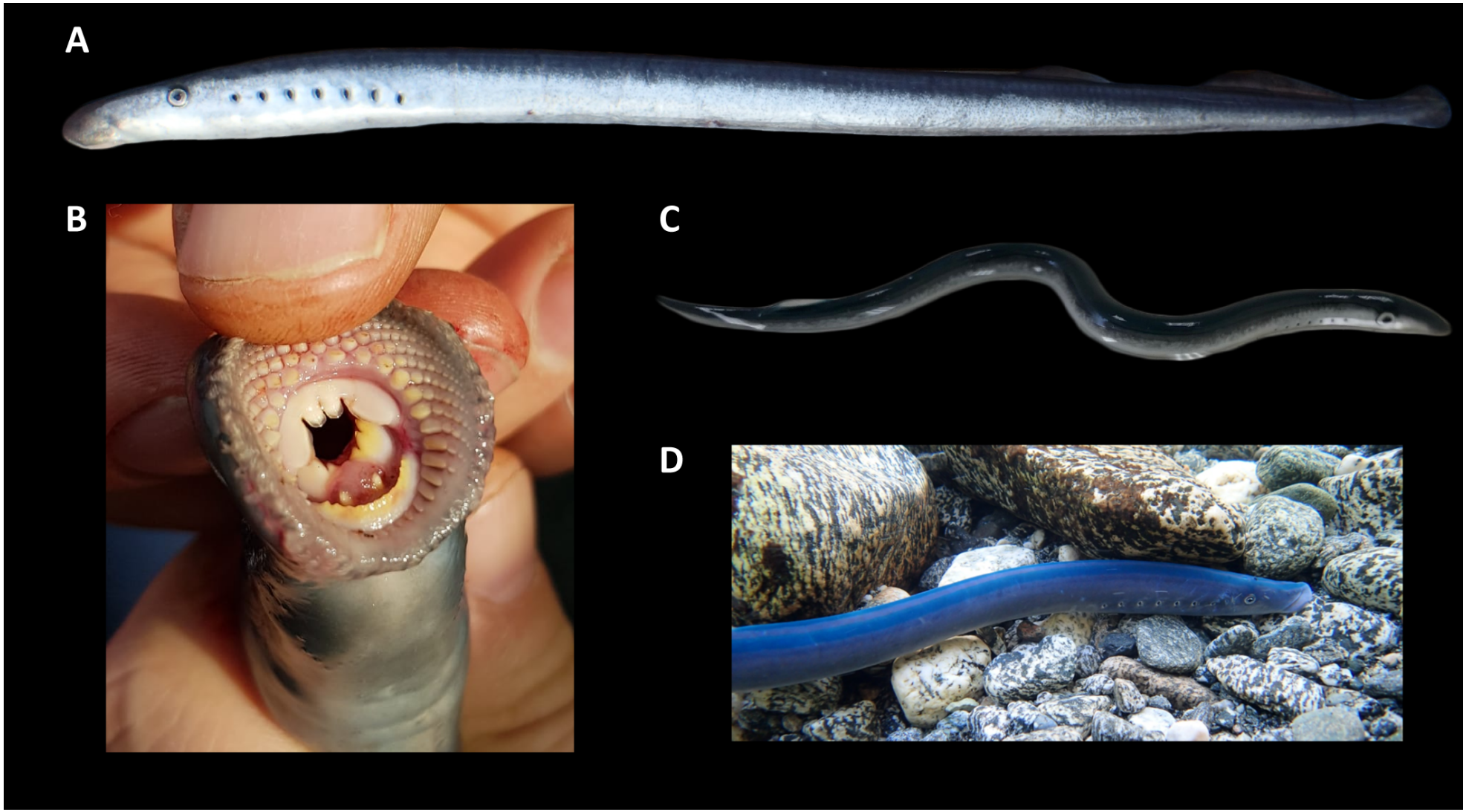


Fig. 2 Aotearoa New Zealand pouched lamprey sampling locations and marine currents (inset). Blue points represent Te Ika-a-Māui/North Island sites. Green points represent Te Waipounamu/South Island sites. Striped blue and striped green points, along with asterisks (*), denote a site sampled on more than one occasion. Numbers below the site names denote the total number of larvae and adults collected from the site. See online version for full colour.

