Genomics-based mixed-stock analysis reveals potential unsampled populations and population differences in intra-lake migration in walleye.

Julie Gibelli

A Thesis

In The Department of

Biology

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Sciences (Biology, Conservation Genomics)

at Concordia University

Montréal, Québec, Canada

November 2023

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Julie Gibelli

Entitled: Genomics-based mixed-stock analysis reveals potential unsampled populations and population differences in intra-lake migration in walleye.

and submitted in partial fulfillment of the requirements for the degree of

Master of Sciences (Biology, Conservation Genomics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____Chair
Dr. David Walsh

_____External Examiner
Dr. David Walsh

_____Examiner
Dr. Selvadurai Dayanandan

_____Examiner
Dr. Grant E. Brown

_____Thesis Supervisor
Dr. Dylan J. Fraser

Approved by:

_____
Dr. Robert Weladji, Graduate Program Director

_____
Dr. Pascale Sicotte, Dean of Arts and Science

_____2023

ABSTRACT

Genomics-based mixed-stock analysis reveals potential unsampled populations and population differences in intra-lake migration in walleye.

Julie Gibelli

Stock contributions to annual harvests provide key insights to conservation, especially in fish species that return to specific spawning sites and may establish genetically distinct populations. In this context, genetic stock identification (GSI) requires reference samples, yet sampling might be challenging as spawning sites could be in remote and/or unknown areas. Thus, any potential missing source population needs to be accounted for in management recommendations. Here, we (i) genotyped 1487 walleye (*Sander vitreus*) samples using a GT-seq panel of 336 single nucleotide polymorphisms and (ii) assessed individual migration distances from GPS records of fish harvested in two neighboring northern Quebec lakes (Mistassini and Mistasiniishish) important to the local Cree community. Samples were assigned to a source population using two methods, one requiring allele frequencies of known populations (*RUBIAS*) and the other without prior knowledge (*STRUCTURE*). Individual assignments to a known population reached 96% consistency between both methods. All five major source populations were identified in Mistassini Lake, but there was evidence of up to three small unsampled populations. Furthermore, Mistassini walleye populations were characterized by large differences in average migration distance with some remaining near their spawning rivers. In contrast, walleye in Mistasiniishish Lake were assigned with very high confidence to two populations with similar distribution throughout the lake. The complex population structure and migration patterns in the larger Mistassini Lake suggest a more heterogenous habitat and thus, greater potential for local adaptation. This study highlights the importance of combining analytical approaches to improve GSI studies for conservation practices.

ACKNOWLEDGEMENTS

This study was part of the FISHES project (Fostering Indigenous Small-scale fisheries for Health, Economy, and food Security) and would not have been possible without the financial support of the following project partners: the Cree Nation Government, the Cree Nation of Mistissini, Niskamoon Corporation, Eeyou Marine Region Wildlife Board, Genome Canada and Genome Québec.

Special thanks to Pamela Macleod and Hubert Petawabano who were key to coordinating sampling efforts and were always there to support the project. Big thanks to the fishers from the community: Norman Neeposh, Jacob Cooncome, Natalya Assance, Andrew Iserhoff, Leslie Mianscum, Charlo Blacksmith, Matthew Shecapio, Tom Shecapio and all the Nibiischii fishers. I would also like to give all my thanks to Nibiischii Corporation which manages the Albanel-Mistassini-and-Waconichi Wildlife Reserve for their help in providing needed samples for this study.

Heartfelt thanks to the Concordia Team: Kia Marin who coordinated field work and was always there to provide support and advice as well as Brett Studden, Badrouyk Chamlian and Shannon Clarke, who were essential to the field work. I would like to extend my sincere thanks to my lab supervisor, Dylan Fraser, who was always supportive and reassuring, and who gave me an opportunity to work on a project that I love. My committee, Dr. Selvadurai Dayanandan, Dr. Grant E. Brown, Dr. David Walsh, has all my gratitude for their support and useful comments. Lastly, I must express my gratitude to the entirety of the Fraser lab: Sozos Michaelides, Hari Won, Hyung-Bae Jeon, Badrouyk Chamlian, Thaïs Bernos, Johnathan Lemay, Brian Gallagher, Alexandra Engler, Raphaël Bouchard and Nicole Yu for being extraordinary colleagues. And an extra special thanks to Sozos Michaelides and Hari Won who were great friends, guides and mentors.

Contribution of Authors

The authors who worked on this project are Julie Gibelli[1], Hari Won[1], Sozos Michaelides[1], Hyung-Bae Jeon[1], Pamela Macleod[2], Hubert Petawabano[2] and Dylan Fraser[1]

[1]Department of Biology, Concordia University, 7141 Sherbrooke St. West, Montreal (QC) H4B 1R6

[2]Cree Nation of Mistissini, 187 Main St, Mistissini (QC) G0W 1C0

Study concept: DF, PM

Local coordinators for sampling effort: PM, HP

Acquisition of data: JG, HW, SM, HBJ, PM, HP, DF

Analysis and interpretation of data: JG, HW, SM

Drafting of manuscript: JG

Critical revision: JG, SM, DF

Table of Contents

List of Figures vii
List of Tables ix
List of Supplementary materials x
INTRODUCTION 1
MATERIAL & METHODS 6
1) Sampling 6
2) GT-seq panel 6
3) DNA extraction 7
4) Library preparation 7
5) Filtering 8
6) Assignment accuracy, mixed-stock harvest assignments and detecting unknown sources in mixed-stock harvest 9
6.A) Population structure 9
6.B) Assignment accuracy and mixed-stock harvest assignments 10
7) Migration distance 11
RESULTS 13
Sequencing Data validation and filtering 13
Population structure 13
Assignment accuracy and mixed-stock harvest assignments in RUBIAS 14
Comparison between RUBIAS and STRUCTURE 15
Detecting unsampled populations (unknown sources) in mixed-stock harvest 15
Migration distance 16
DISCUSSION 18
Mixed-stock harvest and migration 18
Assignment accuracy and unsampled populations 21
Conclusion 23
REFERENCES 25
TABLES AND FIGURES 34
APPENDIX I – Genotyping-in-Thousands by sequencing (GT-seq) panel development 42
APPENDIX II – Supplementary tables and figures 45

List of Figures

Figure 1. Locations (A), Fst (B) and PCA of the samples (C, D) collected in Mistassini Lake (dark blue) and Mistasiniishish Lake (light blue). (A) Walleye spawning rivers (source) markers (coloured crosses) are placed on the river mouth. When needed, mixed-stock sample markers (yellow squares) are grouped to avoid overlaps. (B) The neighbor-joining tree was based on the source population pairwise Fst estimated with the genepop method and 9,999 bootstraps. All populations significantly differed from each other (P<001) except for ICO and PER (P=0.150). Fst values are available in Table S1 (Appendix II). (C, D) Regarding the PCA, Bernoulli was used as the interpolation method for missing values (Interpolated via binomial draw for each allele against minor allele frequency).

Figure 2. Pattern of clustering of source samples. The samples are ordered according to their river of origin starting first with the samples from the 6 rivers flowing into Mistassini Lake and then with the samples for the rivers flowing into Mistasiniishish Lake. The y axis shows the proportion of membership to each cluster (Q values) estimated by STRUCTURE. Runs with the highest log-likelihood were selected to be presented above.

Figure 3. Proportion of Walleye mixed-stock assigned to the source populations in Mistassini (dark blue) and Mistasiniishish (light blue) lakes. Samples without coordinates (Mistassini, n=21 and Mistasiniishish, n=9) are included in the pie charts summarizing the assignments by lake. Few samples in Mistassini (n=12) and Mistasiniishish (n=2) lakes could not be assigned confidently (PofZ<0.8) and are grouped under the unassigned category (grey color).

Figure 4. Pattern of clustering of source and mixed-stock samples ordered according to *RUBIAS* assignments. The samples are presented first by lake and second according to their river of origin (for sources) or their RUBIAS assignments (for mixed-stock). Samples with low RUBIAS PofZ (<0.8) are shown

last with both lakes combined. The y-axis shows the proportion of membership to each cluster (Q values) estimated by STRUCTURE. Runs with the highest log-likelihood were selected to be presented above. Starting at K7, STRUCTURE identified unknown clusters that are not associated with any source population. The location of the samples belonging to these unknown clusters (Q value >= 0.9; diamond shapes) are marked on the map as well as the location of the source population sampled in this study.

Figure 5. Z-scores distribution of source and mixed-stock samples in Mistassini (A) and Mistasiniishish (B) lakes. The number of mixed-stock samples that do not overlap with the normal distribution are indicated on top.

Figure 6. Walleye migration distance measured as the shortest waterway distance between the river of origin (*i.e.,* the source). (A) Migration distances (km) represented as the density for each population and, (B) the interaction between the population and the timing of capture (here in weeks). Because walleye spawn around the end of May up to early June, the data shown above includes the second week of June (week>24) to ensure walleye had left their spawning grounds.

List of Tables

Table 1. Number of samples collected in the rivers (*i.e.,* sources) and the lakes (*i.e.,* mixed-stock harvest). Coordinates were recorded for each mixed-stock sample harvested in the lakes. The analyses included individual migration distances. They were only estimated for mixed-stock samples which had precise GPS coordinates (*i.e.,* exact coordinates or specific areas such as islands, bays or passes where the sample was caught). The second source in Mistasiniishish Lake (TEM) was identified previously (Michaelides et al., in prep) and was named "Temiscamie" (TEM) as hints suggested this river to be a suitable and likely spawning ground for the walleye from Mistasiniishish Lake.

Table 2. Population differences in migration distances during summer months. Predictions from estimated marginal means computed on the best model explaining migration distances. The model included an interaction between the population ID and the timing of capture (measured as the week). Because walleye can spawn up to the end of May – beginning of June, the data before week 24 (first week of June) was not included in the model. In addition, PIP samples were removed from the model as only four of them were assigned this population. Model comparisons and summary of the best model are available in Tables S3 and S4 in Appendix II.

List of Supplementary materials

APPENDIX I – Genotyping-in-Thousands by sequencing (GT-seq) panel development

Figure S1. DAPC (Discriminant Analysis of Principal Components) using the 1000 candidate markers selected for GT-seq panel development.

APPENDIX II – Supplementary tables and figures

Table S1. Pairwise Fst values estimated with the genepop method. The Fst were calculated with 9999 replicates using the snpR function in R. All pairwise Fst were significant (P<0.001) except for ICO/PER (P=0.150).

Table S2. Comparison of z-score average values between *STRUCTURE* clusters at K10. P values were estimated with pairwise t-tests with a Bonferroni correction. Differences in average z-scores are presented in Figure S3. Overall, admixed individuals, unknown clusters 1, 2 and 3 had lower z-scores.

Table S3. Linear model explaining migration distance (km). The covariates included in the models were population ID (pop) and time of capture measured as the week number (week). The best model was retained based of the residual sum of square (RSS). In addition, the models were compared using the *anova* function in R.

Table S4. Summary of the best model. Overall, the model was highly significant (Residual standard error= 1.274, R^2= 0.381, Adjusted R^2=0.366, $F_{11,455}$=25.440, P<0.001). The population ($F_{5,455}$=18.472, p<0.001), the week ($F_{1,455}$=44.158, p<0.001) and the interaction ($F_{5,455}$=28.658, p<0.001) were all significant (these global F and P values were calculated with the *anova* function in R).

Figure S1 Evanno plots for the *STRUCTURE* runs that included the source samples (A-D) or all samples, *i.e.,* sources and mixed-stock (E-H). The log probability of each run (A, E) and their derivatives (B, C, F, G)

are used to estimate ΔK (D, H). The highest ΔK value indicates the best K according to the Evanno method. These plots were created with the *pophelper* R package.

Figure S2. Assignment accuracy using the leave-one-out method in *RUBIAS*. The assignment accuracy to each of the 8 source populations from both lakes was tested either **(A)** with Icon and Perch separated or **(B)** combined as one unit. To note, when **(A)** ICO and PER were on their own, more simulations fell under the threshold of 80% accuracy (PofZ<0.80: ICO=53.2%, PER=62.4%, PIP=5.4%, CHA=3.2%) than when **(B)** both ICO and PER were merged (PofZ<0.80: PIP=5.5%, CHA=3.2%). The parameters used in the simulations were 1000 replicates and 2000 simulated individuals. The number of simulated individuals assigned to each population is indicated on top.

Figure S3. Relationship between *STRUCTURE* assignments (K10) and *RUBIAS* z-scores. Only individuals with P of Z above 0.8 in *RUBIAS* are shown above. All individuals with *STRUCTURE* Q values below 0.8 are included in the admix group. In addition, K10 was chosen because it is the first K where the sampled PIP population forms its own cluster. Fish belonging to the admix group or the unknown clusters (Unk1, Unk2, Unk3) all have lower *RUBIAS* z-scores on average.

Figure S4. Walleye migration paths based on their population of origin. Migration paths were traced to ensure the shortest waterway distance between the river of origin (source) and the point of capture. Only samples with a P of Z above 0.8 (*i.e.,* confident assignments) were included at this step.

Figure S5. Figure S5. Mixed-stock assignment proportions in both lakes (A) with all the data, (B) without extreme z-scores values, (C) without the early season catches and (D-F) across each sampling year.

INTRODUCTION

Intraspecific variation both at the population level and genetic variation within populations are declining due to human impacts (Des Roches et al., 2021). Environmental changes and human disturbances may affect populations differently, especially when these are locally adapted or vary in early life history traits even in sympatry (Schindler et al., 2010; Tamario et al., 2019). This scenario is particularly likely for species that exhibit homing behaviours and return to reproduce at specific sites. In this case, genetic differences between populations would accumulate due to limited gene flow, random genetic drift and/or natural selection. Furthermore, where selection favours specialization to distinct ecological niches, intermediate phenotypes would be at a disadvantage, thereby increasing the benefits of reproductive isolation (Potvin & Bernatchez, 2001). In this context, a higher diversity of discrete populations can increase species resilience in face of environmental changes (Hilborn et al., 2003; Schindler et al., 2010; Willi et al., 2006).

Maintaining population diversity is especially crucial for exploited species that are regularly harvested. For instance, fisheries with a larger number of populations usually remain more stable, sustainable and productive over time (Hilborn et al., 2003). Indeed, high genetic diversity at the individual, population and/or species level generally has a stabilizing effect on species and ecosystems through a portfolio effect (Hui et al., 2017; Schindler et al., 2010, 2015; Waldman et al., 2016). Monitoring and identifying populations that are contributing more or less to mixed-stock harvests is thus key for providing recommendations to support conservation efforts (Begg & Cadrin, 2016). This can be achieved through Genetic Stock Identification (GSI), a technique that uses allele frequencies estimated from genetic markers to assign unknown samples (*i.e.,* mixed-stock harvest) to source populations based on samples collected from spawning or breeding grounds (Araujo et al., 2014; Seeb et al., 2007; Smouse et al., 1990). Panels including hundreds or more genetic markers, usually Single Nucleotide Polymorphisms (SNPs) are increasingly adopted for GSI and related studies on population diversity in a variety of species as such

tools become less costly to implement and a more common application in management and conservation strategies (*e.g.,* Funk et al., 2012).

