OXFORD UNIVERSITY PRESS | Journal of Heredity

Caution and skepticism are critical when analyzing massively parallel sequencing data

Journal:	*Journal of Heredity*
Manuscript ID:	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Miller, Michael; University of California, Animal Science Norgaard, Zachary; University of California, Animal Science Ali, Omar; University of California, Animal Science Amish, Stephen; University of Montana, Biological Sciences O'Rourke, Sean; University of California, Animal Science Prince, Daniel; University of California, Animal Science Luikart, Gordon; University of Montana, Flathead Lake Biological Station Muhlfeld, Clint; US Geological Survey, Northern Rocky Mountain Science Center
Subject Area:	Conservation genetics and biodiversity, Bioinformatics and computational genetics
Keywords:	Massively parallel sequencing (MPS), hybridization, introgression, fisheries, restriction-site associated DNA (RAD), bull trout

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

Norgaard_TableS1.csv
Norgaard_TableS2.csv

Title

Caution and skepticism are critical when analyzing massively parallel sequencing data

Running Head

Caution is critical when analyzing MPS data

Authors and Institutional Affiliations

Zachary K. Norgaard[1], Omar A. Ali[1], Stephen J. Amish[2], Sean M. O'Rourke[1], Daniel J. Prince[1], Gordon Luikart[2,3], Clint C. Muhlfeld[4], Michael R. Miller[1,5]

[1]Genetic Diversity Research Group, Department of Animal Science, University of California, Davis CA 95616, USA
[2]Fish and Wildlife Genomics Group, Division of Biological Sciences, University of Montana, Missoula MT 59812, USA
[3]Flathead Lake Biological Station, University of Montana, Polson MT 59860, USA
[4]US Geological Survey, Northern Rocky Mountain Science Center, Glacier National Park, West Glacier, Montana 59936, USA
[5]Center for Watershed Sciences, University of California, Davis CA 95616, USA

Corresponding Author

Michael R. Miller
Department of Animal Science
University of California
One Shields Avenue
Davis CA 95616
1-530-304-4719
micmiller@ucdavis.edu

Keywords

Massively parallel sequencing (MPS), hybridization, introgression, fisheries, restriction-site associated DNA (RAD), bull trout

Abstract

Massively parallel sequencing (MPS) has revolutionized genomic analysis by providing large scale sequencing data quickly and inexpensively. Due to the immense amount of data produced by MPS, analysis typically relies on preexisting software packages to verify data quality and generate results. Unfortunately, software does not have the ability to report biases or discrepancies it is not designed to identify. Here we present a compelling example of how genetic analysis with MPS can lead to incorrect biological conclusions with significant management consequences unless utmost caution is used. Our original goal was to characterize population structure and demography of native bull trout, *Salvelinus confluentus*, in the Flathead Basin of Montana. Surprisingly, the initial results suggested introgression with introduced lake trout, *Salvelinus namaycush*. We then used established software and methods to confirm these results and were confident introgression and recent hybridization had occurred. Only after directly examining the raw data were we able to conclude our results were caused by sample contamination. If the additional analysis had not been conducted, the positive identification of introgression would have significantly impacted conservation and management of threatened bull trout. The diagnostic allelic contribution (DAC) test developed here should prove a useful tool for characterizing introgression between populations and species observed in MPS data. Furthermore, this example should serve as a lesson to biologists and reinforce the need for caution and skepticism when drawing conclusions from MPS data.

Introduction

Despite massively parallel sequencing (MPS) generating a significant proportion of recently published data, methods for validating results remain scarce. Prior to the advent of MPS, genetic analysis was limited by the amount of data obtainable, and relatively few genetic markers were used to describe phylogenies (Ortí *et al*. 1997; Ritz *et al*. 2000; Petren *et al*. 1999) and construct linkage maps (Kerem *et al*. 1989). Previously, sequencing was slow and expensive, relegated to use after significant effort had been put forth to establish regions of interest. With MPS, millions of reads can be generated at a fraction of the cost associated with previous methods (Metzker 2010; Davey *et al*. 2011). MPS makes the utilization of whole genome sequencing to identify variants cost effective, without conducting a preliminary phenotype screening. The reduction in sequencing cost has also led to the rise of genome wide association studies (GWAS) (Neale *et al*. 2010; Imamura & Maeda 2011; Graham *et al*. 2009).

Overall, MPS has revolutionized our ability to discover and describe genetic variation. MPS is not limited to model organisms and provides a tool to perform genomic analysis in virtually any species, making it an excellent resource for conservation genetics and molecular ecology (Tepolt 2015; Rittmeyer & Austin 2015; Marchant *et al*. 2015; Sharma *et al*. 2012; Sigssgaard *et al*. 2015; Burgar *et al*. 2014). In particular, reduced representation sequencing methods such as RAD-seq (Miller *et al.* 2007; Baird *et al*. 2008; Hohenlohe *et al*. 2010) have revolutionized the field of population genetics for previously uncharacterized natural populations. Prior to the advent of MPS, it was difficult to characterize the diversity of natural populations beyond phenotypes and a handful of genetic markers. With the power of MPS, we can infer demographic history, characterize population structure, detect selection, describe genetic diversity, and discover hybridization (Ekblom & Galindo 2011; Egan *et al.* 2012; Pool *et al*. 2010).

Unfortunately, MPS is subject to many forms of bias or contamination. With previous analyses, the amount of data was relatively small, and verification by manual inspection was practical. With MPS, millions of data points are generated and it is impossible to check every piece of data by manual inspection. Biologists have become reliant on software packages to sort through this copious data, despite lacking the formal bioinformatics training necessary to fully understand the software packages. Most

genomic tools have built in functions to filter reads based on length, depth, and quality scores. However, there is potential for many other forms of bias to be present in the data.

Here we demonstrate how genetic analysis of MPS data has the potential to lead to incorrect biological conclusions with significant management consequences unless the utmost caution is used. Our original objective was to characterize genetic population structure and demography in Flathead Basin (Montana) bull trout. Strikingly, we detected significant introgression between native bull trout, *Salvelinus confluentus,* and lake trout, *Salvelinus namaycush,* within one population. After considerable analysis, we were confident we had detected true introgression. Despite established methods supporting introgression and having confidence in our sample identity and quality, we wanted to formally eliminate the possibility of sample contamination. To achieve this we developed a new method to test for contamination, and ultimately concluded that the original results were not caused by introgression but were a product of contamination. If the additional analysis had not been conducted, we would have reached incorrect biological conclusions that would have had serious conservation and management consequences. This incorrect conclusion could have resulted in wasted resources or even the eradication of a threatened bull trout population that was thought to have widespread introgression with an invasive species. The diagnostic allelic contribution (DAC) test should be used when characterizing hybrids from MPS data, but perhaps more importantly, this study emphasizes the need for caution and skepticism when analyzing MPS data.

Results

RAD-Seq Provides Initial Indication of Introgression

The following genomic analyses focus on bull trout in the Flathead Basin, in and around Glacier National Park, Montana. Lake trout were first introduced to Flathead Lake in 1905 (Hanzel 1969) and have subsequently spread throughout the basin including Glacier National Park (US Department of Interior 2009). This is of primary concern because invasive lake trout populations displace or replace native bull trout populations (Donald & Alger 1993). Bull trout are now listed as threatened throughout their range in the continental United States in large part due to introduced lake trout (US Fish and Wildlife Service 1999). Our original goal in this study was to use genomic methods to characterize the demographic response of

bull trout populations across Flathead Basin to lake trout invasion.

To discover and type genetic variation in bull trout, we performed RAD-sequencing on 160 bull trout, with the majority of samples originating from six Flathead Basin populations, and generated a *de novo* RAD locus assembly (Fig. 1 & Table 1). The mean number of single-end sequence reads per individual was 4.6 million with a standard deviation of 1.5 million. Using the method described by Miller *et al*. (2012), we identified 62,832 loci each consisting of 80 bases, providing 5,026,560 nucleotide sites for comparison (Table S1). This amount of sequence space was expected based on the estimated bull trout genome size (2-3Gb) and the restriction enzyme we used (SbfI). We conclude that this reference RAD locus assembly should provide ample sequence space to discover and genotype thousands of single nucleotide polymorphisms (SNPs) in bull trout.

