Limited evidence of a genetic basis for sex determination in the common creek chub, *Semotilus atromaculatus*

Amanda V. Meuser^{1,2}, Cassandre B. Pyne², Elizabeth G. Mandeville²

¹ Department of Molecular and Cellular Biology, University of Guelph, Guelph, Ontario, Canada

² Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada

Corresponding author: Amanda Meuser
University of Guelph
50 Stone Road East
Guelph, Ontario N1G 2W1, Canada
ameuser@uoguelph.ca

ORCID IDs: Amanda V. Meuser - 0000-0001-9554-2002, Cassandre B. Pyne - 0000-0002-2344-2800, Elizabeth G. Mandeville - 0000-0002-9461-9140

Running title: Sex determination in creek chub

Keywords: Sex determination, sex chromosomes, creek chub, *Semotilus atromaculatus*, genotyping-by-sequencing, RADSex, GEMMA, F_{ST} , linkage disequilibrium

Author contributions: AVM and EGM planned research, designed analysis strategy, and implemented bioinformatics workflows. CBP contributed to analyses. All authors wrote and revised the manuscript.

Conflict of Interest: The authors have no conflict of interest to declare.

Acknowledgements: Thanks to the McCann Lab at the University of Guelph for collection of the Creek Chub samples. Permission to collect was also granted from private landowners and the following conservation areas: Hawkins Trail Conservation Area, Kettle Creek Conservation Area, Watson Conservation Area, Backus Woods Conservation Area. This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund Food From Thought program at the University of Guelph. Computing was accomplished through a RRG Allocation on Compute Canada's Graham cluster to E.G. Mandeville. We thank Daniel Fabrizio for assistance with DNA extractions. This manuscript was improved by comments from Eryn McFarlane, Ben Schultz, Jillian Campbell, Amy Pitura, Eric Sartor, and Erza Dafota.

Limited evidence of a genetic basis for sex determination in the common creek chub, *Semotilus atromaculatus*

1 Abstract

Sexual reproduction is almost universal in the animal kingdom; therefore, each species must have a mechanism for designating sex as male or female. Fish especially have a wide range of sex determining systems. While incompatible systems are thought to increase reproductive isolation, interspecific hybridization is common among groups such as cyprinid minnows, thus, studies such as this can provide insight into hybridization and evolutionary diversification of this clade. In the present study, we aimed to identify a genetic basis for sex determination in the common creek chub (*Semotilus atromaculatus*) using genotyping-by-sequencing (GBS) data. No sex-associated markers were found by RADSex or a GWAS using GEMMA, however, our Weir and Cockerham locus-specific F_{ST} analysis and discriminant analysis of principal components revealed some genetic differentiation between the sexes at several loci. While no explicit sex determination mechanism has been yet discovered in creek chub, these loci are potential candidates for future studies. This study also highlights technical challenges involved in studying sex determination in species with extremely variable mechanisms.

2 Introduction

Incompatible sex determination mechanisms or loci may lead to diversification of species and an increase in biodiversity, as it contributes to reproductive isolation. Conversely, genetically distinct species with compatible sex determination mechanisms may be able to hybridize, which could reduce overall biodiversity if either parent species is lost to genetic or demographic swamping (Dowling & Demarais 1993, O'Neill & O'Neill 2018, Todesco *et al.* 2016). Understanding how sex determination plays a role in preventing or allowing species hybridization can shed light on the evolution of species-rich groups such as cyprinid minnows, as well as how anthropogenic disturbance may influence this evolution through hybridization in the future, as previously isolated species are being newly introduced into sympatry (Grabenstein & Taylor 2018, Todesco *et al.* 2016).

Sex determination mechanisms are extremely diverse, but can be broadly bisected into genetic sex determination (GSD) and environmental sex determination (ESD), for gonochoristic species (those with separate sexes) (Bachtrog *et al.* 2014, Charnov & Bull 1977, Herpin & Schartl 2015, Heule *et al.* 2014). In some cases of GSD, multiple genes or genetic regions across two or more chromosomes work in a threshold-dependent manner to determine sex. These are known as either polygenic sex determination or multiple sex chromosome systems, which may arise from chromosomal fissions and fusions (Bachtrog *et al.* 2014, Bulmer & Bull 1982, Charnov & Bull 1977, Faggion *et al.* 2019, Heule *et al.* 2014, Kitano & Peichel 2012, Kosswig 1964, Vandeputte *et al.* 2007).

More commonly, however, GSD is under the control of a single locus, the master sex determining (MSD) gene (Herpin & Schartl 2015, Heule *et al.* 2014, Pan *et al.* 2019, Pennell *et al.* 2018). An obvious example of GSD with an MSD gene is the heteromorphic XY sex chromosomes in mammals, universal across the clade. The male Y chromosome contains the gene Sry, which acts as a transcription factor for genes required to form the male phenotype (Bachtrog *et al.* 2014, Herpin & Schartl 2015, Kashimada & Koopman 2010). Due to the lack of homology with the X chromosome, this area became isolated from recombination. Over time, mutations and gene loss have eroded the Y chromosome, such that it became cytologically distinct from its X counterpart (Charlesworth 1996, 2017, Furman *et al.* 2020, Heule *et al.* 2014, Pennell *et al.* 2018).