However, accurate assignments with GSI relies on (i) a complete knowledge of the sources contributing to the mixed-stock harvest and (ii) the collection of enough samples to represent each source (Beacham et al., 2020). Challenges can arise due to the need for extensive field surveys with no guarantee of accessing all reproductive or spawning areas where source samples need to be collected (Komoroske et al., 2017; Piovano et al., 2019). One way to resolve this would be to combine traditional GSI analyses with related clustering approaches that vary in their strengths and weaknesses. Notably, comparing results from analyses requiring prior knowledge (*e.g., RUBIAS* R package; Moran & Anderson, 2019) with analyses that do not require an allele frequency baseline for assignments (*e.g., STRUCTURE* software; Pritchard et al., 2000) could provide insight into potential unknown populations. Both methods apply Bayesian inference to assign samples to a given population based on samples' allele frequencies. However, *STRUCTURE* groups samples in a predefined number of clusters while *RUBIAS* assigns a sample to a population using representative samples from each population. Furthermore, metrics such as the z-scores (*i.e.,* standard scores) in *RUBIAS*, allow the assessment of unsampled populations by comparing the assignment likelihood to the normal distribution and have been used to discuss assignment accuracy in previous studies (Bowersox et al., 2023; Colston-Nepali et al., 2020; Marín-Nahuelpi et al., 2022; Musleh et al., 2020; Quinn et al., 2021; Spies et al., 2020). However, to our knowledge, unsampled populations among mixed-stock samples are rarely examined in depth using direct comparisons of assignments methods (but see Kuismin et al., 2020, which compared assignment methods with large SNPs datasets)

Refined mixed-stock assignments can be particularly useful to delineate conservation units for species that are difficult to track and/or located in remote areas. The walleye (*Sander vitreus*) is particularly valuable to fisheries and as such might be impacted by overfishing (Hartman, 2009; Navarroli et al., 2021).

This fish is a broadcast spawner that exhibits spawning site philopatry (*e.g.,* Jennings et al. 1996) with adults returning to their natal river to spawn in spring when the ice thaws (Hansen et al., 2022). This homing behaviour can lead to the rise of discrete populations which has implications for management (Stepien et al., 2018). In addition, walleye can live up to 20 years, reproduce several times, travel long distances of over one hundred kilometers (km) to feeding sites, and can inhabit a large variety of habitats (Bozek et al., 2011; Hansen et al., 2022; Hartman, 2009; Raby et al., 2018). Despite these characteristics, many studies in southern Canada, especially in the Great Lakes, have shown that walleye can be strongly impacted by habitat degradation and overfishing (Euclide et al., 2021; Stepien et al., 2009, 2010; Wilson et al., 2007).

In contrast, few studies related to GSI and population structure have been conducted on walleye in northern Canada (Bowles et al., 2020, 2022; Dupont et al., 2007) where the species is important to the fisheries of Indigenous communities. Notably, the Cree in northern Quebec, Canada, have depended on fishing for a long time, with walleye being a prized catch, for both subsistence fishers and sport fishers at tourism outfitting camps operated by the Cree Nation of Mistissini and Nibischii Corporation's operarion of the Albanel-Mistassini-and-Waconichi Lakes Wildlife Reserve (Bowles et al., 2022; Marin et al., 2017). Mistassini Lake, Quebec's largest lake, is an important part of the Wildlife reserve, with a surface of 2335km^2, a length of 161km and a maximum depth of 183m. Remarkably, Mistassini Lake spans over two climatic zones separated by the 51st parallel, with air temperatures of 816-979 degree-days per year in the north and 979-1141 degree-days in the south (Dupont et al., 2007). Adjacent Lake Mistasiniishish (Albanel) is also a popular destination for walleye anglers. Both lakes are oligotrophic, post-glacial, and considered largely unimpacted by human activities. However, information from local Cree knowledge indicated a reduction in the size and number of walleye caught in southern Mistassini Lake from 2002 to 2017 (Bowles et al., 2020, 2022). Genomic results also suggested a temporal change in population structure that could be sign of harvest-induced evolution (Bowles et al., 2020). Together, these studies

highlight the importance of further investigating the population structure and mixed-stock harvest contributions in both lakes.

The last study on walleye GSI in Mistassini Lake was done 20 years ago (Dupont et al., 2007). The authors identified 4 distinct and temporally stable groups using 10 microsatellite markers. More precisely, the South harboured two groups that originated from the following rivers: Perch-Icon, two populations that formed one unit, and Chalifour. The other two groups were from the Rupert River, the lake's outflow, and from Takwa River, flowing into the northern end of the lake. The Takwa population was the main contributor to the mixed-stock harvest (38% in 2002 and 42% in 2003). The results also suggested that the latter migrated much further than any of the other populations. Dupont *et al.* (2007) hypothesized that the observed dispersal patterns could be due to the harsh thermal conditions for the walleye which thrive at temperatures between 13-21°C while surface temperatures of deep-water areas in Mistassini rarely exceed 15°C. Fish from the northern population would then migrate for longer distances to find warmer and more suitable habitats. It is worth noting that environmental conditions and long-distance migration have been linked to local adaptation and population divergence in the Coho salmon (*Oncorhynchus kisutch*; Rougemont et al., 2023). Both migration distances and population structure can provide insights regarding putative local adaptation either to the lake, to the spawning rivers or to both (*e.g.*, Fraser & Bernatchez, 2005). This information might prove valuable for fine-scale management as any shift in local environmental conditions could impact more locally adapted populations as well as, in the long term, the global species' genetic diversity (Meek et al., 2023)..

The main goals of this thesis, as part of the FISHES project (Fostering Indigenous Small-scale fisheries for Health, Economy, and food Security), are to (i) monitor and assess population contributions to the mixed-stock harvest in both Mistassini Lake and its neighbouring Lake Mistasiniishish, (ii) investigate unsampled populations to refine assignments and pinpoint areas of interest in the lakes for future monitoring and

(iii) assess migration patterns to further understand how walleye use both lakes' habitats. To achieve this, a Genotyping-in-Thousands by sequencing (GTseq) panel of 336 putatively neutral SNPs was developed and adopted to maximize genetic differentiation between populations in the region. Adding to previous genomic work on the walleye in the area (Bowles et al., 2020; Dupont et al., 2007), we were able to do a more extensive sampling to capture as much variation as possible among mixed-stock samples. This is especially meaningful for Mistasiniishish where the genetic structure and mixed-stock harvest composition have not been examined before. We predicted that the combination of GSI analyses will reveal unsampled populations as currently only one population is known in Mistasiniishish Lake while five were sampled in Mistassini Lake. In addition, due to the variation in size, depth, temperature and likely productivity across and between lakes, we expected mixed-stock harvest proportions to differ between locations and, in turn, migration distances as walleye move in search of productive habitats during the summer season.

MATERIAL & METHODS

1) Sampling

To assign harvested fish in the two lakes to their population of origin, we sampled pre-spawning and spawning walleye in rivers where they are known to spawn during the May and early June (n=267). These samples of known origin are hereafter referred to as "source population samples". Six sources were identified in Mistassini Lake, while only one river flowing in Mistasiniishish Lake could be sampled, as the area was difficult to access. However, a low-coverage whole-genome sequencing (lcWGS) project confirmed the presence of a second genetic group among a subset of the samples collected in Mistasiniishish Lake (Michaelides et al., in prep). The lcWGS samples with a proportion of membership above 90% to this second source (n=35; estimated with NGSadmix; (Skotte et al., 2013)) were used in this study as an additional source population for Mistasiniishish Lake (see Table 1 for details). This additional source population will be referred hereafter as "Temiscamie" (TEM) as hints suggested this river to be a suitable and likely spawning ground for the walleye in this lake. Another 1,279 individual samples were collected by Cree partners, recreational fishers and Concordia personnel throughout the summer months across three field seasons (June-August, 2020-2022) in lakes 1 and 2. These samples are hereafter referred to as "mixed-stock samples". Samples were collected non-lethally through caudal fin clips and preserved in 95% ethanol. GPS coordinates or location names were recorded for each sample. Details regarding the sampling design are available in Table 1 and sampling locations can be found in Figure 1.

2) GT-seq panel

The GT-seq panel was developed using ~1,000 candidate SNPs identified in from a Genotyping-By-Sequencing dataset which included samples from four source populations *i.e.,* ICO, PER, CHA, TAK (Bowles et al., 2020). The markers were selected to maximize genetic differentiation between southern

populations in Mistassini Lake (higher Fst), namely CHA and ICOPER (ICO and PER were combined as they formed one genetic groups) which were also known to be close genetically (Bowles et al., 2020). The panel was designed, tested, and validated by the GTseek company (USA). The primers (length = 20bp) were designed to target a product size of 50 – 120 bases with the SNP being within the first 75 bp. A total of 364 SNPs passed all filtering steps. Among those, 28 were monomorphic leaving 336 markers which were used in this paper. More details regarding the panel development are available in Appendix I.

3) DNA extraction

DNA from source and mixed-stock samples was extracted in the lab using either Qiagen kit (DNeasy Blood and Tissue kit) or a salt-extraction method (Aljanabi, 1997). In all cases, the tissues were first cut using scissors and pliers sterilized by passing them through a flame between each sample. The samples were then incubated and digested overnight in a solution containing an extraction buffer (ddH_2O + EDTA + NaCl + Tris), 10% SDS, Proteinase K and RNase A. Regarding salt extractions, the protocol consisted in a series of centrifugation steps during which salt and isopropanol was used to precipitate the DNA into a pellet followed by two washings in 80% ethanol. In the last step, the pellet was dried for 1H and mixed with a 50µL elution buffer (AE from Qiagen).

4) Library preparation

The extracted DNA was prepared for sequencing following Campbell et al. (2015) GTseq protocol with some modifications. It was first quantified using Qubit (Invitrogen broad range kit). If the DNA concentration was above 30ng/µL, it was diluted to 15ng/µL to avoid any amplification biases due to high concentration. Two PCRs were then performed, the first one to amplify the targeted SNPs with the GT-seq primers, and the second one to add the barcodes and capture sequences (P5 and P7) for Illumina sequencing. The second PCR products were pooled before the size selection procedure with AMPure XP

beads (double-sided bead selection 0.5x and 1.2x). Finally, the pools were quantified using Qubit (Invitrogen high sensitivity kit), normalized to 4nM and pooled (up to 6 library-pools, max n=529 samples) for sequencing on three MiSeq v3 lanes (paired-end; 2 x 75 bp) at the Institute of Integrative Biology and Systems (IBIS, Université Laval).

5) Filtering

The quality of SNPs was checked using *GTscore* (McKinney et al., 2020). Thirty-three SNPs were filtered out because their genotype rate was below 50% (*i.e.,* the proportion of genotypes per SNP with non-missing data), leaving in total 303 SNPs for remaining analyses. *GTscore* also removed samples with a high contamination score (n=10). Samples were considered contaminated by *GTscore* when allele ratios of a high proportion of their heterozygous genotypes significantly differed from 1:1 ratios. The threshold used for the contamination score was 0.3 as suggested by McKinney et al. (2020). Adding to that, to avoid any bias due to missing values, we only kept the samples that were genotyped with at least 70% of the GTseq-panel. Finally, two source samples (n=1, CHA and n=1, MAU) were removed because their genotype did not match their population of origin and this could bias the mixed-stock assignments. After these filtering steps, 93% of the samples were retained for downstream analyses (see Table 1).

We tested for departure from the Hardy-Weinberg Equilibrium (HWE) on source samples using the *snpR* package (Hemstrom & Jones, 2023). We applied a Bonferroni correction to control for multiple testing. Only six out of 303 loci significantly departed from HWE – this result did not differ from what should be observed due to random chance (*i.e.,* ~5%). Therefore, all loci were kept for subsequent analyses.

6) Assignment accuracy, mixed-stock harvest assignments and detecting unknown sources in mixed-stock harvest

6.A) Population structure

Fixation index (Fst) values were calculated for each source population pair with *snpR*. We used the genepop method (Rousset, 2008) and 9,999 bootstraps to compute the P-values. To visualize the population structure in both lakes, we performed principal component analyses (PCA) with *snpR* R package using all samples (*i.e.,* source and mixed-stock). Missing data was corrected using Bernoulli interpolation in *snpR i.e.,* binomial draws to estimate minor allele frequencies for each missing data point. A small percentage of the genetic variance was explained by the PCA axes (<5%, see Figure 1C, D) suggesting that few linear combinations of variables (here SNPs) can explain the observed genetic variance. This can be due to the panel being designed to maximize Fst between two of the seven source populations (CHA and ICOPER). Nonetheless, PCA should be interpreted with caution in genomic studies (Elhaik, 2022) and will be used primarily in this study for visualization.

We also ran *STRUCTURE* (Pritchard *et al.* 2000), first on only the sources and then on all samples to (i) refine population structure, (ii) compare with the mixed-stock assignment and (iii) investigate potential sub-clustering or unknown sources in the mixed-stock samples. *STRUCTURE* assigns samples based on their allele frequencies to a number of K-clusters defined by the user. In each *STRUCTURE* run, we used a 100,000 burn-in period to allow convergence toward reliable estimates of the allele frequencies followed by 100,000 *Markov Chain Monte Carlo* (MCMC) repeats. We chose the "admixture model" with correlated allele frequencies allowing the estimation of each individual admixture proportion and is robust to weak population differentiation (Falush et al., 2003). Because the Fst and PCA suggested at least 7 populations across both lakes, we performed 10 iterations for K-values of K2 to K11 to ensure we captured any putative clustering patterns exceeding K=7. We presented the data up to K10 only, as one source (MET) started

splitting into sub-clusters at K11 despite no evidence suggesting sub-structure for this source in any other analyses. The best number of K was determined by (i) combining the results from the Fst, PCA and the Evanno method applied to the *STRUCTURE* outputs using the *pophelper* R package (Evanno et al., 2005; Francis, 2017) and by (ii) taking into account previous studies and what is known about walleye biology in the area (e.g. Dupont et al., 2007; Bowles et al., 2020, 2022).

6.B) Assignment accuracy and mixed-stock harvest assignments

Tests of assignment accuracy and mixed-stock assignments were performed with the functions from the *RUBIAS* R package (Anderson et al., 2008). *RUBIAS* uses Bayesian inference via MCMC to assess mixing proportions, z-scores and individual posterior probabilities of assignment based on source allele frequencies.

First, we evaluated with the function "*assess_reference_loo"* how accurate the source population dataset would be for population assignments using the leave-one-out method. The model makes mixture simulations and estimates the likelihood for a given genotype to belong to its known population of origin after removing it from the allele counts. We carried out the simulations with 1,000 replicates, 2,000 simulated mixture individuals and the default value for mixing proportion (*i.e.,* Dirichlet distribution with alpha=1.5). We also tested different Dirichlet parameters (*i.e.,* alpha=1, 1.5, 2, 2.5 and 3) and realistic mixing proportions based on the mixed-stock assignments (see below).