To elucidate the relationships between Flathead Basin bull trout populations, we aligned reads from each bull trout sample to the reference, discovered and genotyped SNPs, and performed a principal component analysis PCA (see methods, Fig. 2A). The first principal component (PC1) explained 5.88% of total variance in the samples and separated Quartz Lake bull trout from the other collection sites. PC1 also separated three Quartz Lake bull trout from the other Quartz Lake samples. The second principal component (PC2) explained 3.61% of total variance in the samples and differentiated the South Fork, Middle Fork, and the remaining North Fork collection sites from one another. Additionally, PC2 separates the No Name Creek collection site from the Whale-Hallowat Creek collection sites – which remain closely associated. We conclude that our samples represent at least five distinct bull trout populations in the Flathead Basin.

To begin investigating bull trout population specific demographic responses to invasive lake trout, we estimated an unfolded site frequency spectra (SFS) for each bull trout population and used RAD data generated from an outgroup (lake trout) to infer ancestral states (see Materials and Methods). The lake trout sample used as a reference was selected based on read quality from 96 RAD-sequenced lake trout individuals originally processed for an independent study. The mean number of single-end sequence reads per individual was 2.1 million with a standard deviation of 0.5 million. Strikingly, the Quartz Lake SFS showed an enrichment of segregating sites with a high derived allele frequency compared to other

populations (Fig. 2B) and the neutral expectation. This suggested a low frequency of ancestral (lake trout) alleles were present in Quartz Lake bull trout, indicative of introgression between the two species. A history of continuous gene flow between the species is unlikely due to the lack of similar patterns in other populations and the historic geographic separation of the species (Donald & Alger 1993). We conclude that an investigation of possible introgression between native bull trout and invasive lake trout is necessary.

To further examine the relationship between Quartz Lake bull trout and introduced lake trout, we performed a second PCA with the addition of 31 lake trout samples from the Flathead Basin (Fig. 2C). Bull trout and lake trout separate completely on the first principal component (PC1) of the analysis. However, the three individuals from the first PCA (Fig. 2A) were observed to share a significant portion of their variation with lake trout. We conclude that introgression between bull trout and lake trout in Quartz Lake is a strong possibility.

Identified Species Diagnostic Markers Support Wide Spread Introgression

To carefully investigate possible introgression of introduced lake trout into native Quartz Lake bull trout, we used our sample data from across the range of each species to identify species diagnostic SNPs (Table 2). While excluding the potentially introgressed Quartz Lake samples from our analysis, we identified 19,392 species diagnostic SNPs (Table 2, Table S2) from the 5,026,560 sites initially identified. Recent studies have identified diagnostic markers and characterized hybridization using both fewer samples and markers (Amish *et al.* 2012*;* Hasselman *et al*. 2014; Trier *et al.* 2014; Sun *et al.* 2014; Taylor *et al.* 2014). We conclude that 19,392 diagnostic sites are sufficient for the investigation of hybridization and introgression between bull trout and lake trout in Quartz Lake.

To determine if Quartz Lake individuals exhibited genotypes characteristic of introgression, we calculated the proportion of diagnostic sites that were homozygous for the bull trout allele, homozygous for the lake trout allele, and heterozygous for each sample (Fig. 3A & Table 3). Three of the Quartz Lake bull trout (individuals 1, 4, and 11) were heterozygous at approximately 25% of diagnostic sites, consistent with two generations of backcrossing to pure bull trout after an initial hybridization event. Another six Quartz Lake bull trout were heterozygous at >1% of diagnostic sites, indicating they were separated from

the hybridization event by several generations of backcrossing to pure bull trout. We conclude that Quartz Lake bull trout exhibit genotypes characteristic of introgression and these results strongly support wide spread introgression.

To investigate if excluding Quartz Lake individuals while identifying diagnostic sites had somehow biased our results and caused non-diagnostic sites to be miscalled as diagnostic, we re-identified diagnostic sites while excluding additional populations from the analysis. The number of diagnostic sites we identified remained consistent across the different population sets (Table 2). Furthermore, when we examined sample genotypes with the new sets of diagnostic sites, the proportion of sites called homozygous for the bull trout allele, homozygous for the lake trout allele, and heterozygous in Quartz Lake did not change and samples from the other excluded populations (Granite Creek, Hallowat Creek, No Name Creek, Whale Creek, & Wounded Buck Creek) did not have genotypes indicative of introgression (Fig. 3B). We conclude that the observed patterns of introgression cannot be explained by miscalling diagnostic sites.

Diagnostic Allele Contribution (DAC) Test Detects Sample Contamination

Because DNA extraction, RAD library preparation, and Illumina sequencing of the bull trout and lake trout samples occurred months apart and/or in distinct physical locations, we believed sample contamination had not caused the above results. However, we wanted to formally rule out the possibility of contamination causing the observed genotype patterns, but an established software package or method for such a test did not exist. Therefore, we developed the diagnostic allele contribution (DAC) test. The logic behind this test is that truly heterozygous genotypes should have an equal read contribution from both alleles while heterozygous genotype calls resulting from contamination would have allelic contributions skewed towards the more abundant DNA contributor. An example of this expectation for a true BC_2 individual was constructed for comparison (Fig. 4A & Table 3). In a true BC_2 individual, we expected 75% of diagnostic sites to be called homozygous for the bull trout allele and have zero contribution of lake trout alleles. The remaining 25% of sites should be called heterozygous and normally distributed with a mean lake trout allele contribution of 0.5. The DAC test is a qualitative comparison between expected and observed allele contribution distributions.

To further analyze introgression in Quartz Lake, we performed DAC tests for each individual of interest as well as several samples from other populations. As expected, pure bull and lake trout samples were homozygous at diagnostic sites with no contribution from the opposite species (Fig. 4B & Table 3). To our surprise, we observed a striking and consistent decline in the proportion of lake trout allele sequence reads across all diagnostic sites in Quartz Lake bull trout (Fig. 4C). Only sites with a proportion of lake trout alleles above approximately 0.1 were called heterozygous while uncalled genotypes tended to have a lake trout allele contribution between 0.05 and 0.1. These thresholds result from the probabilistic framework used to call genotypes (see Discussion). Lake trout alleles appeared in low frequency throughout the Quartz Lake sample set but rarely reached a contribution of 0.5 (Fig. 4C & Table 3). The patterns of raw allele counts in Quartz Lake bull trout are not what would be expected if these individuals were truly heterozygous at the diagnostic loci. We conclude that the previous evidence of wide spread introgression may be more accurately explained by sample contamination.

Artificial Mixes Verify Sample Contamination Detected by DAC Test

To test if sample contamination could explain our results, we created artificial mixes from pure individuals (see Materials and Methods) and compared the proportion of diagnostic sites called as heterozygous to what was observed in Quartz Lake bull trout samples. The proportions generated by the artificial mixes were comparable to those of the contaminated samples (Fig. 5A). For example, the 4% artificial mix produced approximately 25% of heterozygous genotype calls at diagnostic sites. Most of the Quartz Lake bull trout samples had proportions of diagnostic heterozygous genotype calls less than the observed proportion in the 2% artificial mix. We conclude that the artificial mixes successfully reproduce the Quartz Lake bull trout data with respect to the proportion of diagnostic sites called heterozygous.

To test if sample contamination could reproduce our allele contribution distributions, we performed the DAC test on the artificial mixes and examined the results. Strikingly, we obtained patterns almost identical to those of the Quartz Lake bull trout samples (Fig. 5B & Table 3). The artificial mixes (excluding that with 50% lake trout alleles) had a high number of sites called as homozygous for the bull trout allele with a low contribution of lake trout alleles. The genotype calls are similarly separated by thresholds at 0.05 and 0.1 distinguishing uncalled and heterozygous sites respectively. We conclude that sample

contamination can adequately explain our allele contribution distributions in Quartz Lake bull trout.

