CONAWAPA
GENERATION PROJECT

Report # 15-05
December 2015

Population Genomics of Lake Sturgeon (*Acipenser fulvescens*) in the Churchill, Hayes, and Nelson Rivers

Environmental Studies Program

CONAWAPA

CONAWAPA PROJECT

Environmental Studies Program
Report # 15-05

POPULATION GENOMICS OF LAKE STURGEON (*ACIPENSER FULVESCENS)* IN THE CHURCHILL, HAYES, AND NELSON RIVERS

Draft Report Prepared for Manitoba Hydro

By

T. Gosselin[1], P.A. Nelson, C.A. McDougall and L. Bernatchez[1]

December 2015

[1]IBIS, Département de Biologie Université Laval, Québec, QC G1V 0A6
http://www.bio.ulaval.ca/louisbernatchez/presentation.htm

North/South Consultants Inc.
Aquatic Environment Specialists

83 Scurfield Blvd.
Winnipeg, Manitoba, R3Y 1G4
Website: www.nscons.ca

Tel.: (204) 284-3366
Fax: (204) 477-4173
E-mail: nscons@nscons.ca

Non-Technical Summary

Manitoba Hydro has been investigating the feasibility of developing a hydroelectric generating station at the Conawapa site, located approximately 29 km downstream of the Limestone Generating Station (GS) on the Nelson River. An Environmental Studies Program was developed to provide information that will:

- assist in applying the principles of sustainable development in designing and planning the project; and
- form the baseline for an environmental impact assessment to meet licensing requirements should a decision be made to proceed with the project.

The program includes terrestrial, wildlife, archaeology, and aquatic components and has been developed and conducted with the participation of local First Nations.

The Conawapa aquatic studies program was developed to provide information on the interrelated components of the lower Nelson River aquatic ecosystem. The program includes studies of the physical habitat, water quality, detritus, algae, aquatic macrophytes, aquatic invertebrates, fish and marine mammals. Individual reports are being prepared that focus on specific aquatic components and locations.

This report presents results of a Lake Sturgeon population genetics study using Genotype-by-Sequencing (GBS) to investigate relatedness among groups of fish resident in the Churchill, Hayes, Fox, Gods, and Nelson rivers. The primary objectives of the study were as follows:

- Develop a high-resolution genetics toolkit for Lake Sturgeon, which would enable population assignment tests, parentage, and parentage/sibship inference analyses; and
- Test the methods using samples collected from known parents and hatchery reared progeny.

Using the developed genetic markers, the secondary objectives of this study were to:

- Assess population structure of Lake Sturgeon along the Nelson River, between the Jenpeg GS and Hudson Bay (>500 km) i.e., identify populations and determine the amount of gene flow via population assignment tests; and
- Focusing on juveniles (and where possible a single cohort), use sibship inference to determine the number of contributing females and the degree of relatedness amongst

progeny, and further, determine if siblings of upstream residents exist in downstream locations that are separated by rapids or a generating station.

Previous Lake Sturgeon genetic studies focused on microsatellite markers, which use several relatively long repeat sequences of neutral DNA (i.e., segments of the genetic code not under positive or balancing selection). However, in Lake Sturgeon, microsatellite markers identified to date exhibit a low level of differentiation, which means that the quantity of information that each marker provides in relation to population structuring and gene flow is low. Single nucleotide polymorphisms (SNPs) provide even less information per marker, but they are far easier to identify in a given genome, meaning that the relative amount of information that can be obtained via a GBS approach far exceeds that which can be obtained by microsatellites in this species.

Lake Sturgeon tissue samples were collected from adult Lake Sturgeon captured at known or suspected spawning sites in the Nelson, Fox, Gods, Hayes, and Churchill rivers, as well as from juvenile Lake Sturgeon captured in the various locations. A total of 1,178 Lake Sturgeon were sequenced.

The results revealed genetic differences among many of the putative 'populations'. Given generation time of Lake Sturgeon (~35 years) relative to the timescale of Nelson River hydroelectric development (first station built ~50 years ago), it was clear that structured populations existed along the Nelson River long before the construction of hydroelectric dams; if there was any historical movement of fish past major historical falls/rapids e.g., Grand Rapid [Kelsey GS] and Kettle Rapids [Kettle GS], this would not occur. A variety of clustering methods were employed to assess population structure, but all were consistent in indicating a lack of genetic structure among Weir, Angling and Lower Limestone Rapids spawning groups, suggesting one inter-mixed population.

Based on the inclusion of larval fish produced in the hatchery, sibship analyses were able to resolve known control relationships, thereby validating the method. Sibship results for the cohort from Sea Falls included in the analysis were consistent with expectations of numerous full-sibling relationships; this cohort had previously been generated following rearing the eggs of a single female at the Grand Rapids hatchery. Sibship results also provide information on the quantity of contemporary gene-flow occurring among population within the study area. Specifically, in the context of the secondary objectives, results indicated that 2008 cohort juveniles in Stephens Lake and Long Spruce Reservoir were likely not spawned in those locations, as each individual sequenced was determined to have at least one sibling located further upstream in Gull Lake. Downstream

redistribution of larval Lake Sturgeon must have occurred following spawning at an upstream location.

In conclusion, results of the current study dramatically improve the understanding of both historical and contemporary Lake Sturgeon population structure. Population units can essentially be defined by the presence of contiguous spawn-drift-settle-establish habitat sequences, but the distance over which these units occur varies even along the length of the Nelson River. The upper reaches of the Nelson River flow through the Boreal Shield ecozone where riverine habitats were historically fragmented by repeated rapid/fall-river-lake sequences, and results showed that populations existed historically in relatively small stretches of river, with only one-way (downstream) gene flow occurring. Downstream of Kettle Rapids (Kettle GS), the lower Nelson River flows through the Hudson Plain ecozone and the nature of habitat changes dramatically, which explains the presence of a single inter-mixed population downstream of the Limestone GS, despite utilization of multiple spawning locations.

Acknowledgments

Manitoba Hydro is thanked for the opportunity and resources to conduct this study. Chief and Council of the Fox Lake Cree Nation (FLCN), Tataskweyak Cree Nation (TCN), York Factory First Nation (YFFN), and War Lake First Nation (WLFN) are gratefully acknowledged for support of the program. Don Macdonald from Manitoba Water Stewardship is acknowledged for collecting and supplying tissue and DNA samples from the Landing River area.

The collection of biological samples described in this report was authorized by Manitoba Water Stewardship, Fisheries Branch, under terms of scientific collection permits.

We thank Eric Normandeau, Jeremy Gaudin, Laura Benestan and Charles Perrier for valuable discussions and/or help with the bioinformatics. This report was improved by comments from Stephanie Backhouse and Cam Barth.

Acronyms

Rivers:
SFA: Sea Falls River
JEN: Jenpeg River
LAN: Landing River
GRA: Kelsey/Grass River
BUR: Burntwood River
GUL: Birthday Rapids/Gull Lake
STE: Stephens Lake
LOS: Long Spruce Rapids
LLI: Lower Limestone Rapids
ANG: Angling River
WEI: Weir River
LNR: Lower Nelson Rivers (LLI+ANG+WEI)
FOX: Fox River
HAY: Hayes River
GOD: Gods River
CHU: Churchill River
DU: Designable Units
DU1: Western Hudson Bay
DU2: Saskatchewan River
DU3: Nelson River
DU4: Red-Assiniboine Rivers-Lake Winnipeg
DU5: Winnipeg River- English River
DU6: Lake of the Woods-Rainy River
DU7: Southern Hudson Bay-James Bay
DU8: Great Lakes-and Upper St. Lawrence
GS: Generating Station

Genetic:
GBS: Genotype-by-Sequencing
NGS: Next-Generation Sequencing
SNP: Single Nucleotide Polymorphisms
WGS: Whole Genome Sequencing

CONAWAPA AQUATIC STUDY TEAM 2014

Bella Cook, TCN
Brandy Bone, FLCN
Brianna Wyn, NSC
Cam Barth, NSC
Christian Lavergne, NSC
Clayton Flett, TCN
Clayton Spence, TCN
Cory Beardy, WLFN
Craig Jones, NSC
Craig McDougall, NSC
Daniel Redhead, SFN
Darcy Pisiak, NSC
Darcy Wastesicoot, YFFN
Darwin Flett, WLFN
Dave Szczepanski, NSC
Don MacDonell, NSC
Donovan Flett, YFFN
Douglas Kitchekeesik, TCN
Duane Hudd, NSC
Elena Fishkin, NSC
Eugene Spence, TCN
Evelyn Beardy, YFFN
Franklin Beardy, FLCN
Gail Eaton, NSC
Gary Garson, TCN
Gary Spence, WLFN
Ginger Gill, NSC
Howard Beardy, FLCN
James Lockhart Jr., FLCN
James Redhead, FLCN
Jarod Larter, NSC
Jordna Hill, SFN
Joshua Spence Sr., TCN
Kathleen Dawson, NSC
Kathy Wavey, WLFN
Ken Ambrose, NSC
Kim Mandzy, NSC
Kristine Juliano, NSC
Laura Groening, NSC
Laura Henderson, NSC
Leanne Zrum, NSC
Lee Murray, NSC
Mark Blanchard, NSC
Mark Gillespie, NSC
Matthew Laliberty, WLFN
Matt Martens, NSC
Mike Johnson, NSC
Mike Legge, NSC
Patrick Nelson, NSC
Paul Cooley, NSC
Peter Massan, TCN
Philip Morris, WLFN
Randy Naismith Jr., FLCN
Randy Naismith Sr., FLCN
Raymond Mayham, FLCN
Richard Henderson, FLCN
Robert Beardy, FLCN
Sam Miles, SFN
Stacy Hnatiuk Stewart, NSC
Thomas Nepitabo, WLFN
Victor Spence, TCN
Victoria Henderson, FLCN
Vincent Anderson, FLCN
Warren Bernhardt, NSC

FLCN – Fox Lake Cree Nation,
SFN – Shamattawa First Nation,
TCN – Tataskweyak Cree Nation,
WLFN – War Lake First Nation,
YFFN – York Factory First Nation,
NSC – North/South Consultants Inc

Table of Contents

Page

List of Figures

Page

List of Tables

Page

List of Appendices

Page

1.0 INTRODUCTION

1.1. LAKE STURGEON CONSERVATION STATUS

The Lake Sturgeon (*Acipenser fulvescens*) has suffered steep declines since the late 1800's, described succinctly by Houston (1987) as a "synergistic product of life history factors, exploitation, and environmental change". Today, across most of its historical range in Canada, the Lake Sturgeon is considered endangered by the Committee on the Status of Endangered Wildlife in Canada (COSEWIC) (COSEWIC 2006).

For assessment purposes, the Committee on the Status of Endangered Wildlife in Canada status report (COSEWIC 2006) divided Lake Sturgeon in Canada into eight designable units (DUs). The DUs include: Western Hudson Bay (**DU1** – Churchill River); Saskatchewan River (**DU2**); Nelson River (**DU3**); Red-Assiniboine Rivers-Lake Winnipeg (**DU4**); Winnipeg River- English River (**DU5**); Lake of the Woods-Rainy River (**DU6**); Southern Hudson Bay-James Bay (**DU7**); Great Lakes-and Upper St. Lawrence (**DU8**).

In this study, genetic samples were collected from three DUs: i) DU1, the Churchill River Lake Sturgeon stock, designated as "Endangered"; ii) DU3, the Nelson River stock, designated as "Endangered"; iii) DU7, the Southern Hudson Bay and James Bay tributaries (Hayes River) stock designated as "Special Concern". Some areas within these DUs sustained commercial fisheries until the mid-1900s after which, dramatic declines in landings were observed. Historically, overexploitation was the primary threat; more contemporary threats are harvest, poaching, pollution, and dams (Birstein et al. 1997; Pikitch et al. 2005).

1.2. FISH MOVMENTS AMONG NORTHERN RIVERS OF HUDSON BAY

Recaptures of tagged fish indicate that Lake Sturgeon move among closely situated tributaries of Hudson Bay [e.g., Nelson, Hayes, and Gods rivers (Ambrose et al. 2007) and possibly Churchill and Seal rivers (Keleher and Kooyman 1958)]. Despite potential for inter-river mixing, the extent to which Lake Sturgeon populations are demographically connected through effective dispersal is unknown. Effective dispersal is the movement of individuals between their natal ground and another site, followed by successful reproduction in the latter. Effective dispersal estimation from tagging data requires extensive (and logistically intensive) coverage of the target populations over several generations. In contrast, genetic methods can quickly provide an estimate of the genetic structure of populations and a general understanding of their demographic connectivity. Assessing the genetic structure of Lake Sturgeon populations is therefore

desirable for understanding potential impacts of a project on the species and for developing mitigation strategies. Specifically, issues such as migration routes, spawning site fidelity, and subsequent larval drift and dispersal are key requirements when considering impacts of hydroelectric development.

1.3. PREVIOUS LAKE STURGEON GENETIC STUDIES

Genetic information is valuable for the planning of sound management strategies, especially for supportive breeding and restocking associated with recovery strategy planning. In fact, ill-informed mixing among different genetic populations can have several negative consequences on populations. Specifically, by causing loss of local adaptation or by promoting inbreeding or outbreeding depression, it may decrease individual fitness and survival, leading to population declines or even crashes (Hindar et al. 1991; Waples and Do 1994; Reisenbichler and Rubin 1999; McLean et al. 2004; Ward 2006).

The conservation concerns of most sturgeon species has resulted in the application of population genetic techniques, from molecular phylogenies (Birstein and DeSalle 1998) to microsatellite development (Welsh et al. 2003) and it's concerns for polyploidy (Ludwig et al. 2001; Pyatskowit et al. 2001; Rodzen and May 2002), to genetic assessment of population structure (Drauch et al. 2011; Ferguson and Duckworth 1997; Welsh and May 2006; Welsh et al. 2008; Wozney et al. 2010).

Previous genetic studies on Lake Sturgeon in Manitoba highlighted genetic differentiation at 6 loci (Afug9, Afug68, Afug74, Afug112, Afug122 and Afug160 - see Côté et al. 2011). However, more than half of the significant comparisons were between watersheds (Nelson/Hayes and Churchill River samples). Within watersheds, along the Nelson and Hayes Rivers, less pronounced genetic structure was found. Microsatellite analysis showed no genetic structure between pairs of sites downstream of the Kettle GS (Lower Limestone/Angling/Weir/Hayes), while most sites upstream of Kettle GS were genetically distinct from each other (Côté et al. 2011).

Using microsatellites, of the 9 candidate populations sampled, only 5 could be grouped in genetically distinct populations (LAN, BUR/GRA, GUL, Lower Nelson/Hayes and CHU) (Côté et al. 2011). With an average of 4.7 alleles, the 7 microsatellite loci resolution was not high enough for assignment test and parentage analysis to be performed. However, these markers showed that watersheds and instream historical barriers have shaped the genetic structure of Lake Sturgeon populations in Manitoba.

1.4. STURGEON CONSERVATION GENETICS

Historically, most studies in ecological and conservation genetics have relied upon a small number of putatively neutral molecular markers (e.g., allozymes, microsatellites, AFLPs), covering a very limited subset of the genome (Joop et al. 2009). Many of these studies have been limited to narrow regions of the genome, allowing for limited inferences but making it difficult to generalize about the organisms and their evolutionary history.

Currently, a suite of 12-14 microsatellites is typically used to examine the genetic structure in and between Lake Sturgeon populations (Pyatskowit et al. 2001; McQuown et al. 2002; Welsh et al. 2003; Welsh and May 2006). Research conducted to date has focused on population genetics to delineate stocks (primarily in Great Lakes populations; Welsh et al. 2008)) and define logical management units for conservation (COSEWIC 2006). However, in large river systems (e.g., the Winnipeg and Nelson rivers), more subtle genetic structures may exist, which could be attributed to a number of factors including: flow-influenced larval dispersal patterns, year-round riverine residence, barriers to upstream movement, historical asymmetric (downstream) gene flow due to natural barriers, "learned" spawning site fidelity, and disproportionate contribution of individual spawners to subsequent year-classes. The aforementioned hypotheses have yet to be formally investigated, but there is certainly evidence of population structuring in both the Nelson and Winnipeg rivers that predates hydroelectric development (Cote et al. 2011; McDougall 2011).

Using hatchery reared control samples (known parents and progeny) collected from Winnipeg River Lake Sturgeon, it was recently determined that 12 microsatellite markers were insufficient to accurately resolve sibship (full, partial, unrelated) (McDougall 2011). Two of these markers, known to be polymorphic in Great Lakes populations, were monomorphic and thus uninformative in the Winnipeg River (McDougall 2011), the Nelson/Hayes/Churchill rivers (Côté et al. 2011), and other localities in the Hudson's Bay drainage basin (Welsh and McLeod 2010; Welsh et al. 2008).

While the genetic resolution provided by these microsatellite markers appears sufficient to allow for relatively accurate parentage identification in the Winnipeg River data set (McDougall et al. 2014), the utility of parentage identification is less than sibship identification (Wang 2004). This is in part because potential parents need to be sampled and included in all analyses, which can be difficult or even impossible due to logistic reasons, including harvest complications (i.e., parents may no longer be available for

capture). Furthermore, the need to analyze a large number of parental samples incurs a considerable cost and major logistical constraints.

Consequently, a toolkit that would allow sufficient genetic resolution to: a) allow for identification of fine–scale (subtle) population structuring missed by using the current suite of microsatellite markers, b) facilitate accurate population assignment tests, and iii) accurately resolve sibship in Lake Sturgeon populations is paramount to resolving hypotheses (as listed above) relating to conservation questions in large riverine systems. The applications of the toolkit will undoubtedly extend far beyond the questions related to historical upstream barriers to movement, downstream passage, individual contributions to year-class strength, and rehabilitation stocking programs which we are explicitly concerned with herein.

Recently, Genotyping-by-Sequencing (GBS), a new set of tools that use next-generation sequencing technology (NGS), has been shown to offer major advantages for population genomics by allowing the researcher to screen for thousands of polymorphisms throughout the genome that are subject to the full range of evolutionary histories (Catchen et al. 2013). GBS is part of a family of techniques that uses reduced representation genotyping. GBS enables the researcher to investigate the entire genome of plants and animals by focusing on short sections of the genome highlighted by specific restriction-enzyme anchored positions. Depending on genome size and complexity, GBS allows the discovery of potentially tens or hundreds of thousands of Single Nucleotide Polymorphisms (SNP) spread evenly throughout the genome. GBS approaches have gained popularity among population geneticists because the technique works with species with little or no previous genomic information. Furthermore, because whole genome sequencing (WGS) can become costly for population studies and/or species with large and complex genomes, GBS offers a viable solution for many conservation projects by allowing only a targeted fraction of the genome (a reduced representation library) to be sequenced with NGS (Baird et al. 2008; Davey et al. 2011; Narum et al. 2013).

GBS enables researchers to improve the precision of demographic inferences by greatly increasing the number of neutral markers (Narum et al. 2013). Neutral markers can identify significant differentiation among populations based on limited gene flow or drift, but genomic regions under selection may indicate adaptive similarity that may have been either retained after isolation (Parchman et al. 2013) or evolved in parallel following colonization of new habitats (Hohenlohe et al. 2010). To investigate population structure and parentage/sibship relationships of Nelson River Lake Sturgeon, GBS was used.

In this context, this report presents results of a population genomics study using GBS that focused on Lake Sturgeon in the Churchill, Hayes and Nelson rivers in Manitoba with the following primary objectives:

- develop a high-resolution genetics toolkit for Lake Sturgeon. Based on GBS and SNP markers, this toolkit would enable population assignment tests, parentage and sibship inference analyses, and;

- test the validity of the method using samples collected from known parents and hatchery reared progeny.

Using the developed SNP markers, the secondary objectives of this study were to:

- assess population structure of Lake Sturgeon stocks occurring along the Nelson River, between the Landing River and Weir River tributaries (~340 km). Populations would be identified, levels of contemporary and historical admixture would be quantified, and assignment tests would be conducted; and

- using a single cohort (2008), use sibship inference to determine the number of contributing females and the degree of relatedness within the cohort, and further, determine if siblings of upstream residents exist in downstream localities that are separated by rapids or a generating station.

It is anticipated that recovery of depressed populations, under pressure from hydroelectric development and ongoing subsistence harvest, will depend on stocking. Addressing these objectives for Nelson River Lake Sturgeon populations will provide a better understanding of the population genetics in depressed populations, context to historical structuring patterns, and identify potential genetic impacts of recovery strategies based on stocking.

2.0 METHODS

The methodological steps involved in this project are shown in Figure 1 and are detailed below.

2.1 SAMPLING

Figure 2 and Table 1 provides an overview of the locations where Lake Sturgeon tissue samples were collected in relation to the watersheds and existing hydroelectric developments. A 1-2 cm^2 fragment of pectoral or pelvic fin tissue was removed and preserved in 95% biological grade ethanol. Two daily complete ethanol changes were done prior to storing samples in sealed vials until processed. All samples analyzed in Côté et al. (2011) were chemically and statistically re-analyzed along with additional samples collected since 2007.

Samples were collected from two life stages: 416 adults (>834 mm fork length) and 567 juveniles (107 to 833 mm FL). In general, adult samples were collected during spring spawning periods (2005 - 2012). However, it should be noted that most adults were not in spawning condition at the time of capture. Some opportunistic collections also occurred outside of the spawning period. Juvenile samples were collected primarily during fall (2010 – 2012). In addition, 196 genetic samples were collected from Landing River broodstock (a tributary of the Nelson River located upstream of the Kelsey GS) to be analyzed as controls (sibship and parentage relationships were known). Tissue samples were collected from 4 adults as per above, and a total of ~ 50 larvae from each full-sibling dyad produced were individually preserved (whole) in ethanol as per above.