Second, we used the "*infer_mixture"* function to assign mixed-stock samples based on the source population dataset. We repeated the procedure for mixed-stock samples from Mistassini and Mistasiniishish lakes with the same source population dataset which included all source populations to investigate potential inter-lake migration. We ran the function with 200,000 MCMC iterations, 40,000 burn-ins and 1,000 bootstraps. The assignments with posterior means of group membership (PofZ) above 0.8 were retained for further analyses.

RUBIAS also provides individual z-scores computed from each individual loglikelihood which can be examined to detect potential unsampled populations among mixed-stock samples. Thus, we checked if the mixed-stock z-scores were normally distributed and how they compared to the distribution of source z-score calculated with the "*self_assign*" function. We used a Kolmogorov-Smirnov (KS) test to compare the z-score distributions with the expectation that (i) a deviation from the normal distribution of more than two standard deviations may reflect individuals that originate from unsampled sources (Anderson et al., 2008) and (ii) mixed-stock z-score distribution should also fit the source one if the source population dataset represents well the mixed-stock samples. Lastly, because populations might differ in their assignment accuracy, we also compared the mixed-stock z-scores based on population assignments using pairwise t-tests with a Bonferroni correction (n=11 potential source populations).

7) Migration distance

We were able to record exact GPS coordinates for 93% of the fish caught in Mistassini Lake and 51% of the fish caught in Mistasiniishish Lake (n=860; Table 1). Migration distance was defined as the shortest waterway distance starting from the river mouth of a given population to the point of capture (see locations in Figure 1). The distance travelled by the fish in kilometers (km) was estimated using Google Earth (Google Earth Pro, version 7.3.6.9345). If clouds covered the satellite image, the map view was used instead. Two spawning grounds (rivers), ICO and PER, were geographically close to each other (9.7 km apart) and undistinguishable genetically (this study; Dupont et al 2007; Bowles et al 2020). Consequently, the mean distance from each source was used to estimate the migration distance of ICOPER fish. Lastly, only migration distances of samples caught after week 24, *i.e.,* after the first week of June (n=467) were analysed so that most fish would have had time to leave their spawning grounds. Indeed, source population samples were collected up to early June in previous years, especially in the north (*e.g.,* TAK).

We investigated which combination of covariates, including the interactions, best explained migration distances using linear models. The covariates included population ID (*i.e., RUBIAS* assignments) and the time of capture measured in weeks. We compared the models with the *anova* function in R and selected the one with smaller residual sum of squares (*i.e.,* the best fit). We also computed the estimated marginal means for each population combination using the *emmeans* R package (Lenth et al., 2023) so we would be able to compare population migration distances. We added weights to control for variation in the number of samples from the same population caught at the same locations. This variable was divided by the maximum value so that more weight would be given to migration distances estimated from locations that were more representative for a given population. Because migration distances were not normally distributed, we applied a Box-Cox transformation (Box & Cox, 1964; Venables & Ripley, 2002).

RESULTS

Sequencing Data validation and filtering

Overall, the *GT-score* showed that 90% of the reads per sample were on the targeted SNPs (on-target reads: 29546 ±16071; total reads: 32853 ±17692). The read depth per SNP (i.e., the number of reads at a given position/nucleotide) was on average 82.861 ±68.909 which was enough for GSI (*e.g.,* Bootsma et al., 2020). In addition, sample genotype rate was on average at 0.926 ±0.083 with only 15 samples removed due to being below the 50% threshold. Sample heterozygosity was 0.257 ±0.035 and *GTscore* contamination score was low on average *i.e.,* below the recommended threshold of 0.3 (0.113 ±0.05).

Population structure

Pairwise Fst values among the 8 source population samples ranged from 0.003 to 0.118 (Table S1, Appendix II). Across both lakes, only two sources located close to each other (ICO-PER; Figure 1A), did not significantly differ genetically (Figure 1B, 1C and 2). CHA and PIP appeared to be closely related (Figure 1B, 1C) but their significant pairwise Fst indicated that these populations were distinct (Table S1, Appendix II). In contrast, the population located in the outflow of lake 1 (MAU) was the most differentiated (Figure 1B, 1C). Interestingly, the northern population from lake 1 (TAK) was genetically closer to the populations from lake 2 (MET and TEM; Figure 1B and 2). As highlighted by the PCAs, Lake 1 had a more complex population structure overall with a small subset of mixed-stock samples not overlapping any sampled source population (Figure 1C) while Lake 2 had two clearly differentiated populations (Figure 1D).

Supporting these observations, *STRUCTURE* runs on source samples showed a differentiation between the southern populations (ICOPER), the population in the outflow (MAU) and the northern populations (TAK, MET and TEM) across both lakes (Figure 2). This was apparent as early as K3 which also was the best K according to the Evanno method (Figure S1A-D). However, K3 might be indicative of the population

structure at a broad-scale, with a higher K-value underpinning a finer-scale structure corroborating the results from the PCAs and pairwise Fst. The known source PIP splits from CHA at K7 (Figure 2) suggesting the existence of seven genetically distinct populations among the source populations.

Assignment accuracy and mixed-stock harvest assignments in RUBIAS

We first tested the accuracy of the source population dataset in assigning simulated samples with the leave-one-out method. Merging ICO and PER improved the assignment accuracy across all source population from 84.24% to 98.77% (Figure S1A, B). In addition, changing the Dirichlet parameters or the mixing proportions to match the real mixed-stock proportions consistently yielded a similar degree of accuracy (all simulations >98.73%).

In Mistassini Lake, 97.97% of the mixed-stock samples were confidently assigned by *RUBIAS* to a source population (PofZ>0.8). Most of the low confidence assignments (n=12) belonged to CHA (n=5) while the others were spread among the other populations: MAU (n=3), TAK (n=2), ICOPER (n=1) and TEM (n=1). TAK was the largest population and main contributor (n=285; 48.31%) to the lake's harvest, followed by MAU (n=89; 15.08%) located in the outflow, and ICOPER (n=86; 14.58%) and CHA (n=82; 13.90%) located in the south (Figure 3). PIP was only a minor source (n=7; 1.19%). Interestingly 4.92% of the mixed-stock in Mistassini Lake (n=29) were assigned to Mistasiniishish Lake sources, indicating inter-lake migrations (Figure 3).

In Mistasiniishish Lake, almost all samples (99.67%) were assigned confidently (PofZ >0.8). Most of the mixed-stock harvest in Mistasiniishish Lake originated from the TEM (69.17%), the source that could not be sampled directly during the spawning season, while the remaining fish were assigned to MET (30.51%; Figure 3). There was no indication for bi-directional inter-lake movement *i.e.,* no fish from Mistasiniishish Lake were assigned to sources from Mistassini Lake (Figure 3).

Comparison between RUBIAS and STRUCTURE

Another way to explore whether the assignments were accurate was to compare *RUBIAS* and *STRUCTURE* runs including the source and mixed-stock samples (Figure 4). In this case, the best K suggested by the Evanno method was K5 (Figure S2E-H). However, K6 was more relevant given that, at K5, CHA and PIP were still admixed populations while, at K7, the runs showed splits unrelated to any sampled sources (Figure 4).

The *RUBIAS/STRUCTURE* comparison across both lakes showed that 96.18% (n=1157) of the assignments were consistent at K6, leaving less than 4% (n=46) of the samples lacking clustering correspondence. Of these 46 samples, 14 had both low *RUBIAS* probabilities (PofZ<0.80) and low Q values (*i.e.,* high admixture) in *STRUCTURE* (see low PofZ in Figure 4). Furthermore, all fish assigned to PIP in *RUBIAS* (n=7) were assigned to CHA in *STRUCTURE*. Indeed, the PIP cluster started to differentiate at K10, probably due to the low number of PIP samples compared to the other populations (Figure 4). Note that even at higher K-values, consistency between *RUBIAS* and *STRUCTURE* assignments remained high (K7 to K10, assignment consistency >88.69%). Collectively, this comparison suggests that most fish in the mixed-stock harvest originated from known source populations.

Detecting unsampled populations (unknown sources) in mixed-stock harvest

The mixed-stock z-score distribution from *RUBIAS* generally overlapped with the normal and source z-score distributions in both lakes (Figure 5). Specifically, all sources were likely sampled in Mistasiniishish Lake as the mixed-stock z-scores fitted well both distributions (Figure 5B, normal: D=0.075, P=0.064; source: D=0.070, P=0.330). However, the Kolmogorov-Smirnov tests revealed that the distributions significantly differed in Mistassini Lake (Figure 5A, normal: D=0.169, P<0.001; source: D=0.164, P=0.001). More precisely, 75 samples had z-scores below two standard deviations (z-score<-2). Among those, 70

samples had PofZ>0.8 with 41 being assigned to one of the unknown clusters (at K10), 19 being admixed (Q value<0.8) and only 10 belonging to one of the four major source populations (ICOPER=4, MAU=1, CHA=1, TAK=3, TEM=1). This suggests that some Mistassini Lake mixed-stock samples might not have originated from the known, sampled sources.

Further, *STRUCTURE* runs at higher K-values (*i.e.,* K7 to K10) showed three new, unknown clusters among the mixed-stock samples initially assigned to CHA (unknown 1), TAK (unknown 2) and MAU (unknown 3; Figure 4). The individuals from each unknown cluster were caught in similar locations (Map, Figure 4). In particular, fish from unknown clusters 2 and 3 were caught close to the source to which they were initially assigned (Map, Figure 4). Even with the evidence that all major sources were sampled, it is worth noting that removing the fish belonging to these unknown clusters (with Q values>0.8 at K10) significantly improved the fit of Mistassini Lake mixed-stock to the source z-scores (D=0.089, P=0.123). Furthermore, individual belonging to theses unknown clusters had lower z-scores on average (see Table S2 and Figure S3, Appendix II).

Migration distance

The best fit model suggested that populations varied in their migration distances and that this depended on the time of capture (Table S3, Appendix II). Overall, the population ($F_{5,455}$=18.472, p<0.001, Figure 6A), time of capture ($F_{1,455}$=44.158, p<0.001) and the interaction ($F_{5,455}$=28.658, p<0.001, Figure 6B) were all significant (see Table S4, Appendix II, for all model estimates). CHA (48 ±32 km) and two closely related populations from the north, TAK (53 ±36 km) and TEM (29 ±20 km), were characterized by having the longest migration distance (>120 km). TAK and CHA were however the one migrating further on average (Table 2). MET (28 ±18 km) and MAU (24 ±16 km) migrated less than CHA, TAK and TEM but travel further than PIP (9 ±7 km, but not part of the data used in the model) and ICOPER (15 ±10 km, Table 2). Removal

of the 75 samples that might have been misassigned (z-score<-2), yielded the same results albeit TAK became the one migrating further followed by CHA and TEM (results not shown).

Because of the low number of mixed-stock samples (n=7), PIP samples were not included in the model, but they were consistently caught in close proximity of their spawning ground (see Figure S4, Appendix II for all population migration paths). In addition, the model only included the samples caught after the first week of June as individuals can remain at the spawning grounds up to this time. The removed samples disproportionally originated from northern populations that are also known to spawn later in the season (46% from TAK and 32% from TEM). As an important note, mixed-stock proportions in Mistassini Lake were affected when removing these early season catches. However, TAK remained the main contributor of the annual harvest, albeit by a much small margin (see Figure S5, Appendix II). Other factors, such as the sampling year had also a low impact on the mixed-stock harvest proportions (Figure S5, Appendix II).

DISCUSSION

Using a recently developed SNP panel for walleye, we were able to accurately distinguish at least seven populations, five in Mistassini Lake and two in its neighbouring lake Mistasiniishish. All major populations contributing to the annual harvest were identified, as (i) most of the samples were assigned to a known source with high confidence, (ii) comparison of assignments made using a reference dataset (*RUBIAS*) and a clustering method (*STRUCTURE*) with six clusters assumed (*i.e.,* when all and only the main source populations were represented) showed that 96% of the samples were consistently assigned/clustered to the same source across methods and, (iii) the results were congruent with prior knowledge of the species' structure in this lake system (Bowles et al., 2020; Dupont et al., 2007; Navarroli et al., 2021). We were also able to find evidence of up to three small unsampled sources in Mistassini Lake. The presence of additional genetic groups, even though of minor importance to the mixed-stock harvest, are worth investigating given that high population diversity might bolster species' resilience through a portfolio effect (Schindler et al., 2010). In addition, this observation highlights putative differences in local adaptation between the two studied lakes with implications for management.

Mixed-stock harvest and migration

Two of the main objectives of this study were to assess the contribution of each spawning population to the mixed-stock harvest and the potential for population variation in intra-lake feeding migrations. As in the Dupont et al. (2007) study conducted 20 years ago on Mistassini Lake walleye, we were able to confirm that the same four populations, namely TAK, MAU, CHA and ICOPER, were the main contributors to the annual harvest (Figure 3). Specifically, the northern population (TAK) accounted for the largest proportion (48%), even when removing a set of early catches that may have overlapped with spawning season (Figure S5, Appendix II). Additionally, mixed-stock harvest proportions were minimally affected by the removal of samples that could be misassigned according to z-scores (Figure S5, Appendix II). Overall, Mistassini Lake

harboured diverse populations with large variations in the mixed-stock proportions recorded across the lake (Figure 3). This diversity included 5% of the harvest that originated from the neighbouring Lake Mistasiniishish. The opposite was not true, which suggests that the steep gradient of waterfalls connecting the lakes probably prevented bidirectional movement for walleye toward Mistasiniishish Lake. In contrast to Mistassini, Mistasiniishish had only two populations, with TEM being the largest one (69% of the mixed-stock harvest), distributed in a similar fashion across the lake (Figure 3). Walleye distribution pattern in Mistasiniishish Lake indicates that both populations did not differ in the way they occupy the habitat during the summer growing season, but were likely adapted to their spawning rivers (MET and TEM). These rivers are geographically very close to each other but flow from opposite directions on a North-South axis (Figure 1). Thus, the divergence between Mistasiniishish populations might be explained by differences in spawning time as warmer water from the South might start the spawning season earlier. In a similar way, differences in spawning time has been proposed to explain divergence of the Brook trout (*Salvenilus fontinalis*) population inhabiting the outflow of Mistassini Lake as water cooling down is prolonged before the Fall spawning season (Fraser et al., 2004). Conversely, Mistassini Lake walleye populations have likely diverged based on a combination of both their spawning rivers and habitat choice during the summer growing season in multiple instances. Notable examples include the southern populations where ICOPER spawned in two different rivers and stayed in close proximity while CHA, also a southern population, was characterized by its wide distribution across the lake. In a similar way, the northern population TAK was found everywhere across the lake, but Cree knowledge holders reported that this population spawned later (early June) compared to the southern ones (mid-May; Bowles et al., 2022).