To rule out the possibility that the skewed allele contribution at heterozygous sites could be due to the failure of lake trout reads to properly align to the bull trout reference, we examined the results of the DAC test on the 50% artificial mix. The plot obtained from the 50% lake trout artificial mix provides an approximately normal distribution of reads with a mean of 0.57 (Fig. 5B). The majority of the lake trout reads must have properly aligned to produce this approximately normal distribution. This also indicates it would be difficult to distinguish a 50% contaminated sample from a true F1 hybrid using this test. We conclude that patterns of genetic variation seen in the Quartz Lake bull trout samples are the result of lake trout contamination, not an alignment bias.

Discussion

Potential for Bull-Lake Trout Hybridization

Our initial results strongly suggested that bull and lake trout had hybridized in Quartz Lake. The PCA and SFS generated for the populations gave the initial indication of introgression, and the subsequent identification and testing of species diagnostic sites provided very strong evidence that Quartz Lake bull trout were introgressed with lake trout. Differences in reproductive behavior had led many to consider natural hybridization between the two species unlikely, due to a difference in spawning habitat. Bull trout spawn in streams on gravel substrate (Fraley & Shepard 1989) while lake trout spawn in lakes over rocky bottoms (Deroche 1969). However, field biologists working to remove lake trout from Quartz Lake have reported catching sexually mature bull trout while gill netting for lake trout on their spawning beds. Additionally, Quartz Creek, the tributary that feeds Quartz Lake is very small, and may not provide sufficient water in years with little snow pack to support bull trout spawning. Bull trout might be forced to spawn in the lake these years, leading to hybridization and subsequent introgression. Climate change would exacerbate the problem by increasing the number of years in which the Quartz Creek water level was insufficient for bull trout spawning – resulting in more bull trout spawning events taking place near lake trout and an increased rate of introgression. However, as our results demonstrate, having a logical biological explanation for your results does not make them correct. Alternative explanations for the observed results should be explored.

Management Impacts of Falsely Identifying Hybridized Bull Trout

The lake trout sequence reads observed in Quartz Lake bull trout samples would have remained unexamined, if a lake trout sample had not been used to generate an unfolded SFS. The enrichment of high frequency derived alleles observed in the Quartz Lake population would have been mistaken for an enrichment of low minor allele frequency sites in a folded SFS – characteristic of recent population expansion. When examining the technical aspects of the data, the foreign lake trout reads were of high quality, significant length, and mapped to identified loci. Without the DAC test, this analysis would have supported the existence of hybridized bull trout in Quartz Lake.

Historically, responses to hybridization of threatened species have varied significantly. Some agencies and studies argue hybrids should be protected while others argue for hybrid removal (Allendorf *et al*. 2001). Adding to the controversy, hybridization has been recognized as a cause of extinction (Wolf *et al*. 2001; Rhymer & Simberloff 1996). Gese *et al*. (2015) suggests hybrid removal has been successful in maintaining genetic lineages of the red fox, *Canis rufus*; however, in fisheries management, it is not practical to sample, sequence, and genotype every offspring before determining their fate. Allendorf *et al*. (2001) proposed a classification system that should be implemented when making practical hybridization management decisions. In this situation, the Quartz Lake bull trout population would be considered a potential origin point for widespread introgression, because admixture was detected throughout the population at low levels and in high levels in a select few individuals. Widely introgressed populations are considered to have little conservation value, making removal of the population preferential (Allendorf *et al*. 2001). In the case of Quartz Lake, hybrid removal would occur through poisoning of the lake as outlined by the U.S. Department of the Interior (2009). Poisoning the lake would eliminate all fishes and require approximately 32,000 gallons – $1,900,000 – of rotenone (U.S. Department of the Interior 2009). This merely represents the initial dosage price, detoxifying the lake after hybrid removal and restocking the lake with pure bull trout would significantly increase the cost of the operation.

Genotype Calling Method

We used a maximum likelihood (ML) framework for calling genotypes, because it provides the best statistical method for estimating genotypes if the likelihood function is correct (Nielsen *et al.* 2011).

The probability of observing a particular set of sequence reads given each possible genotype is calculated based on allele counts and quality scores. Genotype likelihoods can then be used for many downstream analyses such as SFS estimation, genotype calling and PCA. Here we used a uniform prior to calculate genotype posterior probabilities and called genotypes for sites that had a posterior greater than or equal to 0.9. Unfortunately, homozygous sites were incorrectly called heterozygous because contamination was not considered when calculating genotype likelihoods. The genotype likelihood function used assumes that all read sequences originate from sample alleles or a sequencing error.

An alternative approach would be to use a heuristic method to call genotypes (Nielsen *et al.* 2011). With such a method, the allele ratio for each genotype call is defined and a minimum number of sequence reads at a site are required before a call is made. For example, we could have required eight sequence reads before making a genotype call, an allele ratio between 0.2 and 0.8 to call a site heterozygous, and an allele ratio between 0 and 0.05 or 0.95 and 1 to call a site homozygous. The remaining sites would remain uncalled. Using this approach, the number of diagnostic sites called heterozygous would have depended on the thresholds chosen. However, the results would have remained qualitatively similar with only the level of introgression for each individual appearing different depending on the thresholds used.

Regardless of the genotype calling method used, cautious and skeptical analysis is critical. Using either method, lake trout alleles would have been observed in bull trout samples. Once contaminated samples are detected, the next concern is what to do with them. Possible options include simply removing all the contaminated samples or perhaps just the sites believed to be a problem from the analysis. However, these samples and generated data could be highly valuable and therefore methods that facilitate the use of data from contaminated samples are needed. To this end, we are developing a probabilistic framework for estimating sample contamination rate and using the estimated rate in a modified genotype likelihood function to calculate accurate genotype likelihoods and calls even in the presence of significant contamination (Miller and Linderoth unpublished). This method should enable accurate genotype information to be extracted from contaminated samples and help prevent the waste of resources.

The DAC Test Detects Sample Contamination

Based on the premise that both alleles from a heterozygous site should contribute equal read counts, the DAC test provides a new way of examining the raw sequence reads. Surprisingly, the DAC test detected sample contamination that had been overlooked when relying on readily available software packages. The DAC test provides a simple method to directly examine raw sequence reads and qualitatively compare the output to expected results. The DAC test should be implemented when admixture between distinct populations is detected. However, on a broader scale, these results provide an example of why it is sometimes necessary to step outside the confines of established software. While these software packages serve an important function, their outputs should certainly be viewed through a lens of caution and skepticism. In its current form, the DAC test is a qualitative comparison of observed and expected distributions. It should also be possible to develop a rigorous mathematical method for comparison of the distributions; however, in many situations a qualitative comparison will suffice. Theoretically it would also be possible to use the DAC test to analyze sites that are not fixed between groups. However, the possibility of contamination sources and samples sharing alleles would make it difficult to distinguish between sampling error and contamination. Truly heterozygous sites with unbalanced allele contributions, due to random sampling, would be indistinguishable from homozygous sites contaminated by another source.

Every step from gathering field samples to the final sequencing process has the potential to introduce contaminant DNA. Today, samples are rarely collected, processed, and analyzed by the same scientists. In such a setting, it becomes imperative to examine every result with care. In this case, contamination likely occurred in the field. Quartz Lake bull trout samples were taken from unintentionally caught fish while gill netting lake trout. Thus, it did not matter that the samples were stored separately and processed months apart. Standard collection, extraction, and sequencing procedures mitigate the impact of technical artifacts (Yirga *et al*. 2012; Casquet *et al*. 2011; Bi *et al*. 2013; Goldberg 2013), but bias and contamination is always a possibility. The current framework for validating MPS samples is often sufficient to identify, and sometimes correct, technical errors (Zagordi *et al*. 2010; Taub *et al*. 2010). However, tools to identify other bias or contamination sources are scarce. With resource intensive management practices centered on the findings of MPS studies, it becomes crucial to ensure the quality of data being generated

and analyzed. In this case, the DAC test was able to detect sample contamination ignored by traditional MPS quality check methods.