2.2 DNA EXTRACTION AND QUANTIFICATION

High molecular weight DNA were extracted from fins and larvae using a standard salt-extraction method (Aljanabi and Martinez 1997) with the additional step of RNase A treatment following the manufacturer's recommended protocols (QIAGEN, Valencia, CA, USA). Extracted genomic DNA (gDNA) was quantified using Quant-iT PicoGreen dsDNA Assay kits (Life technologies, USA) on a Fluoroskan Ascent FL microplate fluorometer and Ascent Software v2.6 (Thermo Labsystems).

2.3 GENOTYPE LIBRARY CONSTRUCTION

To investigate population structure and parentage/sibship relationships of Nelson River Lake Sturgeon, a new genetic approach called Genotype-by-Sequencing (GBS) was used. GBS is part of a family of techniques that uses reduced representation genotyping. GBS enables the researcher to investigate the entire genome of plants and animals by focusing on short sections of the genome highlighted by specific restriction-enzyme anchored positions. Depending on genome size and complexity, GBS allows the discovery of potentially tens or hundreds of thousands of Single Nucleotide Polymorphisms (SNP) spread evenly throughout the genome.

The library construction used in this study followed a modified protocol of Elshire et al. (2011). Detailed step-by-step GBS protocol is provided in Appendix 1. Genome complexity was reduced with the use of two restriction enzymes (*Pst*I and *Msp*I) that digest the genomic DNA into small fragments. The resulting digested DNA is ligated with unique barcoded adapters to identify each individual. The fragments are duplicated by multiple PCR amplification steps. Individual GBS libraries labeled with unique barcodes were multiplex/pooled in equimolar proportions (48 individuals per lane). Single-end sequencing of 48-plex library per flowcell channel was performed on next-generation sequencing technologies (Illumina HiSeq2000, San Diego, USA) at the Genome Quebec Innovation Center (McGill University, Montreal, QC, Canada). Samples from 1178 individual Lake Sturgeon were sequenced.

2.4 BIOINFORMATICS PIPELINE FOR SEQUENCE ANALYSIS

2.4.1 Special concerns for reduced genome de novo assembly

The workflow, software version and options, used in this study are presented in Table 2 and 3. We used cutadapt (Martin 2011) to fully remove the adapter from raw sequences and STACKS *process_radtags* to demultiplex the samples and do the quality trimming (Catchen et al. 2013). Before performing the *de novo* assembly of the short reads into orthologous loci (*ustacks*, Table 3), sequence similarity was explored to find the optimum clustering threshold (Figure 3) with the ploidy-informed empirical procedure developed by Ilut et al. (2014).

Our preliminary STACKS run analysis showed no difference at the catalog level between datasets normalized for the number of individuals and their origins and datasets with all the samples. Consequently, after the individual's de novo assemblies, the catalog

construction used all loci identified across all samples (*cstacks*, Table 3). After, loci from each individual are matched against the catalog to determine the allelic state at each locus in each individual (*sstacks*, Table 3). To improve the quality of the *de novo* assemblies produced in STACKS and reduce the risk of generating nonsensical loci with repetitive sequences and paralogs, we used the correction module *rxstacks* (Table 3). After the correction module, the catalog and individual matches are processed again with the corrected individual's files. The last module of STACKS (*populations,* Table 3) is run with relaxed filtering parameters, because subsequent filtering is undertaken in STACKR (Gosselin and Bernatchez 2015; Table 4).

2.4.2 Ascertainment bias

The landscape covered in our study was characterized by heterogeneous geographical features (e.g., barriers/rapids and watersheds, see Figure 2) that could introduce ascertainment bias in the markers design (see discussion). Therefore, we tested and designed several methods of filtering with *a priori* groupings before choosing the final marker panel. The training dataset consisted of the first series of samples that were analyzed: 8 sampling sites (GRA, BUR, GUL, LLI, ANG, WEI, HAY, GOD) with 301 individuals with a minimum of 1 million reads. Four marker panel (whitelists of loci) were created with the training dataset: i) the 8 sampling sites, ii) all the samples combined into one grouping (1 large population), iii) barrier grouping: the upper Nelson River consisted of sites upstream of Keeyask GS (GRA, BUR, GUL) and all the sites downstream of Limestone GS (the lower Nelson River combined with the Hayes River; LLI, ANG, WEI, HAY, GOD); and iv) watershed grouping (the Nelson and the Hayes Rivers). The whitelists of markers obtained after filtering were used with all the individuals (adults and cohorts) and sampling sites.

2.4.3 Ploidy-based filtering

GBS combined with massive parallel short-read sequencing produce noisy data that requires several bioinformatics filtering steps to remove artifacts. We applied several conservative filtering steps at the individual and population level. Here is an overview of the filtering steps outlined in Table 4: i) remove obvious paralogs (loci with more than 2 alleles), ii) inspect, correct and/or remove loci with excessive coverage, with poor coverage and/or poor genotype likelihood values; iii) individual and populations based filtering to remove underrepresented markers; iv) loci with minor allele frequency are removed to keep informative markers; v) loci with an excess of heterozygosity and with

inbreeding coefficient (based on Hardy-Weinberg equilibrium) overlapping a certain threshold range are removed to prevent genotyping errors, paralogs and *de novo* assembly artifacts; and finally vi) loci with outlier numbers of SNP per haplotypes are removed.

2.4.4 Transposable element

The Basic Local Alignment Search Tool (BLAST, http://www.ncbi.nlm.nih.gov) was used to find sequences similarity between unique reads with high coverage and *Tana*1 (GenBank accession number: JX889425–889438). *Tana*1 is *tc*1-like transposable element newly discovered in sturgeons (Pujolar et al. 2013).

2.4.5 Map-independent imputations

Missing values are intrinsic to GBS approaches consequently the pattern of missingness was inspected after STACKS and prior to filtering. Systematic patterns of missingness were visualized with multidimensional scaling in PLINK (Purcell et al. 2007) identity-by-missingness analysis (IBM). In the interest of understanding the demographic inference impacts of vetting loci based on their level of completeness, varying tolerances for missing data (loci present in ≥ 30%, ≥ 50% and ≥ 70% of individuals) were tested (see Table 1, for number of individuals blacklisted based on missing genotypes).

Missing data were imputed with two methods, implemented in the package STACKR (Gosselin and Bernatchez, 2015). The first method is the most commonly used technique in the software literature, where the most frequent allele of the population replaces the missing value. The second method is more sophisticated and uses the random forest algorithm (Ishwaran and Kogalur, 2015). We use 100 trees to grow a forest, with 10 iterations and 100 random splitting (Ishwaran, 2014; Ishwaran and Kogalur, 2015).

2.5 GENETIC DIVERSITY

Several classes of summary statistics with different theoretical backgrounds have been developed to refine population structure analysis (reviewed in Meirmans and Hedrick 2011). Fis based on AMOVA least squares measure and Nei's heterozygosity-based analogue, Gis, that describes the degree of deviation from Hardy-Weinberg equilibrium. Observed heterozygosity (Ho) and heterozygosity within and across all populations (Hs and Ht, respectively) were inspected. These statistics were also explored with correction for sampling bias. We also looked at a new proxy that measure the realized proportion of the genome that is identical by descent (IBDG), the Fh measure is based on the excess in

the observed number of homozygous genotypes within an individual relative to the mean number of homozygous genotypes expected under random mating (Keller et al. 2011; Kardos et al. 2015). We also report the nucleotide diversity (Pi), the calculations in STACKR based on Nei and Li (1979) include the consensus loci in the catalog (the sequences with no variation between populations).

2.6 DIFFERENTIATION STATISTICS

For distance metrics measures, we explored the use of several parameters and statistics of genetic differentiation, differing in their assumptions about the diversity within and among populations and the mutation regimes involved during populations differentiation (reviewed in Meirmans and Hedrick 2011; Alcala et al. 2014). Some measures tested included correction for sampling bias and standardization: i) the Analysis of Molecular Variance (AMOVA) (Excoffier et al. 1992, Michalakis and Excoffier 1996) with the Infinite Allele Model that uses F-Statistics defined by Weir and Cockerham (1984), ii) Nei's G*st*, that includes a bias-correction for small sample sizes (Nei, 1978), iii) Jost's *D* index of population differentiation (Jost 2008), that is independent of the amount of within-population diversity (H_s, see also Alcala et al. 2014) and iv) the log-likelihood G-statistics with a permutation approach, a robust test when sampling is unbalanced (Goudet et al. 1996). Correlations between distance metrics were measured with the Pearson's product-moment correlation. Corrections for the multiple tests in the pairwise analysis were not applied; instead, we used the AMOVA as an overall test of population differentiation. The AMOVA analysis included 2 tests of hierarchical structure: i) sampling sites nested in watersheds and ii) sampling sites nested in presence/absence of a barrier to migration. The significance was tested with 10 000 permutations. We used GENODIVE (Meirmans and van Tienderen 2004) for all the calculations.

The estimates of genetic diversity and differentiation statistics computed in this study were based on anonymous loci (haplotypes with no physical position in the genome or in a linkage map) and loci under the full range of selection (balanced, directional and neutral).

2.7 CLUSTERING ANALYSIS

The K-Means clustering approach implemented in GENODIVE (Meirmans and Van Tienderen 2004) was used to find the number of cluster (*K*) using the Sums of Squares from an Analysis of Molecular Variance (Excoffier et al. 1992; Meirmans 2012). In GENODIVE, the simulated annealing method that uses a Monte Carlo Markov Chain

(MCMC) was used to prevent the clustering from getting stuck in local optima. The optimal values of K chosen were based on the Bayesian Information Criterion (BIC). The benefit of BIC is that it can be used to determine whether there actually is any population structure at all. The simulated annealing approach usually returns better results, provided that a suitably large number of steps are used. Consequently, the simulated annealing was run with 50,000 steps as a starting point and 400,000 as a stopping rule or before, if no changes were observed. Random start tests ranged from 100 to 1000 and method was repeated several times.

2.8 DISCRIMINANT ANALYSIS OF PRINCIPAL COMPONENTS (DAPC)

For admixture analysis we favored a multivariate analysis that makes no a priori assumptions about the underlying population genetic model. We used the Discriminant Analysis of Principal Components (DAPC) analysis implemented in the r package ADEGENET (Jombart et al. 2010; 2011) to investigate the variance in genetic diversity among individuals between our sampling sites. Group memberships of DAPC, were tested to see how well the genetic clusters described the data. To secure a space with sufficient power for discrimination and avoid apparent perfect discrimination (over-fitting), the a-score, the proportion of successful reassignment corrected for the number of retained PCs, was used to find the optimal number of PCs to retain. The dimension reduction steps uses the average and individual group a-scores computed on DAPC with randomized groups. To validate the robustness of cluster assignments, the predictive capacity of our analysis, stratified cross-validation was employed with a varying numbers of PCs to keep the number of discriminant functions fixed. The data was divided in two datasets by stratified random sampling (training and validation data sets, 90% and 10% of the data, respectively) and at least one member of each group or population in the original data was represented in both datasets. The DAPC was constructed with the training dataset and validated with the hold-out individuals that enable the selection of an optimal number of PCs to retain (100 replicates was used). We selected the number of PCs associated with the lowest root mean squared error. The results are presented for each potential clusters, based on the phylogenetic tree branching. For the tree distance metric we used different parameters and statistics of genetic differentiation measures (see the previous section).

2.9 POPULATION ASSIGNMENT

DAPC also provides membership probabilities of each individual, interpreted as proximities of individuals to the different clusters, based on the retained discriminant functions. In addition, the likelihood approach of Paetkau's (1995), implemented in GENODIVE (Meirmans and van Tienderen 2004), was used to assess the population membership of individuals (Manel et al. 2005). For each potential source populations, a distribution of genotypes was generated by Monte Carlo simulations based on allele frequencies. The expected distribution of genotypes for the population is constructed by resampling with permutation (n = 10,000). We used the recommended alpha-value values of 0.002 (equal to < 1 type I error, sample size * 0.002), as a compromise between power to detect migrant and type I errors (Paetkau et al. 2004). The threshold value, used to infer population of origin, is pre-defined by the alpha-value and was calculated separately for every population. Population assignment tests requires a dataset with no missing values, we use the dataset imputed by the random forest algorithm. The power to detect population of origins is influenced by the test statistics used (reviewed in Paetkau et al. 1995; 2004). The likelihood ratio statistic (Lh/Lmax) and the home likelihood statistic (Lh), assumes or not that all source populations for putative immigrants were sampled, respectively. Both statistics are presented in this report

Power to detect F_0 immigrants was calculated with the package STACKR using the mean genotype likelihood ratio distance (D_{LR}) for each pair of populations (see Paetkau et al. 1995; 2004). This measurement of genetic distance is better at estimating statistical power than F_{ST} (Paetkau et al. 1997) and is independent of the statistics used (L_h or L_h/L_{max}).

The allele frequencies of adults were used for the assignment test of the juveniles captured from SFA, JEN, LAN, GRA, BUR, GUL, STE, LOS, and LNR (LNR : LLI, ANG and WEI). The juvenile samples were treated as a separate population to perform the analysis in GENODIVE.

2.10 RECONSTRUCTING RELATIONSHIPS

Relationships of juvenile cohorts with unknown relationship were reconstructed with the full-pedigree likelihood approach of COLONY (Jones and Wang 2010, Wang 2013). We ran nine groups of juveniles; upstream reach group (SFA-JEN-LAN-GRA-BUR), middle reach group (GUL-STE-LOS), and a downstream group LNR, to infer full-sib/half-sib pairs and the number of putative parents (sexes are arbitrarily set by COLONY). The two

groups of cohorts (upstream and middle) were run to highlight connectivity patterns among 'populations': upstream-reach group extended from Lake Winnipeg to Split Lake (Jenpeg GS - Kelsey GS, see Figure 2) and middle-reach group extentended from Clark Lake to Long Spruce Reservoir (Kelsey GS to Long Spruce GS, see Figure 2).

Cohorts run in groups were analyzed with adults of the same sampling sites, as putative parents. Adults with known sexes were used to develop markers with sex-specific alleles (~ 800 loci). Correct assignment in 60% of the markers was required for gender identification, below this threshold, individuals remained with an unknown gender in COLONY.

Because it is not possible to distinguish paternal and maternal contributions to half-sib dyads, Wang clearly stresses that "*caution should be exercised in interpreting and using the information about paternal and maternal families from a sibship assignment analysis*" (Wang 2009). However, since the sturgeon mating system is polygynandry (reviewed in Bruch and Binkowski 2002), conclusions, regarding sexes attribution in COLONY, are more reliable than in a monogamous mating system (see Jones and Wang 2010, Wang 2013). The technique was validated with the four experimental families (four parent crosses, 50 progeny per family) of sturgeon from Landing River broodstock. COLONY analysis was run with no priors on sibship, fathers or mothers and the mating system inferred for sturgeon was polygynandrous. All subsequent analyses and interpretations are based on the full range of probabilities of observing the relationship; consequently the numbers provided are conservative estimates. The influence of the number of markers on the relationship inferences was explored by using 5 different COLONY datasets with 500, 1000, 2000, 3000 and all the markers.

2.11 DATA VISUALIZATION AND COMPUTER HARDWARE

Data tidying and visualization were performed with STACKS web-based interface (mysql database) and R (R Development Core Team, 2015) packages: dplyr (Wickham 2011; 2014) and ggplot2 (Wickham 2010). The custom STACKS pipeline used in this study is available here (https://github.com/enormandeau). Most of the software was executed with an Apple retina MacBook Pro (16 GB memory) or with a Mac Pro (64 GB memory) and SSD flash storage disks. Between 2 and 4 TB of external disk storage were necessary to complete the analysis. When more computer power was necessary, cloud computing with Amazon Elastic Cloud Compute (EC2) was used. The pipeline used in this study (Gosselin and Bernatchez, 2015) is available in a tutorial-workflow (http://gbs-cloud-tutorial.readthedocs.org).

3.0 RESULTS

Overall, a toolkit of 5637 haplotype loci (8848 SNP) was developed for the Sturgeon. These markers show great promises for demographic inference, assignment and parentage analysis. Below, are the detailed results that led to these numbers.

3.1 *DE NOVO* ASSEMBLY AND GENOTYPING

After adapter trimming, demultiplexing and quality trimming of raw reads, final sequences of biological interest were 80 bp long. The overall quality score per reads was >30 (Illumina 1.9 encoding) and the mean GC content ranged between 50 - 51%. To avoid introducing ascertainment bias and variance in our data set, the distribution of the number of reads per individual was evaluated, globally, by sampling site and/or cohorts. No sampling sites or cohort bias was observed. The median number of reads per individuals for the 416 adult and the 567 juvenile sturgeons is 2 605 220 and 2 845 300, respectively.

Based on figure 3, showing the clusters of similar sequences built for a series of similarity thresholds, the *de novo* genome assemblies were optimized by allowing a maximum of 5 differences (in mismatches) between reads within a cluster (-M in ustacks, Table 3). This threshold optimizes the clustering of alleles by minimizing over- and under-splitting of allele that might have contrasting ancestry. Using an optimal mismatch threshold between reads during *de novo* genome assemblies reduces paralogs and artifactual loci (Ilut et al. 2014; Harvey et al. 2015).

After STACKS pipeline of the training dataset, we recovered a total of 85985 putative loci, haplotype markers, corresponding to approximately 160000 Single Nucleotide Polymorphisms (SNP).

3.2 ASCERTAINMENT BIAS, FILTERING AND MISSING DATA

The identity-by-missingness analysis (IBM) did not reveal *a priori* bias or pattern of correlation between hierarchical grouping considered in this study (lanes/chips, sequencing machines, sampling sites, barriers and watersheds). This test is the baseline prior to any filtering steps (after STACKS last module) to ensure no batch effects are introduced by missing data and later by the filtering pipeline. During this step, missingness was tracked by making blacklists of individuals with a maximum of 30%, 50%, and > 70 % of missing data (Table 1).

To avoid creating ascertainment bias, the process of filtering the GBS data was iterative, stepwise and required regular inspection of loci statistics distribution before and after filtration. No bias in the downstream analysis was observed using the different *a priori* groupings with the training dataset. Consequently, the figures, tables and numbers presented below refer to the marker panel developed with the 8 sampling sites as grouping (whitelist no.1). The number of markers before and after each filters of the training dataset is reported in Table 4. Density distribution and box plots of individual and/or sampling sites raw data statistics (mean, median, min, max, diff : min-max) before and after filters revealed no particular bias. After each filtering steps, the pipeline was carried up to the population assignment. Knowing what to expect enabled the acquisition of a general overview of the data to weight the filtering threshold.

Not all filters have the same power to prune the markers, as highlighted by Table 4. Within STACKS and subsequently in STACKR, the genotype likelihood (that incorporate coverage information), the individual and population filters remove most of the markers (Table 4). For coverage, most of the loss in marker number is at the low end of the sequencing depth spectrum, but accurate heterozygosity calls requires depth of coverage (we used 7 as the minimum). Legitimate loci are found at high sequencing depth, but the maximum coverage threshold value represent a compromise required to exclude unique reads with high coverage (putative paralogs and transposable elements). Out of the 382 unique reads with high coverage (> 100), 84% aligned to Tana1, a newly discovered transposable element of 1 595 pb in Lake Sturgeons (Pujolar et al. 2013). The genotype likelihood filter guarantees accurate genotyping before using the remaining filters and estimating genetic diversity (Table 4).

Overall the filtering steps reduce the amount of missing data inside each sampling sites and cohorts. The number of individuals blacklisted per sampling sites is presented in Table 1. These blacklists were tested for potential bias in assignment and demographic inference (see below). Using the most frequent allele to impute missing values showed bias that was creating impossible migration and connectivity scenarios, whereas the random forest algorithm showed more realistic scenarios. Consequently, the figures and tables presented in the report with the different context of missing genotypes tested shows map-independent imputations with the random forest algorithm.

3.3 GENETIC DIVERSITY AND DIFFERENTIATION

The lower polymorphism found in JEN and CHU, is consistent with the lower sample number (more monomorphic loci) and not surprisingly this influenced the remaining

estimates (Table 5). No other biases were observed. The new proxy that measures the realized proportion of the genome that is identical by descent (Fh) was very close to zero for all sampling sites. Using different threshold values of heterozygosity and Fis filter, range from -0.3 to -1 and from 0.3 to 1, for the lower and upper bracket respectively, did not remove many loci, because the genome-wide pattern of Fis, shows that the majority of loci within each sampling site are zero or nearly so, indicating a lack of pervasive overall structure within each population. However, for several populations, a small fraction of loci exhibit values of <-0.3, including some loci >0.3. The strongly negative Fis values represent an excess of heterozygosity over that expected by HWE. Loci beyond the Fis range threshold were all passing the other statistics used for filtering, consequently we decided to keep them in the dataset.

Pairwise values of estimates of genetic differentiation were mostly statistically significant (using 10 000 permutations) and they differ in magnitude substantially among pairwise comparisons (Table 6). Only 2 pairwise comparisons (LLI-ANG and LLI-WEI) showed p-values higher than 0.0001 (but lower than 0.01, in all tests). Pairwise estimates with CHU, the most northern site in this study, and Nelson and Hayes tributaries showed elevated values of genetic differentiation (ranging from 0.047 to 0.079). The lowest values (Gst = 0.001) was observed between the tributaries of the lower Nelson River, downstream of Limestone GS (LLI, ANG and WEI). Overall, the measure of genetic differentiation was 0.027, 0.026 and 0. 005 for Gst, Fst and Jost D, respectively. Jost D pairwise estimates were consistently lower, the metric is independent of the amount of within-population diversity (see Jost 2008; Alcala et al. 2014), but the three distance metrics were all highly correlated across pairwise estimates (corr > 0.99; $p < 0.0001$). Contrasting differences between imputed and non-imputed data were particularly pronounced with Churchill and Jenpeg estimates, where imputations could double (or more) the values. Variance between the two methods is discussed below. Log-likelihood G-test did not provide different information from the other three metrics. LLI, ANG and WEI, from the lower Nelson River, were combined for the AMOVA that showed, for both the barrier and watershed hierarchical analysis, significant differences at the two levels: i) sampling sites, ii) sampling sites nested in watersheds and barrier to migration ($p < 0.001$).