The complex distribution of walleye populations was reflected by migration distances. During summer, walleye populations either travelled large distances of hundreds of kilometres (TAK, CHA, TEM), average distances (MAU, TEM) or small distances (PIP, ICOPER). MAU, ICOPER and PIP populations might have

stayed near their spawning grounds because their spawning rivers entered bays and littoral areas where suitable walleye habitats occur. In comparison, the TAK spawning river in Mistassini Lake connects almost directly to deep and likely cooler basins (>20m); conditions that are usually not preferred by walleye and might encourage migration (Matley et al., 2020). In a similar way, MAU fish would encounter deep basins near the outflow entrance and probably faced stronger currents toward Mistassini Lake limiting their presence outside of the outflow despite having average migration distances (Figure 3). Dupont et al., (2007) results using microsatellites also suggested that walleye from the Mistassini Lake outflow were more isolated; however, they found an indication of longer migrations for ICOPER. Indeed, the ICOPER population accounted for approximatively 10% of the harvest above the 51st parallel while in this study, only one ICOPER fish was reported in the same area. Although this could be due to the increased resolution of the SNP panel compared to microsatellites (84% assignment accuracy on average), this opens the possibility of a shift in migration patterns in the last 20 years. It is worth noting that this southern population also exhibited recent signs of harvest-induced evolution, which included a reduction in body size (Bowles et al., 2020). Thus, (i) the probability of detecting ICOPER fish would decrease as the distance increases and/or (ii) migration might be too energetically costly if they do not grow to a larger size. The latter might be consistent with the pace of life syndrome where some individuals might trade off faster growth rate, food intake and reproductive output at the cost of energetical expenditure and predation risk (Réale et al., 2010). Following this idea, McKee et al., (2022) found in Lake Superior that migratory walleye achieved a larger body size and grew at a faster rate than resident fish. In any case, we would suggest investigating the impact of harvesting, especially given that this population is located in Mistassini Lake in one the most targeted areas by recreational and Cree subsistence fishers.

Consistent with McKee et al., (2022), not all individuals adopted a long-distance migratory strategy, even in populations migrating further on average (Figure 6). We hypothesized that these differences could be sex-biased even though this remains to be tested in these northern lakes. Indeed, females generally travel

greater distances in search for productive habitats and resources (Matley et al., 2020; McKee et al., 2022; Raby et al., 2018). In contrast, males usually arrive first and leave the spawning grounds last to maximize reproductive opportunities (Bade et al., 2019; Matley et al., 2020; Pritt et al., 2013). Alternatively, walleye may even skip reproduction to spend more time growing. They might experience slower growth in these cold lakes which are on the lower end of the walleye thermal optimum (Bozek et al., 2011). Supporting this idea, migration distances and time of capture were negatively linked for some populations, suggesting that a large fraction of the fish were far from their source river near spawning time (Figure 6B). It was the case for both Mistasiniishish Lake populations but any conclusion regarding this result should be taken with caution due to uneven sampling. Overall, walleye population differences in migratory strategies might reflect a combination of individual life history traits, population idiosyncrasies (locally adaptations) and habitat heterogeneity.

Assignment accuracy and unsampled populations

RUBIAS simulations on the source population dataset using different priors were consistently above 98% accuracy. However, this accuracy also relies on whether all sources were sampled. Mistassini and Mistasiniishish lakes are both large lakes which have hundreds of rivers with many areas along them that could be suitable for walleye spawning. Indeed, unsampled source populations were identified in both lakes because they were in remote locations, too close to known spawning sites or too small to be easily detected.

For example, the Temiscamie river (TEM), which is Mistasiniishish Lake's main contributor to the mixed-stock harvest, was not directly sampled. Only MET (*i.e.,* the Metawashish river) was accessible and generally well known to harbour walleye spawning at the time of sample collection. TEM 'source' samples were instead identified among Mistasiniishish Lake mixed-stock. Indeed, multiple evidence supported the existence of two genetic clusters (PCA, Figure 1D; *STRUCTURE*, Figure 2 and 4). The genomic structure of

walleye in Mistasiniishish Lake was further supported by a parallel study conducted on the same samples using higher number of markers through low-coverage Whole Genome Sequence (Michaelides et al., in prep). Even though Temiscamie river would be an ideal candidate (large river, suitable habitat) for walleye to spawn in Mistasiniishish Lake, the geographic origin of the main contributor to the mixed-stock harvest in this lake remains to be elucidated.

In contrast, Mistassini Lake mixed-stock population structure (Figure 1C) was not as straightforward as the one in Mistasiniishish Lake (Figure 1D) and thus, identifying unknown sources proved to be more challenging. As expected, *RUBIAS* z-scores computed from the assignment loglikelihood showed that all sources were likely identified in Mistasiniishish Lake while it was not the case in Mistassini Lake (Figure 5). Seventy of 590 mixed-stock samples had z-scores with more than two standard deviations from the normal distribution (despite PofZ's higher than 0.8) indicating that they were misassigned. In addition, more than one quarter of them were also admixed individuals (0.2 < q-value <0.8). Given that other metrics (such as the PofZ) would otherwise suggest that these samples were confidently assigned to a known source, examining the q-values from *STRUCTURE* should also be considered (Marín-Nahuelpi et al. 2022). Interestingly, 60% (42/70) of these potentially misassigned samples belonged to one of the three unknown clusters identified in *STRUCTURE* runs between K7 and K9. Further, these clusters had more extreme z-score values on average (see Figure S3, Appendix II). Together, these observations open the question regarding what the true number of clusters K would be.

One possibility would be that these unknown clusters represent fine-scale structure in two regions (*i.e.,* the north and the outflow) where the habitat is more complex favouring locally adapted subpopulations. This is supported for the unknown clusters 2 and 3 that were caught near the known source populations they were most related to (Figure 4). The fish from unknown cluster 2 were caught near an island located next to the mouth of the Takwa river they were originally assigned to (TAK). In a similar way, individuals

belonging to the unknown cluster 3 were all caught in the Rupert River that feeds into the de Maurès River (MAU; also the initial assignment of these fish). In fact, Dupont et al., (2007) sampled the Rupert River for their study and based our results, this indicates the presence of two separate spawning locations in close proximity. Still, we did not necessarily expect differentiation such as between the neighbouring rivers Icon (ICO) and Perch (PER) in the south which were known to form one homogeneous genetic group (Bowles et al., 2020; Dupont et al., 2007).

Lastly, unknown cluster 1 was the most likely candidate for a new, unsampled source. Indeed, the samples of this cluster were located in the western part of the outflow, far from their original assignment (CHA) and remarkably, it was well differentiated with high pairwise Fst (>0.08). It should be noted however that *STRUCTURE* is sensitive to unbalanced sampling (Puechmaille, 2016; Wang, 2017) and the low number of walleye belonging to these unknown clusters could have over- or under-estimated the true population's genetic structure in the larger and more complex Mistassini Lake. Nevertheless, these results would help pinpoint regions important for monitoring walleye populations. For instance, the outflow of Mistassini Lake where only one source population was sampled might be home to up to three populations instead.

Conclusion

We were able to highlight large differences in mixed-stock harvest proportions between the studied lakes with Mistassini Lake having a more complex structure and distribution of its populations across the lake likely driven by habitat heterogeneity. In both lakes, however, one population contributed more to the annual harvest, and both were characterized by migrating long distances. These results contribute to a growing literature highlighting intraspecific variation in migration patterns which might be a driver of local adaptation (Rougemont et al., 2023) as well as emphasize the value of genetic stock identification for inferring the spatial ecology of species that are difficult to track otherwise. By combining population genetic analyses, we were able to identify up to three potentially unsampled populations. Samples

belonging to these populations did not alter the mixed-stock harvest proportions, but their identification drew attention to areas of interest in lake. For instance, two of the potential unsampled population were linked to the outflow known for its unique characteristics and generally harbouring distinct fish populations (Fraser et al., 2004). Further studies should seek to monitor these areas as well as investigate the location of the main source population in Mistasiniishish Lake. Taken together, our results highlight the importance of combining analytical tools and data (*e.g.,* genomics, gps coordinates) for planning future surveys and delineating conservation units.

REFERENCES

Aljanabi, S. (1997). Universal and rapid salt-extraction of high quality genomic DNA for PCR- based techniques. *Nucleic Acids Research*, *25*(22), 4692–4693. https://doi.org/10.1093/nar/25.22.4692

Anderson, E. C., Waples, R. S., & Kalinowski, S. T. (2008). An improved method for predicting the accuracy of genetic stock identification. *Canadian Journal of Fisheries and Aquatic Sciences*, *65*(7), 1475–1486. https://doi.org/10.1139/F08-049

Araujo, H. A., Candy, J. R., Beacham, T. D., White, B., & Wallace, C. (2014). Advantages and Challenges of Genetic Stock Identification in Fish Stocks with Low Genetic Resolution. *Transactions of the American Fisheries Society*, *143*(2), 479–488. https://doi.org/10.1080/00028487.2013.855258

Bade, A. P., Binder, T. R., Faust, M. D., Vandergoot, C. S., Hartman, T. J., Kraus, R. T., Krueger, C. C., & Ludsin, S. A. (2019). Sex-based differences in spawning behavior account for male-biased harvest in Lake Erie walleye (*Sander vitreus*). *Canadian Journal of Fisheries and Aquatic Sciences*, *76*(11), 2003–2012. https://doi.org/10.1139/cjfas-2018-0339

Beacham, T. D., Wallace, C., Jonsen, K., McIntosh, B., Candy, J. R., Rondeau, E. B., Moore, J.-S., Bernatchez, L., & Withler, R. E. (2020). Accurate estimation of conservation unit contribution to coho salmon mixed-stock fisheries in British Columbia, Canada, using direct DNA sequencing for single nucleotide polymorphisms. *Canadian Journal of Fisheries and Aquatic Sciences*, *77*(8), 1302–1315. https://doi.org/10.1139/cjfas-2019-0339

Begg, G. A., & Cadrin, S. X. (2016). Stock Identification. In *Fish Reproductive Biology* (pp. 252–278). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118752739.ch6

Bootsma, M. L., Gruenthal, K. M., McKinney, G. J., Simmons, L., Miller, L., Sass, G. G., & Larson, W. A. (2020). A GT-seq panel for walleye (Sander vitreus) provides important insights for efficient development and implementation of amplicon panels in non-model organisms. *Molecular Ecology Resources*, *20*(6), 1706–1722. https://doi.org/10.1111/1755-0998.13226

Bowles, E., Jeon, H. B., Marin, K., Macleod, P., & Fraser, D. J. (2022). Freshwater fisheries monitoring in northern ecosystems using Indigenous ecological knowledge, genomics, and life history: Insights for community decision-making. *Facets*, *7*(1), 1214–1243. https://doi.org/10.1139/facets-2021-0049

Bowles, E., Marin, K., Mogensen, S., MacLeod, P., & Fraser, D. J. (2020). Size reductions and genomic changes within two generations in wild walleye populations: Associated with harvest? *Evolutionary Applications*, *13*(6), 1128–1144. https://doi.org/10.1111/eva.12987

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Bozek, M. A., Haxton, T. J., & Raabe, J. K. (2011). Walleye and sauger habitat. In *Biology, management and culture of walleye and sauger* (pp. 133–197).

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, *15*(4), 855–867. https://doi.org/10.1111/1755-0998.12357

Des Roches, S., Pendleton, L. H., Shapiro, B., & Palkovacs, E. P. (2021). Conserving intraspecific variation for nature's contributions to people. *Nature Ecology and Evolution*, *5*(5), 574–582. https://doi.org/10.1038/s41559-021-01403-5

Dupont, P. P., Bourret, V., & Bernatchez, L. (2007). Interplay between ecological, behavioural and historical factors in shaping the genetic structure of sympatric walleye populations (Sander vitreus). *Molecular Ecology*, *16*(5), 937–951. https://doi.org/10.1111/j.1365-294X.2006.03205.x

Elhaik, E. (2022). Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports*, *12*(1), 14683. https://doi.org/10.1038/s41598-022-14395-4

Euclide, P. T., Robinson, J., Faust, M., Ludsin, S. A., MacDougall, T. M., Marschall, E. A., Chen, K.-Y., Wilson, C., Bootsma, M., Stott, W., Scribner, K. T., & Larson, W. A. (2021). Using Genomic Data to Guide Walleye Management in the Great Lakes. In *Yellow Perch, Walleye, and Sauger: Aspects of Ecology, Management, and Culture* (pp. 115–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-80678-1_5

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. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics*, *164*(4), 1567–1587. https://doi.org/10.1093/genetics/164.4.1567

Francis, R. M. (2017). pophelper: An R package and web app to analyse and visualize population structure. *Molecular Ecology Resources*, *17*(1), 27–32. https://doi.org/10.1111/1755-0998.12509

Fraser, D. J., & Bernatchez, L. (2005). ADAPTIVE MIGRATORY DIVERGENCE AMONG SYMPATIRIC BROK CHARR POPULATIONS. *Evolution*, *59*(3), 611–624. https://doi.org/10.1111/j.0014-3820.2005.tb01020.x

Fraser, D. J., Lippé, C., & Bernatchez, L. (2004). Consequences of unequal population size, asymmetric gene flow and sex-biased dispersal on population structure in brook charr (Salvelinus fontinalis). *Molecular Ecology*, *13*(1), 67–80. https://doi.org/10.1046/j.1365-294X.2003.02038.x

Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution*, *27*(9), 489–496. https://doi.org/10.1016/j.tree.2012.05.012

Hansen, G. J. A., Ruzich, J. K., Krabbenhoft, C. A., Kundel, H., Mahlum, S., Rounds, C. I., Van Pelt, A. O., Eslinger, L. D., Logsdon, D. E., & Isermann, D. A. (2022). It's Complicated and It Depends: A

Review of the Effects of Ecosystem Changes on Walleye and Yellow Perch Populations in North America. *North American Journal of Fisheries Management*, *42*(3), 484–506. https://doi.org/10.1002/nafm.10741

Hartman, G. F. (2009). *A Biological Synopsis of Walleye (Sander vitreus)*. Fisheries and Oceans Canada, Science Branch, Pacific Region Pacific Biological Station.