Materials and Methods

Samples

A total of 160 bull trout samples were collected by gill net and hook and line angling. Additionally, 96 lake trout samples were collected as part of a separate study. Bull trout DNA extractions were performed at the University of Montana in Missoula using a simple ethanol extraction protocol. Lake trout DNA extractions were performed at the Flathead Lake Biological Station using the DNeasy extraction protocol (Qiagen). RAD-seq was performed as previously described (Miller *et al*. 2012). Bull trout Illumina libraries were prepared at the University of Oregon before lake trout DNA was ever present in the laboratory. Lake trout Illumina libraries were prepared at the University of Oregon several months later. All samples were sequenced using single-end Illumina sequencing.

Data Preparation

Using a subset of the 160 bull trout samples, sequence reads were processed to identify 80 base pair loci using the method described by Miller *et al.* (2012). Loci were identified by aligning sequences from the subset of bull trout samples to each other and filtering based on quality scores as well as read counts. Using the identified loci, a fasta reference file was constructed. Subsequently, sequences reads for all individuals were quality trimmed to 80 base pairs using a simple perl script. Alignments to the fasta reference file were performed using Novoalign (novocraft.com), allowing three mismatches per alignment, and output in SAM format. The SAM files were converted to BAM files using SAMtools (Li *et al*. 2009).

Principal Component Analyses

A principal component analysis (PCA) was performed using ANGSD and the method outlined by Korneliussen *et al*. (2014). Major alleles were inferred from genotype likelihoods using the method described by Skotte *et al*. (2012). Genotype posteriors were calculated assuming a uniform prior, because we expect alleles to have different frequencies in each population. Allele frequency estimates were calculated from genotype likelihoods as described by Kim *et al*. (2011), but using the EM algorithm described by Korneliussen *et al*. (2014). Sites were filtered to only include those with SNPs using a p-

value less than or equal to 1x10e-6. Data were also filtered using a minimum base quality score of 10, a minimum map quality score of 10 and reads were present in at least 70 individuals. A binary file of posterior probabilities for genotypes was output for use with ngsPopGen. The correlation matrix between individuals was calculated with called genotypes using ngsPopGen (Fumagalli *et al*. 2013; Fumagalli 2013). The PCA was visualized using the script provided by ngsPopGen (Fumagalli *et al*. 2013; Fumagalli 2013). A second PCA was constructed for comparison of bull trout and lake trout samples collected in Flathead Basin using the same method. To prevent skewing of the PCA based on read count, lake trout with less than one million raw reads and half a million mapped reads were excluded from the analysis (FIL22, 23, and 29). All bull trout samples were above this threshold.

Site Frequency Spectra

A site frequency spectrum (SFS) was estimated for each bull trout population in the Flathead Basin with ANGSD by estimating the allele frequency of every SNP in a given population using the method outlined by Korneliussen *et al*. (2014). SFS were generated using ten random individuals from each population for ease of comparison. Allele frequency likelihoods at each nucleotide site were calculated using individual genotype likelihoods and assuming Hardy-Weinberg equilibrium (Korneliussen *et al*. 2014). A lake trout sample with high coverage (SuL19) was designated as the ancestral reference to infer ancestral state and produce an unfolded SFS. Major alleles were inferred from genotype likelihoods using the method described by Skotte *et al*. (2012). Allele frequency estimates were calculated from genotype likelihoods with ANGSD as described by Kim *et al*. (2011), but using the EM algorithm described by ANGSD (Korneliussen *et al.* 2014). Reads were filtered using a minimum base quality score of 10, a minimum map quality score of 10 and must have been recorded in at least 7 individuals. The site allele frequency likelihoods were output to a binary file. This file was used to generate an estimate of the real SFS and output in logarithmic scale. The proportion of sites in each bin was modified surrounding a derived allele frequency of 0.5 due to an artificial enrichment caused by paralogs. The middle four bins of each SFS were replaced with a linear decline of the flanking bins. This allows for a crude visualization of the SFS. The visualization code provided by ANGSD (2014) was used to generate the SFS plots.

Diagnostic Sites

Diagnostic sites were identified using a perl script that parses files generated by ANGSD. Using ANGSD, we called genotypes for each species separately (lake trout and bull trout). Major alleles were again inferred from genotype likelihoods using the method described by Skotte *et al*. (2012). Genotype posteriors were calculated assuming a uniform prior, because we expected alleles to have different frequencies in each population. Allele frequency estimates were calculated from genotype likelihoods as described by Kim *et al*. (2011), but using the EM algorithm described by Korneliussen *et al*. (2014). Reads were filtered using a minimum base quality score of 10 and a minimum map quality score of 10. Sites were not filtered for number of individuals to include as many sites as possible in the analysis. Additionally, sites were not filtered for SNPs, because fixed differences would not be labeled as SNPs in individual species. Genotypes were called at a low threshold (0.6), to minimize the chance of miscalling fixed sites, and output as a set of bases (AA, AC, AG, …). This ANGSD command also generated species wide major and minor allele at each site output in a separate file.

Using the previously generated files, we identified fixed differences between the species to be used as diagnostic sites. At each site, individuals of each species were evaluated for homozygosity for that species' major allele. If all individuals were homozygous for the species' major allele and less than 10% of the individuals were missing genotype calls, the site was determined to be a fixed difference and diagnostic site. Quartz Lake samples were excluded from this step, because they were not expected to be homozygous at diagnostic sites if they had been introgressed. When generating additional sets of diagnostic sites to test for the miscalling of diagnostic sites, additional individuals were excluded from the analysis (Table 2). The diagnostic site position, bull trout allele, and lake trout allele are output for use in the subsequent analysis. To prevent miscalled genotypes from influencing the proportion of heterozygous diagnostic sites, genotype calls were repeated in ANGSD using a higher threshold of 0.9. Genotype calls for samples were then evaluated at diagnostic sites to determine the genotype at each position.

Diagnostic Allelic Contribution (DAC) Test

Using the SAMtools mpileup command, we generated a pileup file restricted to diagnostic sites for each Quartz Lake sample (Li *et al.* 2009). From the pileup file, the number of reads attributable to each allele are determined and cross referenced with the genotype file to associate a proportion with each

genotype call. The proportions and associated genotype calls were then used to generate a simple box plot depicting the proportion at which genotype calls were made (Fig. 4). The observed plots were then qualitatively compared to expected plots containing a normal distribution of heterozygous sites centered on a proportion of 0.5 and homozygous sites with proportions of zero or one.

Artificial Mixes

Artificial mixes of bull trout and lake trout sequence reads were created for analysis by sampling raw reads from a lake trout (FIL16) and bull trout (NoB07) with high read counts and quality using Seqtk. Artificial mixes were created using different proportions (0.01, 0.02, 0.04, 0.08, & 0.5) of lake trout reads to approximate a variety of contamination levels. The proportions of lake trout reads used to create the artificial mixes were chosen to reflect the mean proportion of lake trout reads present across all diagnostic sites in Quartz Lake bull trout samples. The artificial mixes were then analyzed using the methods described previously for the empirical samples.

Funding

This work was supported by the United States Geological Survey and the Great Northern Landscape Conservation Cooperative.

Acknowledgments

The Genetic Diversity Research Group at UC Davis in particular Ismail Saglam for coding help and Ryan Peek for providing the excellent map. Vin D'Angelo, Pat DeHaan, Chris Downs, Jeff Olsen, and John Wenburg for providing samples. Glacier National Park for logistical support.

References

Allendorf FW, Leary RF, Spruell P, & Wenburg JK (2001) The problems with hybrids: setting conservation guidelines. Trends in Ecology & Evolution, 16(11), 613-622.

Amish SJ, Hohenlohe PA, Painter S, Leary RF, Muhlfeld C, Allendorf FW, Luikart G (2012) RAD sequencing yields a high success rate for westslope cutthroat and rainbow trout species-diagnostic SNP assays. *Molecular Ecology Resources*, 12(4), 653-660.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE* 3(10), e3376.