In contrast to the example of heteromorphic sex chromosomes, species with homomorphic sex chromosomes have the same sex determination genes in both males and females and instead possess different alleles which trigger genes for the development of distinct phenotypes (Furman *et al.* 2020, Heule *et al.* 2014). These sex chromosomes share a greater percentage of homology, are able to recombine to a greater extent, are subject to lower rates of degradation through gene loss, and remain cytologically indistinguishable from their counterpart (Charlesworth 1996, Heule *et al.* 2014, Pan *et al.* 2019). Rates of linkage disequilibrium (LD), the measure of non-random assortment of alleles, can therefore be used to support evidence for potential sex determination regions, as markers present in these non-recombining areas would be expected to have high values of LD (Charlesworth 1996, Kamiya *et al.* 2012, Pan *et al.* 2019). For example, tiger pufferfish (*Takifugu rubripes*) possess homomorphic sex chromosomes with alleles that differ by only a single nucleotide and recombine freely throughout this region of the genome (Kamiya *et al.* 2012). Thus, homomorphic sex chromosomes are often more arduous to detect than heteromorphic, especially in fish, which vary greatly in sex determination systems from species to species (Bachtrog *et al.* 2014, Natri *et al.* 2019, Pan *et al.* 2019).

In this study, we used data from individuals of known sex to investigate a potential genetic basis of sex determination in *Semotilus atromaculatus* (the common creek chub), a cyprinid fish. While hermaphroditism is ten times more common in fish than other animal clades, gonochorism has been exclusively observed in creek chub, and most cyprinid fish are believed to have genetically determined sex (Bachtrog *et al.* 2014, Pennell *et al.* 2018, Copes 1978). Thus, only the mechanisms of sex determination for gonochoristic species were investigated. Several complementary genomic analyses were used on genotyping-by-sequencing (GBS) data, with the objective of determining areas of the genome that are associated with sex in creek chub.

3 Materials and methods

3.1 Experimental model and subject details

Creek chub fish were collected from 14 agricultural streams across Southern Ontario in the summer of 2019. Fish were euthanized using an overdose of MS-222, then fin tissue was

sampled and preserved in 95% ethanol. 79 individuals were used in this study; 38 females and 41 males. Sex was determined through dissection.

Sampling was done under the consent of Animal Use Protocol #3682, approved by the University of Guelph Animal Care Committee (F. Laberge PI) and Ministry of Natural Resources Licence to Collect Fish for Scientific Purposes (K. McCann PI).

3.2 DNA isolation and genomic sequencing

We extracted DNA from each of the samples using the QIAGEN DNeasy Blood & Tissue Kit (Qiagen, Inc.), according to the manufacturer’s instructions. Following DNA extraction, we prepared genomic libraries using a reduced-representation genotyping by sequencing method (Parchman *et al.* 2012) to prepare samples for high-throughput sequencing. DNA was digested with EcoRI and MseI restriction enzymes, then adaptors containing the adaptor sequence, barcode, cut site, and a protector base were ligated onto the fragments, before being amplified with Illumina PCR primers added onto either end (Parchman *et al.* 2012). After amplification, the DNA samples were sent to the Genomic Sequencing and Analysis Facility at the University of Texas at Austin. There, the samples were size selected for fragments 300-400bp in length using the Sage Science Blue Pippin, then sequenced with an Illumina NovaSeq 6000. All computing was accomplished via an allocation on Compute Canada’s high-performance computing cluster, Graham, with the exception of data visualization in RStudio on a local computer.

3.3 Primary investigation of all GSD types

We used several complementary methods to investigate the genetic basis of sex determination in creek chub. We first demultiplexed the raw FASTQ file using sabre (<https://github.com/najoshi/sabre>), resulting in one FASTQ file for each individual fish. We performed no additional quality control on the reads. While there may be some unintended additional sequences remaining in the data (e.g. primers, adaptors), it is not expected to bias our results. Initial analysis of sex-associated markers in the creek chub genome was performed using RADSex, a computational workflow intended as an “all-in-one” GSD detection program and visualizer, with the assistance of the RStudio package `sgtr` (Feron *et al.* 2021). RADSex treats all markers as unassociated genotypes, rather than alleles of the marker, and searches for markers that are found in one sex exclusively, at a statistically significant level (Feron *et al.* 2021). Thus, it is theoretically able to detect heteromorphic or homomorphic sex chromosomes, or polygenic sex determination loci, using only one workflow. Additionally, it does not require a genome for alignment, making it ideal for use with non-model organisms, and does not require imputation before use as it allows for missing data.

RADSex takes an input of a FASTQ file of demultiplexed reads for each individual and a file with a list of each individual and their sex. First, the software creates a table containing each marker and its read depth, then the table is filtered such that only markers with a certain read depth are retained (Feron *et al.* 2021). To fit the low-coverage GBS data used for the present study, we set the depth threshold to one read. Next, RADSex identifies makers that

are significantly associated with sex, using Pearson’s chi-squared test of independence (Feron *et al.* 2021). We visualized the output in RStudio using the `radsex_distrib` function from the accompanying R package, `sgtr`, which creates a tile plot (Fig. 1).

3.4 Identification of variant sites and filtering of VCF files

As there is not yet a reference genome for creek chub, we constructed an artificial reference genome with CD-HIT (as described in LaCava *et al.* 2019). We aligned FASTQ files to this reference and sorted into contigs using the Burrows-Wheeler Alignment tool (BWA) (Li & Durbin 2009, version 0.7.17). We identified variable genetic loci using `bcftools mpileup` (Li & Durbin 2009, version 1.11) and converted the resulting BCF to a VCF file using SAMtools (Li *et al.* 2009, version 1.12). We then filtered the resulting VCF using VCFtools (Danecek *et al.* 2011). We produced two different filtered VCFs to fit our analyses.