Hemstrom, W., & Jones, M. (2023). snpR: User friendly population genomics for SNP data sets with categorical metadata. *Molecular Ecology Resources*, *23*(4), 962–973. https://doi.org/10.1111/1755-0998.13721

Hilborn, R., Quinn, T. P., Schindler, D. E., & Rogers, D. E. (2003). Biocomplexity and fisheries sustainability. *Proceedings of the National Academy of Sciences*, *100*(11), 6564–6568. https://doi.org/10.1073/pnas.1037274100

Hui, C., Fox, G. A., & Gurevitch, J. (2017). Scale-dependent portfolio effects explain growth inflation and volatility reduction in landscape demography. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(47), 12507–12511. https://doi.org/10.1073/pnas.1704213114

Komoroske, L. M., Jensen, M. P., Stewart, K. R., Shamblin, B. M., & Dutton, P. H. (2017). Advances in the Application of Genetics in Marine Turtle Biology and Conservation. *Frontiers in Marine Science*, *4*, 156. https://doi.org/10.3389/fmars.2017.00156

Kuismin, M., Saatoglu, D., Niskanen, A. K., Jensen, H., & Sillanpää, M. J. (2020). Genetic assignment of individuals to source populations using network estimation tools. *Methods in Ecology and Evolution*, *11*(2), 333–344. https://doi.org/10.1111/2041-210X.13323

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (1.8.9) [Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html

Marin, K., Coon, A., & Fraser, D. J. (2017). Traditional ecological knowledge reveals the extent of sympatric lake trout diversity and habitat preferences. *Ecology and Society*, *22*(2). https://doi.org/10.5751/ES-09345-220220

Matley, J. K., Faust, M. D., Raby, G. D., Zhao, Y., Robinson, J., MacDougall, T., Hayden, T. A., Fisk, A. T., Vandergoot, C. S., & Krueger, C. C. (2020). Seasonal habitat-use differences among Lake Erie's walleye stocks. *Journal of Great Lakes Research*, *46*(3), 609–621. https://doi.org/10.1016/j.jglr.2020.03.014

McKee, G., Hornsby, R. L., Fischer, F., Dunlop, E. S., Mackereth, R., Pratt, T. C., & Rennie, M. (2022). Alternative migratory strategies related to life history differences in the Walleye (Sander vitreus). *Movement Ecology*, *10*(1), 10. https://doi.org/10.1186/s40462-022-00308-7

McKinney, G. J., Pascal, C. E., Templin, W. D., Gilk-Baumer, S. E., Dann, T. H., Seeb, L. W., & Seeb, J. E. (2020). Dense SNP panels resolve closely related chinook salmon populations. *Canadian Journal of Fisheries and Aquatic Sciences*, *77*(3), 451–461. https://doi.org/10.1139/cjfas-2019-0067

Meek, M. H., Beever, E. A., Barbosa, S., Fitzpatrick, S. W., Fletcher, N. K., Mittan-Moreau, C. S., Reid, B. N., Campbell-Staton, S. C., Green, N. F., & Hellmann, J. J. (2023). Understanding Local Adaptation to Prepare Populations for Climate Change. *BioScience*, *73*(1), 36–47. https://doi.org/10.1093/biosci/biac101

Michaelides, S., Jeon, H.-B., Marin, K., MacLeod, P., Euclide, P., Kuhl, H., & Fraser, D. J. (in prep). *Ecological, contemporary, and historical factors shaping the genetic structure of sympatric and allopatric walleye (Sander vitreus) populations in Quebec lakes.*

Moran, B. M., & Anderson, E. C. (2019). Bayesian inference from the conditional genetic stock identification model. *Canadian Journal of Fisheries and Aquatic Sciences*, *76*(4), 551–560. https://doi.org/10.1139/cjfas-2018-0016

Navarroli, G., Koumrouyan, R. A., Marin, K., & Fraser, D. J. (2021). *Population status, life history and conservation of Mistassini, Albanel and Waconichi Lake brook trout and walleye populations*.

Piovano, S., Batibasaga, A., Ciriyawa, A., LaCasella, E. L., & Dutton, P. H. (2019). Mixed stock analysis of juvenile green turtles aggregating at two foraging grounds in Fiji reveals major contribution from the American Samoa Management Unit. *Scientific Reports*, *9*(1), 3150. https://doi.org/10.1038/s41598-019-39475-w

Potvin, C., & Bernatchez, L. (2001). Lacustrine spatial distribution of landlocked Atlantic salmon populations assessed across generations by multilocus individual assignment and mixed-stock analyses. *Molecular Ecology*, *10*(10), 2375–2388. https://doi.org/10.1046/j.0962-1083.2001.01374.x

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). *Inference of Population Structure Using Multilocus Genotype Data*. http://www.stats.ox.ac.uk/pritch/home.html.

Pritt, J. J., DuFour, M. R., Mayer, C. M., Kocovsky, P. M., Tyson, J. T., Weimer, E. J., & Vandergoot, C. S. (2013). Including independent estimates and uncertainty to quantify total abundance of fish migrating in a large river system: Walleye (*Sander vitreus*) in the Maumee River, Ohio. *Canadian Journal of Fisheries and Aquatic Sciences*, *70*(5), 803–814. https://doi.org/10.1139/cjfas-2012-0484

Puechmaille, S. J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*(3), 608–627. https://doi.org/10.1111/1755-0998.12512

Raby, G. D., Vandergoot, C. S., Hayden, T. A., Faust, M. D., Kraus, R. T., Dettmers, J. M., Cooke, S. J., Zhao, Y., Fisk, A. T., & Krueger, C. C. (2018). Does behavioural thermoregulation underlie seasonal movements in Lake Erie walleye? *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(3), 488–496. https://doi.org/10.1139/cjfas-2017-0145

Réale, D., Garant, D., Humphries, M. M., Bergeron, P., Careau, V., & Montiglio, P.-O. (2010). Personality and the emergence of the pace-of-life syndrome concept at the population level. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1560), 4051–4063. https://doi.org/10.1098/rstb.2010.0208

Rougemont, Q., Xuereb, A., Dallaire, X., Moore, J., Normandeau, E., Perreault-Payette, A., Bougas, B., Rondeau, E. B., Withler, R. E., Van Doornik, D. M., Crane, P. A., Naish, K. A., Garza, J. C., Beacham, T. D., Koop, B. F., & Bernatchez, L. (2023). Long-distance migration is a major factor driving local adaptation at continental scale in Coho salmon. *Molecular Ecology*, *32*(3), 542–559. https://doi.org/10.1111/mec.16339

Rousset, F. (2008). genepop'007: A complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, *8*(1), 103–106. https://doi.org/10.1111/j.1471-8286.2007.01931.x

Schindler, D. E., Armstrong, J. B., & Reed, T. E. (2015). The portfolio concept in ecology and evolution. *Frontiers in Ecology and the Environment*, *13*(5), 257–263. https://doi.org/10.1890/140275

Schindler, D. E., Hilborn, R., Chasco, B., Boatright, C. P., Quinn, T. P., Rogers, L. A., & Webster, M. S. (2010). Population diversity and the portfolio effect in an exploited species. *Nature*, *465*(7298), 609–612. https://doi.org/10.1038/nature09060

Seeb, L. W., Antonovich, A., Banks, M. A., Beacham, T. D., Bellinger, M. R., Blankenship, S. M., Campbell, M. R., Decovich, N. A., Garza, J. C., Guthrie, C. M., Lundrigan, T. A., Moran, P., Narum, S. R., Stephenson, J. J., Supernault, K. J., Teel, D. J., Templin, W. D., Wenburg, J. K., Young, S. F., & Smith, C. T. (2007). Development of a Standardized DNA Database for Chinook Salmon. *Fisheries*, *32*(11), 540–552. https://doi.org/10.1577/1548-8446(2007)32[540:doasdd]2.0.co;2

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating Individual Admixture Proportions from Next Generation Sequencing Data. *Genetics*, *195*(3), 693–702. https://doi.org/10.1534/genetics.113.154138

Smouse, P. E., Waples, R. S., & Tworek, J. A. (1990). A Genetic Mixture Analysis for use with Incomplete Source Population Data. *Canadian Journal of Fisheries and Aquatic Sciences*, *47*(3), 620–634. https://doi.org/10.1139/f90-070

Stepien, C. A., Murphy, D. J., Lohner, R. N., Haponski, A. E., & Sepulveda-Villet, O. J. (2010). Status and delineation of walleye (Sander vitreus) genetic stock structure across the Great Lakes. In E. Roseman, P. Kocovsky, & C. Vandergoot (Eds.), *Status of walleye in the Great Lakes: Proceedings of the 2006 symposium: Vol. Technical Report 69* (pp. 189–223). MI: Great Lakes Fishery Commission.

Stepien, C. A., Murphy, D. J., Lohner, R. N., Sepulveda-Villet, O. J., & Haponski, A. E. (2009). Signatures of vicariance, postglacial dispersal and spawning philopatry: Population genetics of the walleye Sander vitreus. *Molecular Ecology*, *18*(16), 3411–3428. https://doi.org/10.1111/j.1365-294X.2009.04291.x

Stepien, C. A., Snyder, M. R., & Knight, C. T. (2018). Genetic Divergence of Nearby Walleye Spawning Groups in Central Lake Erie: Implications for Management. *North American Journal of Fisheries Management*, *38*(4), 783–793. https://doi.org/10.1002/nafm.10176

Tamario, C., Sunde, J., Petersson, E., Tibblin, P., & Forsman, A. (2019). Ecological and Evolutionary Consequences of Environmental Change and Management Actions for Migrating Fish. *Frontiers in Ecology and Evolution*, *7*. https://doi.org/10.3389/fevo.2019.00271

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fouth). Springer. https://www.stats.ox.ac.uk/pub/MASS4/

Waldman, J., Wilson, K. A., Mather, M., & Snyder, N. P. (2016). Une approche résiliente peut améliorer le rétablissement des populations de poissons anadromes. *Fisheries*, *41*(3), 116–126. https://doi.org/10.1080/03632415.2015.1134501

Wang, J. (2017). The computer program structure for assigning individuals to populations: Easy to use but easier to misuse. *Molecular Ecology Resources*, *17*(5), 981–990. https://doi.org/10.1111/1755-0998.12650

Willi, Y., Van Buskirk, J., & Hoffmann, A. A. (2006). Limits to the adaptive potential of small populations. *Annual Review of Ecology, Evolution, and Systematics*, *37*, 433–458. https://doi.org/10.1146/annurev.ecolsys.37.091305.110145

Wilson, C. C., Lavender, M., & Black, J. (2007). Genetic assessment of walleye (Sander vitreus) restoration efforts and options in Nipigon Bay and Black Bay, Lake Superior. *Journal of Great Lakes Research*, *33*(SUPPL. 1), 133–144. https://doi.org/10.3394/0380-1330(2007)33[133:GAOWSV]2.0.CO;2

TABLES AND FIGURES

Table 1. Number of samples collected in the rivers (*i.e.,* sources) and the lakes (*i.e.,* mixed-stock harvest). Coordinates were recorded for each mixed-stock sample harvested in the lakes. The analyses included individual migration distances. They were only estimated for mixed-stock samples which had precise GPS coordinates (*i.e.,* exact coordinates or specific areas such as islands, bays or passes where the sample was caught). The second source in Mistasiniishish Lake (TEM) was identified previously (Michaelides et al., in prep) and was named "Temiscamie" (TEM) as hints suggested this river to be a suitable and likely spawning ground for the walleye from Mistasiniishish Lake.

Lake	Type	Location (river/lake)	Abbr.	Year	Samples available (n)			
					Total	after filtering	with GPS coord.	with precise GPS coord.
Mistassini Lake	Source	Icon	ICO	2016	**40**	36	-	-
	Source	Perch	PER	2016	**40**	26	-	-
	Source	Pipounichouane	PIP	2022	**22**	21	-	-
	Source	De Maurès	MAU	2020/22	**42**	38	-	-
	Source	Chalifour	CHA	2017	**40**	30	-	-
	Source	Takwa	TAK	2020	**41**	36	-	-
	Mixed-Stock	**Mistassini**	-	2020-22	**638**	590	569	547
Mistasinii-shish Lake	Source	Metawashish	MET	2022	**42**	40	-	-
	Source	**Mistasiniishish (likely Tem.)**	TEM*	2021	**35**	35	-	-
	Mixed-Stock	**Mistasiniishish**	-	2020-21	**641**	613	604	313

Table 2. Population differences in migration distances during summer months. Predictions from estimated marginal means computed on the best model explaining migration distances. The model included an interaction between the population ID and the timing of capture (measured as the week). Because walleye can spawn up to the end of May – beginning of June, the data before week 24 (first week of June) was not included in the model. In addition, PIP samples were removed from the model as only four of them were assigned this population. Model comparisons and summary of the best model are available in Tables S3 and S4 in Appendix II.

Result	Estimate	SE	Df	t ratio	p value
ICOPER < MAU	-1.344	0.370	455	-3.636	**0.004**
ICOPER < CHA	-3.897	0.514	455	-7.586	**<0.001**
ICOPER < TAK	-5.121	0.420	455	-12.178	**<0.001**
ICOPER < MET	-1.046	0.600	455	-1.745	0.503
ICOPER < TEM	-2.410	0.335	455	-7.203	**<0.001**
MAU < CHA	-2.552	0.495	455	-5.152	**<0.001**
MAU < TAK	-3.776	0.398	455	-9.490	**<0.001**
MAU = MET	0.298	0.584	455	0.510	0.996
MAU < TEM	-1.066	0.306	455	-3.485	**0.007**
CHA = TAK	-1.224	0.534	455	-2.291	0.200
CHA > MET	2.850	0.684	455	4.165	**0.001**
CHA > TEM	1.486	0.470	455	3.165	**0.020**
TAK > MET	4.074	0.617	455	6.599	**<0.001**
TAK > TEM	2.710	0.365	455	7.416	**<0.001**
MET = TEM	-1.364	0.562	455	-2.425	0.150