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C (2013) Unlocking the vault: next-generation museum population genomics. Molecular Ecology, 22(24), 6018-6032.

Burger JM, Murray DC, Craig MD, Haile J, Houston J, Stokes V, Bunce M (2014) Who's for dinner? High-throughput sequencing reveals bat dietary differentiation in a biodiversity hotspot where prey taxonomy is largely undescribed. Molecular Ecology, 23(15), 3605-3617.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM Baxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics,

12, 499-510.
Deroche SE (1969) Observations on the Spawning Habits and Early Life of Lake Trout. *The Progressive Fish-Culturist*, 31(2), 109-113.
Donald DB, Alger DJ (1993) Geographic distribution, species displacement, and niche overlap for lake trout and bull trout in mountain lakes. Canadian Journal of Zoology, 71(2), 238-247.
Egan AN, Schlueter J, Spooner DM (2012) Applications of Next-Generation Sequencing in Plant Biology. American Journal of Botany, 99(2), 175-185.
Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107, 1-15.
Fraley JJ, Shepard BB (1989) Life History, Ecology and Population Status of Migratory Bull Trout (*Salvelinus confluentus*) in the Flathead Lake and River System, Montana. *Northwest Science*, 63(4), 133-143.
Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. PLoS One, 8(11):e79667.
Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R (2013) Quantifying population genetic differentiation from next-generation sequencing data. Genetics, 195(3), 979-92.
Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics, 30(10), 1486-1487.
Gese EM, Knowlton FF, Adams JR, Beck K, Fuller TK, Murray DL, Steury TD, Stoskopf MK, Waddell WT, Waits LP (2015) Managing hybridization of a recovering endangered species: The red wolf Canis rufus as a case study. Current Zoology, 61(1), 191-205.
Graham RR, Hom G, Ortmann W, Behrens, TW (2009) Review of recent genome□wide association scans in lupus. Journal of internal medicine, 265(6), 680-688.
Hanzel DA (1969) Flathead Lake, investigation of its fish populations and its chemical and physical characteristics. Project F-33-R-3, job no. I, final report. Montana Department of Fish and Game, Kalispell, MT.
Hasselman DJ, Argo EE, McBride MC, Bentzen P, Schultz TF, Perez-Umphrey AA, Palkovacs EP (2014) Human disturbance causes the formation of a hybrid swarm between two naturally sympatric fish species. Molecular Ecology, 23, 1137-1152.
Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population Genomics of Parallel Adapatation in Threespine Stickleback using Sequenced RAD Tags. PloS Genetics, 6(2), e1000862.
Imamura M, Maeda S (2011) Genetics of type 2 diabetes: the GWAS era and future perspectives. Endocrine journal, 58(9), 723-739.
Kerem BS, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, Buchwald M, Tsui LC (1989) Identification of the cystic fibrosis gene: genetic analysis. Science, 245(4922), 1073-1080.
Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, Jorgensen T, Hansen T, Pedersen O, Wang J, Nielsen R (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics, 35, 231.
Korneliussen T, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics, 15(1):356.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-2079.
Marchant A, Mougel F, Almeida C, Jacquin-Joly E, Costa J, Harry M (2015) De novo transcriptome assembly for a non-model species, the blood-sucking bug Triatoma brasiliensis, a vector of Chagas disease. Genetica, 143(2), 225-239.
Metzker ML (2010) Sequencing Technologies – the next generation. Nature Reviews Genetics, 11, 331-46.
Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17, 240-248.
Miller MR, Brunelli JP, Wheeler PA, Liu S, Rexroad CE III, Palti Y, Doe CQ, & Thorgaard GH (2012) A

conserved haplotype controls parallel adaptation in geographically distant salmonid populations. Molecular Ecology, 21(2), 237-249.
Neale BM, Medland SE, Ripke S, Asherson P, Franke B, Lesch KP, ... McGough J (2010) Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. Journal of the American Academy of Child & Adolescent Psychiatry, 49(9), 884-897.
Nielsen R, Paul JS, Albreschtsen A, Song YS (2011) Gentpe and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* 12, 443-451.
Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from New-Generation Sequencing data. PLoS One, 7(7):e37558.
Novocraft.com: Novoalign short read mapper (http://www.novocraft.com/main/downloadpage.php)
Ortí, G, Pearse DE, Avise JC (1997) Phylogenetic assessment of length variation at a microsatellite locus. Proceedings of the National Academy of Sciences, 94(20), 10745-10749.
Petren K, Grant BR, Grant PR (1999). A phylogeny of Darwin's finches based on microsatellite DNA length variation. Proceedings of the Royal Society of London. Series B: Biological Sciences, 266(1417), 321-329.
Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. Genome Research, 20, 291-300.
Rittmeyer EN, Austin CC (2015) Combined next-generation sequencing and morphology reveal fine-scale speciation in Crocodile Skinks (Squamata: Scincidae: Tribolonotus). Molecular Ecology, 24(2), 466-483.
Ritz LR, Glowatzki□Mullis ML, MacHugh DE, Gaillard C (2000) Phylogenetic analysis of the tribe Bovini using microsatellites. Animal genetics, 31(3), 178-185.
Sharma R, Goossens B, Kun-Rodrigues C, Teixeira T, Othman N, Boone JQ, Jue NK, Obergfell C, O'Neill RJ, Chikhi L (2012) Two Different High Throughput Sequencing Approaches Identify Thousands of De Novo Genomic Markers for the Genetically Depleted Bornean Elephant. PLoS One, 7(11), e49533.
Sigsgaard EE, Carl H, Moller PR, Thomsen PF (2015) Monitoring the near-extinct European weather loach in Denmark based on environmental DNA from water samples. Biological Conservation, 183, 46-52.
Singhal S (2013) De novo transcriptomic analyses for non-model organims: an evaluation of methods across a multi-species data set. Molecular Ecology Resources, 13, 403-416.
Skotte L, Korneliussen TS, Albrechtsen A(2012) Association testing for next-generation sequencing data using score statistics. Genetic Epidemiology, 36(5), 430-437.
Sun Y, Abbott RJ, Li L, Li L, Zou J, Liu J (2014) Evolutionary history of Purple cone spruce (Picea purpurea) in the Qinghai-Tibet Plateau: homoploid hybrid origin and Pleistocene expansion. Molecular Ecology, 23, 343-359.
Taub MA, Corrada Bravo H, Irizarry RA (2010). Overcoming bias and systematic errors in next generation sequencing data. Genome Medicine, 2(12), 87.
Taylor SA, White TA, Hochachka WM, Ferretti V, Curry RL, Lovette I (2014) Climate-Mediated movement of an Avian Hybrid Zone. Current Biology, 24, 671-676.
Tepolt CK (2015) Adaptation in marine invasion: a genetic perspective. Biological Invasions, 17(3), 887-903.
Trier CN, Hermansen JS, Sætre GP, Bailey RI (2014) Evidence for Mito-Nuclear and Sex-Linked Reproductive Barriers between the Hybrid Italian Sparrow and Its Parent Species. PLoS Genetics, 10(1), e1004075.
US Fish and Wildlife Service (1999) Endangered and threatened wildlife and plants: Determination of threatened status for bull trout in the coterminous United States.
US Department of the Interior (2009) Large-Scale Removal of Lake Trout in Quartz Lake Environmental Assessment.
Zagordi O, Geyrhofer L, Roth V, Beerenwinkel N (2010) Deep Sequencing of a Genetically Heterogeneous Sample: Local Haplotype Reconstruction and Read Error Correction. Journal of Computational Biology, 17(3), 417-428.

Figure and Table Legends

Fig. 1 Map of bull trout sample locations in the Flathead Basin.

Fig. 2 Initial population genetic analysis of Flathead Basin bull trout. (A) Principal component analysis of Flathead Basin bull trout. Colors and symbols indicate sampling location. Three Quartz Lake individuals are labeled. (B) Derived allele frequency spectra of Flathead Basin bull trout populations. (C) Principal component analysis of Flathead Basin bull trout and lake trout. Colors indicate species. Three Quartz Lake individuals are labeled.