First, we produced a higher-coverage VCF. This VCF originally contained 616,477 markers, but was filtered to require that retained loci have data for 40% or more of individuals and a minor allele frequency of 0.01. Markers were thinned such that only one marker on each contig was kept, so that all markers being analyzed were completely unlinked. This filtered VCF file contained 3052 markers, had a mean read depth of 33 reads per marker, and a median read depth of 20 reads per marker.

However, the filter initially applied for missing data may have excluded markers that were associated with sex. A sex-associated marker may only be present in about 40% or less of the sample population, as the marker may not be present in one sex or every individual of that sex due to stochasticity in library preparation and sequencing. After re-filtering to adjust the missing data proportion to 0.2 (markers present in only 20% or fewer individuals were excluded) the new VCF file contained 16020 markers, had a mean depth of 11 reads per marker, and a median depth of 8 reads per marker.

We also aligned raw data to a draft *de novo* genome assembly for the northern redbelly dace (*Chrosomus eos*; Schultz & Mandeville, unpublished data). Although this reference genome is more distantly related to creek chub and we therefore receive a lower alignment rate than aligning to the artificial creek chub reference genome (95% alignment to the artificial reference genome and 73% to the dace genome), aligning to a reference with larger scaffolds might be preferable for this analysis, as it could allow better identification of physically linked loci associated with sex determination (Gopalakrishnan *et al.* 2017). We performed the alignment and created a VCF file with the same methodology as above, with a missing data allowance of 40%. This VCF file originally contained 1,233,215 markers, which dropped to 1132 after filtering, and had a mean read depth of 25 and a median read depth of 15.

3.5 Investigation of homomorphic sex chromosomes and polygenic sex determination

As there is some evidence that RADSex might work less well in datasets with lower average sequencing depth (as detailed in the methods paper describing RADSex; Feron *et al.* 2021),

we also applied a number of other analyses to identify potential sex-specific or sex-associated loci. We next performed a GWAS using GEMMA, a single-locus GWAS software (Zhou & Stephens 2012). Essentially, GEMMA treats each marker as a polymorphic allele and searches for two genotypes present exclusively in either sex. Therefore, GWAS would be able to identify markers from homomorphic sex chromosomes or regions from polygenic sex determination, but not heteromorphic sex chromosomes.

GEMMA requires genotype data at all markers, therefore, imputation of missing genotypes must be done beforehand. This increases statistical power compared to eliminating markers without genotypes for all individuals, which would drastically reduce the total number of markers used in the analysis (Guan & Stephens 2008). We used the software entropy for imputation of missing genotypes, as it estimates posterior genotype probabilities at each locus, for each individual (Gompert *et al.* 2014, Shastry *et al.* 2021). The phenotype file contains a numeric character for the trait of each individual (Zhou & Stephens 2012). As the trait being examined, sex, is binary, 0 represented male and 1 represented female. The annotation file contained the information for each marker: the marker ID, base position, and contig number, sorted in the same order as the genotype file (Zhou & Stephens 2012). A relatedness matrix is first created that shows a measure of how related each of the individuals are to one another, then this file is used to run the univariate linear mixed model association test (Zhou & Stephens 2012).

GEMMA results were visualized in RStudio, using a custom R script to create a Manhattan plot (Figure 2). The threshold for this plot is set at $-\log_{10}(5e-6)$. Loci with p-values above this threshold on the Manhattan plot are considered associated with sex at a statistically significant level and could indicate regions of the genome that are part of either a homomorphic sex chromosome or a polygenic sex determination system.

3.6 Analysis of genetic differentiation

Regions of the genome associated with either a homomorphic sex chromosome or polygenic sex determination are expected to be highly differentiated between males and females of a species relative to other regions of the genome. To complement GWAS analyses, we calculated Weir and Cockerham locus-specific F_{ST} using VCFtools (Danecek *et al.* 2011, Weir & Cockerham 1984).

We used Rosner's test to identify significant outliers from the Weir and Cockerham F_{ST} values (Dan & Ijeoma 2013). While the test is innately two-tailed and the outliers being examined only exist in the upper tail, the test is still suitable as any negative values returned will be excluded because they are not biologically relevant. As well, the data set fits the test's assumption of a normal distribution. The test was performed for all three data sets in RStudio, using the `rosnerTest` command from the package, `EnvStats`. The value of alpha (p-value) was set at the default 0.05. The output is described in the relevant section of Results.

3.7 Discriminant analysis of principal components

We completed an additional analysis, a discriminant analysis of principal components (DAPC) (Jombart *et al.* 2010), through the R package *adegenet* (Jombart 2008) to identify loci that contribute strongly to genetic differences between male and female individuals (similar to Junker *et al.* 2020). We first identified sex-associated loci using DAPC, then examined patterns of heterozygosity at the loci most strongly associated with differentiation between sexes.

3.8 Linkage disequilibrium of F_{ST} and DAPC outlier loci

Linkage disequilibrium was quantified by calculating the squared coefficient of correlation (r^2) between pairwise comparisons of all markers using VCFtools (Danecek *et al.* 2011, Pritchard & Przeworski 2001). Using RStudio, we then took subsets of the data that only contained markers deemed outliers from DAPC and the Rosner's Test on F_{ST} values. A Whitney-Mann U test was performed on the data, using the function `wilcox.test`, to assess whether the means of the subsets were significantly different from those of the entire relevant dataset, therefore undergoing less recombination than the rest of the genome (Mann & Whitney 1947, Wilcoxon 1945). This was done for all 3 VCF files, resulting in 6 calculations that are summarized in Table 1.