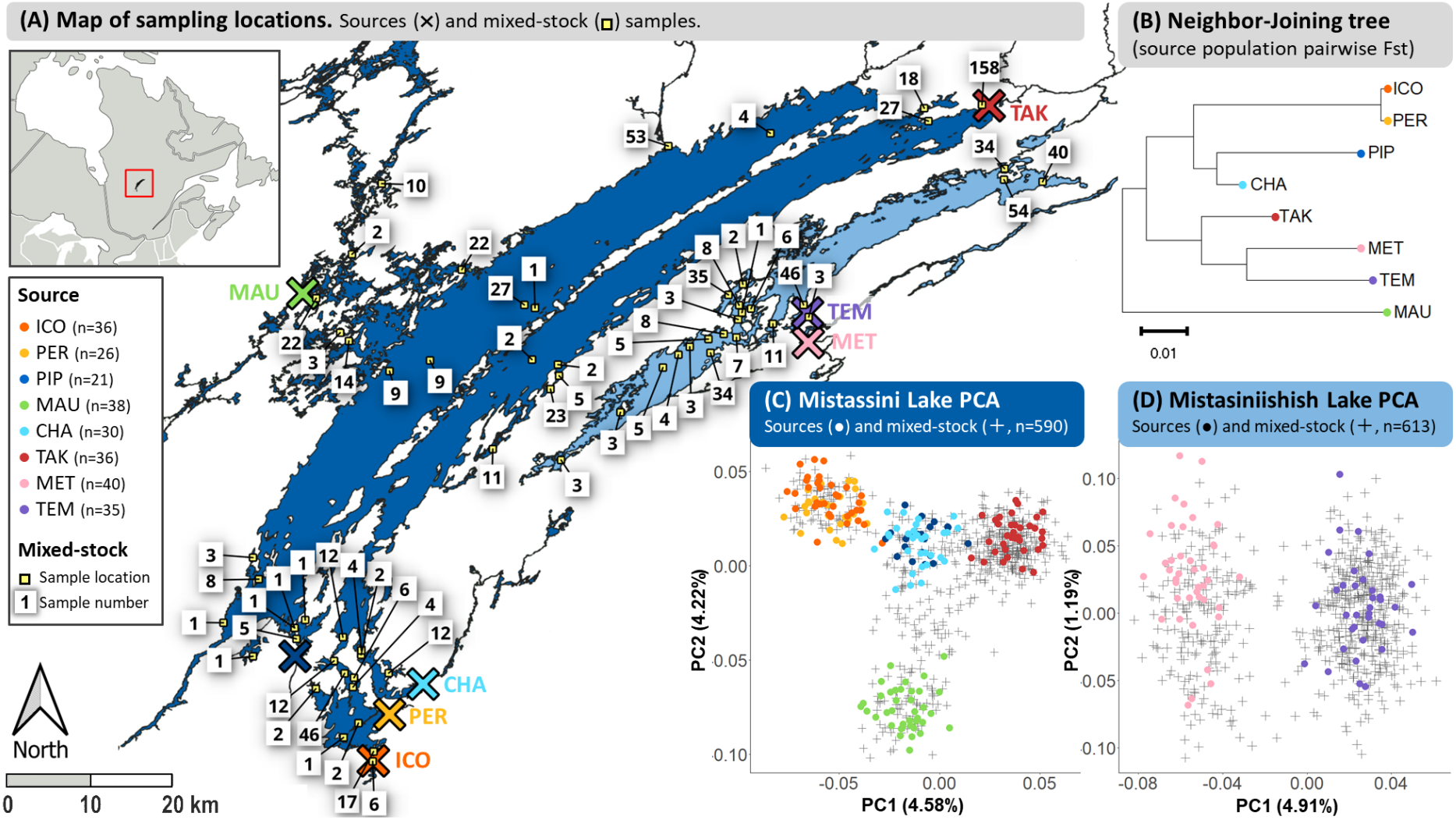

Figure 1. Locations (A), Fst (B) and PCA of the samples (C, D) collected in Mistassini (dark blue) and Mistasiniishish (light blue) lakes. (A) Walleye spawning rivers (source) markers (coloured crosses) are placed on the river mouth. When needed, mixed-stock sample markers (yellow squares) are grouped to avoid overlaps. **(B)** The neighbor-joining tree was based on the source population pairwise Fst estimated with the genepop method and 9,999 bootstraps. All populations significantly differed from each other (P<001) except for ICO and PER (P=0.150). Fst values are available in **Table S1** (**Appendix II**). **(C, D)** Regarding the PCA, Bernoulli was used as the interpolation method for missing values (Interpolated via binomial draw for each allele against minor allele frequency).

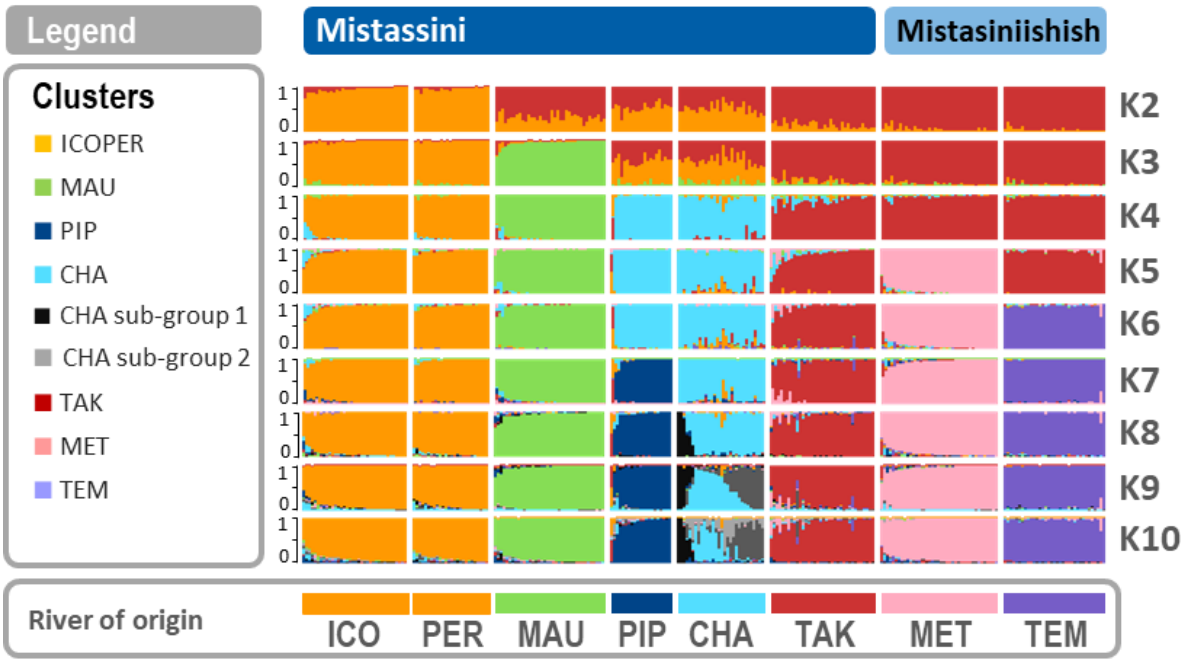

Figure 2. Pattern of clustering of source samples. The samples are ordered according to their river of origin starting first with the samples from the 6 rivers flowing into Mistassini Lake and then with the samples for the rivers flowing into Mistasiniishish Lake. The y axis shows the proportion of membership to each cluster (Q values) estimated by *STRUCTURE*. Runs with the highest log-likelihood were selected to be presented above.

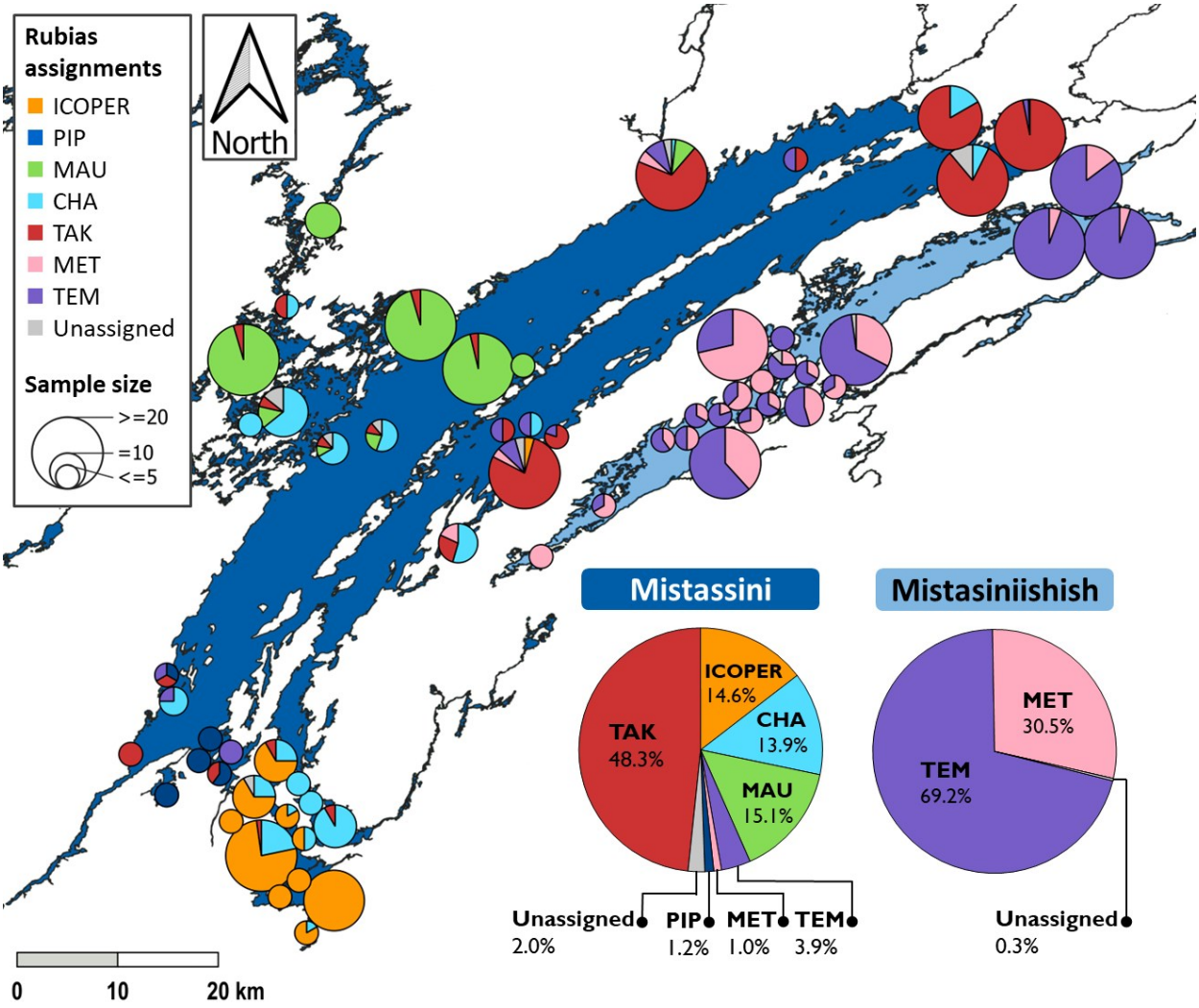

Figure 3. Proportion of Walleye mixed-stock assigned to the source populations in Mistassini (dark blue) and Mistasiniishish (light blue) lakes. Samples without coordinates (Mistassini, n=21 and Mistasiniishish, n=9) are included in the pie charts summarizing the assignments by lake. Few samples in Mistassini Lake (n=12) and 2 (n=2) could not be assigned confidently (PofZ<0.8) and are grouped under the unassigned category (grey color).

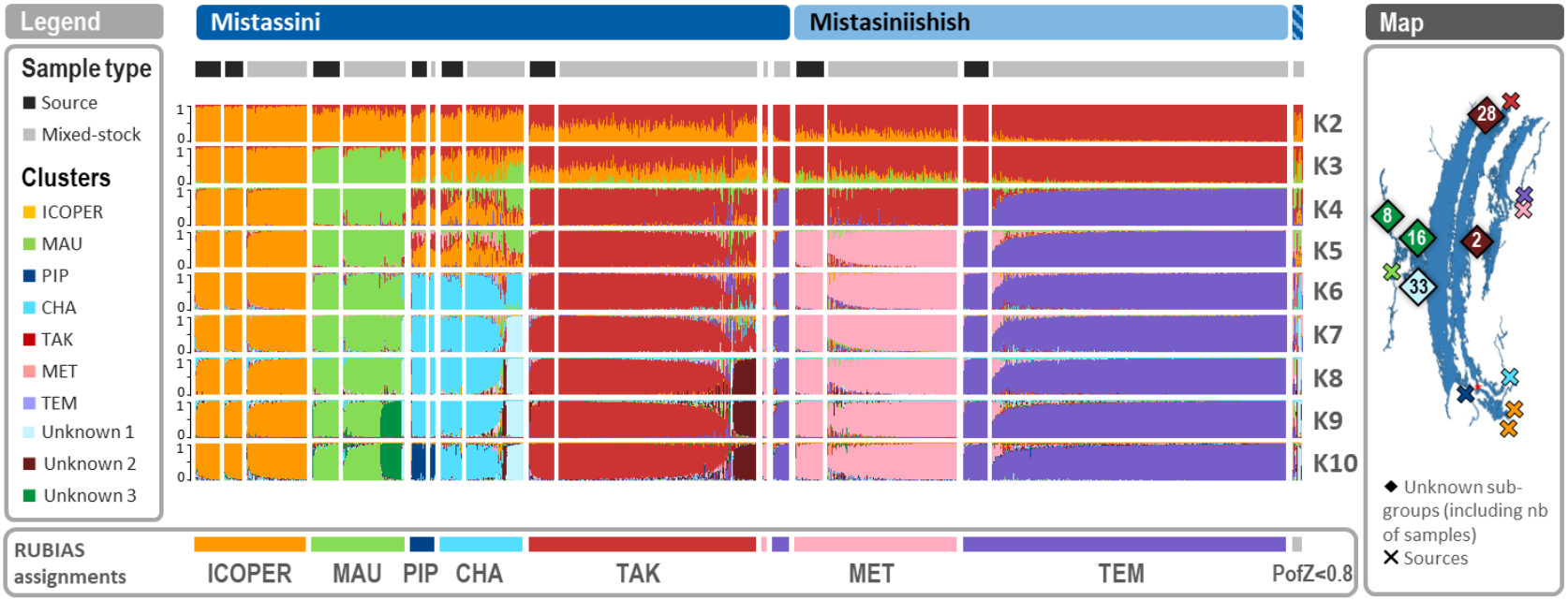

Figure 4. Pattern of clustering of source and mixed-stock samples ordered according to *RUBIAS* assignments. The samples are presented first by lake and second according to their river of origin (for sources) or their *RUBIAS* assignments (for mixed-stock). Samples with low *RUBIAS* PofZ (<0.8) are shown last with both lakes combined. The y-axis shows the proportion of membership to each cluster (Q values) estimated by *STRUCTURE*. Runs with the highest log-likelihood were selected to be presented above. Starting at K7, *STRUCTURE* identified unknown clusters that are not associated with any source population. The location of the samples belonging to these unknown clusters (Q value >= 0.9; diamond shapes) are marked on the map as well as the location of the source population sampled in this study.

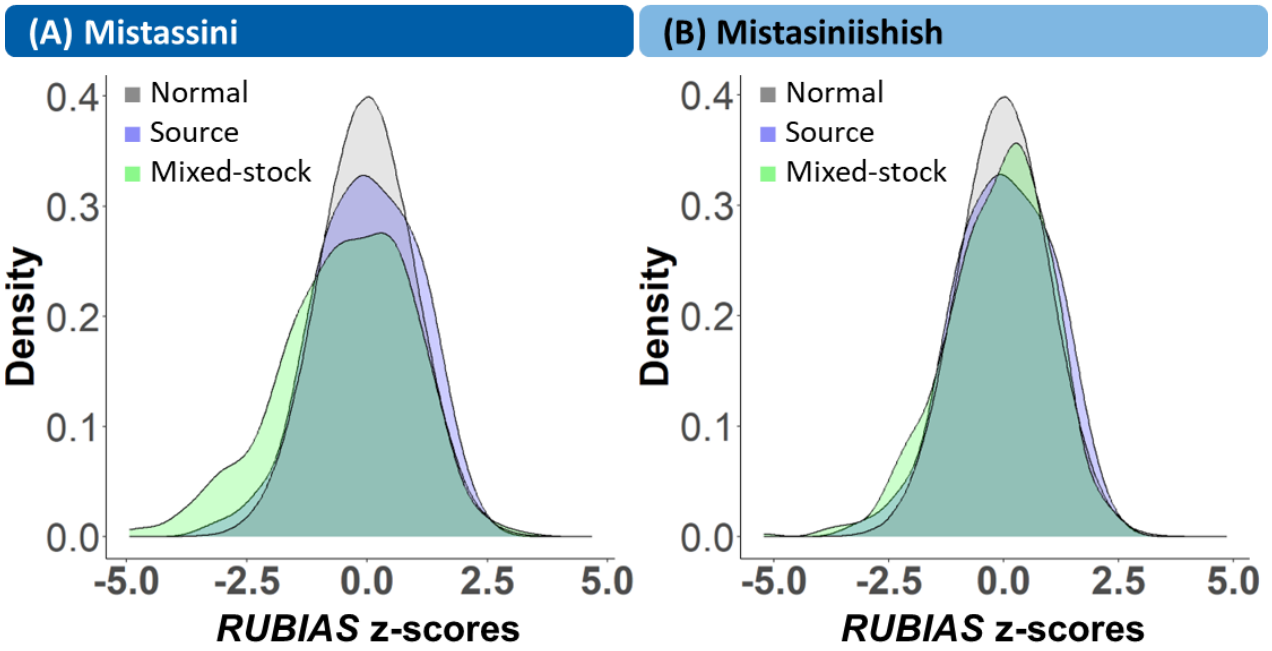

Figure 5. Z-scores distribution of source and mixed-stock samples in Mistassini (A) and Mistasiniishish (B) lakes. The number of mixed-stock samples that do not overlap with the normal distribution are indicated on top.