Fig. 3 (A) Percent bar plot for genotype calls at diagnostic sites established using diagnostic set 1 in Quartz Lake bull trout and lake trout. (B) Percent bar plot for genotype calls at diagnostic sites established using diagnostic set 2 in Granite Creek bull trout, Quartz Lake bull trout, and Quartz Lake lake trout. Shades indicate genotype calls.

Fig. 4 Stacked bar plot showing the number of sites with a particular genotype call for each proportion of lake trout reads in 0.05 increments. Shades indicate genotype calls. (A) An expectation of a two generation backross with 25% heterozygous and 75% homozygous bull trout diagnostic sites. (B) Bull and lake trout samples from outside Quartz Lake. (C) Bull trout samples from Quartz Lake.

Fig. 5 (A) Percent bar plot for genotype calls at diagnostic sites established using diagnostic set 1 in select Quartz Lake bull trout and artificial mixes. (B) Stacked bar plot showing the number of sites with a particular genotype call for each proportion of lake trout reads within 0.05 increments for each artificial mix.

Table 1 Species and number of samples collected at each location.

Table 2 Number of diagnostic sites obtained from each diagnostic set.

Table 3 Number and proportion of 19392 diagnostic sites given particular genotype calls for a subset of bull trout, lake trout, expected BC and artificial mixes.

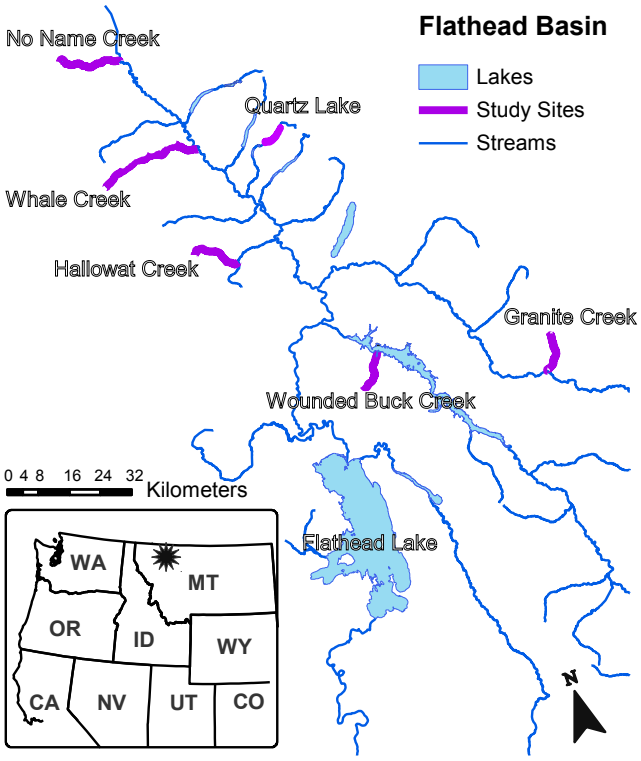
Flathead Basin
Lakes
Study Sites
Streams
No Name Creek
Quartz Lake
Whale Creek
Hallowat Creek
Granite Creek
Wounded Buck Creek
Flathead Lake
0 4 8 16 24 32
Kilometers
WA
MT
OR
ID
WY
CA
NV
UT
CO
N

Fig. 1

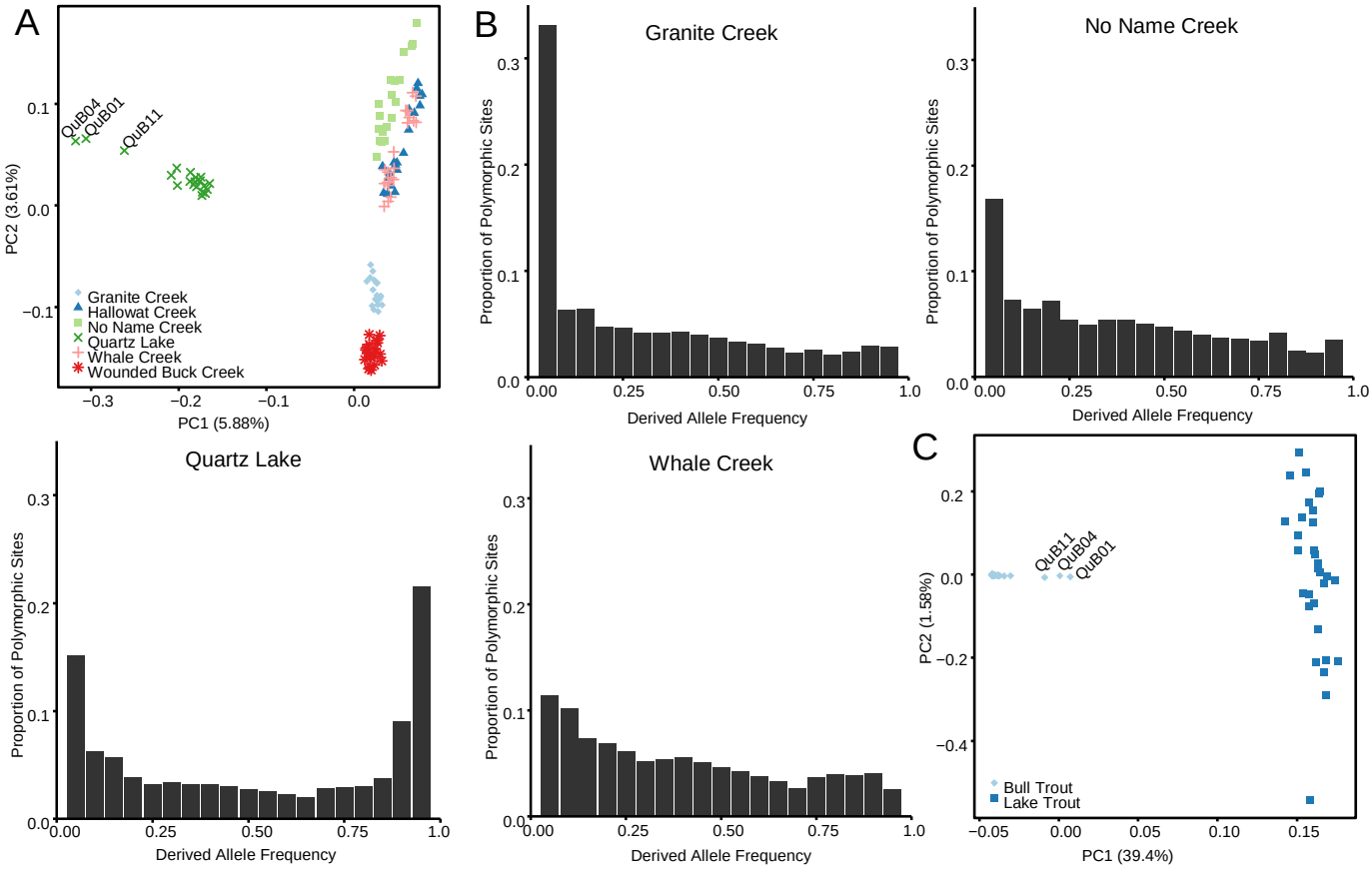
A
QuB04
QuB01
QuB11
PC2 (3.61%)
PC1 (5.88%)
Granite Creek
Hallowat Creek
No Name Creek
Quartz Lake
Whale Creek
Wounded Buck Creek
B
Granite Creek
No Name Creek
Quartz Lake
Whale Creek
Proportion of Polymorphic Sites
Derived Allele Frequency
C
QuB11
QuB04
QuB01
PC2 (1.58%)
PC1 (39.4%)
Bull Trout
Lake Trout