4 Results

4.1 Using RADSex and GWAS, no markers were found to be associated with sex

Using Radsex and GEMMA we found no loci significantly associated with sex (Fig. 1 and 2). Additionally, for RADSex, there were no markers with alleles shared by more than 19 of the 39 females or 21 of the 41 males.

4.2 Weir and Cockerham F_{ST} analysis revealed some locus-specific genetic differentiation between sexes

The 99th quantiles for each the 1132-, 3052-, and 16020-marker data sets were 0.136, 0.079 and 0.112 respectively, while the average F_{ST} values were all approximately 0 (Fig. 3). A Rosner's test performed on the 16,020-marker data set showed that the markers with the largest 97 F_{ST} values, ranging in value from 0.130 to 0.379, were statistical outliers that deviated from a normal distribution of F_{ST} , with p -value < 0.05 . Similarly, for the 3052- and 1132-marker F_{ST} data sets, the 20 highest F_{ST} values, ranging in value from 0.087 to 0.224, and the 13 highest F_{ST} values, ranging in value from 0.134 to 0.300, respectively, were found to be statistical outliers which deviated from a normal distribution of F_{ST} , p -value < 0.05 .

4.3 Values from F_{ST} and GWAS had a weak, negative relationship

A scatter plot was created for all three data sets to examine potential correlation of F_{ST} and GWAS p-values. Ideally, some of the markers with F_{ST} values determined to be outliers may have correlated positively with the markers with higher scoring, albeit non-significant, values from GEMMA. This was not the case, as the values were weakly correlated, often in the negative direction, when a simple linear regression was performed using R (Fig. ??). The 1132-, 3052-, and 16,020-marker data sets had a regression line with a slope of 0.298, -0.246, and -0.193, a p-value of 0.335, 0.496, and 0.036, and 1127, 2008, and 15961 degrees of freedom, respectively. While one of these findings is statistically significant, they are all unlikely to be biologically significant.

4.4 Discriminant analysis of principal components

DAPC analyses identified a small number of loci that are differentiated between sexes in each data set. For the alignment to the artificial reference genome, 128 loci were significantly associated with sex in the 16,020 locus data set, and 27 loci were significantly associated with sex in the 3,052 locus data set. For the data set aligned to the dace reference genome (1132 markers), only 5 loci were significantly associated with sex. These loci were spread across 5 different scaffolds, rather than co-localized on a single large scaffold, and are therefore not consistent with placement on a differentiated sex chromosome or in proximity to a master sex determining gene.

4.5 Linkage disequilibrium was slightly higher among F_{ST} and DAPC outlier loci

In each data set, a small number of overlapping loci were identified as both F_{ST} outliers and outliers in the DAPC analyses (Fig. 4). The subsets of outlier markers from F_{ST} analysis had similar read depth to the average for the entire set: 11, 44, and 27 read depths on average for the subsets compared to 11, 33, and 25 for the whole 16,020-, 3052-, and 1132-marker data sets, respectively. Whereas the subsets of DAPC outliers had much higher mean read depths of 41, 119, and 35, respectively.

Linkage disequilibrium among loci identified as sex-associated by F_{ST} and DAPC analyses was slightly higher than for the entire data set on average, suggesting that these loci are more closely linked and may be physically closer together in the genome or otherwise inherited together. The r^2 values for all data sets follow a right-skewed, exponential distribution. For the original 16,020-marker data set, the mean r^2 value for the F_{ST} outliers was 0.103 and the mean of the entire data set was 0.054. The Whitney-Mann U test resulted in p-value = 0.003. The mean r^2 value for the DAPC outliers was 0.068 and the p-value from the Whitney-Mann U test was $< 2.2e-16$ (Table 1). Similarly, for the 3052-marker data set, the mean r^2 of the F_{ST} subset is 0.206, the mean of the entire data set is 0.031 and the Whitney-Mann U test resulted in p-value $< 2.2e-16$. The mean r^2 value for the DAPC outliers was 0.035 and the p-value from the Whitney-Mann U test was $3.745e-13$ (Table 1).

Therefore, while the means were not large, they were still all significantly higher for both data sets aligned to our artificial reference genome. For the 1132-marker data set, the mean overall r^2 value is 0.0449, the mean r^2 for significant DAPC values is 0.063, and the mean r^2 for F_{ST} outlier values is 0.102. The difference was significant for the F_{ST} outliers (p-value = 1.21×10^{-12}), but not for the outliers from DAPC (p-value = 0.711) (Table 1). Again, the LD values are quite low, but taken all together, LD tended to be higher for the significant F_{ST} and DAPC markers when compared to other markers in each data set.

5 Discussion

The study of sex determination presents significant challenges, because many clades, including some vertebrates, feature a diversity of sex determination systems (Bachtrog *et al.* 2014). While mammals and birds follow a straightforward heteromorphic sex chromosome system, other vertebrate clades may use much more subtle means that remain elusive to our research methods (Charlesworth 1996). Fish are particularly diverse in their systems of sex determination, and such systems are believed to have rapid evolutionary turnover in these lineages (Heule *et al.* 2014, Pennell *et al.* 2018).