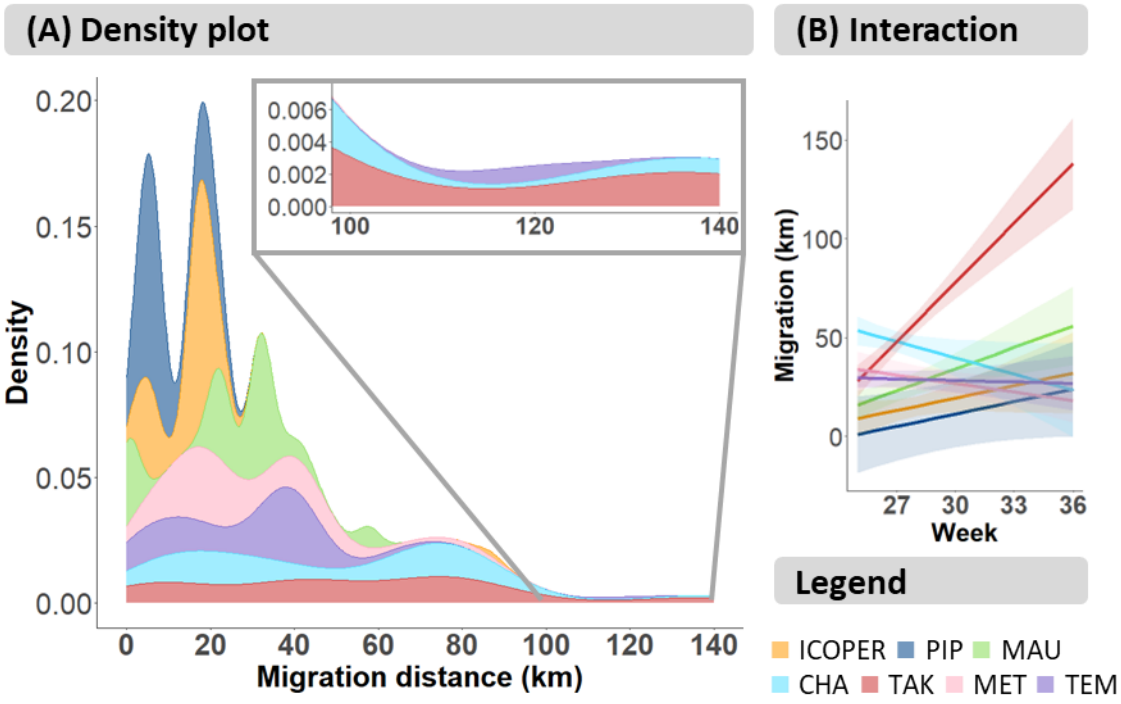

Figure 6. Walleye migration distance measured as the shortest waterway distance between the river of origin (*i.e.,* the source). (A) Migration distances (km) represented as the density for each population and, **(B)** the interaction between the population and the timing of capture (here in weeks). Because walleye spawn around the end of May up to early June, the data shown above includes the second week of June (week>24) to ensure walleye had left their spawning grounds.

APPENDIX I – Genotyping-in-Thousands by sequencing (GT-seq) panel development

Genome data

Individual-based genotyping by sequencing (GBS) data from a previous study on Mistassini walleye was used to develop the panel **(Bowles et al. 2020)**. The data included samples collected in 4 of the main rivers in which walleye come back to spawn in Mistassini Lake: Icon (ICO), Perch (PER) and Chalifour (CHA) in the south, and Takwa (TAK) in the north.

Marker selection

A total of 1000 candidate markers (*i.e.,* Single Nucleotide Polymorphisms (SNPs)) were identified using the GBS data. Marker selection was performed following the stack analyses pipeline published by Eric Normandeau on GitHub **(Normandeau 2021)**. We retained candidate markers with the highest Fst values between two closely related populations *i.e.,* CHA and ICO/PER (combined as one unit since they were genetically identical). We then confirmed that these markers could discriminate between walleye populations (Figure S1). Furthermore, the reads were aligned to the yellow perch reference genome **(INRA Fish Physiology and Genomics laboratory 2019)** so putative adaptive SNPs could be identified with ***PCadapt*** and removed. Additionally, the SNPs needed to be biallelic without insertion or deletion.

Primer design and validation

The GTseek company designed the primers for the GT-seq panel **(Campbell et al. 2015)**. They aimed for 20 bp long primers and a product size of 50 to 120 bp. The targeted SNPs also needed to be within the first 75 bp. The sequences flanking the SNPs was identified using the high quality and well annotated genome from the yellow perch used in the marker selection step. Then, the sequences were mapped to a non-annotated walleye genome **(Auburn University 2019)** so that flanking primers would be based on walleye sequence data.

As a result, 505 loci could be accurately mapped to the walleye genome and thus could be used in the next steps. Among these 505 loci, 417 passed all the filters for primer design from the GTseek company. After testing, 53 loci were removed due to high level of off-target reads leaving 364 loci. Lastly, the performance of the panel was tested using a test plate (n=96 samples). The average genotyping call rate was 91.4% for these samples with an average on-target rate of 61%.

FIGURES

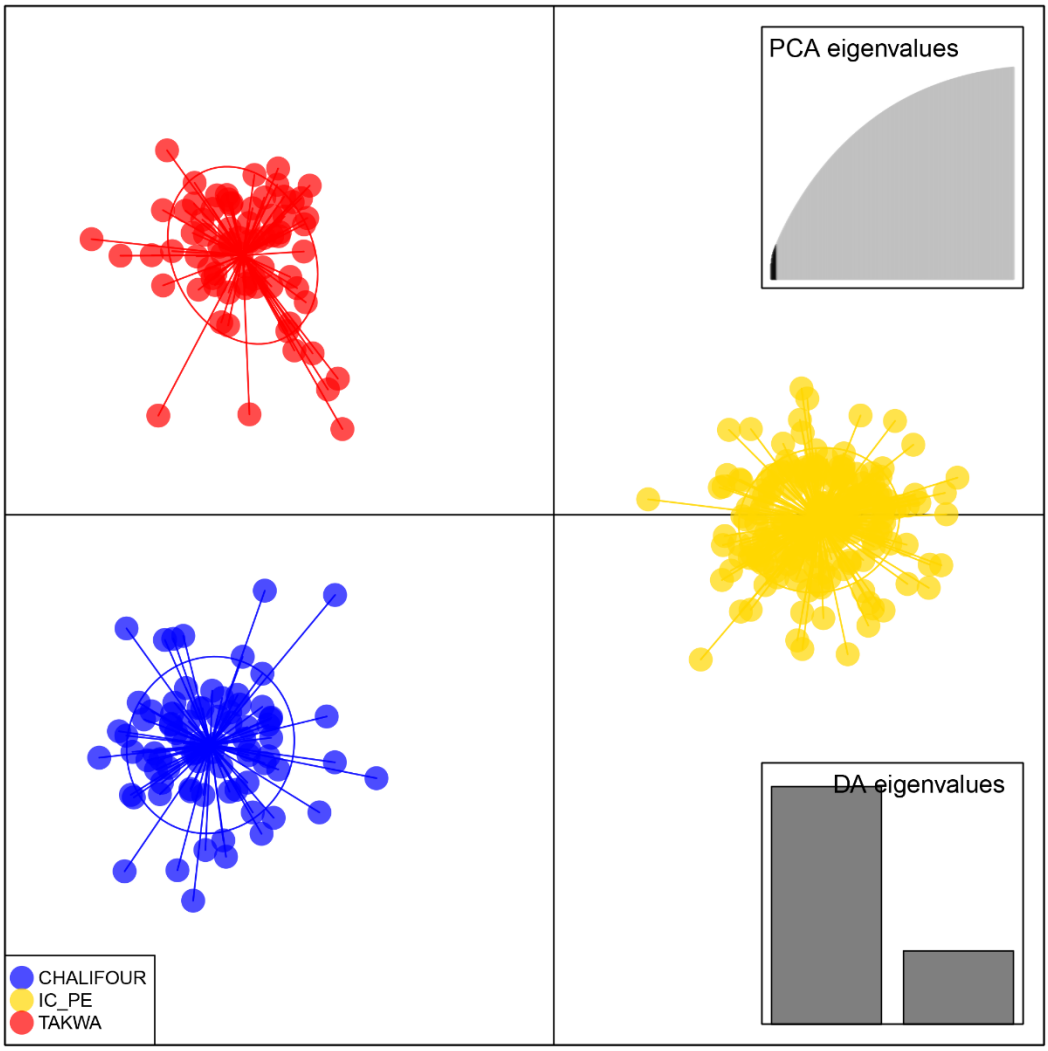

Figure S1. DAPC (Discriminant Analysis of Principal Components) using the 1000 candidate markers selected for GT-seq panel development.

REFERENCES

Auburn University. (2019). Sander Vitreus ASM919308v1 assembly (GenBank accession GCA_009193085.1) [dataset]. National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_009193085.1/

Bowles, E., Marin, K., Mogensen, S., MacLeod, P., & Fraser, D. J. (2020). Size reductions and genomic changes within two generations in wild walleye populations: Associated with harvest? Evolutionary Applications, 13(6), 1128–1144. https://doi.org/10.1111/eva.12987

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. Molecular Ecology Resources, 15(4), 855–867. https://doi.org/10.1111/1755-0998.12357

INRA Fish Physiology and Genomics laboratory. (2019). Perca flavescens PFLA_1.0 assembly (GenBank accession GCF_004354835.1) [dataset]. National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_004354835.1/

Normandeau, E. (2021). RADseq workflow built around STACKS (v2.5.4) [Computer software]. https://github.com/enormandeau/stacks_workflow

APPENDIX II – Supplementary tables and figures

Table S1. Pairwise Fst values estimated with the genepop method. The Fst were calculated with 9999 replicates using the snpR function in R. All pairwise Fst were significant (P<0.001) except for ICO/PER (P=0.150).

	ICO	PER	PIP	MAU	CHA	TAK	MET	TEM	Unk1	Unk2
PER	0.003									
PIP	**0.082**	**0.085**								
MAU	**0.120**	**0.114**	**0.120**							
CHA	**0.058**	**0.058**	**0.040**	**0.092**						
TAK	**0.087**	**0.086**	**0.076**	**0.101**	**0.049**					
MET	**0.112**	**0.117**	**0.098**	**0.109**	**0.069**	**0.058**				
TEM	**0.120**	**0.118**	**0.101**	**0.118**	**0.072**	**0.052**	**0.056**			
Unk1	**0.120**	**0.109**	**0.108**	**0.095**	**0.081**	**0.095**	**0.108**	**0.117**		
Unk2	**0.086**	**0.089**	**0.092**	**0.104**	**0.059**	**0.042**	**0.082**	**0.074**	**0.107**	
Unk3	**0.144**	**0.136**	**0.142**	**0.058**	**0.110**	**0.120**	**0.119**	**0.131**	**0.113**	**0.118**

Table S2. Comparison of z-score average values between *STRUCTURE* clusters at K10. P values were estimated with pairwise t-tests with a Bonferroni correction. Differences in average z-scores are presented in Figure S3. Overall, admixed individuals, unknown clusters 1, 2 and 3 had lower z-scores.

	Admix	Unk1	Unk2	Unk2	ICOPER	MAU	PIP	CHA	TAK	MET
Unk1	0.075									
Unk2	**<0.001**	1								
Unk3	**0.029**	1	1							
ICOPER	**<0.001**	**<0.001**	**<0.001**	**<0.001**						
MAU	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**					
PIP	1	1	0.327	0.972	1	0.213				
CHA	**<0.001**	**<0.001**	**<0.001**	**<0.001**	1	1	1			
TAK	**<0.001**	**<0.001**	**<0.001**	**<0.001**	1	**0.018**	1	1		
MET	**0.001**	**<0.001**	**<0.001**	**<0.001**	1	**<0.001**	1	**0.009**	**<0.001**	
TEM	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.030**	0.314	1	1	1	**<0.001**

Table S3. Linear model explaining migration distance (km). The covariates included in the models were population ID (pop) and time of capture measured as the week number (week). The best model was retained based of the residual sum of square (RSS). In addition, the models were compared using the *anova* function in R.

Model	Covariates	Residual Df	RSS	Df	F	P value
m1	Pop	461	1043.16			
m2	Pop + week	460	971.46	1	44.158	<0.001
m3	Pop*week	455	738,79	5	28.658	<0.001

Table S4. Summary of the best model. Overall, the model was highly significant (Residual standard error= 1.274, R^2= 0.381, Adjusted R^2=0.366, $F_{11,455}$=25.440, P<0.001). The population ($F_{5,455}$=18.472, p<0.001), the week ($F_{1,455}$=44.158, p<0.001) and the interaction ($F_{5,455}$=28.658, p<0.001) were all significant (these global F and P values were calculated with the *anova* function in R).