Fig. 2

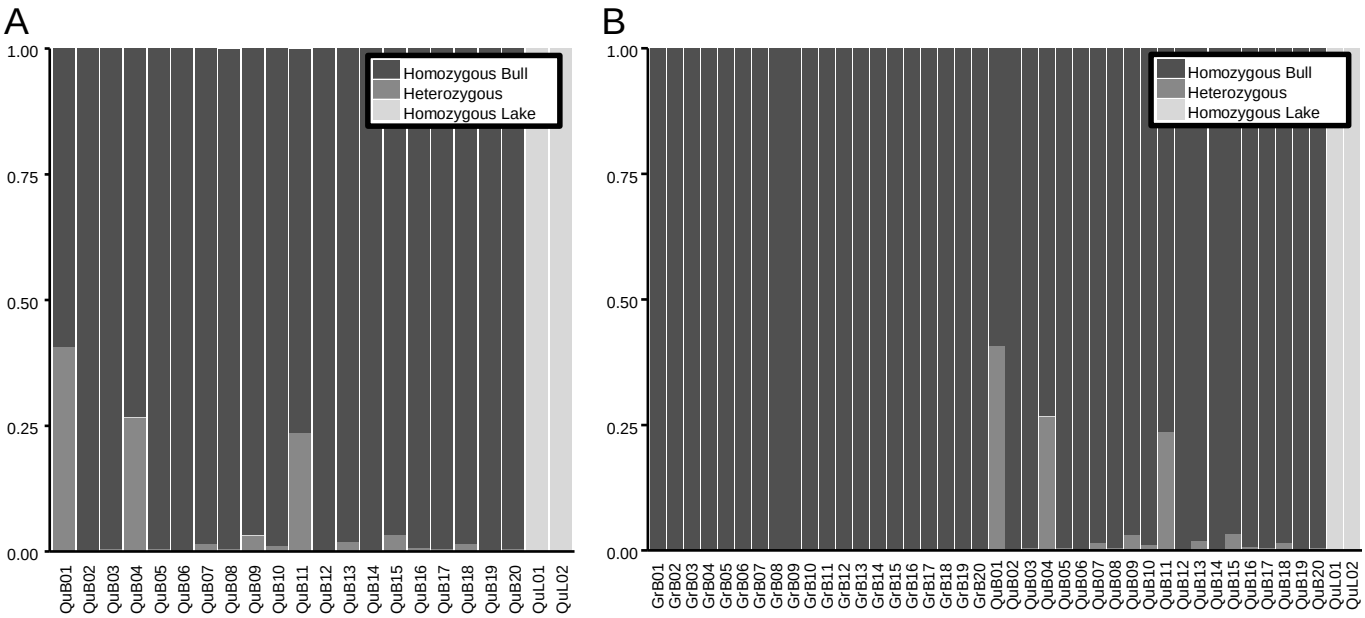
A
B
1.00
0.75
0.50
0.25
0.00
Homozygous Bull
Heterozygous
Homozygous Lake
QuB01
QuB02
QuB03
QuB04
QuB05
QuB06
QuB07
QuB08
QuB09
QuB10
QuB11
QuB12
QuB13
QuB14
QuB15
QuB16
QuB17
QuB18
QuB19
QuB20
QuL01
QuL02
GrB01
GrB02
GrB03
GrB04
GrB05
GrB06
GrB07
GrB08
GrB09
GrB10
GrB11
GrB12
GrB13
GrB14
GrB15
GrB16
GrB17
GrB18
GrB19
GrB20
QuB01
QuB02
QuB03
QuB04
QuB05
QuB06
QuB07
QuB08
QuB09
QuB10
QuB11
QuB12
QuB13
QuB14
QuB15
QuB16
QuB17
QuB18
QuB19
QuB20
QuL01
QuL02

Fig. 3

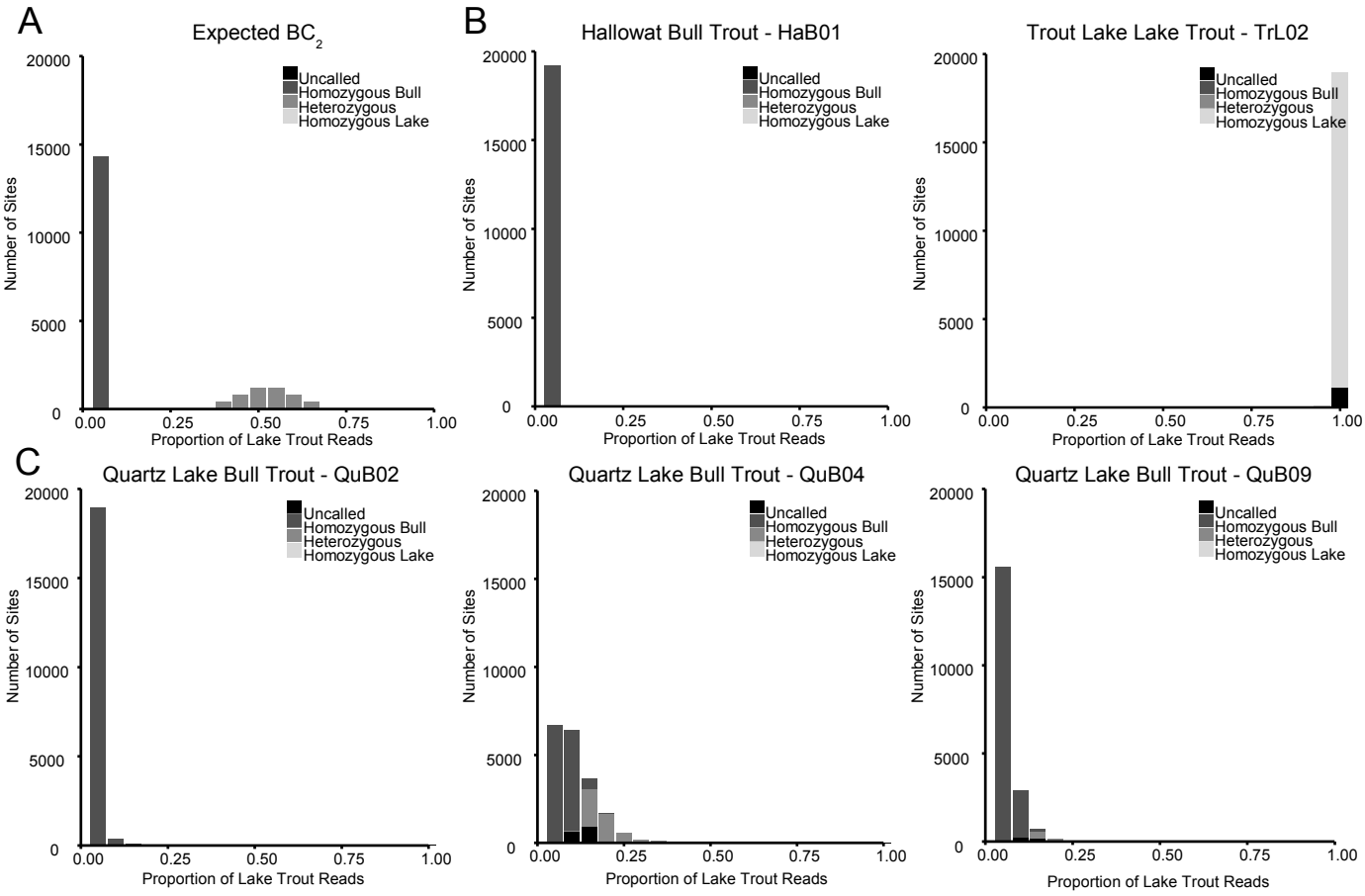
A
Expected BC2
B
Hallowat Bull Trout - HaB01
Trout Lake Lake Trout - TrL02
C
Quartz Lake Bull Trout - QuB02
Quartz Lake Bull Trout - QuB04
Quartz Lake Bull Trout - QuB09
Uncalled
Homozygous Bull
Heterozygous
Homozygous Lake
Number of Sites
Proportion of Lake Trout Reads
0
5000
10000
15000
20000
0.00
0.25
0.50
0.75
1.00