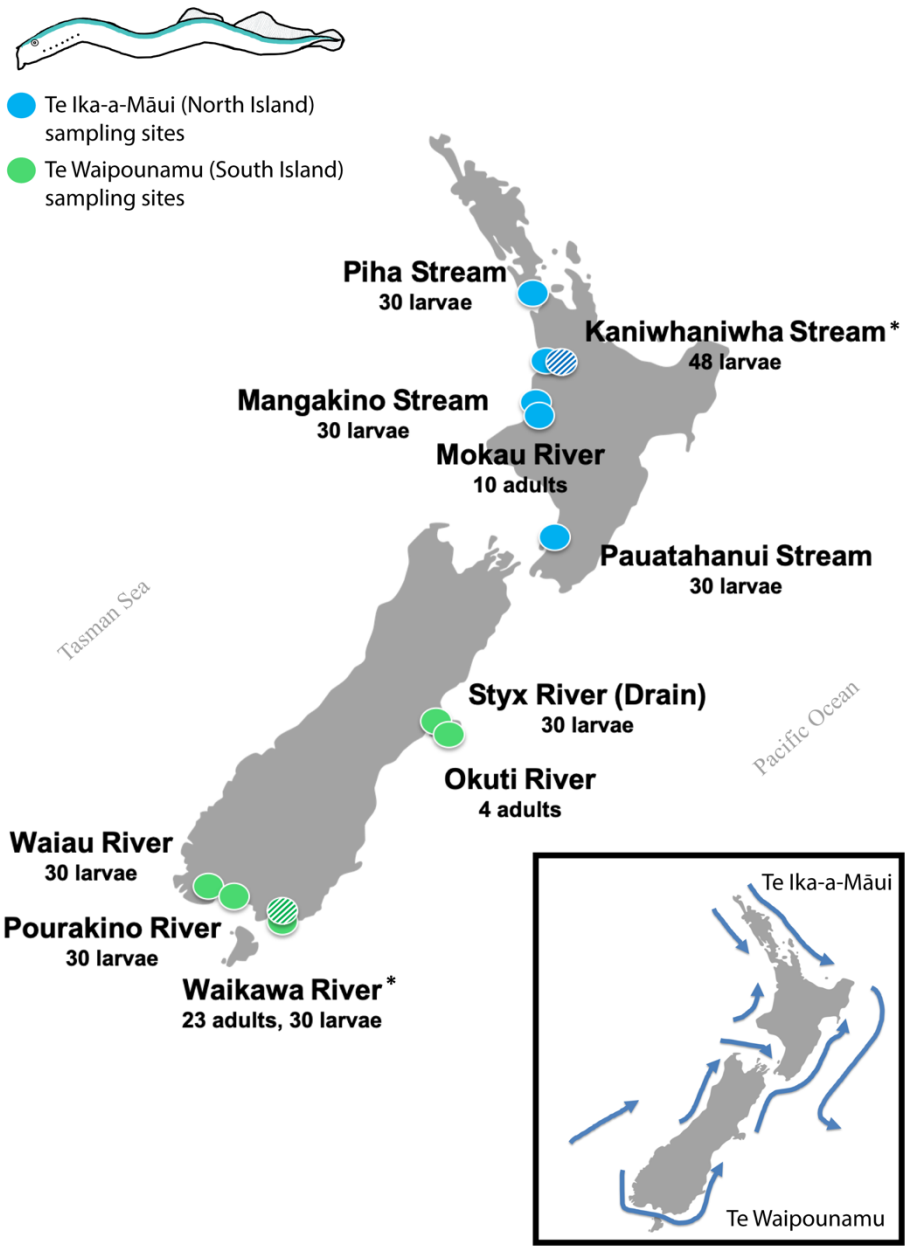


Fig. 3 Median joining haplotype network (created with PopArt) and pouched lamprey sampling locations. Haplotype network: colours correspond to locations, sizes of circles correspond to the frequency of the sequences belonging to a specific haplotype, numbers correspond to the number of fixed substitutions (mutations) from the large Aotearoa New Zealand haplotype, and the Aotearoa New Zealand haplotype acquired from mitochondrial data (NC_029404) is coloured in grey-blue. All of the Aotearoa New Zealand samples represented in the haplotype networks, except for the mitogenome haplotype, were collected during the current study. All of the Australian and Chilean samples were acquired from GenBank. See online version for full colour.