In this context, it is not entirely unexpected that analyses of potential sex determining regions in this study produced equivocal results. Overall, while no statistically significant sex-associated markers were found in the analyses with RADSex and GEMMA, some loci identified by analyses of F_{ST} outliers and discriminant analysis of principal components may be candidates for involvement in sex determination in creek chub. Based on these results, and because most North American cyprinid fish with known sex determination have some form of genetic sex determination, we believe that it is still likely that there is a genetic basis for sex determination in creek chub (Ashman *et al.* 2014). Based on our analysis, it is likely that sex determination in creek chub may take the form of homomorphic sex chromosomes with a small sex determining region, or polygenic sex determination.

5.1 The sex determination region may be small or minimally differentiated

Based on initial results, it is unlikely that creek chub possesses highly differentiated, heteromorphic sex chromosomes similar to an XX/XY or ZW/ZZ sex determination system. An area of the genome as large and differentiated as a heteromorphic sex chromosome would have likely been detected by at least a small number of markers using RADsex (Fig. 1), had greater genetic differentiation between sexes (Fig. 3), or exhibited greater loadings or involved more loci in the DAPC analyses.

The creek chub genome has 25 chromosomes, so on average, our data set would have contained over 100 markers per chromosome in the 3052-marker data set and over 600 markers per chromosome in the 16,020-marker data set (Gold & Amemiya 1987). While this is still only a small fraction of the genome, as is typical of reduced-representation sequencing strategies like genotyping by sequencing and RADseq, it is unlikely that a chromosome

as differentiated between the sexes as a heteromorphic sex chromosome would have gone completely undetected.

Creek chub may instead have homomorphic sex chromosomes; however, the sex determining region would likely be quite small or scarcely differentiated, as it was not detected by GWAS and only a few loci were sex-associated in Fst and DAPC analyses. The same sentiment exists for polygenic sex determination, where there may be multiple, small sex determination regions across two or more chromosomes that evaded targeting by the sequenced markers. The latter is quite probable, as the loci of interest were scattered throughout the genome rather than localizing to one scaffold in the dace genome alignment.

Additionally, these few, scattered sex-associated loci may suggest that sex determination systems are not a reproductive barrier for cyprinid fish. Sex chromosomes can change in fashion quickly and species with heteromorphic sex chromosomes are faster at reaching reproductive isolation than those with other forms (Pennell *et al.* 2018). Haldane's rule also states in cases where hybrid individuals are inviable or infertile, it is more likely to be the heterogametic sex that is negatively affected (Haldane 1922). Given that we find no evidence for heterogamy in creek chub, and as interspecific hybridization and production of fertile offspring is known to occur among cyprinid fishes including creek chub and a suite of closely related species, this suggests that their sex determination system is not a barrier to reproduction and that Haldane's rule is unlikely to be involved in the maintenance of reproductive isolation between cyprinid minnows (Broughton *et al.* 2011, Dowling & Demarais 1993, Ross & Cavender 1981). This further supports our evidence which suggests creek chub likely have small or minimally differentiated homomorphic sex chromosomes or a polygenic sex determination system.

5.2 Assessing the likelihood of environmental sex determination in creek chub

Creek chub could also determine sex using external environmental influences, or a combination of environmental and genetic sex determination (as in sea bass *Palaiokostas et al.* 2015). While most species with a known sex determination mechanism are exclusively subject to GSD, a fraction of reptile and fish species may undergo either ESD or GSD with effects of temperature on sex ratios (GSD + TE) (Bachtrog *et al.* 2014, Charnov & Bull 1977, Muralidhar & Veller 2018, Shen & Wang 2019). In this process, external factors, such as temperature, pH, photoperiod, or social factors, induce epigenetic modification or hormonal targeting of sex determination genes (Heule *et al.* 2014, Shen & Wang 2019). This reduces gene expression, promoting development of an alternate phenotypic sex (Heule *et al.* 2014, Shen & Wang 2019). In the case of environmental sex determination in creek chub, there would be little to no sex-related genomic differences between males and females; therefore, all high LD, F_{ST} , and DAPC outliers values identified would just be coincidental.

However, to the best of our current knowledge, ESD is almost non-existent in the Cyprinidae family and seen in only 15% of bony fishes (Ashman *et al.* 2014, Bachtrog *et al.* 2014). While turnover of various GSD mechanisms is quite common within fish families (for example in sticklebacks Natri *et al.* 2019), an entire switch to ESD is not (Heule *et al.* 2014,

Pan *et al.* 2019, Pennell *et al.* 2018). Thus, while the possibility exists for creek chub to undergo ESD, particularly GSD + TE or a mixed system, it is most likely still the case that GSD is the mechanism of sex determination in these fish (Muralidhar & Veller 2018, Shen & Wang 2019), but that either the genetic architecture of sex determination is too complex to conclusively identify it, or our dataset did not contain the relevant genomic regions.

This sort of ambiguity about environmental vs. genetic sex determination is not unique to creek chub. Sex determination in lampreys, for example, has remained a mystery despite over 20 years of study. The mechanism has been suggested to be environmental, as there has been no large sex chromosomes found and there is some evidence for density-dependent sex determination (Docker & Beamish 1994, McCauley *et al.* 2015). Thus, while studies with equivocal results about the basis of sex determination may seem less satisfying than those that confidently identify a clear genetic basis of sex determination, studying more challenging species provides a unique opportunity to learn more about non-standard sex determination systems.