Terms	Estimate	Std. Error	t value	p value
Intercept	-10.364	3.383	-3.064	0.002
Pop MAU	-6.489	4.335	-1.497	0.135
Pop CHA	19.812	6.025	3.288	0.001
Pop TAK	-56.709	8.218	-6.900	**<0.001**
Pop MET	13.559	4.543	2.985	0.003
Pop TEM	17.809	3.818	4.665	**<0.001**
Week	0.526	0.117	4.510	**<0.001**
Pop MAU*week	0.284	0.152	1.862	0.063
Pop CHA*week	-0.577	0.218	-2.642	0.009
Pop TAK*week	2.241	0.302	7.413	**<0.001**
Pop MET*week	-0.454	0.150	-3.021	0.003
Pop TEM*week	-0.558	0.134	-4.165	**<0.001**

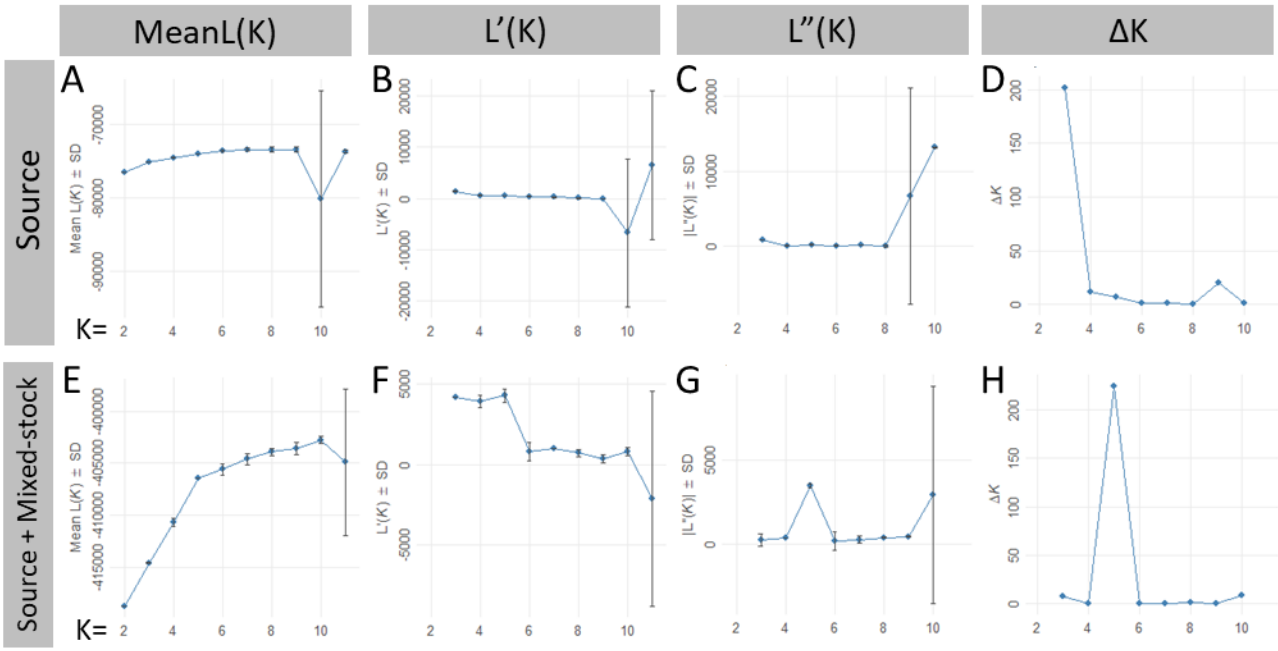

Figure S1 Evanno plots for the *STRUCTURE* runs that included the source samples (A-D) or all samples, *i.e.,* sources and mixed-stock (E-H). The log probability of each run (A, E) and their derivatives (B, C, F, G) are used to estimate ΔK (D, H). The highest ΔK value indicates the best K according to the Evanno method. These plots were created with the *pophelper* R package (Francis, 2017).

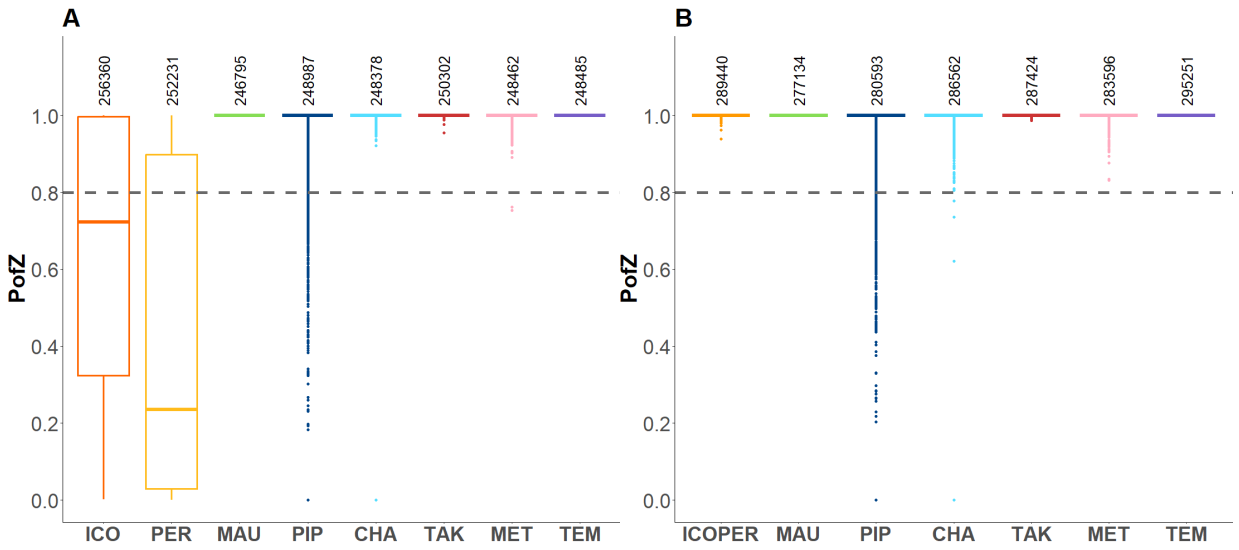

Figure S2. Assignment accuracy using the leave-one-out method in *RUBIAS*. The assignment accuracy to each of the 8 source populations from both lakes was tested either **(A)** with Icon and Perch separated or **(B)** combined as one unit. To note, when **(A)** ICO and PER were on their own, more simulations fell under the threshold of 80% accuracy (PofZ<0.80: ICO=53.2%, PER=62.4%, PIP=5.4%, CHA=3.2%) than when **(B)** both ICO and PER were merged (PofZ<0.80: PIP=5.5%, CHA=3.2%). The parameters used in the simulations were 1000 replicates and 2000 simulated individuals. The number of simulated individuals assigned to each population is indicated on top.

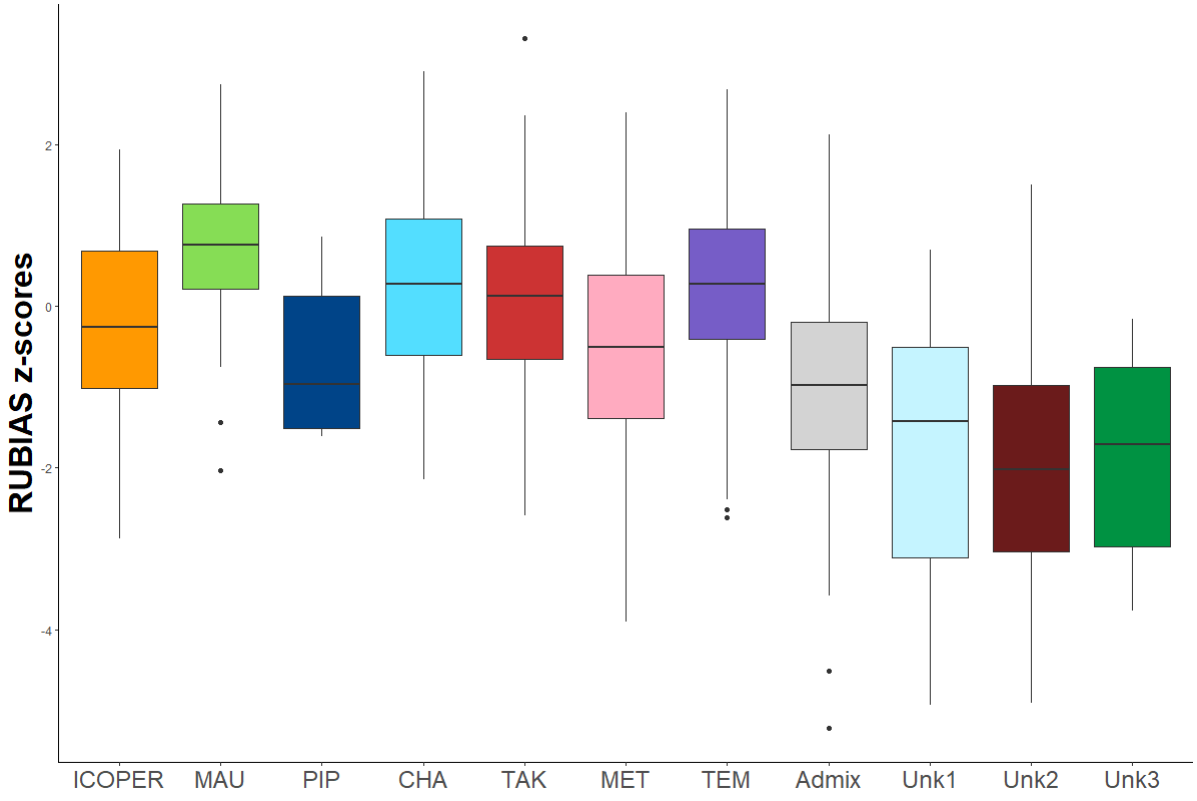

Figure S3. Relationship between *STRUCTURE* assignments (K10) and *RUBIAS* z-scores. Only individuals with P of Z above 0.8 in *RUBIAS* are shown above. All individuals with *STRUCTURE* Q values below 0.8 are included in the admix group. In addition, K10 was chosen because it is the first K where the sampled PIP population forms its own cluster. Fish belonging to the admix group or the unknown clusters (Unk1, Unk2, Unk3) all have lower *RUBIAS* z-scores on average.

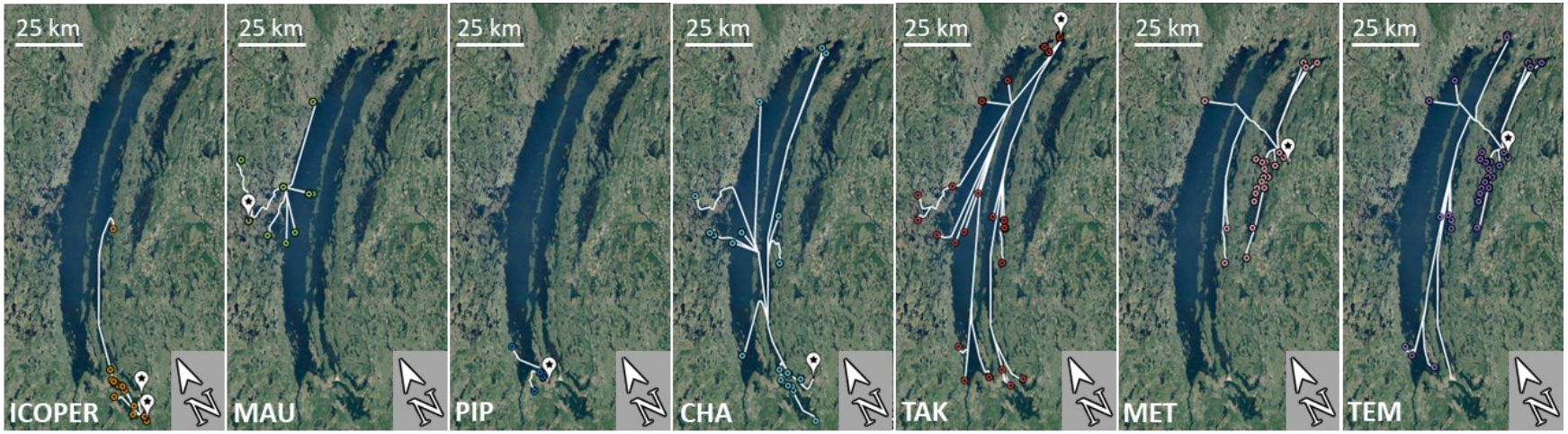

Figure S4. Walleye migration paths based on their population of origin. Migration paths were traced to ensure the shortest waterway distance between the river of origin (source) and the point of capture. Only samples with a P of Z above 0.8 (*i.e.,* confident assignments) were included at this step.

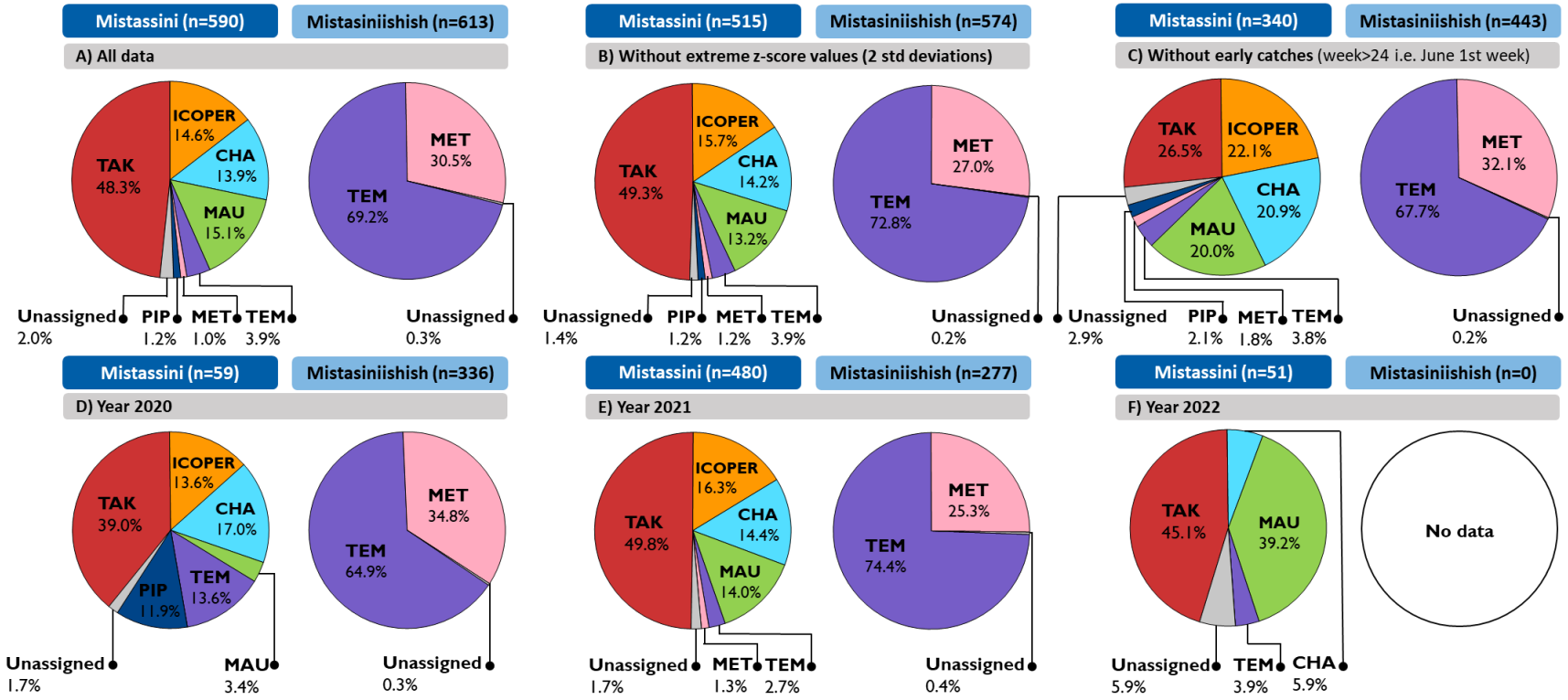

Figure S5. Mixed-stock assignment proportions in both lakes (A) with all the data, (B) without extreme z-scores values, (C) without the early season catches and (D-F) across each sampling year.