Compromises between the different pairwise estimates are best summarized with a UPGMA tree. Trees are also useful to highlight differences between distance metrics. Figure 4 show the tree build with Nei's Gst distance metrics. Branches rearrangements were observed between the three different genetic differentiation statistics explored (Fst from the AMOVA, Nei's G*st*, and Jost's *D*), but were confined to the least differentiated

sites (the lower Nelson GOD and HAY clusters). Using imputation did not change the highly distinct position of CHU from the Nelson and Hayes Watersheds. Clustered together with small branches are the sampling sites located downstream on the Nelson River (ANG, LLI and WEI). Whereas tributaries upstream of Limestone GS are separated by relatively longer branches among themselves and from other sites (Figure 4). Hayes tributaries are interesting, as HAY and GOD cluster together and far from FOX (Figure 4).

3.4 CLUSTERING ANALYSIS

With the K-Means clustering approach implemented in GENODIVE, the lowest value of Bayesian Information Criterion suggested a K = 10. The results were the same during testing of the simulated annealing using a wide range of steps and random start.

3.5 DISCRIMINANT ANALYSIS OF PRINCIPAL COMPONENTS (DAPC)

The DAPC was conducted with different values of K (from 2 to 12). The first two differentiated clusters are the Churchill River and the Nelson and Hayes watersheds combined Figure 5). At K = 3, Jenpeg, the most meridional sampling sites in our study shows some distinctiveness. The Hayes watershed shows a unique signature at K = 4, with the Fox River (FOX). From K = 5 and onward, differentiation is building upstream of Keeyask GS and inside Hayes Watershed, to a lesser degree (Figure 5). Downstream of Limestone GS, for all values of K, the lower Nelson tributaries are very similar. For sampling sites showing admixed origin, assignment analysis will reveal if the admixture proportion is unique enough to create a population signature.

The number of principal components (PC) was optimized for each value of K with the alpha-score to avoid perfect discrimination (over-fitting). Between 7 and 15 PC were used to explain the distribution of K in Figure 5. Membership of individuals was also visualized by using for K = 12 the individual's coordinates into the dimension of 4 pairs of combinations of principal components (PC 1 to 4, Figure 6). PC 1 and 2 shows the clear division between the upper and lower Nelson, the Hayes and Churchill Watersheds (Figure 6).

3.6 POPULATION ASSIGNMENT

The reassignment analysis using the home likelihood (Lh) and likelihood ratio (Lh/Lmax) statistics shows an overall discrimination power of 77% and 85%, respectively (Figure 7).

When HAY and GOD are combined and JEN and CHU removed, several sampling sites showed reassignment > 90 % (LAN, GUL, LNR, HAY). When all potential sources of populations are sampled, likelihood ratio statistics did not increase power. The home likelihood statistic is the preferred method to use when not all source populations have been sampled. However the detection power for migrant site origin is usually reduced (see below).

The genotype likelihood ratio distance (Dlr) showed high power to detect migrants, with little impact of the imputation scope on the mean value across sampling sites (Dlr ~ 7.7, 7.9 and 9.5, for dataset of individuals missing a maximum of 30%, 50% and 70% genotypes, respectively). Combining LLI, ANG and WEI increased the Dlr to 7.8 for the dataset with a maximum of 30% missing values. The greatest power was found between JEN and CHU with a Dlr ~67, this number provides almost no chance (> 1 billion times more likely) of mis-assigning individuals between the 2 sampling sites. With sampling sites showing re-assignment in the 80%-90% range (e.g. BUR-GRA, Gst = 0.016 and Dlr = 3), the assignment plot shows the discrimination power between the two sites (Figure 8). The lowest Dlr, close to 0, was found between the lower Nelson River cluster (LLI-ANG-WEI, Gst ~ 0.001-0.002), providing no power inside that cluster to reassign individuals to their population of origin (Figure 8).

3.7 RELATIONSHIP RECONSTRUCTION

Relationship reconstruction using the full-likelihood implementation in COLONY performed better than assignment analysis to reveal the fine-scale structure of juveniles, with and between cohorts. Table 7 presents the number of offspring that could be used in each case, the number and proportion of offspring with half-sib and full-sib relationships, the estimated number of mothers and fathers, and the number of mating pairs. In interpreting these results, one must keep in mind that the number of parents discovered will depend on the number of available offspring available for analysis. As can be seen in Table 7, these numbers varied considerably among sites. The first numbers interpreted are the reported numbers of putative mothers and fathers. As mentioned in the methods, sexes were assigned arbitrarily by COLONY so it is really the total number of putative parents and number of mating pairs that matters. When considering only the parents with probabilities > 0.75, the total number of parents varied from 13 for SFA up to 271 for GUL-STE-LOS (Table 7). Those values would represent the minimal number of putative parents that produced the progeny analyzed. The total number of putative parents increases substantially when considering the full range of probabilities. Those values would represent the highest potential number of parents involved in reproduction.

Increase was particularly important for GUL because the probability of correctly identifying contributing parents was lower than for most other sites. Combining GUL, STE and LOS and using a very long run, so that estimates converge, the probability with > 0.75 increased substantially increasing at the same time the number of putative parents (Table 7).

Another way the data were examined was to use the estimate of n_{50} values, which estimates the number of putative adults that contributed 50% of all offspring assigned (Table 8). This shows that generally, relatively few parents were responsible for producing 50% of the progeny at each location (Table 8). For example, the analysis indicates that for SFA cohort, that was stocked, 2 adults contributed to 50% of the offspring assigned (Table 8). Naturally reproducing cohorts varied considerably, ranging from 13 (BUR) to 92 (STE) all were considerably higher than SFA. These results reflect the fact that the reproductive contribution of individual parents was disproportionate (Figure 9); most contributing parents had very few offspring assigned to them, while a few putative parents were linked to a high proportion of the juveniles included in analysis (Figure 9).

It should be reiterated that the number of parents discovered will partly depend on the sampling strategy and the number of offspring analyzed. In order to explore the influence of the number of sampled offspring on the discovery of parents, 999 resampling bootstrap iterations were used. The likelihood of discovering more parents with a greater number of offspring showed a similar trend for 5 sampling locations (Figure 10). The simulations indicated that for STE and GUL (Figure 10) the data are close to reaching the asymptotic number of parents, while more data are required for BUR and LNR. However, this must be interpreted cautiously, these slopes suggest that GUL and STE may reach a plateau at about 150 reproductive adults, LNR would be above that, possibly >200, while BUR would plateau at lower values, likely between 100 and 150.

Sibship inference for hatchery reared controls (known relationship and parents) make the most complex and dense cluster of this study, since full- and half-sibs relationships between families are found for each progeny (Figure 11). The sequential numbering of the full and half sibling samples produce characteristic triangles for full-siblings (above the diagonal) and half siblings (below the diagonal). Sibship inference of cohorts with COLONY revealed comparable half-sib relationship structure among the different sampling sites with proportions of half-sibs generally above 90%, except for LAN-JEN, GRA and LNR. However, LNR samples represent multiple cohorts and therefore further data would be required to validate any initial conclusions. Full-sib relationship analysis

revealed a wider range of inference, with 5% to 79% of individuals having the same two parents (Table 7) the significance of these values relative to each other depends on the numbers of samples. When the full range of probabilities of observing the relationship is taken into account, most individuals are interconnected through one of their two parents forming family clusters that vary in complexity and numbers (figures not shown). Indeed, the tendency for horizontal and/or vertical patterning for high probability siblings and half-siblings in the heat maps (figures 12-14), indicate that Lake Sturgeon are not only siblings and half-siblings, but that they are captured together in the field years after being spawned. GUL-STE-LOS was one complex and large family cluster with several small substructures of independent clusters (figure not shown). The mating clusters in all sites/cohort revealed by the relationship reconstruction in COLONY is consistent with the reproductive behavior of Lake Sturgeon, even though sexes are arbitrarily set by COLONY.

Relationship reconstruction also provides an estimate of the ideal (Wright-Fisher) or effective number of breeders (N_b) in a population that will result in the same amount of genetic drift as in the actual population being considered (Table 9). N_b can be used to predict the genetic changes in a cohort, accounting for factors such as unequal sex ratio and the variable contribution of parents. Because it is impossible to distinguish paternal and maternal sibship in the analysis conducted, caution should be exercised in interpreting and using the information about paternal and maternal families from a sibship assignment analysis. The average inbreeding coefficient of the cohort is expected to increase at a rate of $1/(2N_b)$, compared with that of the parents who produced the cohort. Cohorts from the different populations show a range of N_b values, which is expected as the samples represent different sample sizes, large single cohorts, and multiple cohorts. The SFA population has an N_b of 4, reflecting a single cohort of stocked Lake Sturgeon (Table 9). The upper Nelson River JEN-LAN populations have an N_b of 187 reflecting a 'moderate' population of adults (Table 9). The GRA population has a moderate N_b of 162, which reflects the level of admixture among LAN-GRA-BUR and a mixed cohort of juveniles (Table 9). The BUR population and GUL-STE-LOS group have N_b values of 59 and 91 respectively and reflect well the smaller adult populations in these areas (Table 9). The LNR population has the highest N_b value at 312, which reflect both the sample size of multiple cohorts and it is also the largest population in the study area. Perhaps more importantly, the LNR cohort was produced in an area where population assignment/differentiation analyses suggest extensive mixing among multiple spawning sites.

4.0 DISCUSSION

This report presents results of a population genetic study that focused on Lake Sturgeon (*Acipenser fulvescens*) in the Churchill, Hayes and Nelson Rivers in Manitoba using GBS, a molecular technique that looks for polymorphisms (e.g., SNP) throughout the genome. The study objectives were:

- to develop SNP markers that would improve population differentiation to allow assignment tests to be performed;
- to reconstruct sibship of the strong 2008 cohort to determine the origin of Stephens Lake juveniles; and
- to provide the baseline genetic data to help in the planning of recovery strategies that may involve stocking.

4.1 DEMOGRAPHIC INFERENCE AND POPULATION ASSIGNMENT OF GENOMIC DATA

Genotyping-by-Sequencing, a technique that samples the entire genome for geneomic polymorphims using a reduced representation library, allowed us to develop and screen thousands of markers. Because of the noisy nature and quantity of GBS data, numerous filtering steps were required to obtain quality markers, which would enable us to conduct assignment and population structure analyses and relationship reconstruction. Having a high quality marker panel was also mandatory because Lake Sturgeon are functional tetraploid fish, with a large genome (>5 pg) and 2n = 262 chromosomes (Blacklidge and Bidwell 1993; Fontana et al. 2004; Pyatskowit et al. 2001; Welsh and May 2006). Genome duplication can complicate population genetic studies because many of the genetic analyses assume disomic segregation of the alleles. Consequently, the filters applied also helped to screen out tandem-duplicated paralogous regions and homeologous duplicate chromosomal regions (Hohenlohe et al. 2013; Seeb et al. 2011). Using a random forest approach instead of the traditional most frequent alleles (or software default) for missing data imputation provided more reliable demographic estimates. The random forest algorithm is a promising method in genome-wide studies and was highlighted recently in several studies in ecology and evolution, epidemiology and GBS (Rutkoski et al. 2013; Huang et al. 2014; Jarquín et al. 2014; Penone et al. 2014; Shah et al. 2014).

In a previous genetic study using the same populations, five distinct populations were identified: Landing, Kelsey/Burntwood, Birthday/Gull, Lower Nelson/Hayes, and Churchill (Côté et al. 2011). However, performing assignment tests with the microsatellite loci was not possible due to the low discrimination power of available markers coupled with weak population differentiation typical of Lake Sturgeon (see Côté et al. 2011). The marker panel comprising over 5,000 loci (8,000 SNP) developed in this study enabled fine-scale discrimination of genetic/demographic patterns. Five putative populations (JEN, LAN, GRA, BUR, GUL) were revealed downstream of Jenpeg GS, upstream of Kettle GS (Figure 16). Until more samples are analyzed for JEN, its uniqueness remains somewhat ambiguous. The sampling sites on the lower Nelson River (LLI, ANG, WEI) form the sixth population that, based on the current study, can now be discriminated from Churchill and Hayes watersheds (Figure 16). Hayes River tributaries have somewhat conflicting levels of differentiation; FOX status is clearly differentiated, but HAY and GOD were considered a single population by DAPC (Figure 5) and assignment analysis (Figure 7) due to high levels of admixture with reduced discrimination power. The Churchill River (CHU) formed the 9th population.

To our knowledge, there is no empirical data or experiments showing the weakness or strength of next-generation sequencing and assignment analysis using known migrants or individuals with known mixed ancestries. The inference of migrant origin is thought to be real time, F_0 immigrants (Paetkau et al. 2004). However, the high resolution provided by GBS could enable the detection of effective dispersal through past ancestry of F_{1+} offspring (see also Gagnaire et al. 2015), resulting from mating between fish with different population origins. Sampling different genome ancestries would increases the noise of the inference. This deeper sampling of genomic region, could also explain the small discrepancy between the assignment (Figure 7) and DAPC (Figure 5) results.

Furthermore, the Dlr metric used to give an overall power to detect migrants in this study has not been cited with genome-wide data sets. In order to have sufficient power to detect migrants there must be strong differentiation, even with thousands of loci, but for differentiation to develop among populations migrants need to be rare. Typically, you would need to sample hundreds of individuals to find a single migrant (Meirmans, 2014), however, it is clear from this study that immigrants can be detected with reduced sample sizes.

Overall, clustering analysis, admixture, and population assignment analyses were congruent in defining the same population units. Moreover, we could define in detail the direction of individual straying among the different populations. The marker panel had an

average individual assignment accuracy of 85%, providing an unprecedented tool for delineating management units and recovery strategies for Lake Sturgeon. In general, analyses showed that effective dispersal ranged from low (GUL) to moderate (LAN-GRA-BUR) to extensive (GUL-STE-LOS and LLI-ANG-WEI and HAY-GOD). The different levels of admixture illustrated in Figure 5, represent the genetic makeup from historical to contemporary admixture. Simulations would be required to understand just how far back in time, the demographic history is revealed by the different analysis.

However, it is evident that until more studies are conducted, the 9 population units should be considered in the context of any future management and conservation plan, the assumption being that preservation of genetic integrity and trajectory is desirable. The method will also be applicable to other riverine and dendritic systems where population structure may be missed using lower resolution genetic techniques (e.g., the current suite of microsatellites).

4.2 RELATIONSHIP RECONSTRUCTION AND POTENTIAL DRIVERS OF RECRUITMENT PATTERNS

Relationship reconstruction, simulation of discovery of the number of parents, and contribution of parents provided estimates of families, sibship/parentage, and number of spawners. With the exception of SFA (stocked), results from remaining sites suggest that the number of adults that produced the cohorts analyzed varied depending on specific river reach. These numbers are based on several assumptions, the most critical are i) individuals are sampled at random (with respect to kinship) from a single cohort of the population and ii) parentage includes the full range of probability of observing the relationship, the numbers therefore should be used as a relative measure and interpreted with caution. Parentage accuracy is influenced by the complexity and density of the family cluster. With complex and dense half-sib clusters (e.g., GUL-STE-LOS), the resulting number of breeding adults vary more compared with simple clusters (e.g., BUR and GRA). Complex and dense clusters indicate a highly polygamous mating system generated these cohorts. Consequently, if a sample of individuals was taken at random (with respect to kinship), sampling effort was controlled, and sampling year and sites are indeed representative, then previous observations of Lake Sturgeon mating system can offer several pointers for interpreting the observed family patterns.

During mating season, both sexes may spawn in numerous locations, depending on availability and proximity of habitat (Bruch and Binkowski 2002; LaHaye et al. 1992). However, it has often been suggested that Lake Sturgeon will move upstream as far as

passable, typically spawning in swift water below dams and rapids that provide oxygen to developing embryos. Female sturgeon will spawn in short and repetitive bouts (few seconds) during which several males will mate with a single female by producing a cloud of sperm. Male reproductive tactics to maximize their reproductive success are numerous and include defending spawning grounds, locating a female and pounding the female's abdomen with their tails and caudal peduncles while ejaculating (Bruch and Binkowski 2002). Authors have suggested that swarm formation and mating behaviors in sturgeon maximize the genetic diversity of the offspring (Bruch and Binkowski 2002).

The complex and dense relationship clusters observed (e.g., GUL-STE-LOS) could be the result of spawning swarms where males and females are spawning simultaneously. The operational sex ratio (OSR), and the number, quality, and distance between available spawning habitats will be important factors driving the complexity of the relationship observed in the juvenile cohorts. These factors are all potential drivers of recruitment patterns in Lake Sturgeon. Additionally, the high contribution of just a few parents (the n50 estimates) as estimated in this study could result in pulses of cohorts over the years which may act in conjunction with, or in opposition to, natural environmental factors. Certainly, erratic recruitment patterns of Lake Sturgeon in Boreal Shield rivers appear to be the contemporary normal (McDougall et al. 2014). Life history traits (frequency, size, and condition dependent), that alternately attenuate and exacerbate recruitment patterns, particularly in depressed populations, likely include: female and male sexual dimorphism (age at maturity, longevity and size), female and male reproductive intervals (3-6 years and 2-3 years, respectively) and lastly, polygynandry. Having a polygynandrous mating system that generates multiple paternities has many advantages and disadvantages for long-lived and endangered species like Lake Sturgeon, the most important being that it will slow down the manifestation of inbreeding compared to a monogamous mating system. Furthermore, it is unlikely that exploited Lake Sturgeon populations would ever manifest decreased genetic diversity prior to being extirpated, due to the life history traits above. Conversely, with mating swarms it is easy to think that mating and reproductive success is random in Lake Sturgeon, but the last three decades of parentage analysis in the wild has revealed very different scenarios (reviewed in Parker and Birkhead 2013). Polyandry and multiple paternity generates patterns of variance in male and female reproductive success, such as the observed reproductive contribution of both sexes in our data (Table 7), and is expected to reduce the effective population size (Karl 2008). Consequently, probabilities and confidence intervals of estimates are necessary uncertainties associated with the number of spawners estimated herein.

4.3 LAKE STURGEON BIOLOGY AND BIOGEOGRAPHY

Suppositions regarding Lake Sturgeon ecology have given rise to questionable conclusions about the species demographic and biogeographic patterns, as well as the impediments to population recovery. SNP marker development has afforded the level of genetic resolution necessary to conduct population assignment tests, yielding a significant contribution to the biological understanding Lake Sturgeon populations, and ultimately providing strong evidence of the role of spatial habitat in determining biogeographic patterns and phylogeographic processes.

The first supposition refuted by the current study stems from Auer's (1996) paper, which stresses the importance of migration to sturgeons. Migration distance of anadromous sturgeon species was correlated with mean body size, yielding the conclusion that Lake Sturgeon populations (which are potamodromous) would require barrier free sections of lake/river habitat, at least 250 - 300 km and perhaps 750 – 1000 km in length (see Figure 2 Auer 1996). Broadly accepted by the scientific community, sweeping generalizations have followed; for example: "*Lake Sturgeon undergo extensive migrations for spawning, and large open distances (250 to 1000 km) are needed to support self-sustaining populations* (Auer 1996)" (DFO January 2014).

Caution must be taken when reading Auer's paper (1996) that lacked the integration of genetics and the role of habitat (driven by geomorphology and hydrology) in potential delineation of populations in larger riverine systems outside of the Great Lakes area. In recent years, two studies have revealed evidence of historically structured populations in both the Nelson (Côté et al. 2011) and Winnipeg rivers (McDougall 2011). The results of the current study provide strong evidence for a complex and historically structured meta-population and more specifically, no evidence of migrations spanning the entire length of the river.

The second supposition (related to the first) is that Lake Sturgeon populations in the Nelson River have been fragmented by hydroelectric development, described in a DFO report as follows: "*The construction of hydroelectric dams, beginning in 1960, fragmented the distribution of Lake Sturgeon and isolated the species into a series of reservoirs, particularly between Kettle and Limestone generating stations*… (DFO 2014)" with the implicit assumption that restricted gene flow has also resulted. In fact, the results from genetic studies on the Nelson River (Côté et al. 2011; present study) suggest populations were naturally restricted to one-way gene flow by geomorphic control points at Kelsey GS (Grand Rapid 6.1 m drop [Denis and Challies 1916]) and Kettle Rapids (23.8 m drop [Denis and Challies 1916]). Furthermore, there is no

documented evidence that Lake Sturgeon were abundant year-round in the Nelson River between Kettle Rapids and Long Spruce Rapids, a reach that dropped ~52 m over 32 km (Denis and Challies 1916), as the river transitioned from the Boreal Shield ecozone to Hudson Plain ecozone (Figure 15); it seems unlikely that a self-sustaining Lake Sturgeon population could have existed in this reach. The results of the current study provide unprecedented discriminatory power for Lake Sturgeon and do not reveal the presence of any transitional genotypes, indicating gene flow between Gull Lake and the lower Nelson River is a one way trip in the downstream direction. Furthermore, the results of the current study indicate that there has been minimal upstream to downstream gene flow between the Gull Lake and Lower Nelson populations historically (adult data) and contemporarily (adult and juvenile data). When one considers the historic gradient profile of the Nelson River from Gull Rapids through Long Spruce Rapids (Denis and Challies 1916), it is likely that the impoundment of the Nelson River post-Kettle GS (i.e., the creation of Stephens Lake) has increased the population level habitat suitability downstream of Gull Rapids, and allowed development of a Lake Sturgeon stock, which is notably comprised of the Gull Lake genotype.