5.3 Technical limitations associated with the identification of sex determining regions

Thus far, we have focused on biological interpretations of our results, but technical limitations may also contribute to the inability to detect the genetic basis of sex determination in some studies, including this one. Specifically, reduced-representation library preparation and sequencing methods like GBS, ddRAD, and RAD sequence only a small fraction of the genome, despite identifying thousands of single nucleotide polymorphisms (SNPs), so in any study like this, the loci of interest may be excluded (Lowry *et al.* 2017). This becomes particularly relevant when attempting to identify the genetic basis traits that are controlled by many loci of small effect, like a polygenic sex determination system might be.

As well, the power of this study may have been somewhat limited by our sample size of 79 individuals – 39 males and 41 females. In a study on quantitative trait loci, which work in similar fashion to polygenic sex determination loci, power analyses on GWAS done by Kardos *et al.* 2016 show that it may be difficult to identify loci that each work in small effect, without thousands of individuals. However, that is not to say that this sample size would be ineffective at identifying loci with large effect sizes, especially for a binary trait like sex in our sampled fish. For the binary trait of survival, Margres *et al.* 2018 used a similar number of individuals for GEMMA – 41 males and 69 females – but were able to find 5 SNPs in females and 37 in males that explained 76% and 43% of phenotypic variance, respectively. Therefore, while our sample size still may not have been strong enough to detect multiple small polygenic regions, it would have likely wielded enough power to detect a large sex determining region or chromosome.

Alternatively, our data set may have contained sex-associated loci that we were unable to detect statistically due to insufficient sequencing depth. Marker depth was extremely variable across loci, and many marker depths were relatively low, particularly in the 16,020-marker data set (Fig. S2 B). High marker depth was associated with results from DAPC, but not from Weir and Cockerham F_{ST} , suggesting that the results from DAPC may be

somewhat biased towards markers with better coverage. While DAPC might be affected by read depth, F_{ST} doesn't seem to be, so those markers that stick out from both DAPC and F_{ST} analyses might be more reliable than just the ones from DAPC.

Additionally, there were no markers in the RADSex output tile plot that were shared by all individuals, likely because the read depth was low and variable enough that not all markers had even one read for all individuals (Fig. 1). This was not the case for the any of the 7 species with shown tile plots in the methods paper describing the software (Feron *et al.* 2021), therefore, RADSex may not have been able to accurately assess the creek chub sequencing data, leading to a null result. Issues with sequence depth at individual loci may have arisen with GEMMA as well. However, methods like DAPC have been successfully used in to identify sex-associated loci in similar data sets (e.g. in Lake Tanganyika sardines Junker *et al.* 2020), so it is difficult to parse how much data quantity/quality contributes to this result relative to actual complexity of sex determination systems. It is clear that methods for the detection of sex determination systems are still in need of development, especially for studies of non-model organisms without reference genomes and lower-coverage data sets that prioritize sampling more individuals over sequencing at higher coverage, as is the case for many conservation-associated studies.

5.4 Future directions

While GSD is the most likely method of sex determination in creek chub, to date, no studies have been done on this species to rule out the effects of temperature or other environmental influences on sex ratios of creek chub offspring. ESD is unlikely to be the only mechanism behind sex determination in this species, however, temperature effects may play a role (Muralidhar & Veller 2018, Shen & Wang 2019). Additionally, development of a creek chub genome assembly would allow for genome-wide linkage mapping and assessment of gene ontology in areas of recombination suppression, as done by Kamiya *et al.* 2012, as MSD genes are often transcription factors involved in development and maintenance of reproductive organs (Herpin & Schartl 2015, Pan *et al.* 2019). Further research into sex determination in creek chub and other minnows in the *Cyprinidae* family will likely yield insight into the evolutionary diversification and current dynamics of interspecific hybridization in this species-rich group of fishes.

Data accessibility

Upon the acceptance of this manuscript, data and scripts used for analysis will be made publicly available in Data Dryad and the NCBI Sequence Read Archive. These materials will be made available to reviewers upon request.

References

- Ashman TL, Bachtrog D, Blackmon H, *et al.* (2014) Tree of Sex: A database of sexual systems. *Scientific Data*, **1**, 1–8.
- Bachtrog D, Mank JE, Peichel CL, *et al.* (2014) Sex Determination: Why So Many Ways of Doing It? *PLoS Biology*, **12**, 1–13.
- Broughton RE, Vedala KC, Crawl TM, Ritterhouse LL (2011) Current and historical hybridization with differential introgression among three species of cyprinid fishes (genus *Cyprinella*). *Genetica 2011 139:5*, **139**, 699–707.
- Bulmer MG, Bull JJ (1982) Models of Polygenic Sex Determination and Sex Ratio Control. **36**, 13–26.
- Charlesworth B (1996) The evolution of chromosomal sex determination and dosage compensation. *Current Biology*, **6**, 149–162.
- Charlesworth D (2017) Evolution of recombination rates between sex chromosomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **372**.
- Charnov EL, Bull J (1977) When is sex environmentally determined? *Nature*, **266**, 828–830.
- Copes FC (1978) Ecology of the Creek Chub. *Reports on the Fauna and Flora of Wisconsin*, **12**, 1–24.
- Dan ED, Ijeoma OA (2013) Statistical Analysis/Methods of Detecting Outliers In A Univariate Data In A Regression Analysis Model. *International Journal of Education and Research*, **1**, 24.
- Danecek P, Auton A, Abecasis G, *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
- Docker MF, Beamish FW (1994) Age, growth, and sex ratio among populations of least brook lamprey, *Lampetra aepyptera*, larvae: An argument for environmental sex determination. *Environmental Biology of Fishes*, **41**, 191–205.
- Dowling TE, Demarais BD (1993) Evolutionary significance of introgressive hybridization in cyprinid fishes. **362**, 444–446.
- Faggion S, Vandeputte M, Chatain B, Gagnaire PA, Allal F (2019) Population-specific variations of the genetic architecture of sex determination in wild European sea bass *Dicentrarchus labrax* L. *Heredity*, **122**, 612–621.
- Feron R, Pan Q, Wen M, *et al.* (2021) RADSex: A computational workflow to study sex determination using restriction site-associated DNA sequencing data. *Molecular Ecology Resources*.
- Furman BL, Metzger DC, Darolti I, *et al.* (2020) Sex Chromosome Evolution: So Many Exceptions to the Rules. *Genome Biology and Evolution*, **12**, 750–763.