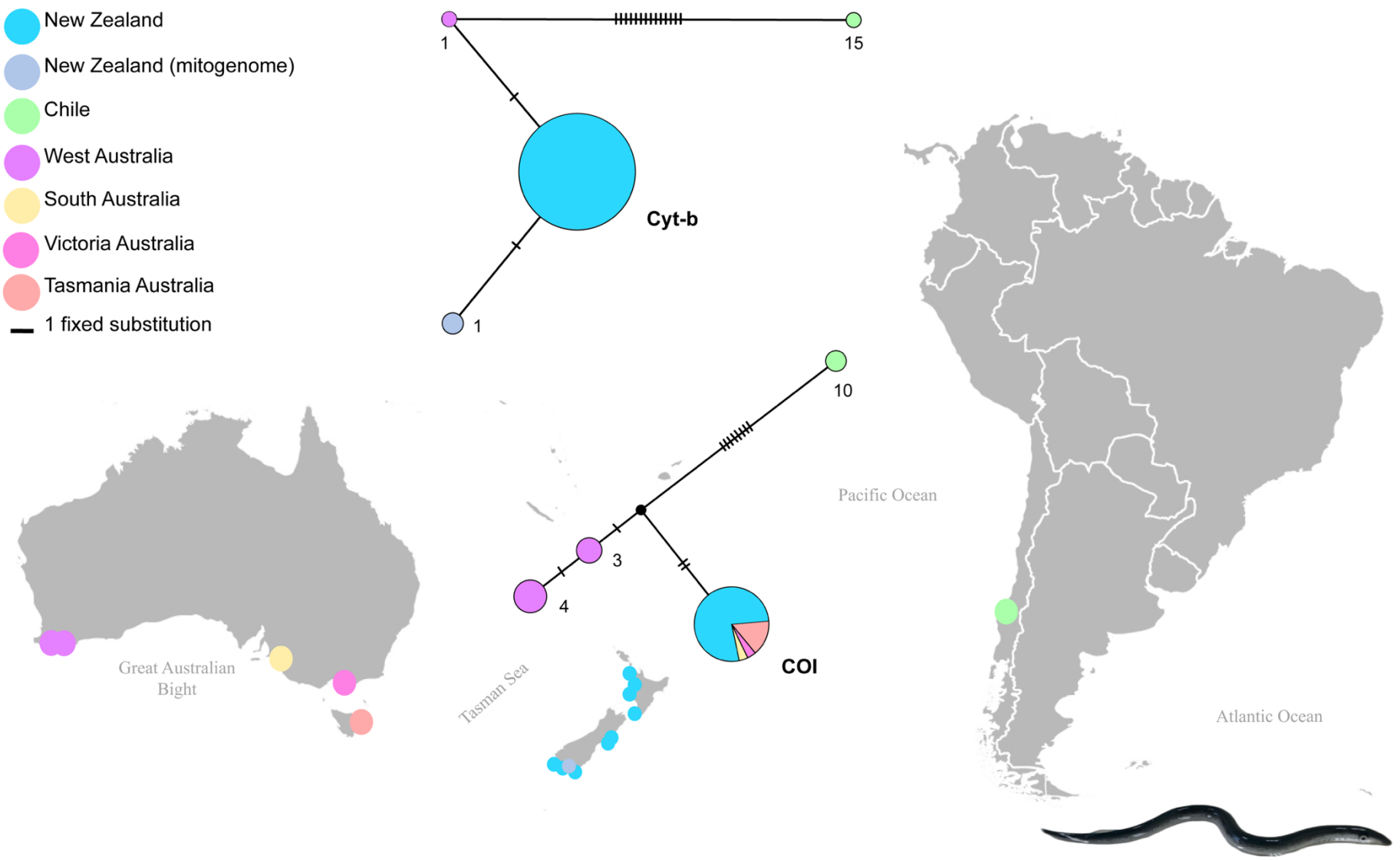


Fig. 4 STRUCTURE v2.3.4 (Pritchard *et al*, 2000) plots of the final filtered Stacks2-assembled RADSeq dataset (see text) for pouched lamprey from 10 sites in Aotearoa New Zealand. Plots for each of the seven genetic clusters (K) are displayed as rows and the ten sampling sites are displayed as columns. Colours represent different genetic clusters. "N" and "S" denote North and South Island sites respectively. K=1 (not shown) had the highest likelihood and therefore the Evanno method could not be implemented.