The most parsimonious explanation for structuring in Lake Sturgeon populations occurring in large riverine and dendritic systems (and likely elsewhere) is that it is controlled by the spawn-drift-settle-establish habitat sequence, ultimately resulting in spatial isolation of populations. Essentially, Lake Sturgeon spawn, eggs hatch, and larvae disperse downstream with the distance dictated by water velocity gradients. In Great Lakes tributaries, this might mean distances exceeding 40 km (Auer and Baker 2002), but in stepped gradient Boreal Shield systems, where falls/rapids transition quickly into natural lakes (or man-made reservoirs), the distances may be less. Juveniles show a tendency to resist downstream redistribution, residing in discrete sections of river (Barth et al. 2011; McDougall et al. 2013). When maturity is reached, adults move back to spawning locations (falls, rapids, or tributaries). The cycle continues, albeit at a very slow pace due to the life history characteristics of the species (late age of maturity, infrequent spawning intervals, etc.), ultimately giving rise to structured and locally adapted populations.

The spawn-drift-settle-establish rationale is evident in GUL-STE-LOS sibship analysis results for the 2008 cohort which indicated that 2008 cohort juveniles in STE and LOS were likely not spawned in those locations, as each individual sequenced was determined to have at least one sibling located further upstream in GUL (Figure 13). Downstream redistribution of larval Lake Sturgeon must have occurred following spawning at an upstream location. Conversely, in the moderate-gradient Hudson Plain habitat of the

lower Nelson River, where larvae generated at multiple spawning sites seem to only establish in the a couple areas (i.e., Nelson River mainstem downstream of Angling River or the estuary), low levels of genetic differentiation among adults captured at the Lower Limestone Rapids, Angling River, and Weir River spawning sites were observed. Indeed, the differences among adults captured at each of these spawning sites are so small that their biological significance is questionable; for all intents and purposes, the results of various clustering analyses described herein considered Lake Sturgeon from the lower Nelson River to be a single population. Field samples were ordered in Figures 12 and 13 from left to right, producing horizontal and vertical patterning when siblings are in sequence, providing evidence that siblings are captured in the field together. Combine this with the fact that adults captured in spawning grounds are often recaptured together several years later, supports the hypothesis that Lake Sturgeon may spend their lives in closely related 'groups' with inbreeding mitigated by a combination of asynchronous mating behavior and longevity.

4.4 MANAGEMENT CONSIDERATIONS

The results of this study support the previous conclusions on population genetic structure (Coté et al. 2011) and have implications for conservation and recovery planning. The following should be considered with regards to recovery and mitigation strategies:

- Management Units should be delineated based on spawn-drift-settle-establish habitat sequences. The length of unrestricted reaches varies naturally, and is dependent on instream habitat specific to individual locations; and
- Strategies that involve stocking should give priority to maintaining genetic integrity and trajectory patterns, as the neutral markers and markers under selection indicate the differences are both real and adaptive.

Nine 'groups' exist for consideration (Figure 16) in developing stocking programs that would maintain the 'status quo' for northern Manitoba Lake Sturgeon populations:

- Jenpeg - presumably maintained by natal philopatry (i.e., fidelity and/or retention) with contemporary exchange with Landing River area;
- Landing River - presumably maintained by natal philopatry (i.e., fidelity and/or retention) with one-way transfer to populations downstream of Kelsey GS;
- Kelsey/Grass River - presumably maintained by natal philopatry exhibiting immigration from Landing River area and contemporary exchange with Burntwood,

and limited exchange with Gull Lake, despite the absence of barriers with evidence of rare transfer to populations downstream of Kettle GS;

- Burntwood River - presumably maintained by natal philopatry exhibiting contemporary exchange with Kelsey/Grass and little historical exchange with Gull Lake, despite absence of barriers with no evidence of transfer to populations downstream of Kettle GS;
- Gull/Stephens Lakes - presumably maintained by natal philopatry exhibiting little historical exchange with Kelsey/Burntwood despite absence of true barriers, limited historical downstream exchange with lower Nelson River, and limited contemporary transfer to populations downstream of Limestone GS. The contemporary transfer from GUL-STE-LOS provides some evidence that LOS juveniles are immigrants from upstream;
- Lower Nelson River - high mobility, natal philopatry appears to be more plastic with extensive exchange among lower Nelson River groups;
- Lower Hayes/Gods rivers - high mobility, natal philopatry possibly more plastic with extensive exchange between lower Hayes and Gods groups facilitated through larval drift stage;
- Fox river - high mobility, natal philopatry possibly more plastic with limited contemporary exchange with Hayes and Gods group; and
- Lower Churchill River - isolated by historical biogeographic pattern of watershed isolation starting 7500 years ago during the formation of the Tyrrell Sea (300 generations ago).

From a population genetic structure perspective, results from the current study provide strong evidence that hydroelectric development on the Nelson River has not fragmented a previously panmictic population, as the population structure revealed using GBS methods cannot be attributed to contemporary processes. Upstream gene flow appears to have been historically restricted at Kelsey GS and Kettle GS. Samples analyzed from the Long Spruce Reservoir (as well as Stephens Lake) are consistent with the Gull Lake genetic signature indicating contemporary gene flow from GUL to LOS, a pattern that is not evident when looking at historical structure i.e., GUL and LNR adult genetics indicate upstream to downstream gene flow was very rare. In addition, it also appears that the vast majority of Lake Sturgeon that previously utilized the now backwatered Nelson River mainstem (i.e., Long Spruce and Limestone reservoirs) likely were displaced downstream into the lower Nelson River following impoundment.

5.0 REFERENCES

ALCALA, N., J. GOUDET, and S. VUILLEUMIER. 2014. On the transition of genetic differentiation from isolation to panmixia: What we can learn from Gst and *D*. Theoretical Population Biology. 93: 75–84.

ALJANABI, S.M., and I. MARTINEZ. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Research. 25(22): 4692–4693.

AMBROSE, K.M., C.A. MCDOUGALL, P.A. NELSON, L. MURRAY, and D.S. MACDONELL. 2007. Results of the 2005 fish community investigations focusing on lake sturgeon in the Conawapa study area. A report prepared for Manitoba Hydro by North/South Consultants Inc., Winnipeg, Manitoba. 176 pp. #5613.05-08.

AUER, N.A. 1996. Importance of habitat and migration to sturgeons with emphasis on lake sturgeon. Canadian Journal of Fisheries and Aquatic Sciences. 53(1): 152-160.

AUER, N.A. and E.A. BAKER. 2002. Duration and drift of larval lake sturgeon in the Sturgeon River, Michigan. Journal of Applied Ichthyology. 18: 557-564.

BAIRD, N.A., P.D. ETTER, T.S. ATWOOD, M.C. CURREY, A.L. SHIVER, Z.A. LEWIS, E.U. SELKER, W.A. CRESKO, and E.A. JOHNSON. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3:e3376.

BARTH, C.C., W.G. ANDERSON, L.M. HENDERSON, and S.J. PEAKE. 2011. Home range size and seasonal movement of juvenile lak esturgeon in a large river in the Hudson Bay drainage basin. Transactions of the American Fisheries Society. 140(6): 1629-1641.

BIRSTEIN, V.J., and R. DeSALLE. 1998. Molecular phylogeny of Acipenserinae. Molecular Phylogenetics and Evolution. 9(1): 141–155. doi:10.1006/mpev.1997.0443.

BIRSTEIN, V.J., W.E. BEMIS, and J.R. WALDMAN. 1997. The threatened status of acipenseriform species: a summary. Environmental Biology of Fishes. 48(1-4): 427–435. doi:10.1023/A:1007382724251.

BLACKLIDGE, K.H., and C.A. BIDWELL. 1993. Three Ploidy Levels Indicated by Genome Quantification in Acipenseriformes of North America. Journal of Heredity. 84: 427–430.

BRUCH, R.M., and F.P. BINKOWSKI. 2002. Spawning behavior of lake sturgeon (*Acipenser fulvescens*). Journal of Applied Ichthyology. 18(4-6): 570–579. doi:10.1046/j.1439-0426.2002.00421.x

CATCHEN, J.M., P.A. HOHENLOHE, S. BASSHAM, A. AMORES, and W.A. CRESKO. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology. 22(11): 3124–3140. doi:10.1111/mec.12354

COSEWIC. 2006. COSEWIC assessment and update status report on the lake sturgeon, Acipenser fulvescens, in Canada. Committee on the Status of Endangered Wildlife in Canada. Ottawa. xi + 107 pp.

CÔTÉ, G., P.A. NELSON, and L. BERNATCHEZ. 2011. Population genetics of lake sturgeon from northern Manitoba. A report prepared for Manitoba Hydro by Université Laval and North/South Consultants Inc., Winnipeg, Manitoba. 67p. #5671.08-08.

DAVEY, J.W., P.A. HOHENLOHE, P.D. ETTER, J.Q. BOONE, J.M. CATCHEN, and M.L. BLAXTER. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics. 12(7): 499–510. doi:10.1038/nrg3012

DENIS, L.G. and J.B. CHALLIES. 1916. Water powers of Manitoba, Saskatchewan, and Alberta. Commission of Conservation Canada. Warwick Bro's & Rutter Limited, Toronto, Canada. 334p.

DFO 2014. Review of lake sturgeon analyses for the proposed Keeyask generating station. Canadian Science Advisory Secretariat Science Response 2014/008.

DRAUCH SCHREIER, A., D. GILLE, B. MAHARDJA, and B. MAY. 2011. Neutral markers confirm the octoploid origin and reveal spontaneous autopolyploidy in white sturgeon, *Acipenser transmontanus*. Journal of Applied Ichthyology. 27: 24–33.

ELSHIRE, R.J., J.C. GLAUBITZ, Q. SUN, J.A. POLAND, K. KAWAMOTO, E.S. BUCKLER, and S.E. MITCHELL. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. (R.J. Elshire, J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S. Buckler, & S.E. Mitchell, Eds.) PLoS ONE. 6(5): e19379. doi:10.1371/journal.pone.0019379.g006

EXCOFFIER, L., P.E. SMOUSE, and J.M. QUATTRO. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 131(2): 479–491.

FERGUSON, M.M., and G.A. DUCKWORTH. 1997. The status and distribution of lake sturgeon, *Acipenser fulvescens*, in the Canadian provinces of Manitoba, Ontario and Quebec: a genetic perspective. Environmental Biology of Fishes. 48(1-4): 299–309. doi:10.1023/A:1007367818353.

FONTANA, F., R.M. BRUCH, F.P. BINKOWSKI, M. LANFREDI, M. CHICCA, N. BELTRAMI, and L. CONGIU. 2004. Karyotype characterization of the lake sturgeon, Acipenser fulvescens (Rafinesque 1817) by chromosome banding and fluorescent in situ hybridization. Genome. 47(4): 742–746. doi:10.1139/g04-028

GAGNAIRE, P.A., T. BROQUET, D. AURELLE, F. VIVARD, A. SOUISSI, F. BONHOMME, S. ARNAUD-HAOND, and N. BIERNE. 2015. Using neutral, selected and hitchhiker loci to assess connectivity of marine populations in the genomic era. Evolutionary Applications. 8(8): 1–35. doi:10.1111/eva.12288.

GOSSELIN, T., and L. BERNATCHEZ. 2015. Stackr: a R Package for the Analysis of GBS/RAD Data. doi: 10.5281/zenodo.30371

GOSSELIN, T., and L. BERNATCHEZ. 2015. Genotype-by-Sequencing analysis in the Cloud. doi: 10.5281/zenodo.30271

GOUDET, J., M. RAYMOND, T. DE MEEÛS, and F. ROUSSET. 1996. Testing differentiation in diploid populations. Genetics. 144: 1933–1940.

HARVEY, M.G., C.D. JUDY, G.F. SEEHOLZER, J.M. MALEY, G.R. GRAVES, and R.T. BRUMFIELD. 2015. Similarity thresholds used in short read assembly reduce the comparability of population histories across species. PeerJ. 3: e895. doi:10.7717/peerj.895

HINDAR, K., N RYMAN, and F. UTTER. 1991. Genetic effects of cultured fish on natural fish populations. Canadian Journal of Fisheries and Aquatic Sciences. 48(5): 945-957.

HOHENLOHE, P.A., S. BASSHAM, P.D. ETTER, N. STIFFLER, E.A. JOHNSON, and W.A. CRESKO. 2010. Population Genomics of Parallel Adaptation in Threespine Stickleback using Sequenced RAD Tags. (P.A. Hohenlohe, S. Bassham, P.D. Etter, N. Stiffler, E.A. Johnson, & W.A. Cresko, Eds.) PLoS Genetics. 6(2): e1000862. doi:10.1371/journal.pgen.1000862.t003

HOHENLOHE, P.A., M.D. DAY, S.J. AMISH, M.R. MILLER, N. KAMPS-HUGHES, M.C. BOYER, C.C. MUHLFELD, F.W. ALLENDORF, E.A. JOHNSON, and G. LUIKART. 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end RAD sequencing. Molecular Ecology. 22(11): 3002-3013 doi: 10.1111/mec.12239.

HOUSTON, J.J. 1987. Status of the Lake Sturgeon, *Acipenser fulvescens*, in Canada. The Canadian Field-Naturalist. 101(2): 171–185.

HUANG, H., and L.L. KNOWLES. 2014. Unforeseen Consequences of Excluding Missing Data from Next-Generation Sequences: Simulation Study of RAD Sequences. Systematic Biology, DOI: 10.1093/sysbio/syu046.

ILUT D.C., M.L. NYDAM, and M.P. HARE. 2014. Defining Loci in Restriction-Based Reduced Representation Genomic Data from Nonmodel Species: Sources of Bias and Diagnostics for Optimal Clustering. BioMed Research International. 2014, 1–9.

ISHWARAN, H. 2014. The effect of splitting on random forests. Machine Learning. 99: 75–118.

ISHWARAN, H., and U.B. KOGALUR. 2015. Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.6.1.

JARQUÍN , D., K. KOCAK, L. POSADAS, K. HYMA, J. JEDLICKA, G. GRAEF, and A. LORENZ. 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC genomics. 15(1): 740.

JOMBART, T., and I. AHMED. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 27**:** 3070–3071.

JOMBART, T., S. DEVILLARD, and F. BALLOUX. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics. 11(1): 94.

JONES, O.R., and J. WANG. 2010. COLONY: a program for parentage and sibship inference from multilocus genotype data. Molecular Ecology Resources. 10(3): 551–555. doi:10.1111/j.1755-0998.2009.02787.x

JOOP OUBORG, N., F. ANGELONI, and P. VERGEER. 2009. An essay on the necessity and feasibility of conservation genomics. Conservation Genetics. 11(2): 643–653. doi:10.1007/s10592-009-0016-9

JOST, L. 2008. G_{ST} and its relatives do not measure differentiation. Molecular Ecology. 17(18): 4015-4026.

KARDOS, M., G. LUIKART, and F.W. ALLENDORF. 2015. Measuring individual inbreeding in the age of genomics: marker-based measures are better than pedigrees. Heredity. 115: 63–72.

KARL, S. 2008. The effect of multiple paternity on the genetically effective size of a population. Molecular Ecology. 17(18): 3973–3977.

KELEHER, J.J., and B. KOOYMAN. 1958. Supplement to Hinks "The fishes of Manitoba." Department of Mines and Natural Resources, Manitoba, Canada.

KELLER, M.C., P.M. VISSCHER, and M.E. GODDARD. 2011. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics. 189: 237–249.

LAHAYE, M., A. BRANCHAUD, M. GENDRON, R. VERDON, and R. Fortin. 1992. Reproduction, early life history, and characteristics of the spawning grounds of the lake sturgeon (*Acipenser fulvescens*) in Des Prairies and L'Assomption rivers, near Montréal, Quebec. Canadian Journal of Zoology. 70(9): 1681–1689. doi:10.1139/z92-234

LUDWIG, A., N.M. BELFIORE, C. PITRA, V. SVIRSKY, and I. JENNECKENS. (2001). Genome Duplication Events and Functional Reduction of Ploidy Levels in Sturgeon (*Acipenser*, *Huso* and *Scaphirhynchus*). Genetics. 158(3): 1203–1215.

MANEL, S., O. GAGGIOTTI, and R. WAPLES. 2005. Assignment methods: matching biological questions with appropriate techniques. Trends in Ecology and Evolution. 20(3): 136–142. doi:10.1016/j.tree.2004.12.004

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17: 10–12.

MCDOUGALL, C.A. 2011. Investigating downstream passage of lake sturgeon, *Acipenser fulvescens*, through a Winnipeg River generating station. M.Sc. Thesis. University of Manitoba, Winnipeg, Manitoba. x + 175 pp.

MCDOUGALL, C.A., C.C. BARTH, J.K. AIKEN, L.M. HENDERSON, M.A. BLANCHARD, K.M. AMBROSE, C.L. HRENCHUK, M.A. GILLESPIE, and P. NELSON. 2014. How to sample juvenile Lake Sturgeon (*Acipenser fulvescens*, Rafinesque 1817), in Boreal Shield rivers using gill nets, with an emphasis on assessing recruitment patterns. Journal of Applied Ichthyology. 30(6): 1402-1415.

MCDOUGALL, C.A., P.J. BLANCHFIELD, S.J. PEAKE, and W.G. ANDERSON. Movement patterns and size-class influence entrainment susceptibility of Lake Sturgeon in a small hydroelectric reservoir. Transactions of the American Fisheries Society. 142(6): 1508-1521.

MCLEAN, M.F., C.A. SIMPFENDORFER, M.R. HEUPEL, M.J. DADSWELL, and M.J.W. STROKESBURY. 2014. Diversity of behavioural patterns displayed by a summer feeding aggregation of Atlantic sturgeon in the intertidal region of Minas Bay, Bay of Fundy, Canada. Marine Ecology Progress Series. 496: 59-69.

MCQUOWN, E., C.A., GALL, and B. MAY. 2002. Characterization and inheritance of six microsatellite loci in Lake Sturgeon. Transactions of the American Fiosheries Society. 131(2): 299-307.

MEIRMANS, P.G. 2012. AMOVA-based clustering of population genetic data. Journal of Heredity. 103(5): 744-750.

MEIRMANS, P.G. 2014. Nonconvergence in Bayesian estimation of migration rates. Molecular Ecology Resources. 14: 726–733.

MEIRMANS, P.G., and P.W. HEDRICK. 2011. Assessing population structure: F(ST) and related measures. Molecular Ecology Resources. 11(1): 5–18. doi:10.1111/j.1755-0998.2010.02927.x

MEIRMANS, P.G., and P.H. VAN TIENDEREN. 2004. Genotype and genodive: two programs for the analysis of genetic diversity of asexual organisms. Molecular Ecology Notes. 4(4): 792–794. doi:10.1111/j.1471-8286.2004.00770.x

MICHALAKIS, Y., and L. EXCOFFIER. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics. 142(3): 1061–1064.

NARUM, S.R., C.A. BUERKLE, J.W. DAVEY, M.R. MILLER, and P.A. HOHENLOHE. 2013. Genotyping-by-sequencing in ecological and conservation genomics. Molecular Ecology. 22(11): 2841–2847. doi:10.1111/mec.12350

NEI, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 89(3): 583–590.

NEI M, and W.H. LI. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America. 76: 5269–5273.

PAETKAU, D., W. CALVERT, I. STIRLING, and C. STROBECK. 1995. Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology. 4: 347–354.

PAETKAU, D., L.P. WAITS, P.L. CLARKSON, L. CRAIGHEAD, and C. STROBECK. 1997. An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. Genetics. 147: 1943–1957.

PAETKAU, D., R. SLADE, M. BURDEN, and A. ESTOUP. 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Molecular Ecology. 13(1): 55–65.

PARCHMAN, T.L., Z. GOMPERT, M.J. BRAUN, R.T. BRUMFIELD, D.B. MCDONALD, J.A.C. UY, G. ZHANG, E.D. JARVIS, B.A. SCHLINGER, and C.A. BUERKLE. 2013. The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. Molecular Ecology. 22(12): 3304-3317. doi:10.1111/mec.12201

PARKER, G.A., and T.R. BIRKHEAD. 2013. Polyandry: the history of a revolution. Philosophical Transactions of the Royal Society B: Biological Sciences. 368(1613): 20120335–20120335. doi:10.1098/rstb.2012.0335

PENONE C., A.D. DAVIDSON, K.T. SHOEMAKER, M. DI MARCO, C. RONDININI, T.M. BROOKS, B.E. YOUNG, C.H. GRAHAM, and G.C. Costa. 2014. Imputation of missing data in life-history trait datasets: which approach performs the best? (R Freckleton, Ed,). Methods in Ecology and Evolution. 5: 961–970.

PIKITCH, E.K., P. DOUKAKIS, L. LAUCK, P. CHAKRABARTY, and D.L. ERICKSON. 2005. Status, trends and management of sturgeon and paddlefish fisheries. Fish and fisheries. 6(3): 233–265. doi:10.1111/j.1467-2979.2005.00190.x

PYATSKOWIT, J.D., C.C. KRUEGER, H.L. KINCAID, and B. MAY. 2001. Inheritance of microsatellite loci in the polyploid lake sturgeon (*Acipenser fulvescens*). Genome. 44(2): 185–191.