- 453 Gold JR, Amemiya CT (1987) Genome size variation in North American minnows
454 (Cyprinidae). II. Variation among 20 species. *Genome / National Research Council*
455 *Canada*, **29**, 481–489.
- 456 Gompert Z, Lucas LK, Buerkle CA, Forister ML, Fordyce JA, Nice CC (2014) Admixture
457 and the organization of genetic diversity in a butterfly species complex revealed through
458 common and rare genetic variants. *Molecular Ecology*, **23**, 4555–4573.
- 459 Gopalakrishnan S, Samaniego Castruita JA, Sinding MHS, *et al.* (2017) The wolf reference
460 genome sequence (*Canis lupus lupus*) and its implications for *Canis* spp. population ge-
461 nomics. *BMC Genomics*, **18**.
- 462 Grabenstein KC, Taylor SA (2018) Breaking Barriers: Causes, Consequences, and Experi-
463 mental Utility of Human-Mediated Hybridization. *Trends in Ecology and Evolution*, **33**,
464 198–212.
- 465 Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. *PLoS*
466 *Genetics*, **4**, 1–11.
- 467 Haldane JB (1922) Sex ratio and unisexual sterility in hybrid animals. *Journal of Genetics*,
468 **12**, 101–109.
- 469 Herpin A, Schartl M (2015) Plasticity of gene-regulatory networks controlling sex determi-
470 nation: of masters, slaves, usual suspects, newcomers, and usurpators. *EMBO reports*, **16**,
471 1260–1274.
- 472 Heule C, Salzburger W, Böhne A (2014) Genetics of sexual development: An evolutionary
473 playground for fish. *Genetics*, **196**, 579–591.
- 474 Jombart T (2008) Adegnet: A R package for the multivariate analysis of genetic markers.
475 *Bioinformatics*, **24**, 1403–1405.
- 476 Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components:
477 A new method for the analysis of genetically structured populations. *BMC Genetics*, **11**,
478 1–15.
- 479 Junker J, Rick JA, McIntyre PB, *et al.* (2020) Structural genomic variation leads to genetic
480 differentiation in Lake Tanganyika’s sardines. *Molecular Ecology*, **29**, 3277–3298.
- 481 Kamiya T, Kai W, Tasumi S, *et al.* (2012) A trans-species missense SNP in *Amhr2* is
482 associated with sex determination in the tiger Pufferfish, *Takifugu rubripes* (Fugu). *PLoS*
483 *Genetics*, **8**.
- 484 Kardos M, Husby A, Mcfarlane SE, Qvarnström A, Ellegren H (2016) Whole-genome rese-
485 quencing of extreme phenotypes in collared flycatchers highlights the difficulty of detecting
486 quantitative trait loci in natural populations. *Molecular Ecology Resources*, **16**, 727–741.
- 487 Kashimada K, Koopman P (2010) Sry: The master switch in mammalian sex determination.
488 *Development*, **137**, 3921–3930.