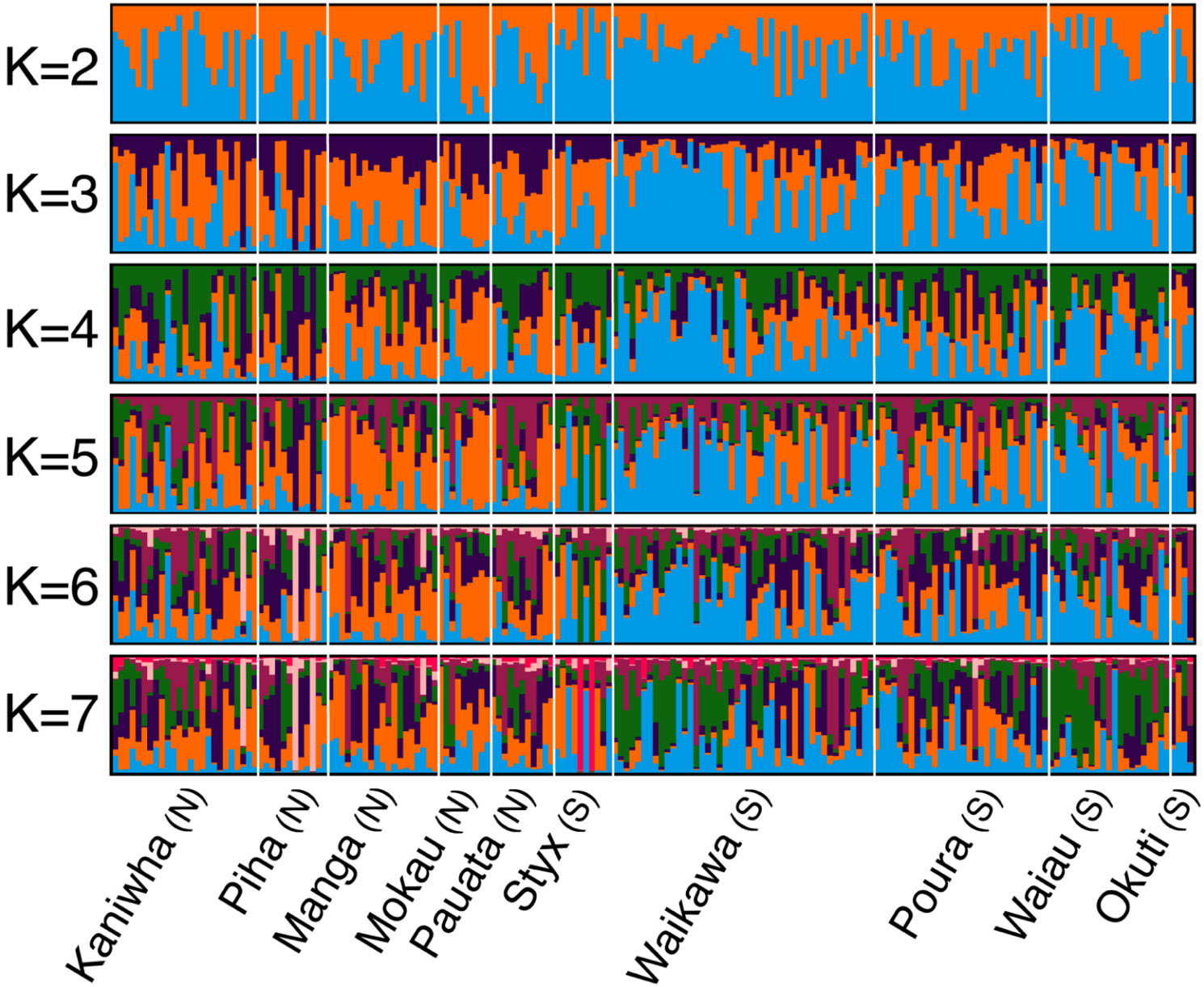


Fig. 5 Principal component analysis (PCA) plots of pouched lamprey from 10 sites in Aotearoa New Zealand coloured by A. sampling location B. latitudinal position C. longitudinal position. Symbols "(N)" and "(S)" represent South Island and North Island respectively. Ellipses = 95% confidence. See online version for full colour.