PUJOLAR, J.M., L. ASTOLFI, E. BOSCARI, M. VIDOTTO, A. BRUSON, and L. CONGIU. 2013. Tana1, a new putatively active Tc1-like transposable element in the genome of sturgeons. Molecular Phylogenetics and Evolution. 66: 223–232.

PURCELL, S., B. NEALE, K. TODD-BRAUN, L. THOMAS, M.M. FERREIRA, D. BENDER, J. MALLER, P. SKLAR, P.I. DE BAKKER, M.J. DALY, and P.C. SHAM. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. American Journal of Human Genetics. 81(3): 559-575.

REISENBICHLER, R.R., and S.P. RUBIN. 1999. Genetic changes from artificial propagation of Pacific salmon affect the productivity and viability of supplemented populations. ICES Journal of Marine Science: Journal du Conseil. 56(4): 459-466.

RODZEN, J.A., and B. MAY. 2002. Inheritance of microsatellite loci in the white sturgeon (Acipenser transmontanus). Genome. 45(6): 1064-1076.

RUTKOSKI, J.E., J. POLAND, J.L. JANNINK, and M.E. SORRELLS. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3: Genes| Genomes| Genetics. 3(3): 427–439.

SEEB, L.W., W.D. TEMPLIN, S. SATO, S. ABE, K. WARHEIT, J.Y. PARK, and J.E. SEEB. 2011. Single nucleotide polymorphisms across a species' range: implications for conservation studies of Pacific salmon. Molecular Ecology Resources. 11(s1): 195–217. doi:10.1111/j.1755-0998.2010.02966.x

SHAH, A.D., J.W. BARTLETT, J. CARPENTER, O. NICHOLAS, and H. HEMINGWAY. 2014. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. American Journal of Epidemiology. 179: 764–774.

WANG, J. 2004. Sibship reconstruction from genetic data with typing errors. Genetics. 166(4): 1963-1979.

WANG, J. 2009. A new method for estimating effective population sizes from a single sample of multilocus genotypes. Molecular Ecology. 18(10): 2148–2164. doi:10.1111/j.1365-294X.2009.04175.x

WANG, J. 2013. A simulation module in the computer program colony for sibship and parentage analysis. Molecular Ecology Resources. 13(4): 734-739. doi:10.1111/1755-0998.12106.

WAPLES, R.S., and C. DO. 1994. Genetic risk associated with supplementation of Pacific salmonids: captive broodstock programs. Canadian Journal of Fisheries and Aquatic Sciences. 51(S1): 310-329.

WARD, R.D. 2006. The importance of identifying spatial population structure in restocking enhancement programs. Fisheries Research. 80(1): 9-18.

WEIR, B.S., and C.C. COCKERHAM. 1984. Estimating F-statistics for the analsus of population structure. evolution. 1: 1358-1370.

WELSH, A., M. BLUMBERG, and B. MAY. 2003. Identification of microsatellite loci in lake sturgeon, *Acipenser fulvescens*, and their variability in green sturgeon, *A. medirostris*. Molecular Ecology Notes. 3(1): 47–55. doi:10.1046/j.1471-8286.2003.00346.x

WELSH, A., and B. MAY. (2006). Development and standardization of disomic microsatellite markers for lake sturgeon genetic studies. Journal of Applied Ichthyology. 22(5): 337–344.

WELSH, A.B., and D.T. MCLEOD. 2010. Detection of natural barriers to movement of Lake Sturgeon (Acipenser fulvescens) within the Namakan River, Ontario. Canadian Journal of Zoology. 88(4): 390-397.

WELSH, A., T. HILL, H. QUINLAN, C. ROBINSON, and B. MAY. 2008. Genetic assessment of lake sturgeon population structure in the Laurentian Great Lakes. North American Journal of Fisheries Management. 28(2): 572–591.

WICKHAM, H. 2010. A Layered Grammar of Graphics. Journal of Computational and Graphical Statistics. 19: 3–28.

WICKHAM, H. 2011. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software. 40: 1–29.

WICKHAM H. 2014. Tidy data. Journal of Statistical Software, 59.

WOZNEY, K.M., T.J. HAXTON, S. KJARTANSON, and C.C. WILSON. 2010. Genetic assessment of lake sturgeon (*Acipenser fulvescens*) population structure in the Ottawa River. Environmental Biology of Fishes. 90(2): 183–195. doi:10.1007/s10641-010-9730-x

TABLES AND FIGURES

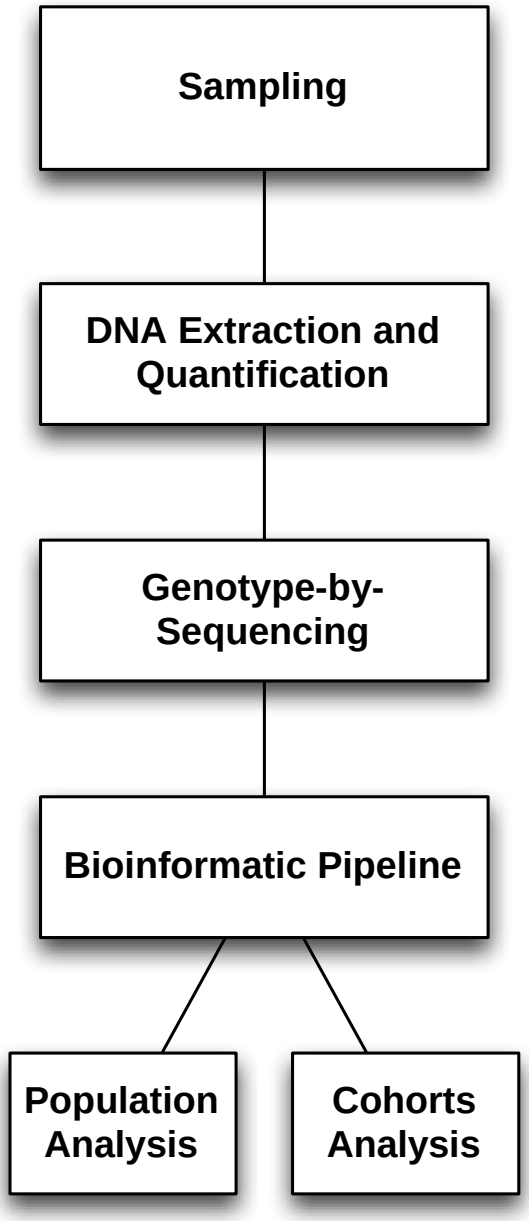
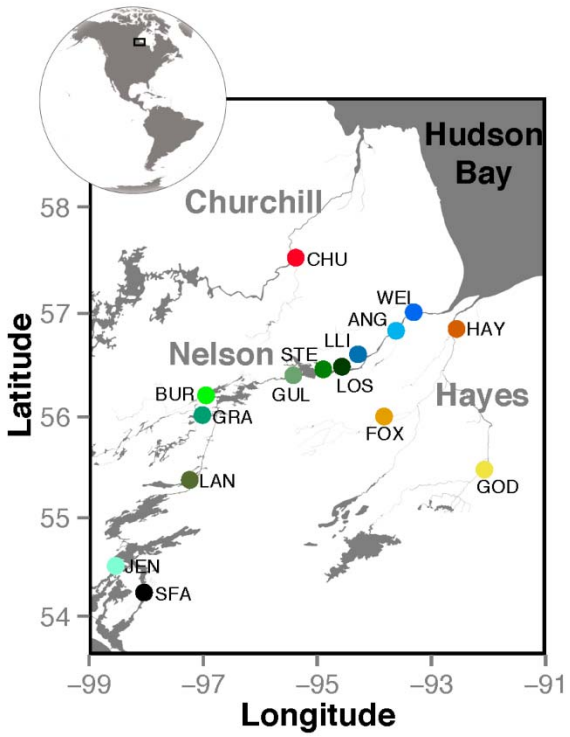


Figure 1. Flow chart of methodological steps involved in this study.



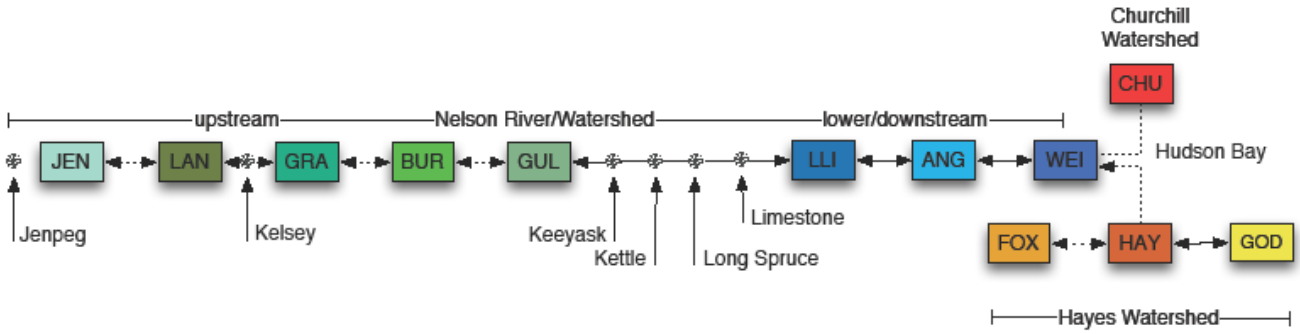


Figure 2. Map of Manitoba of sampling locations

Map and flow chart of Manitoba sampling locations showing the hierarchical structure of the potential populations in relation to the generating stations.

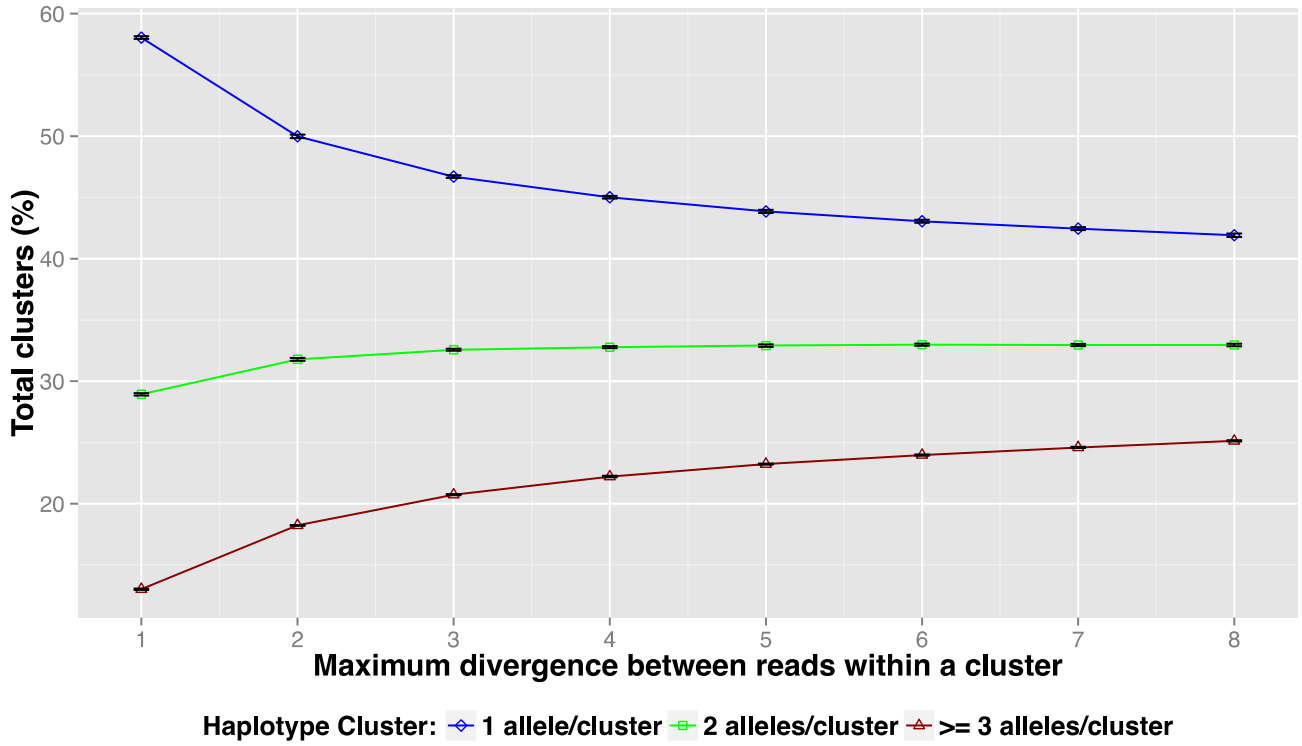


Figure 3. Clustering mismatch threshold series.

This is the -M parameter in the *ustacks* module of STACKS. The divergence is based on the sturgeons 80 pb reads. The Y-axis represents the percentage of total clusters at a given mismatch value and the X-axis represents the maximum differences (in mismatches) allowed between reads within a cluster. Single haplotype clusters (putative homozygous loci) are represented by a solid blue line and diamonds, two-haplotype clusters (putative heterozygous loci) are represented by a solid green line and squares, and three or more haplotype clusters (combined alleles from 2 or more paralogous loci) are represented by a solid red line and triangle.

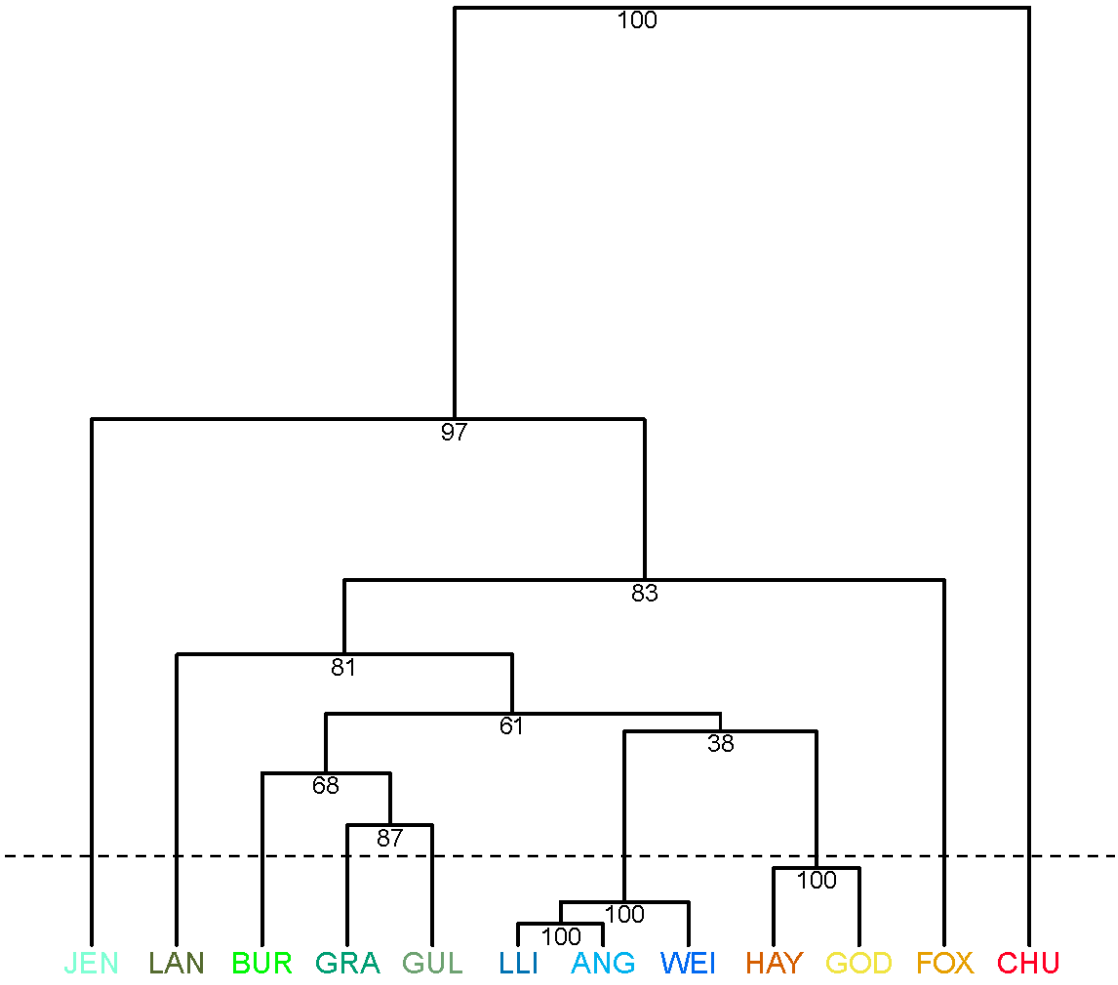


Figure 4. Distances between sampling sites using an UPGMA tree.

Nei's *G*st distance estimates visualized with a UPGMA tree that highlight splits between all sturgeons sampling sites in Manitoba. The measure includes a bias-correction for sampling a limited number of samples and populations (Meirmans & Hedrick, 2011). Only sites with adults were used for the pairwise distances. Bootstrap percentage values are below nodes (10 000 replicates). Dataset was imputed with individuals missing less than 50% genotypes.

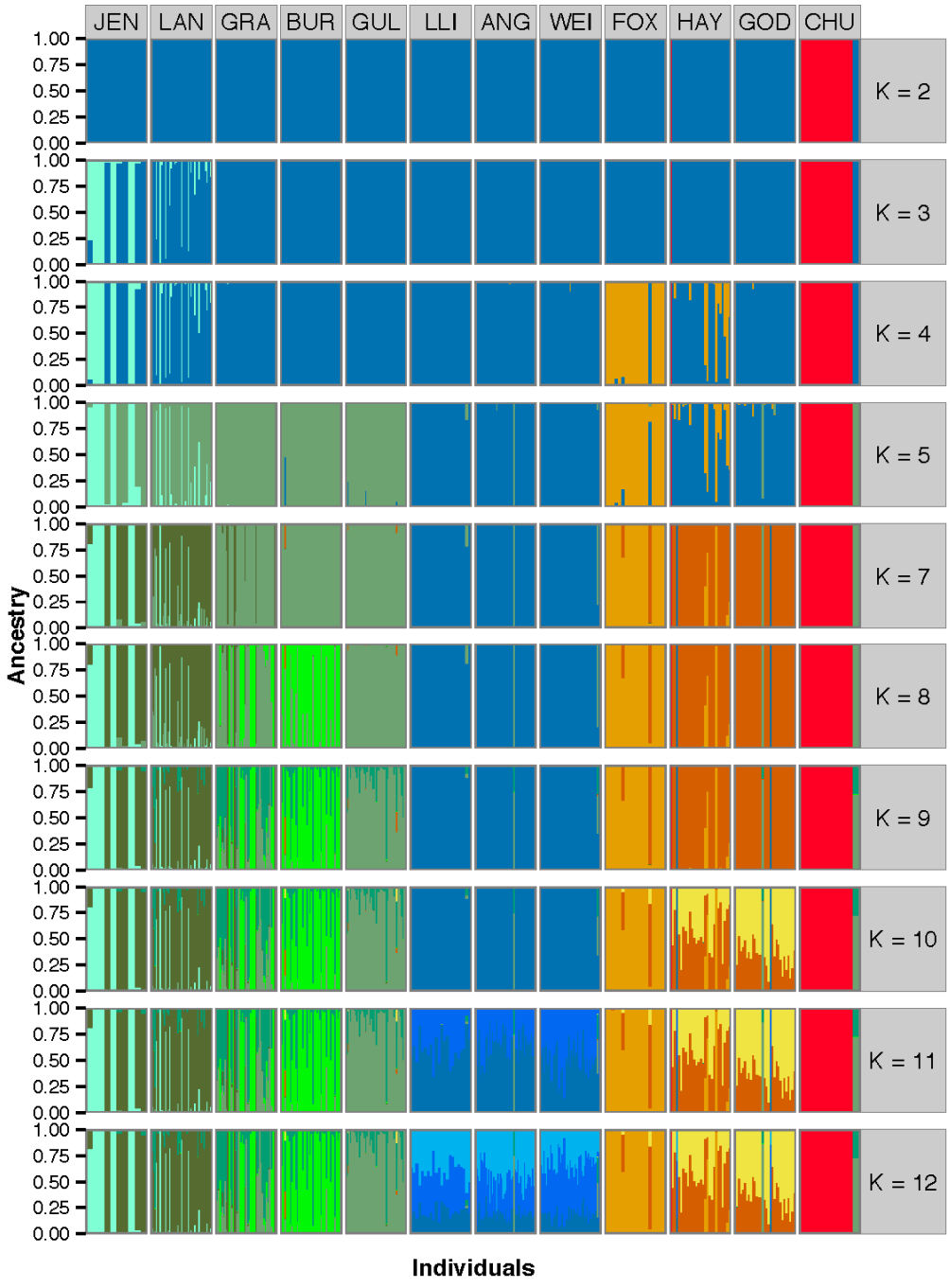


Figure 5. Discriminant Analysis of Principal Components (DAPC)

The DAPC figure shows the clustering of sampling sites for different K values. Dataset was imputed with individuals missing less than 30% genotypes. At K = 2, Churchill become a distinct cluster from the remaining Watersheds. At K = 3, Jenpeg, shows some distinctiveness. The Hayes watershed shows a unique signature at K = 4, with the Fox River (FOX). The different admixture proportions of the upper Nelson tributaries are revealed at K = 5 and after. For all values of K, downstream of the Limestone GS, there is no distinct population along the Nelson tributaries.