- 489 Kitano J, Peichel CL (2012) Turnover of sex chromosomes and speciation in fishes. *Environ-*
490 *mental Biology of Fishes*, **94**, 549–558.
- 491 Kosswig C (1964) Polygenic sex determination. *Experientia*, **20**, 190–199.
- 492 LaCava ME, Aikens EO, Megna LC, Randolph G, Hubbard C, Buerkle CA (2019) Accuracy
493 of de novo assembly of DNA sequences from double-digest libraries varies substantially
494 among software. *Molecular Ecology Resources*, **20**, 360–370.
- 495 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler trans-
496 form. *Bioinformatics*, **25**, 1754–1760.
- 497 Li H, Handsaker B, Wysoker A, *et al.* (2009) The Sequence Alignment/Map format and
498 SAMtools. *Bioinformatics*, **25**, 2078–2079.
- 499 Lowry DB, Hoban S, Kelley JL, *et al.* (2017) Breaking RAD: an evaluation of the utility
500 of restriction site-associated DNA sequencing for genome scans of adaptation. *Molecular*
501 *ecology resources*, **17**, 142–152.
- 502 Mann HB, Whitney DR (1947) On a Test of Whether one of Two Random Variables is
503 Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, **18**, 50–60.
- 504 Margres MJ, Jones ME, Epstein B, *et al.* (2018) Large-effect loci affect survival in Tasmanian
505 devils (*Sarcophilus harrisii*) infected with a transmissible cancer. *Molecular Ecology*, **27**,
506 4189–4199.
- 507 McCauley DW, Docker MF, Whyard S, Li W (2015) Lampreys as Diverse Model Organisms
508 in the Genomics Era. *BioScience*, **65**, 1046–1056.
- 509 Muralidhar P, Veller C (2018) Sexual antagonism and the instability of environmental sex
510 determination. *Nature Ecology and Evolution*, **2**, 343–351.
- 511 Natri HM, Merilä J, Shikano T (2019) The evolution of sex determination associated with a
512 chromosomal inversion. *Nature Communications*, **10**.
- 513 O'Neill MJ, O'Neill RJ (2018) Sex chromosome repeats tip the balance towards speciation.
514 *Molecular Ecology*, **27**, 3783–3798.
- 515 Palaikostas C, Bekaert M, Taggart JB, *et al.* (2015) A new SNP-based vision of the genet-
516 ics of sex determination in European sea bass (*Dicentrarchus labrax*). *Genetics Selection*
517 *Evolution*, **47**, 1–10.
- 518 Pan Q, Feron R, Yano A, *et al.* (2019) Identification of the master sex determining gene
519 in Northern pike (*Esox lucius*) reveals restricted sex chromosome differentiation. *PLoS*
520 *Genetics*, **15**.
- 521 Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA (2012)
522 Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecol-*
523 *ogy*, **21**, 2991–3005.

- 524 Pennell MW, Mank JE, Peichel CL (2018) Transitions in sex determination and sex chro-
525 mosomes across vertebrate species. *Molecular Ecology*, **27**, 3950–3963.
- 526 Pritchard JK, Przeworski M (2001) Linkage Disequilibrium in Humans: Models and Data.
527 *The American Journal of Human Genetics*, **69**, 1–14.
- 528 Ross MR, Cavender TM (1981) Morphological Analyses of Four Experimental Intergeneric
529 Cyprinid Hybrid Crosses. *Copeia*, **1981**, 377–387.
- 530 Shastry V, Adams PE, Lindtke D, *et al.* (2021) Model-based genotype and ancestry estima-
531 tion for potential hybrids with mixed-ploidy. *Molecular Ecology Resources*.
- 532 Shen ZG, Wang HP (2019) Environmental Sex Determination and Sex Differentiation in
533 Teleosts - How Sex Is Established. *Sex Control in Aquaculture*, **1**, 85–115.
- 534 Todesco M, Pascual MA, Owens GL, *et al.* (2016) Hybridization and extinction. *Evolutionary*
535 *Applications*, **9**, 892–908.
- 536 Vandeputte M, Dupont-Nivet M, Chavanne H, Chatain B (2007) A polygenic hypothesis for
537 sex determination in the European sea bass *Dicentrarchus labrax*. *Genetics*, **176**, 1049–
538 1057.
- 539 Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Struc-
540 ture. Tech. Rep. 6.
- 541 Wilcoxon F (1945) Some Uses of Statistics in Plant Pathology. *Biometrics Bulletin*, **1**, 41–45.
- 542 Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association
543 studies. *Nature Genetics*, **44**, 821–824.

6 Figures

Figure 1: Tile plot depicting the output from RADSex. The number of females with a given marker are displayed on the y-axis and number of males with a given marker on the x-axis, with number of markers held by that number of males and females indicated by darkening colour. Tiles containing markers that are found in only one sex at a statistically significant level would be shown surrounded by a red box, however, no such markers were found.

Figure 2: Manhattan plots displaying output from Wald Chi-Squared Test implemented in GEMMA. The $-\log_{10}(\text{p-value})$ for each marker's association with sex is displayed on the y-axis. The significance threshold is represented by a solid line, at $-\log_{10}(5e-6)$. No significant markers were found in any of the three data sets. A: 3052-marker data set. B: 16,020-marker data set. C: 1132-marker data set with chromosome numbers along the bottom for the 25 largest contigs. Chromosomes 1 to 4 and 7 are not shown as no markers mapped to these regions. Chromosome 0 contains markers from all small, excess contigs that did not map to any other chromosome.

Figure 3: Weir and Cockerham F_{ST} values. Solid line indicates 0 and dashed line indicates 99th quantile. A: Histogram of the 3052-marker data set. B: Scatter plot of the 3052-marker data set. C: Histogram of the 16,020-marker data set. D: Scatter plot of the 16,020-marker data set. E: Histogram of the 1132-marker data set. F: Scatter plot of the 1132-marker data set.

Figure 4: Statistically significant outliers from Fst analysis and DAPC. Markers that overlap between the Fst outliers and DAPC outliers have been identified as loci of interest for future analyses. A: 3052-marker data set. B: 16,020-marker data set. C: 1132-marker data set.

Table 1: Linkage disequilibrium expressed as mean r^2 values and one standard deviation for the 3 data sets and subsets of outlier markers from Fst and DAPC. Values that are statistically different from the mean of the entire data set is denoted with a star.

Data set	Overall mean	Fst outlier loci	DAPC outlier loci
3052-marker	0.0313 ± 0.0891	$0.2064^* \pm 0.1671$	$0.0350^* \pm 0.0487$
16020-marker	0.0541 ± 0.1373	$0.1030^* \pm 0.1497$	$0.0678^* \pm 0.1373$
1132-marker	0.0449 ± 0.0821	$0.1021^* \pm 0.1055$	0.0629 ± 0.0842