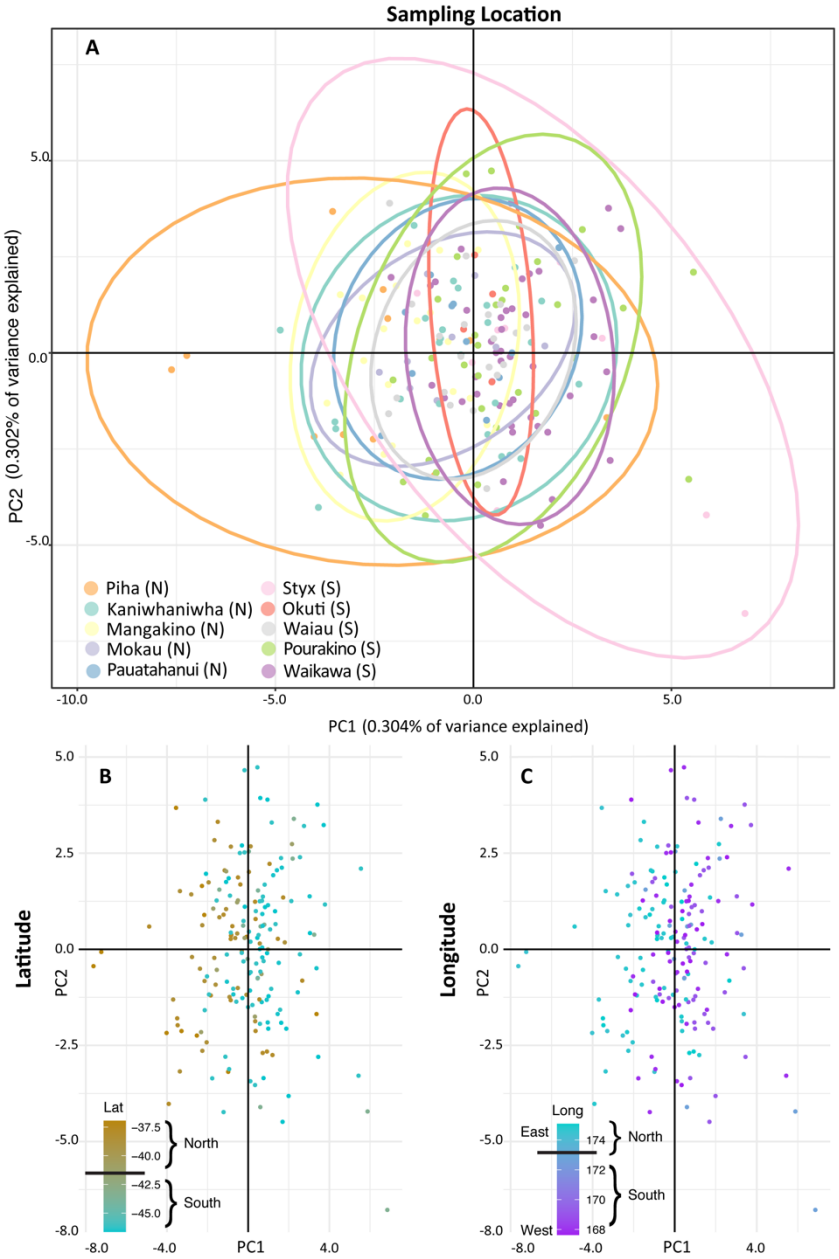


Fig. 6 Discriminant analysis of principle components (DAPC) plot of pouched lamprey from 10 sites in Aotearoa New Zealand coloured by sampling location. A plot of the cumulated variance explained by the eigenvalues is positioned in the top left corner. A plot of the eigenvalues that were retained is positioned in the bottom right corner. Sampling sites are listed in a latitudinal manner with the most northern site at the top left and the most southern site at the bottom right. See online version for full colour.