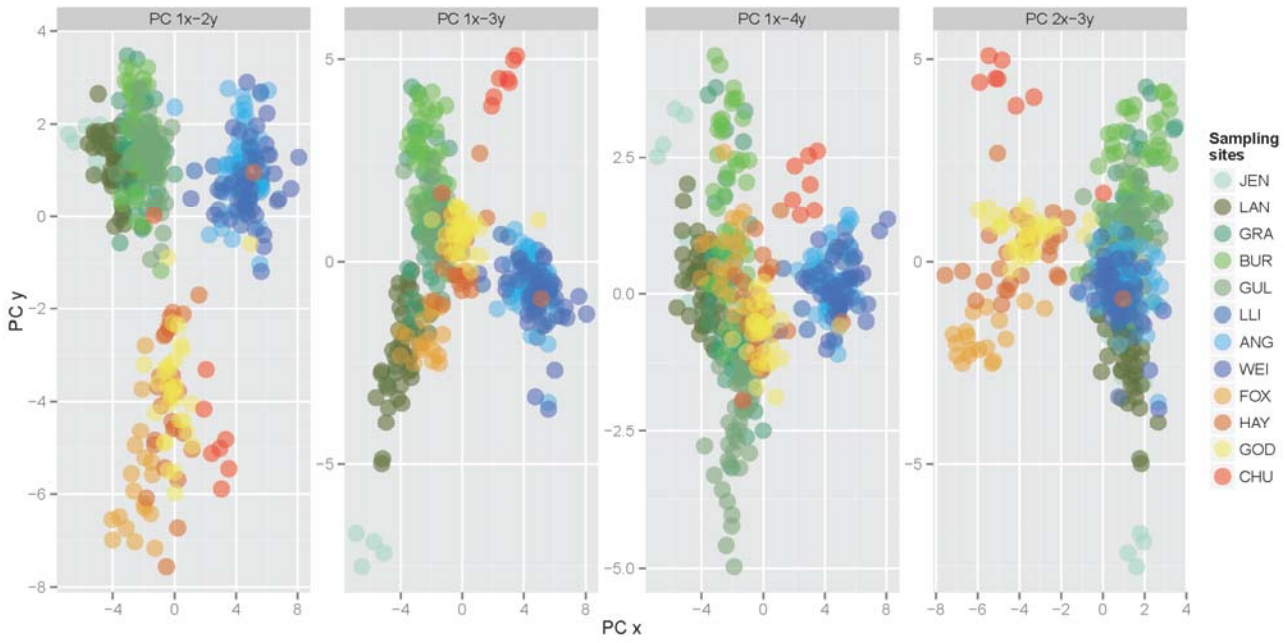


Figure 6. Scatter plot of Discriminant Analysis of Principal Components (DAPC)

The DAPC figure shows the clustering of sampling sites for a K = 12 with different combination of principal components (1 to 4). Circles represent the individuals' coordinate and colors represent the 12 sampling sites. Dataset was imputed with individuals missing less than 30% genotypes.

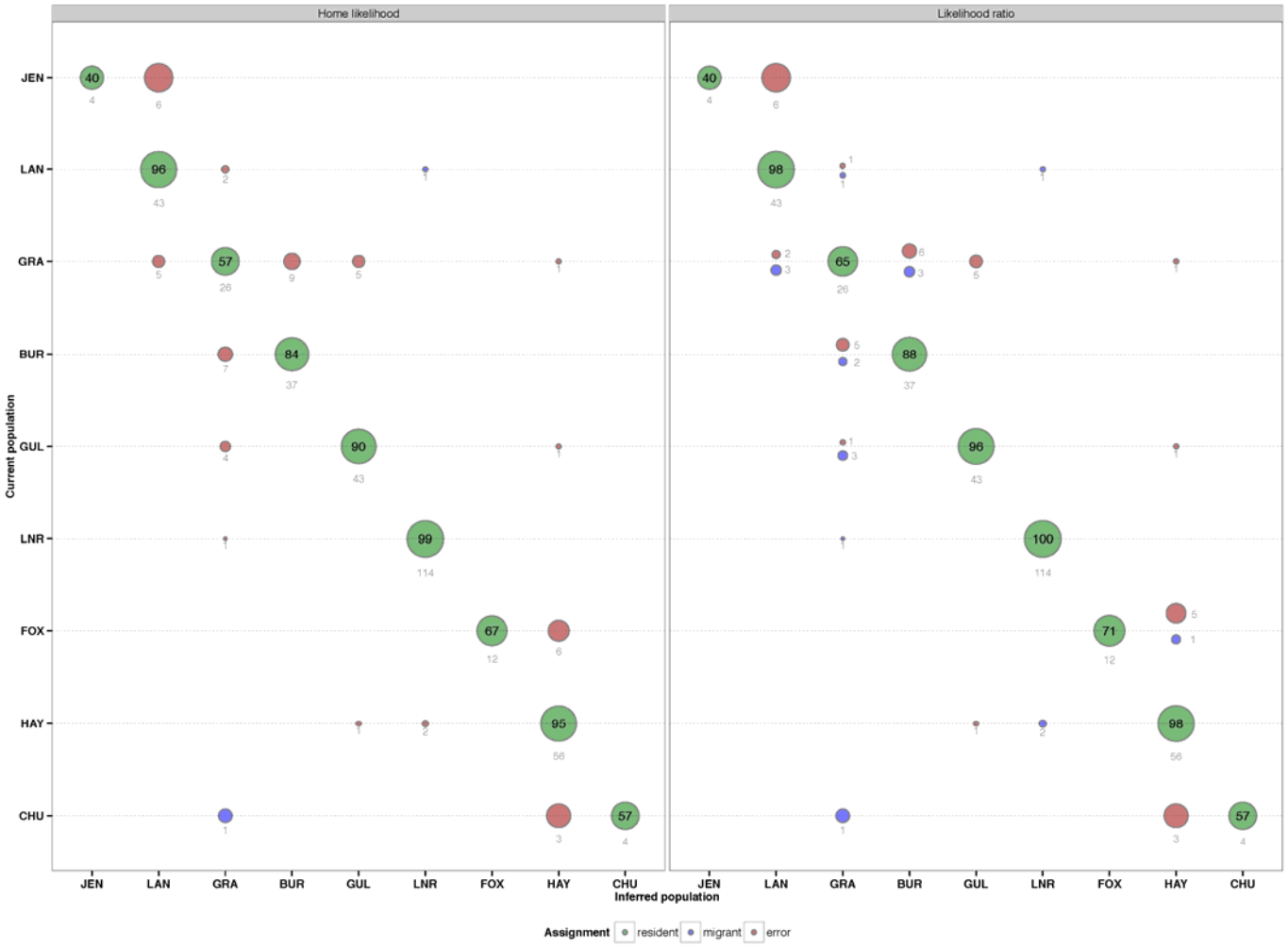


Figure 7. Assignment probabilities

Assignment probabilities of individuals based on home likelihood (Lh, left panel) and likelihood ratio (Lh/Lmax, right panel). Population discrimination power in black and number of individuals assigned in grey. Number of migrants is in blue and when current and inferred populations are identical or different, green or red are used, respectively. The overall discrimination power ranged from 77 to 85% with Lh and Lh/Lmax, respectively. When JEN and CHU are removed (sample size < 15) and HAY and GOD combined, the overall discrimination power > 90%. This assignment test used filtered and imputed data.

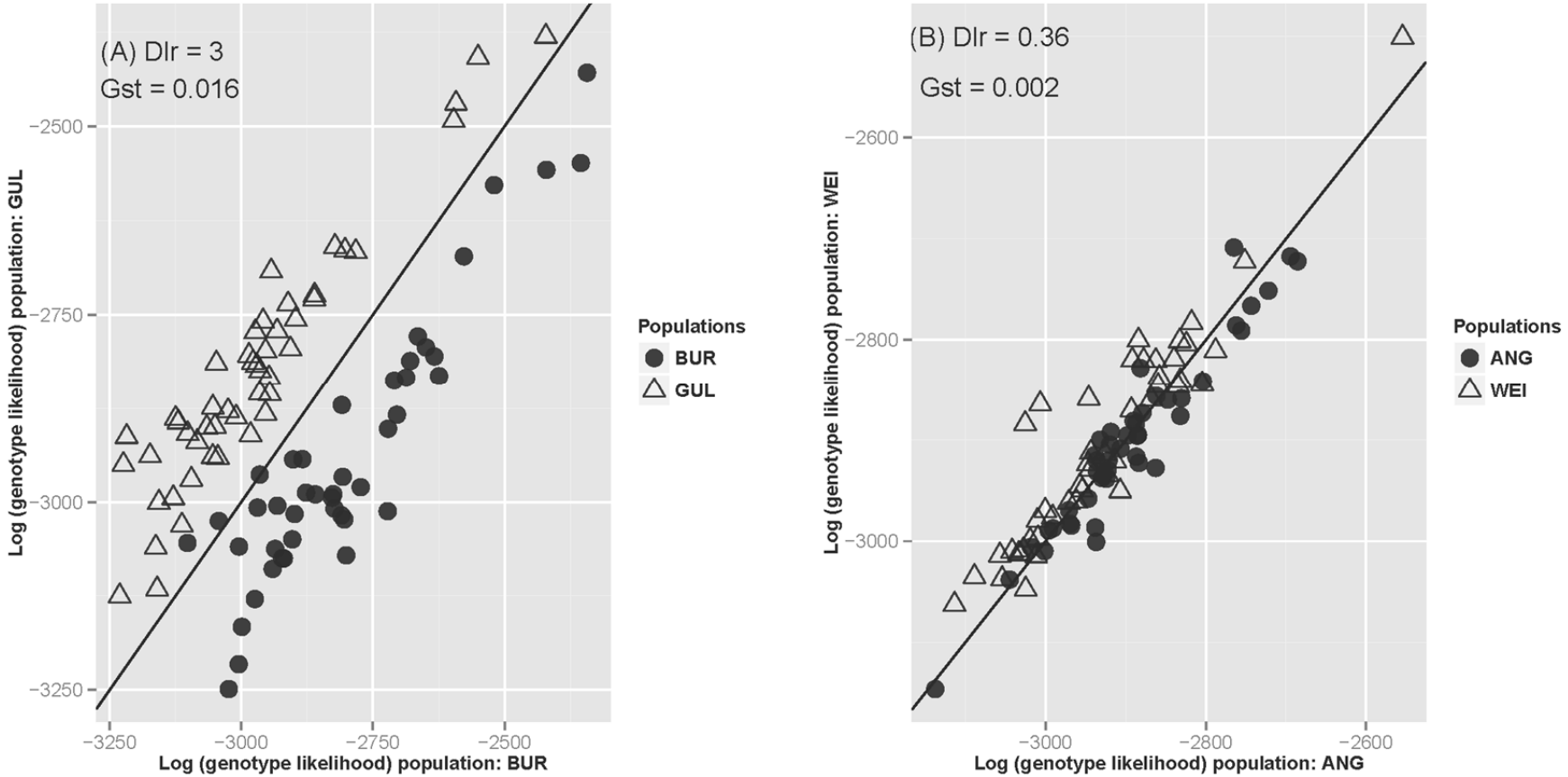


Figure 8. Assignment plot of BUR-GUL and ANG-WEI.

Assignment plot for two pairs of sampling sites: BUR-GUL (left) and ANG-WEI (right) where the power to identify F_0 immigrants is related to the mean genotype likelihood distance ratio (Dlr). DLR is the distance of individuals from the diagonal center line. Also shown on the figure, the G_{st} values.

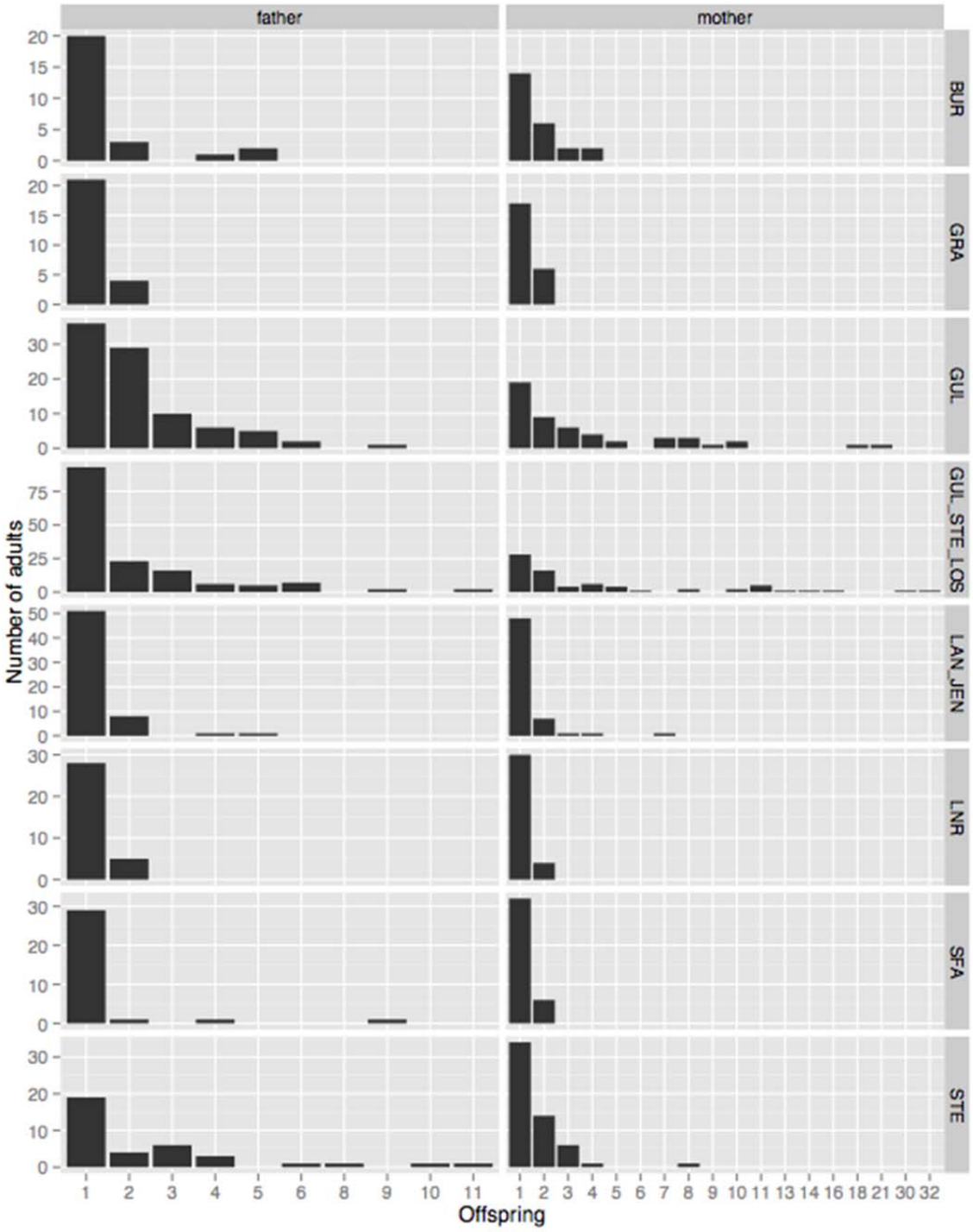


Figure 9. Reproductive success of putative parents

Histogram showing the number of adults (fathers and/or mothers) having one or more offspring. Results are by sampling site and sex. Sexes are arbitrarily set by COLONY and cannot be combined for this figure. Most mothers and/or fathers had one offspring in the sample while few have more than two offspring.

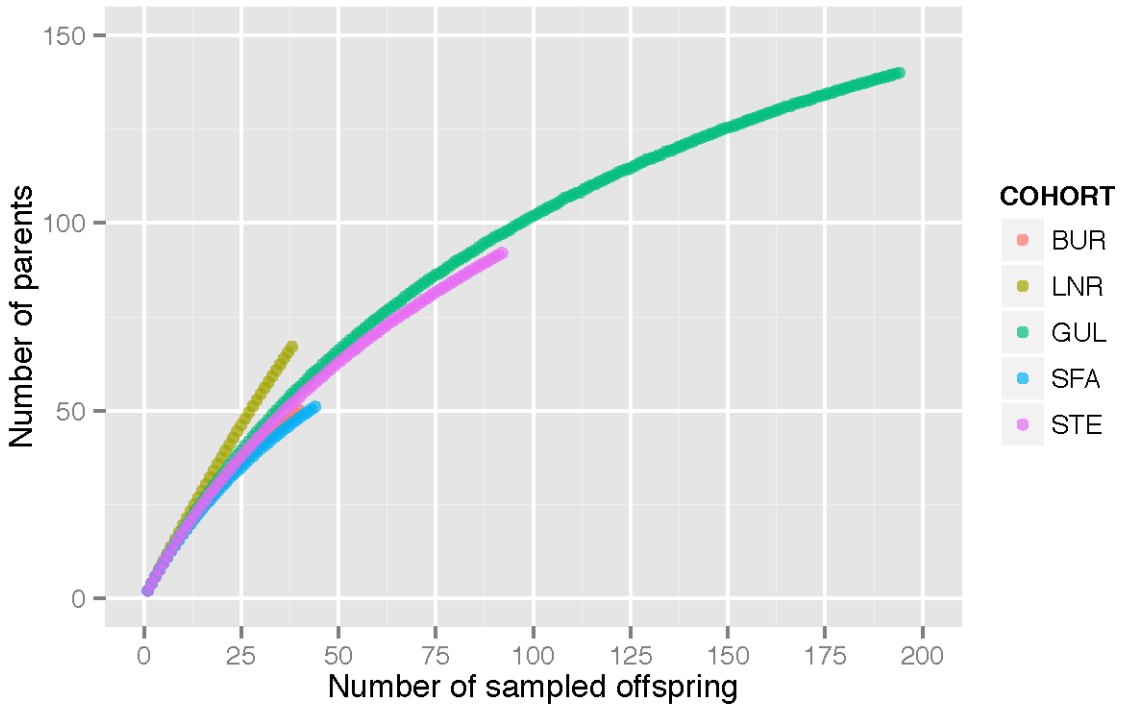


Figure 10. Simulations of discovery of the putative number of parents

Simulations of discovery of the number of unique parents influenced by the number of sampled offspring. Each point represent the mean bootstrap (n=999) on number of sampled offspring for 5 cohorts.

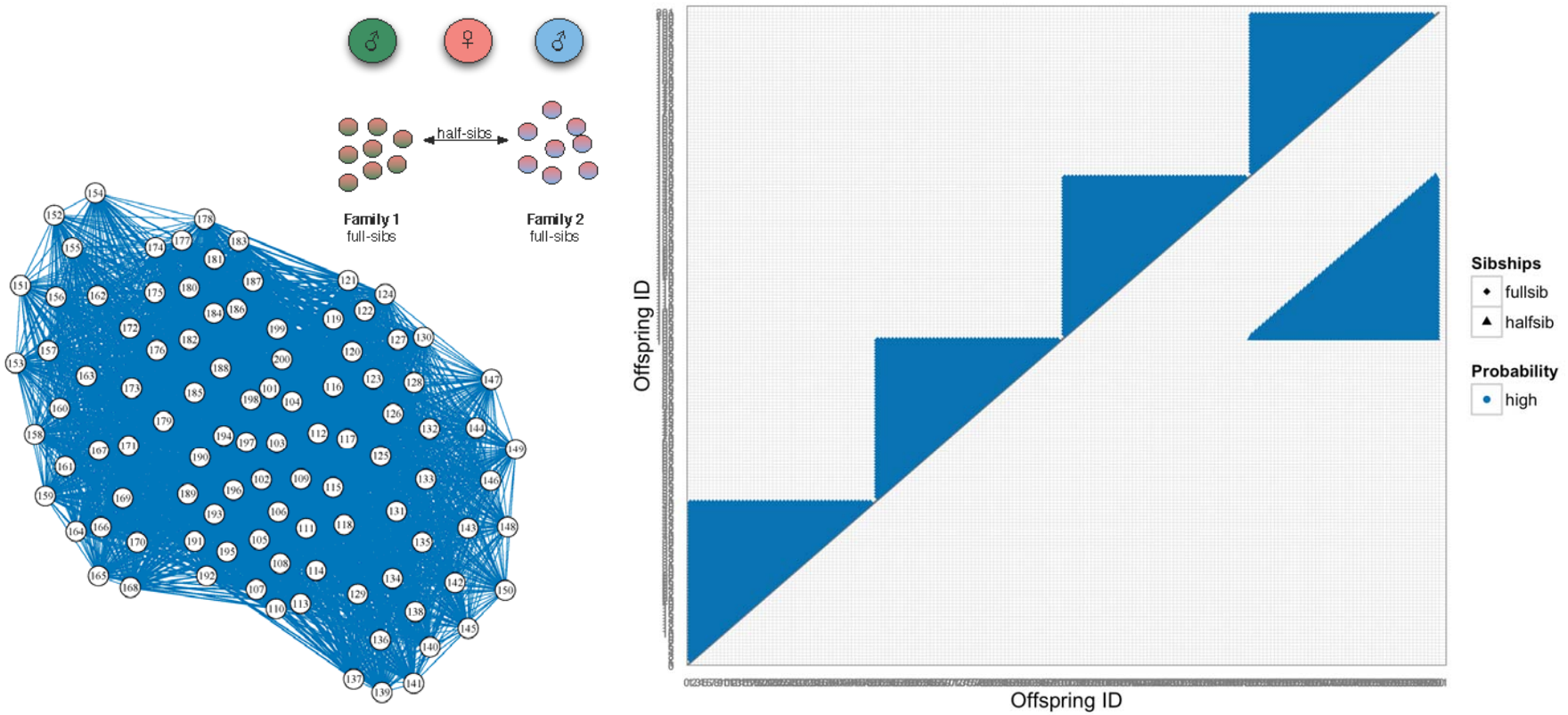


Figure 11. Network graph (left) and heat map (right) for reared progeny with known parents

Relationship network of two families (family 1: 151-200; family 2: 101-150) of offspring that share the same mother but different fathers (full-sibs within a family, half-sibs between families). Probability of relationship inferred by COLONY is divided in three groups and represented by color (blue: >= 0.75; gray: 0.20 to 0.75 and red: <= 0.20). Offspring relationship (full- and half-sib, diamond and triangle shape, above and below diagonal, respectively) and probability of relationship inferred by COLONY.