Fig. 4

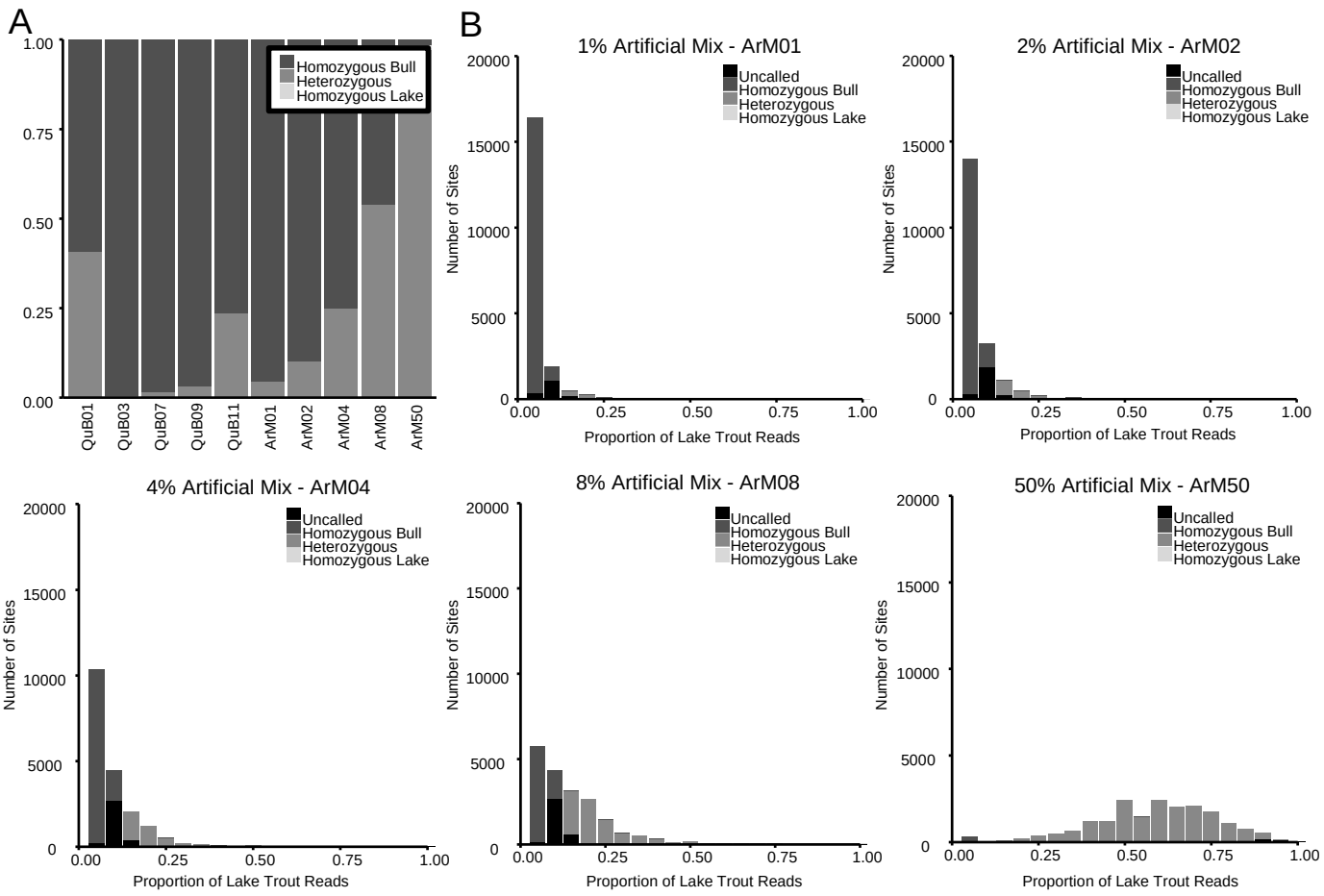
A
B
Homozygous Bull
Heterozygous
Homozygous Lake
QuB01
QuB03
QuB07
QuB09
QuB11
ArM01
ArM02
ArM04
ArM08
ArM50
1% Artificial Mix - ArM01
2% Artificial Mix - ArM02
4% Artificial Mix - ArM04
8% Artificial Mix - ArM08
50% Artificial Mix - ArM50
Uncalled
Number of Sites
Proportion of Lake Trout Reads

Fig. 5

Table 1

Species	Region	Location	ID	Number of Samples
Bull Trout	Montana	Granite Creek	GrB	20
	Montana	Hallowat Creek	HaB	20
	Montana	No Name Creek	NoB	19
	Montana	Quartz Lake	QuB	20
	Montana	Whale Creek	WhB	20
	Montana	Wounded Buck Creek	WoB	20
	Montana	Bitterroot River	BiB	8
	Montana	Swan Lake	SwB	4
	Montana	Meadow Creek	McB	4
	Montana	Whitefish Lake	WlB	5
	Montana	Whitefish River	WrB	5
	Idaho	Clearwater River	ClB	4
	Nevada	Jarbidge River	JaB	3
	Oregon	Metolius River	MrB	3
	Oregon	North Fork Sprague River	SpB	2
	Washington	Hoh River	HoB	1
	Washington	Skokomish River	SkB	2
Lake Trout	Alaska	Fielding Lake	FiL	2
	Alaska	Hidden Lake	HiL	2
	Alaska	Lake Schrader	ScL	2
	Alaska	Ugashik Lake	UgL	2
	Idaho	Lake Pend Oreille	PeL	2
	Michigan	Rush Lake	RuL	2
	Minnesota	Lake Superior	SuL	32
	Montana	Cosely Lake	CoL	2
	Montana	Flathead Lake	FlL	32
	Montana	Quartz Lake	QuL	2
	Montana	Saint Mary's Lake	SaL	2
	Montana	Swan Lake	SwL	2
	Montana	Yellowstone Lake	YeL	2
	North West Territory	Great Bear Lake	GrL	2
	Ontario	Hawley Lake	HaL	2
	Ontario	Lake Opiongo	OpL	2
	Ontario	Lake of Woods	WoL	2
	Wisconsin	Trout Lake	TrL	2

Table 2

Diagnostic Set	Populations Excluded	Diagnostic Sites
1	Quartz Lake	19392
2	Quartz Lake & Granite Creek	19454
3	Quartz Lake & Hallowat Creek	19408
4	Quartz Lake & No Name Creek	19416
5	Quartz Lake & Whale Creek	19405
6	Quartz Lake & Wounded Buck Creek	19437

Table 3

	Genotypes			Proportion Lake Trout Reads		
Sample ID	Homo. Bull	Het.	Homo. Lake	Homo. Bull	Het.	Homo. Lake
HaB01	19387	0	0	0.00	NA	NA
GrB02	19258	0	0	0.00	NA	NA
SuL29	0	0	18129	NA	NA	1.00
TrL02	0	0	17917	NA	NA	1.00
BC_2	14534	4845	0	0.00	0.50	NA
QuB01	10413	7155	2	0.04	0.18	1.00
QuB02	19259	65	0	0.00	0.16	NA
QuB03	19169	109	0	0.00	0.15	NA
QuB04	13034	4746	5	0.05	0.17	1.00
QuB05	19203	106	0	0.00	0.15	NA
QuB06	19375	0	0	0.00	NA	NA
QuB07	18843	291	0	0.01	0.15	NA
QuB08	19152	117	0	0.00	0.14	NA
QuB09	18347	604	2	0.02	0.15	1.00
QuB10	18949	216	0	0.01	0.14	NA
QuB11	13674	4236	3	0.03	0.17	1.00
QuB12	19365	2	0	0.00	0.10	NA
QuB13	18755	362	0	0.01	0.15	NA
QuB14	19214	71	0	0.00	0.15	NA
QuB15	18300	627	0	0.01	0.15	NA
QuB16	19146	140	0	0.01	0.16	NA
QuB17	19202	85	0	0.00	0.15	NA
QuB18	18840	281	0	0.01	0.15	NA
QuB19	19288	58	0	0.00	0.16	NA
QuB20	19069	102	0	0.00	0.15	NA
ArM01	16935	808	2	0.01	0.24	0.97
ArM02	15110	1746	3	0.01	0.21	0.91
ArM04	11968	3970	5	0.01	0.20	0.96
ArM08	7293	8524	6	0.02	0.22	0.94
ArM50	279	18367	283	0.01	0.57	0.97