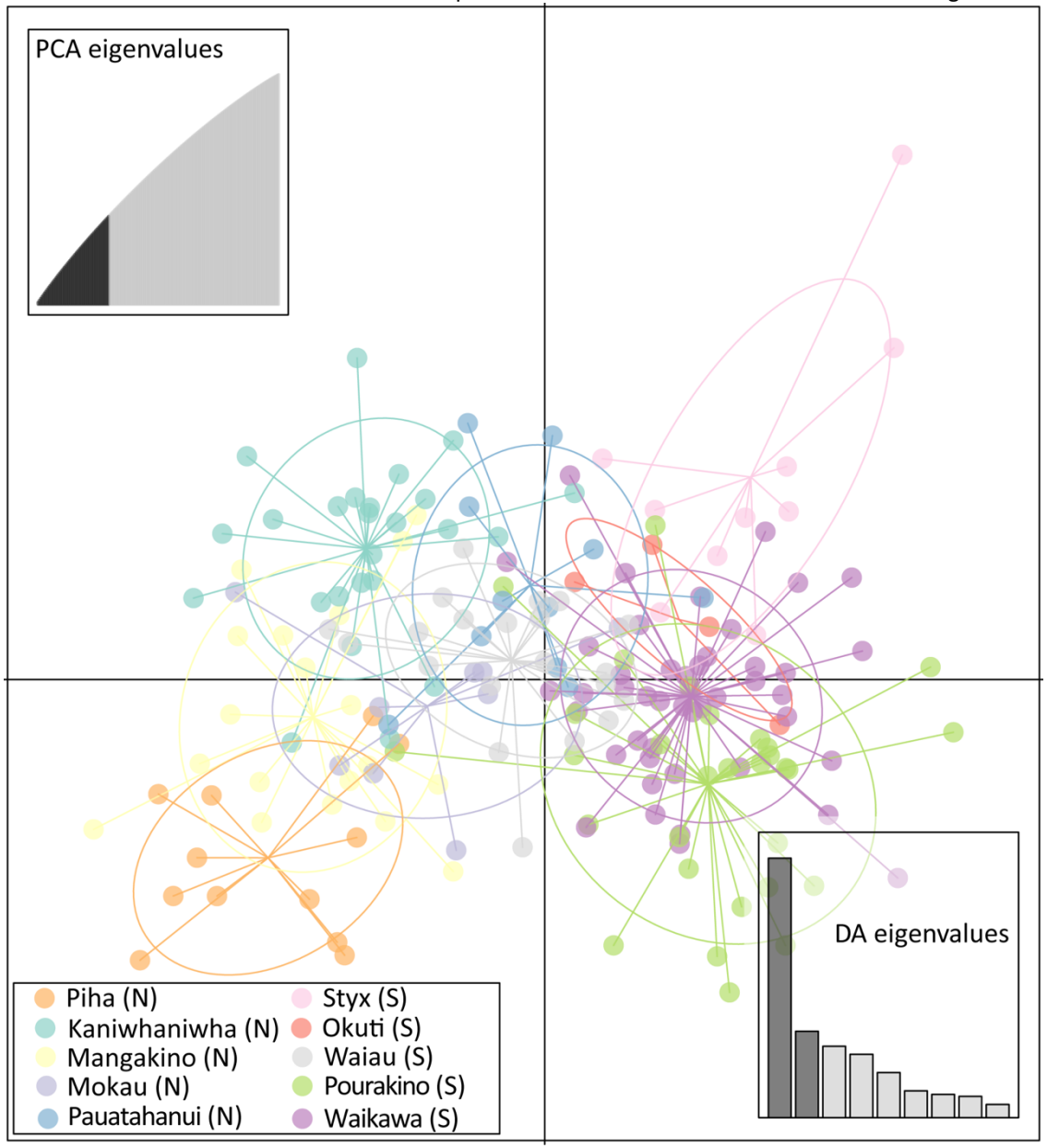


Fig. 7 EEMS posterior mean migration rates m (on the log10 scale) for pouched lamprey in Aotearoa New Zealand. Green indicates higher than average migration and brown indicates lower than average migration. Black circles represent deems consisting of one or more sampling sites and are proportional in size to the number of isolates in the deme. Major marine currents are drawn in blue. See online version for full colour.