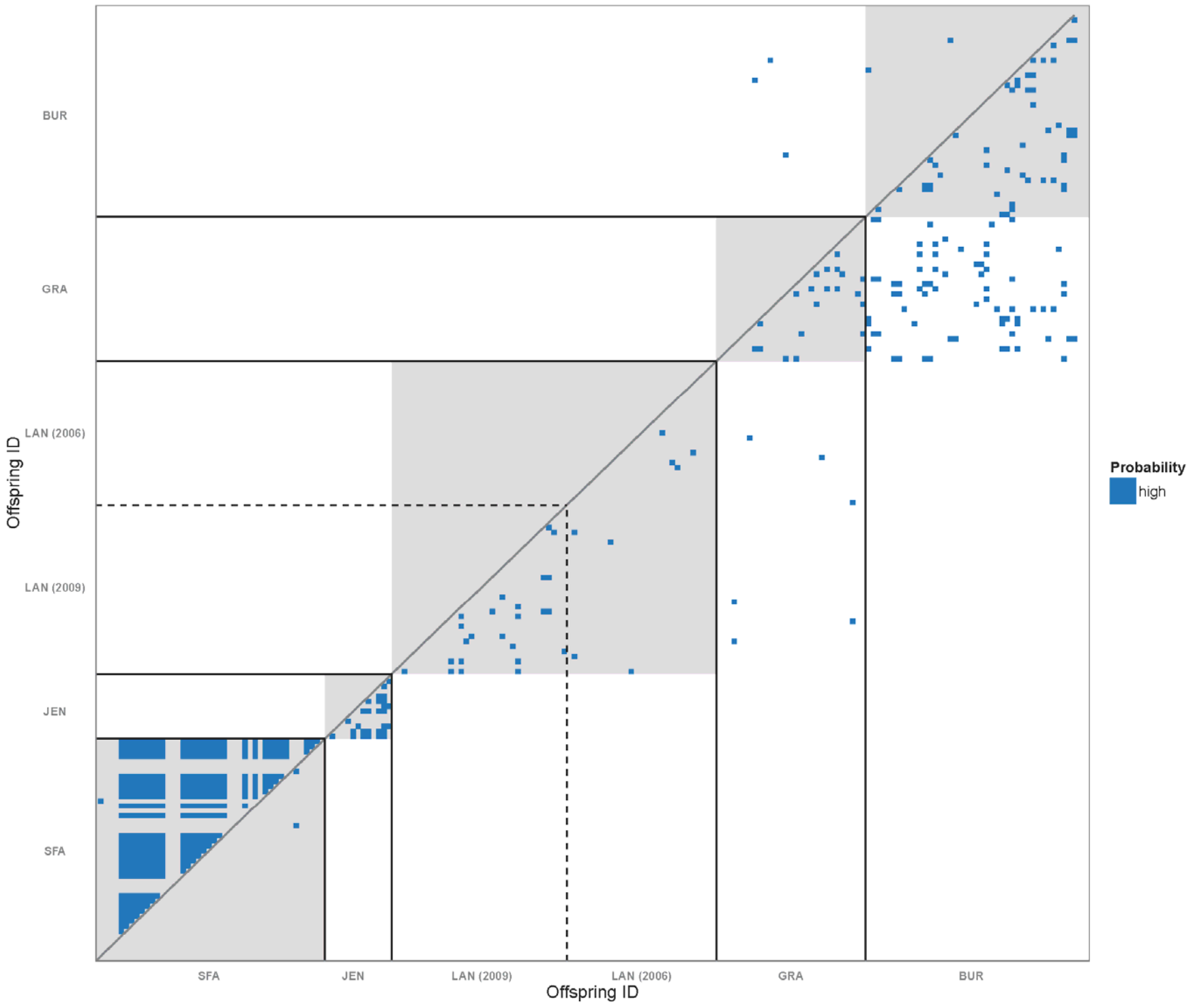


Figure 12. Heat map for pairwise relationships for upper Nelson River sites

Probability of relationship inferred by COLONY (blue: >= 0.75; gray: 0.20 to 0.75 and red: <= 0.20). Full-sibling pairwise relationships are shown above the diagonal while half-sibling pairwise relationships are below diagonal. Cohorts from different sampling sites are separated by solid black line. LAN site as 2 cohorts separated by a dash line. Large shaded gray squares highlight pairwise relationships from the same sample sites, while pairwise relationships observed outside gray shaded squares represent between sample sites i.e., gene flow.

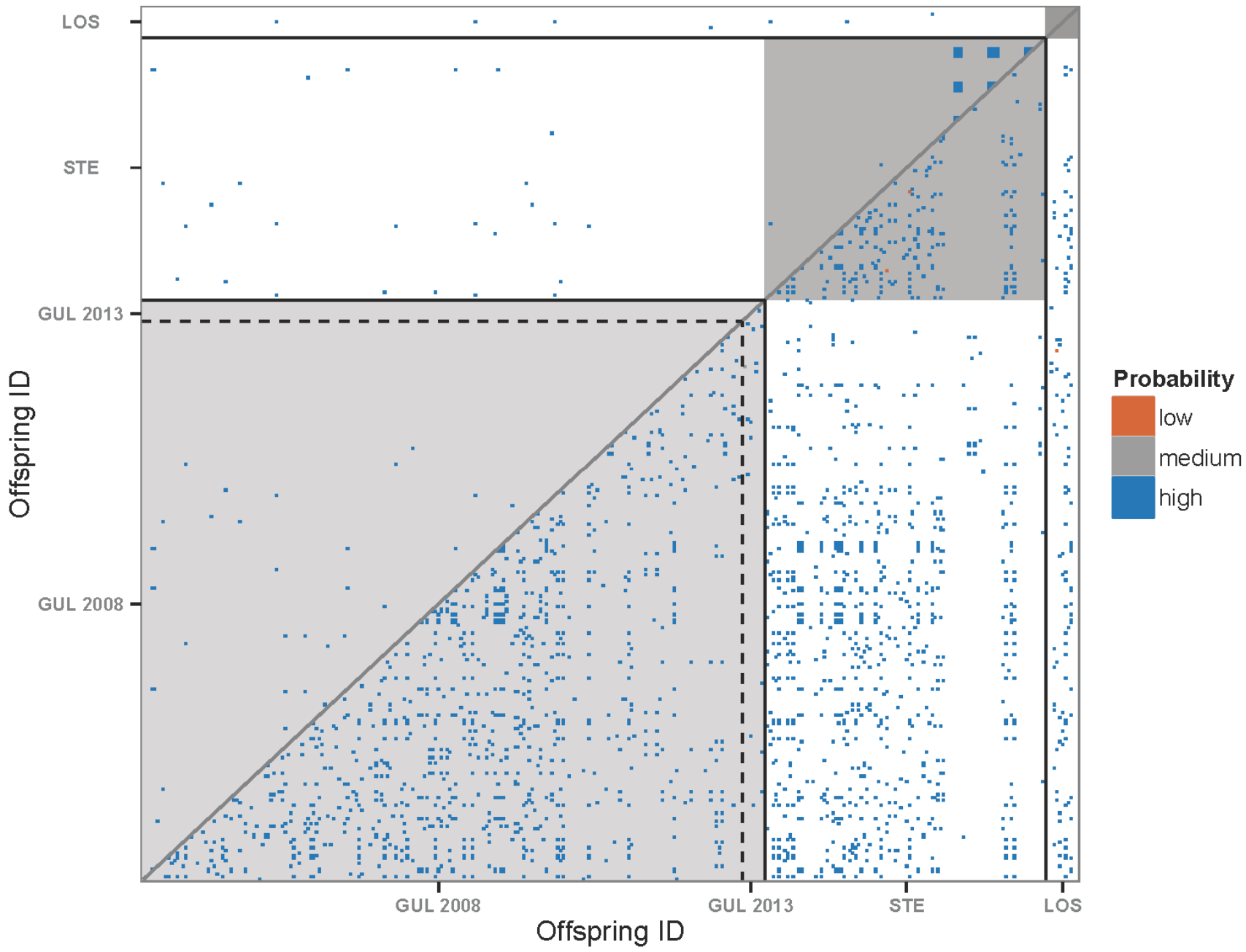


Figure 13. Heat map for pairwise relationships for GUL-STE-LOS.

Probability of relationship inferred by COLONY (blue: >= 0.75; gray: 0.20 to 0.75 and red: <= 0.20). Full-sibling pairwise relationships are shown above the diagonal while half-sibling pairwise relationships are below diagonal. Cohorts from different sampling sites are separated by solid black line. GUL sampling site as 2 cohorts separated by a dash line. Large shaded gray squares highlight pairwise relationships from the same sample sites, while pairwise relationships observed outside gray shaded squares represent between sample sites i.e., gene flow.

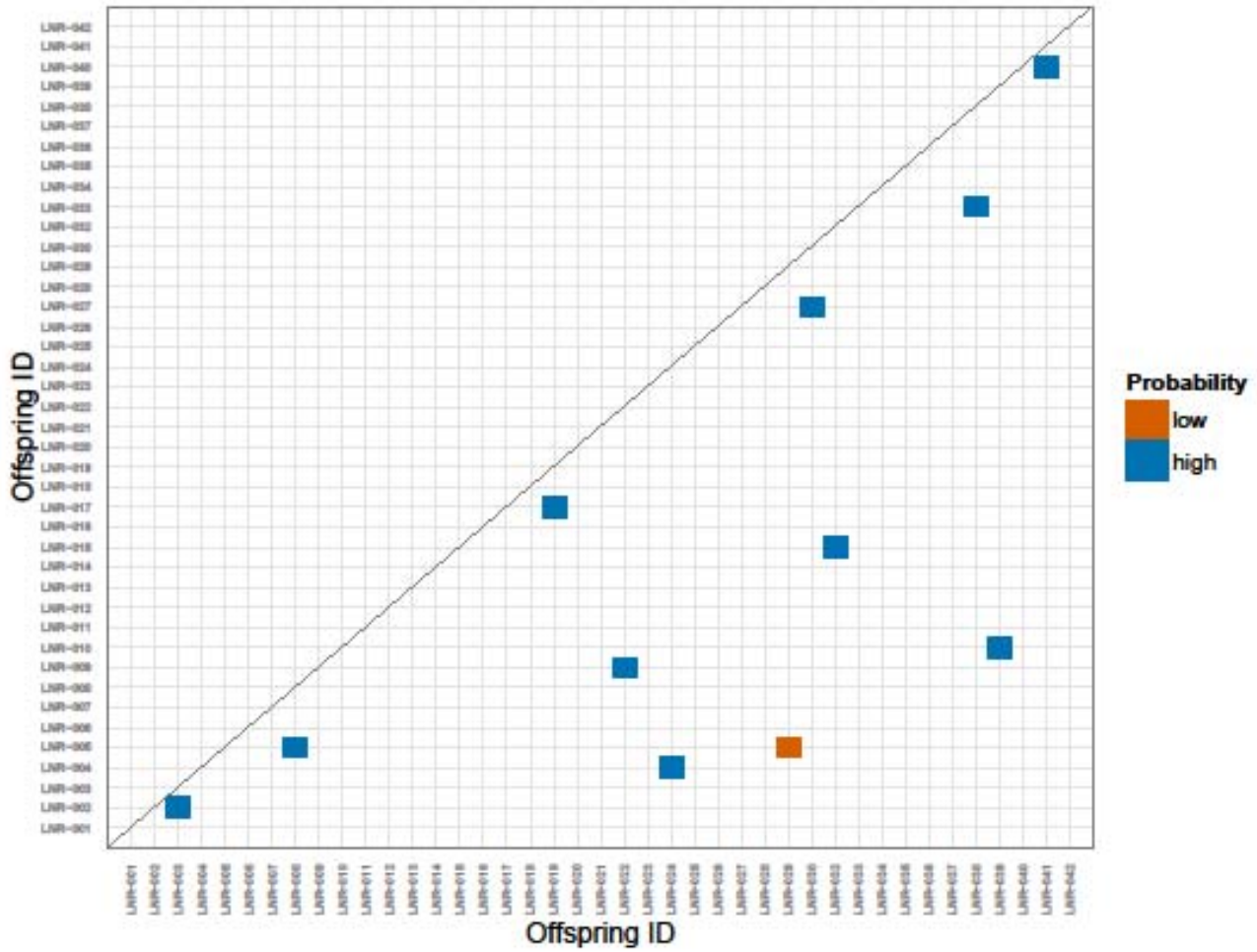


Figure 14. Heat map for pairwise relationships for LNR.

Probability of relationship inferred by COLONY (blue: >= 0.75; gray: 0.20 to 0.75 and red: <= 0.20). Full-sibling pairwise relationships are shown above the diagonal while half-sibling pairwise relationships are below diagonal.

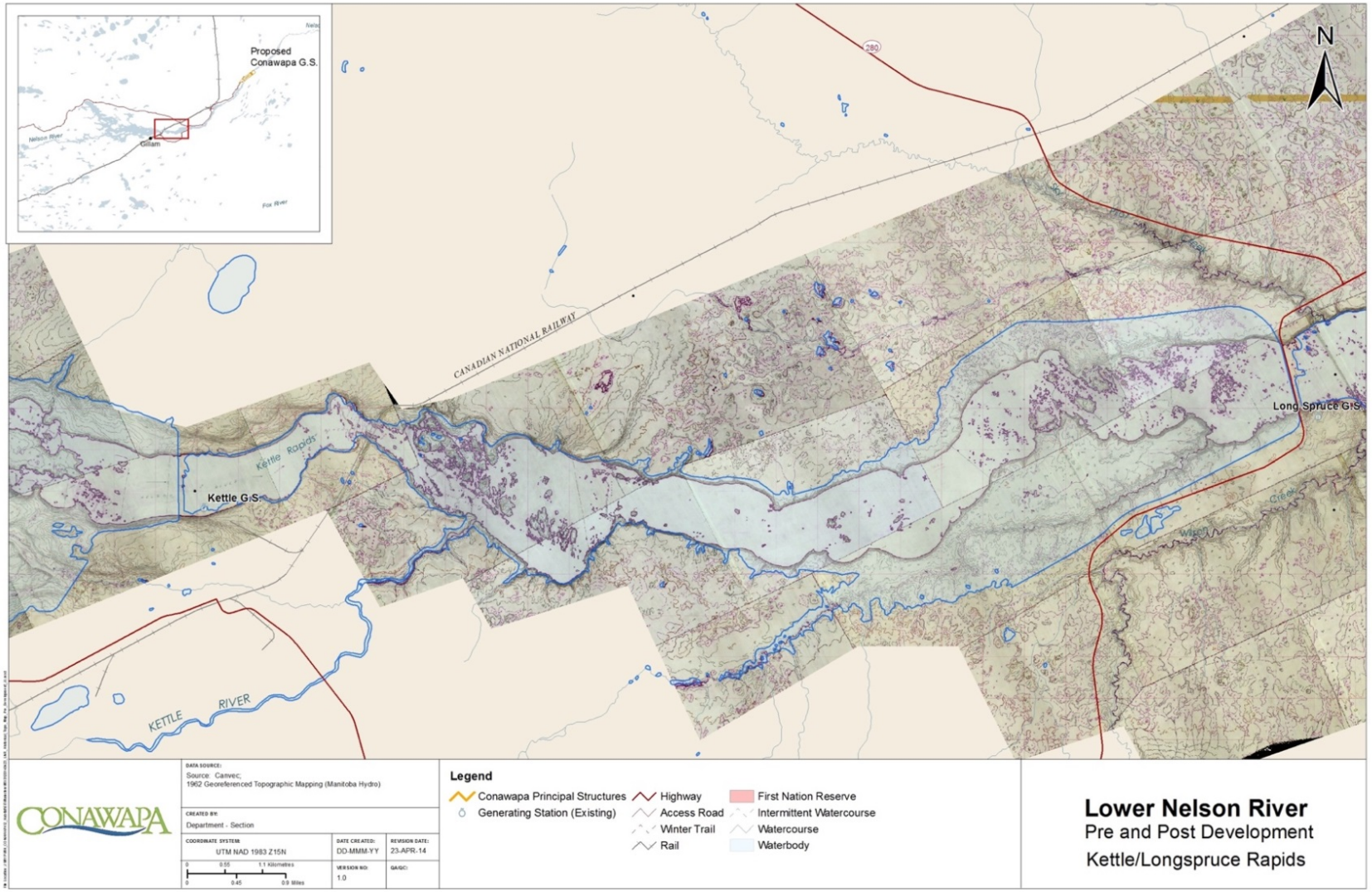


Figure 15. Nelson River between Kettle GS and Long Spruce GS predevelopment

Nelson River mainstem as it transitions from Boreal Shield ecozone into Hudson Plain ecozone. Genetic data suggests this reach represented a historical barrier to gene flow.

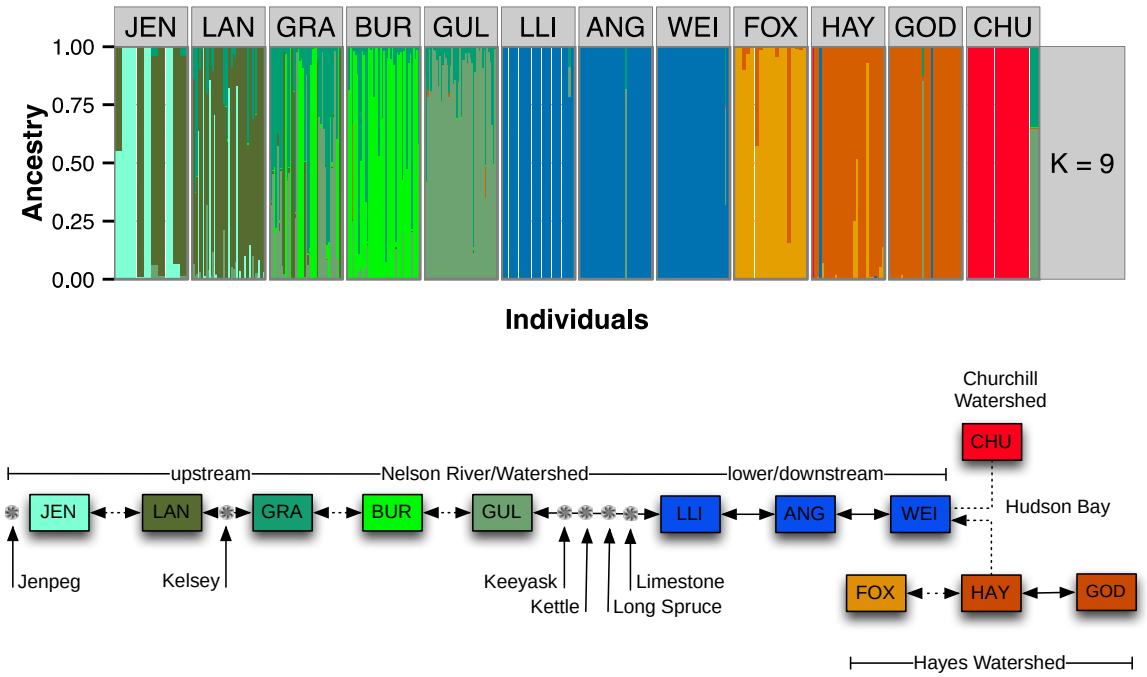


Figure 16. Clustering of sampling sites for K = 9

Clustering of sampling sites for K = 9 using Discriminant Analysis of Principal Components (DAPC; top), and a flow chart of Manitoba sampling locations showing the hierarchical structure of the Lake Sturgeon populations in relation to the generating stations for K = 9. The upper Nelson River (Boreal Shield) between Jenpeg GS and Keeyask GS, shows significant structure. The lower Nelson River (Hudson Plain) downstream of the Limestone GS, no structure exists among the spawning sites. Similar structure exists in the Hayes watershed Fox River (Boreal Shield) shows a different population signature than the Gods and Hayes (Hudson Plain) which show no structure.

Table 1. Sampling sites summary.

Number of individuals per sampling sites and groups (adults, juveniles/cohorts and larvae) with number of individuals blacklisted based on three categories of missing genotypes used after filtering (30%, 50% and 70%).

Sites	Groups	N	Missing genotypes		
			30%	50%	70%
SFA	JUVENILES	46	2	2	2
JEN	ADULTS	10	0	0	0
	JUVENILES	13	0	0	0
LAN	ADULTS	46	0	0	0
	JUVENILES	62	0	0	0
	LARVAE	196	2	2	2
GRA	ADULTS	48	2	1	1
	JUVENILES	72	2	0	0
BUR	ADULTS	48	4	2	1
	JUVENILES	43	1	1	1
GUL	ADULTS	48	0	0	0
	JUVENILES	219	4	3	1
STE	JUVENILES	100	4	4	4
LOS	JUVENILES	10	1	0	0
LLI	ADULTS	23	1	0	0
ANG	ADULTS	46	0	0	0
WEI	ADULTS	48	1	1	1
LNR	JUVENILES	72	0	0	0
FOX	ADULTS	19	1	1	1
HAY	ADULTS	30	2	2	1
GOD	ADULTS	33	2	2	2
CHU	ADULTS	17	9	8	4
TOTAL		1179	38	29	21

Table 2. GBS bioinformatics pipeline.

Description of the different steps in the bioinformatics GBS pipeline with software used, version number and references.

Steps	Description	Software, version and reference
1	Raw reads are inspected for overall quality and presence of adapters	FASTQC v.0.11.3[1] FQGREP[2]
2	Adapters are removed from raw reads	CUTADAPT v.1.9 (Martin, 2011)
3	Reads are cleaned and demultiplexed by barcodes (STACKS *process_radtags*)	STACKS v.1.30 (Catchen et al., 2013)
4	Reads are inspected for overall quality	FASTQC v.0.11.3[1] FQGREP[2]
5	Data from each individual are grouped into loci, and polymorphic nucleotide sites are identified (STACKS *ustacks* for *de novo*).	STACKS v.1.30 (Catchen et al., 2013)
6	Loci are grouped together across individuals and a catalog of loci is written (STACKS *cstacks*).	STACKS v.1.30 (Catchen et al., 2013)
7	Loci from each individual are matched against the catalog to determine the allelic state at each locus in each individual (STACKS *sstacks*).	STACKS v.1.30 (Catchen et al., 2013)
8	Genotype and haplotype calls in individual samples are corrected based on population-wide data (STACKS *rxstacks*).	STACKS v.1.30 (Catchen et al., 2013)
9	Allelic states are converted into a set of *de novo* stack formation (STACKS *populations*).	STACKS v.1.34 (Catchen et al., 2013)
10	SNP visualization, filtering, D_{lr} and figures	STACKR v.0.1.3[3]
11	*F*-Statistics and population assignment analysis	GENODIVE v.2.0b27 (Meirmans & van Tienderen 2004)
12	Tree visualization	GGTREE v.1.07[4]
13	Principal Component Analysis, admixture and assignment analysis	ADEGENET v.2.0.0 (Jombart et al. 2011)
14	Parentage analysis	COLONY v.2.0.5.9 (Jones & Wang 2010, Wang 2013)

[1] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[2] https://github.com/indraniel/fqgrep

[3] https://github.com/thierrygosselin/stackr

[4] https://github.com/GuangchuangYu/ggtree

Table 3. GBS bioinformatics settings.

Bioinformatics steps, options and values used in the GBS pipeline.

Steps	Options	Value
adapter removal	CCGAGATCGGAAGAGCG (a)	
	error tolerance (e)	0.2
	reads shorter than *N* bases are discarded	80
process_radtags	Clean data by removing any read with an uncalled base (c)	yes
	Discard reads with low quality scores (q)	yes
	Truncate final read length to this value (t)	80
	Set the size of the sliding window as a fraction of the read length (w)	0.15
	Score limit within the sliding window drops below this value (s)	10
ustacks	Minimum depth of coverage required to create a stack (m)	4
	Maximum nucleotides distance allowed between stacks (M)	5
	Maximum distance secondary reads to primary stacks (N)	7
	Disable calling haplotypes from secondary reads (H)	yes
	Enable the removal algorithm (r)	yes
	Enable the Deleveraging algorithm (d)	yes
	Maximum locus stacks	3
	Bounded model with an alpha	0.05
	Lower and upper bound epsilon	0-0.15
cstacks	Number of mismatches (n)	1
sstacks	default	
rxstacks	Log likelihood filtering	yes
	Minimum log likelihood threshold	-10
	Prune haplotypes	yes
	Filter confounded loci	yes
	Confounded threshold	0.75
	Bounded model with an alpha	0.1
	Lower and upper bound epsilon	0-0.1
populations	Minimum percentage of individuals/population (r)	0.50
	Minimum number of populations (p)	6
	Specify a minimum stack depth required for individuals at a locus (m)	7

Table 4. Filters statistics.

The number of markers (SNP and loci) discarded and kept after each filter, including the minor allele frequency (MAF), observed heterozygosity (Het) and inbreeding coefficient (Fis). Values of filters threshold are shown in numbers or percentages (%). (Loci = 85985; SNP = 163070).

Steps	Filter	Number of markers blacklisted		Number of markers after filter	
		SNP	LOCI	SNP	LOCI
1	**Consensus sequences**	-	18574	-	85985
2	**Paralogs** (loci > 2 alleles)	-	4309	140566	62758
3	**Genotype likelihood & max coverage** read.depth.threshold = 7 allele.depth.threshold = 7 allele.imbalance.threshold = 0.15 read.depth.max.threshold = 100 gl.mean.threshold = 20 gl.min.threshold = 5 gl.diff.threshold = 100 gl.pop.threshold = 50%	105613	44613	34953	18145
4	**Individuals** ind.threshold = 65%	4912	2237	30041	15908
5	**Populations** pop.threshold = 8	0	0	30041	15908
6	**MAF** local.maf.threshold = 0.02 global.maf.threshold = 0.1 maf.pop.threshold = 1 pop	11466	5026	18575	10882
7	**Het** het.threshold = 0.5 het.diff.threshold = 0.5 het.pop.threshold = 5 pop	8984	4913	9591	5969
8	**Fis** fis.min.threshold = -0.3 fis.max.threshold = 0.3 fis.diff.threshold = 0.5 fis.pop.threshold = 5 pop	637	318	8954	5651
9	**SNP number per haplotypes** max.snp.number = 6 pop.threshold = 100%	106	14	**8848**	**5637**

Table 5. Genetic diversity summary statistics.

These statistics for adult samples in each potential population (Sites) include the number of individuals genotyped in at least 70% of the locus (N), the total number of loci in the catalog (Loci), the proportion of mono- (Mono), polymorphic (Poly) and consensus loci (Con). The genetic diversity measure are presented for the raw and imputed dataset (in bold), the average observed heterozygosity (Ho), the heterozygosity within population (Hs), Nei's heterozygosity-based G_{is}, analogue to Wright's inbreeding coefficient (F_{IS}). The nucleotide diversity (Pi) presented here consider the consensus loci in the catalog.

Sites	N	Loci	Mono/Poly/Con	Ho	Hs	Gis	Pi
JEN	10	22591	2062 / 3573 / 16956	0.1751 **0.1715**	0.1726 **0.1663**	-0.0142 **-0.0312**	0.0012
LAN	46	23554	881 / 4754 / 17919	0.1867 **0.1831**	0.1797 **0.1755**	-0.0393 **-0.0433**	0.0013
GRA	47	22490	700 / 4936 / 16854	0.1890 **0.1809**	0.1831 **0.1757**	-0.0324 **-0.0293**	0.0013
BUR	47	19147	815 / 4821 / 13511	0.1773 **0.1655**	0.1767 **0.1660**	-0.0039 **-0.0028**	0.0013
GUL	48	21369	731 / 4905 / 15733	0.1821 **0.1748**	0.1790 **0.1722**	-0.0174 **-0.0149**	0.0013
LLI	23	23390	1057 / 4579 / 17754	0.1948 **0.1894**	0.1850 **0.1797**	-0.0528 **-0.0540**	0.0013
ANG	46	22398	727 / 4909 / 16762	0.1944 **0.1892**	0.1844 **0.1796**	-0.0542 **-0.0536**	0.0013
WEI	47	24301	808 / 4828 / 18665	0.2073 **0.2033**	0.1894 **0.1857**	-0.0948 **-0.0949**	0.0014
FOX	18	21232	1463 / 4171 / 15598	0.1754 **0.1697**	0.1767 **0.1697**	0.0076 **0.0000**	0.0013
HAY	29	22040	904 / 4732 / 16404	0.1975 **0.1911**	0.1878 **0.1817**	-0.0517 **-0.0517**	0.0014
GOD	31	22659	774 / 4862 / 17023	0.2064 **0.2003**	0.1950 **0.1891**	-0.0587 **-0.0593**	0.0015
CHU	13	11359	2225 / 3409 / 5725	0.1373 **0.1068**	0.1648 **0.1295**	0.1668 **0.1752**	0.0012
ALL	405	25408	0 / 5636 / 19772	0.1853 **0.1771**	0.1811 **0.1726**	-0.0230 **-0.0262**	0.0014

Table 6. Genetic differentiation.

Genetic differentiation of lake sturgeon highlighted using pairwise estimates of Nei's Gst with a correction for a bias that stems from sampling a limited number of populations and Jost D, above and below diagonal, respectively. P-values of all the comparisons were highly significant at $p<0.0001$, except for LLI-ANG and LLI-WEI ($p = 0.0030$ and $p = 0.0130$, respectively). Grey cells indicate comparison between watersheds. Bold indicate presence of barrier. Example: ANG-BUR: significant Gst of 0.020 in presence of barrier in the same watershed; ANG-HAY: significant Gst of 0.014, in the absence of a barrier and in a different watershed. The significance was tested with 10,000 permutations.

Gst	JEN	LAN	GRA	BUR	GUL	LLI	ANG	WEI	FOX	HAY	GOD	CHU
JEN	—	0.020	**0.027**	**0.037**	**0.034**	**0.039**	**0.039**	**0.041**	**0.043**	**0.036**	**0.039**	**0.079**
LAN	0.004	—	**0.011**	**0.022**	**0.019**	**0.023**	**0.024**	**0.025**	**0.028**	**0.022**	**0.025**	**0.062**
GRA	**0.005**	**0.002**	—	0.006	0.007	**0.014**	**0.013**	**0.015**	**0.026**	**0.012**	**0.014**	**0.048**
BUR	**0.006**	**0.004**	0.001	—	0.016	**0.021**	**0.020**	**0.023**	**0.033**	**0.019**	**0.022**	**0.054**
GUL	**0.006**	**0.003**	0.001	0.003	—	**0.015**	**0.014**	**0.017**	**0.028**	**0.014**	**0.016**	**0.051**
LLI	**0.007**	**0.004**	**0.003**	**0.004**	**0.003**	—	0.001	0.002	0.031	0.013	0.015	0.053
ANG	**0.007**	**0.004**	**0.002**	**0.004**	**0.003**	0.000	—	0.002	0.030	0.014	0.017	0.052
WEI	**0.007**	**0.005**	**0.003**	**0.004**	**0.003**	0.000	0.000	—	0.032	0.014	0.017	0.055
FOX	**0.008**	**0.005**	**0.005**	**0.006**	**0.005**	0.006	0.005	0.006	—	0.014	0.022	0.064
HAY	**0.007**	**0.004**	**0.002**	**0.004**	**0.003**	0.002	0.003	0.003	0.003	—	0.005	0.047
GOD	**0.007**	**0.005**	**0.003**	**0.004**	**0.003**	0.003	0.003	0.003	0.004	0.001	—	0.050
CHU	**0.014**	**0.011**	**0.008**	**0.009**	**0.009**	0.010	0.009	0.010	0.011	0.008	0.009	—

Table 7. Relationship analysis summary.

Summary of COLONY relationship analysis with cohorts and cohort groups, offspring used in the analysis (N), the resulting numbers and percentage, in parenthesis, of offspring with half-sib and/or full-sib relationship. The number of unique candidate mothers and fathers when full range of probabilities is used and probability > 0.75, in parenthesis. Number of mating pairs based on full probabilities with the mean probability ± SE and range is given (some fathers and mothers are found in more than one pair, explaining the large numbers in last column). Sexes are arbitrarily set by COLONY.

Cohorts	N	HS n (%)	FS n (%)	Mother n (0.75)	Father n (0.75)	Mating pairs n:mean prob ± SE [min-max]
SFA	44	3 (7)	31 (70)	14 (14)	14 (14)	13: 1.00 ± 0.00 [1-1]
LAN-JEN	76	46 (61)	60 (79)	58 (58)	61 (61)	75:0.97 ± 0.01 [0.76-1]
GRA	29	18 (62)	0 (0)	23 (23)	25 (25)	29: 0.98 ± 0.01 [0.85-1]
BUR	40	37 (93)	6 (15)	24 (21)	26 (22)	37: 0.94 ± 0.02 [0.74-1]
SFA-JEN-LAN-GRA-BUR	190	115 (61)	41 (22)	100 (100)	125 (125)	156: 1.00 ± 0.00 [1-1]
GUL	194	194 (100)	53 (27)	51 (6)	89 (10)	166: 0.15 ± 0.02 [0.07-1]
STE	92	83 (90)	16 (17)	56 (19)	36 (17)	81: 0.49 ± 0.04 [0.21-1]
GUL-STE-LOS	318	300 (94)	68 (21)	73 (70)	154 (152)	271: 0.86 ± 0.01 [0.21-1]
LNR	38	30 (79)	2 (5)	34 (27)	33 (25)	38: 0.86 ± 0.03 [0.50-1]

Table 8. Reproductive contribution.

Number (n50) and percentage of the total number of putative adults (father/mother) based on full probability estimates that contributed to 50% of offspring present in the relationship analysis. Sexes are arbitrarily set by COLONY.

Cohorts	Offspring		Father		Mother		Parents	
	tot	n_{50}	tot	n_{50} (%)	tot	n_{50} (%)	tot	n_{50} (%)
SFA	44	29	14	1 (7)	14	1 (7)	28	2 (7)
LAN-JEN	76	38	61	23 (37)	58	20 (34)	119	43 (36)
GRA	29	20	25	11 (44)	23	9 (39)	48	20 (42)
BUR	40	14.5	26	6 (23)	24	7 (29)	50	13 (26)
GUL	194	97	89	23 (26)	51	9 (18)	140	32 (23)
STE	92	46	36	7 (19)	56	16 (29)	92	23 (25)
GUL-STE-LOS	318	159	154	33 (21)	73	10 (14)	227	43 (19)
LNR	38	19	33	14 (42)	34	15 (44)	67	29 (43)

Table 9. Effective number breeder (Nb).

The number of breeders in the population (Nb) estimated by COLONY is given under scenario of random and non-random mating for five sturgeon cohorts/rivers. The 95% confidence intervals are obtained from bootstrapping.

Cohorts	Random mating				Non-random mating			
	Alpha	Nb	CI95(L)	CI95(H)	Alpha	Nb	CI95(L)	CI95(H)
SFA	0.00	4	2	12	-0.02	5	2	20
LAN-JEN	0.00	187	134	263	0.00	187	137	273
GRA	0.00	162	94	416	-0.03	146	92	293
BUR	0.00	59	40	95	0.08	56	35	91
GUL	0.00	82	61	113	-0.07	93	69	125
STE	0.00	67	47	97	-0.06	74	54	100
GUL-STE-LOS	0.00	91	69	121	0.00	91	68	123
LNR	0.00	312	176	963	-0.05	237	154	420

APPENDIX 1

APPENDIX 1: GLOSSARY

Glossary of terms found in this report and associated readings.

Admixture

A composite gene pool in which at least some individuals (F_1, F_2, F_x and/or backcross) can trace ancestry to more than one population.

Alignment

Arrangement of sequence based on similarity.

Assembly

Gathering together sequences into their correct chromosomal positions.

Assignment analysis

Statistical methods that use genetic information to ascertain population membership of individuals or groups of individuals.

Barcode

A molecular barcode is an individual-specific nucleotide of known sequence varying length inserted adjacent to the genomic sequence read and used for tracking an individual sample in multiplexed next-generation sequencing libraries.

Bayesian statistics

Statistical framework in which the parameters of the models are treated as random variables, allowing expression of the probability of parameters, given the data; this is called the posterior. The posterior probability is obtained by Bayes' rule, and it is proportional to the likelihood times the prior.

Cloud computing

The use of computing resources distributed in a network (typically the Internet) to store, manage and analyze data, rather than doing so on a local server or personal computer.

Cluster

Two or more reads varying by a small number of mismatches between them that are grouped or "stacks" with clustering software based on sequence similarity.

Clustering

Method for decomposing a mixture into its component parts (e.g., gene pools or populations) in the absence of information to characterize the units *a priori*.

Conservation genomics

Defined broadly as the use of new genomic techniques to solve problems in conservation biology.

CPU

Central processing unit within a computer. Multicore computer processing units (Multicore CPUs). Single computing processors with two or more independent computing units (called cores). Running multiple instructions on multiple cores at the same time can increase the overall speed of programs.

***de novo* assembly**

Process of aligning and merging together individual sequence reads to form long contiguous sequences (contig) in the absence of a reference sequence to compare with (i.e., with no prior information).

Depth of coverage (read or sequencing depth)

Total number of a genomic unit (base, reads, etc.) represented.

Discriminant function

A linear combination of variables that maximizes the contrast among different groups of interest. Unknowns (e.g. individuals) are classified into one of the groups (e.g., populations) based on their score on the discriminant function. An AT is a type of discriminant function.

Frequentist methods

Statistical methods that test hypotheses about an event based on the expected frequency of that event happening over a large number of trials (frequency distribution). If no such information is available (e.g., from a theoretical frequency distribution), randomization techniques are used to generate an empirical frequency distribution.

F_{IS}

Wright's inbreeding coefficient measuring the level of correlation between two genes drawn from an individual relative to two genes drawn from the population. Also defined as the probability that two alleles in an individual are both descended from a single allele in an ancestor.

F_{ST}

A measure of populations subdivision that indicates the proportion of genetic diversity found between populations (S : for subpopulation) relative to the amount within populations (T : for total).

Haplotype

A DNA sequence with a number of closely linked loci on a chromosome that is inherited through generations as a single unit without being changed by crossing-over or other recombination mechanisms. In STACKS an *haplotype* represents the configuration of SNPs in a short-read locus.

Homeologs

A special case of paralogy resulting from polyploidy

Homolog

A member of a chromosome pair in diploid organisms or a gene that has the same origin and function in two or more species.

Library

Collection of DNA fragments

Likelihood

The probability of obtaining the observed data under a certain model or hypothesis. The assignment index can be viewed as the likelihood of an individual occurring in the population in which it was sampled. An ML approach finds or approximates the parameter values that maximize the likelihood.

Linkage

The proximity of two or more markers on a chromosome; the closer the markers, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in prokaryotes, mitosis or meiosis in eukaryotes), and hence the greater the probability that they will be inherited together.

Linkage disequilibrium

The non-random association of alleles at different gene loci. A standard index of linkage disequilibrium, D, is defined as the difference between observed and expected frequencies of a two-locus gamete. If the two loci are independent, the expected frequency of a gamete is the product of the frequencies of the two alleles.

Markers (also genetic markers)

In STACKS, haplotypes and define alleles in each locus are called genetic markers. A genetic marker (locus or loci [plural]) is a segment of an organism's DNA that usually varies in its composition (i.e., its nucleotide sequence) and thus presents different variants or alleles. An individual carries two copies of each marker, which are either identical (same allele) or different. These two copies together form its genotype for that marker. Two individuals differ from each other with respect to their genotype if they have either only one or no alleles in common.

Marker-assisted selection

The use of genetic markers to predict and increase the response to selection by favouring reproduction of individuals with a certain allele or genotype. The marker is closely linked to a quantitative trait locus.

Markov Chain Monte Carlo (MCMC)

A simulation technique to generate samples from a probability distribution of interest. MCMC can estimate complex multi-variate distributions that cannot be generated by standard simulation methods. It has also been used to approximate likelihood surfaces in the context of ML methods.

Minor allele frequency

The frequency of the less frequent allele at a biallelic genetic locus.

Missingness

Is a term for missing data. Imputation models require that data have a specific distribution to their missing values, also known as their missingness pattern.

Mixture

Proportion of individuals from different source populations that contribute to a genetic mixture or admixture.

Ne and Nb

The effective population size (Ne) of a population is the size of the ideal (Wright-Fisher) population (N) that will result in the same amount of genetic drift as in the actual population being considered. For sturgeon, because of overlapping generations, COLONY *Ne* is actually *Nb*, the effective number of breeders.

Next generation sequencing (NGS)

Highly parallel DNA sequencing where hundreds of thousands or millions of reads (sequences) are produced in one run.

Operational Sex Ratio (OSR)

The ratio of receptive males to receptive females at one time and site.

Orthologs

Homologous genes related by speciation.

Paralogues

Homologous genes related by duplication.

Parentage analysis

A classification method for determining the parents of an individual or group of individuals.

Pipelines

Semi-autonomous script to automate analyses.

Principle component analysis (PCA)

A tool for transforming a set of observations with correlated variables into a set of linearly independent variables called principle components, making sure that the first principle component accounts for the largest variability of the data.

Polyploidy

The heritable condition of possessing more than two complete sets of chromosomes. *Allopolyploidy* arise from fusion of nuclei from different species, whereas *autopolyploidy* arise when the different sets of chromosomes are derived from the same species.

RAD site

A site cut by a restriction enzyme (RE)

RAD-tags

Restriction-site Associated DNA tags is a method for typing large numbers of SNPs on the Illumina Genome Analyser. Fragments are cut by a restriction enzyme and sequenced, those fragments are over-represented in the sequence reads, and so genotypes at polymorphic sites can be reliably called. This approach differs from other SNP typing methods in that SNPs do not need to be discovered beforehand and because SNP identification and estimates of allele frequencies are obtained simultaneously, saving time and money.

Random forest

A statistical technique that are used to control for over-fitting bias in decision trees. In general an ensemble learning method for classification, regression, and other tasks, that operate by building many decision trees during a training period and outputting a mode of the classes (classification) or a mean prediction (regression) of the individual trees.

Read

A genomic unit consisting of an individual short piece of DNA sequence output from an Next-Generation Sequence (NGS) platform

Restriction enzyme

A protein that recognizes and cuts specific short nucleotide sequences.

RRL

Reduced representation library; a DNA library created from only a certain fraction of the genome

Scaffolds

A genomic unit composed of assembled contigs with gaps in between.

Single-nucleotide polymorphisms (SNPs)

Sites in the DNA in which there is variation that occur when a single nucleotide (A, T, C, or G) sequence is altered.

Whitelist

A method used for screening out data that fit *a priori* criteria acceptable for analysis.