Single nucleotide polymorphism genotypes and ploidy estimates for ploidy variable species generated with massively parallel amplicon sequencing

Stuart Willis[1], Thomas Delomas[2], Blaine Parker[3], Donella Miller[4], Paul Anders[5], and Shawn Narum[3]

[1]Texas A&M University Corpus Christi
[2]Pacific States Marine Fisheries Commission
[3]Columbia River Inter-Tribal Fish Commission
[4]Yakama Nation Fisheries
[5]Cramer Fish Sciences

August 28, 2020

Abstract

Polyploidization has played a critical role in the evolution of several major organism groups, including vertebrates, but much of our knowledge of the evolution of polyploids comes from allopolyploid and often rediploidized lineages, which partly reflects the difficulty of obtaining genotype data from polysomic genomes. We combined several contemporary methods to develop markers for single nucleotide polymorphisms compatible with simultaneous ploidy estimation and high throughput genotyping, and analyzed these data with recent software developments that accept polysomic data. We demonstrate the utility of this combination to develop genetic resources for polysomic species by applying it to the ploidy-variable and polysomic white sturgeon (*Acipenser transmontanus*), an imperiled species under conservation management in the Pacific Northwest. We introduce a primer and probe set for 325 SNP markers for use with the 'Genotyping-by-thousands' (GT-seq) method, and provide updated scripts that incorporate a function to estimate ploidy from each individual using read count data. We examine the reliability of tetrasomic inheritance in a large sample of paleo-octoploid individuals and the expected Mendelian inheritance patterns in known cross families. We then demonstrate our ability to use these data to infer parentage, relatedness, and other population genetic parameters. Our combined process thus improves the accessibility of genetic information to facilitate future investigations of white sturgeon and is expected to be widely applicable to other polyploid species.

Introduction

Whole genome duplication is hypothesized to have played a fundamental role in evolution (Dufresne, Stift, Vergilino, & Mable, 2014; Soltis, Visger, Blaine Marchant, & Soltis, 2016), including of vertebrates (Dehal & Boore, 2005; Holland, Garcia-Fernandez, Williams, & Sidow, 1994), and in particular in fishes (Crow, Stadler, Lynch, Amemiya, & Wagner, 2006; Meyer & Van De Peer, 2005). Despite this, there are relatively few extant vertebrate species that are known to be polysomic (exhibiting multivalent chromatids) (Comai, 2005), which stems in part from the processes of diploidization that occur following most polyploidization events (Lynch & Conery, 2000; Ohno, 1971; Wendel, 2000; Wolfe, 2001). Select lineages however, including some vertebrates, appear to be prone to episodic polyploidization and prolonged polysomism (Dufresne et al., 2014).

Despite the obvious differences, our understanding of polyploid evolution has largely come via study of allopolyploids, those that arise from combination of two ancestral genomes, usually through hybridization, rather than their autopolyploid counterparts, which arise from the doubling of a single ancestral genome, usually through fertilization of unreduced gametes (Dufresne et al., 2014; Soltis et al., 2016). In part this stems from several methodological challenges to developing genetic insights from polyploids, which are often more significant in auto- than allopolyploids. Developing reliable genetic markers for polyploids has been impeded by both the presence of co-amplifying homeologs whose signals cannot be discriminated, as well as true polysomic segregation of those homeologs, with the true somy obscured by homeolog co-amplification. For example, while microsatellites have often been the standard marker for population genetics because of their ease of discovery and high allelic diversity, many studies of polyploids have found mixed inheritance patterns that could reflect true mixed-somy segregation or variable amplification of homeologs from each ancestral genome (Dufresne et al., 2014).

While allopolyploids may often exhibit disomy of the ancestral genomes soon or immediately following polyploidization (Spoelhof, Soltis, & Soltis, 2017), in which case developing diploid markers is a matter of identifying ancestral genome-specific primers or probes (Dufresne et al., 2014), true polysomy in autopolyploids and segmental allopolyploids (those formed from merger of partially divergent ancestral genomes) presents additional challenges. In polysomes, determining the dosage (count or ratio) of microsatellite alleles in an individual's genotype may be difficult when the genotyping technology is not quantitative, and the presence of null alleles can impede this further. Moreover, while the estimators for many population genetic parameters can be extended to include polysomic inheritance (Meirmans & Van Tienderen, 2013; Ronfort, Jenczewski, Bataillon, & Rousset, 1998), until recently there has been relatively little interest in incorporating these extensions into popular genetic software, the majority of which permit only diploid data. Several recent software packages that were updated to permit polyploid data (e.g. Genodive (Meirmans, Liu, & Van Tienderen, 2018), the R package adegenet (Jombart, 2008), and others, reviewed in (Dufresne et al., 2014)) or were designed specifically for polyploids (EBG; (Blischak, Kubatko, & Wolfe, 2018); Polygene; (Kang Huang, Dunn, Ritland, & Li, 2020)) make progress on this front, but these require that the ploidy and either the allelic phenotype (dosage blind genotype) or full ploidy-aware genotype be provided for each individual. For species that vary in ploidy, this generally requires separately assessing ploidy from genotyping/allelic phenotyping, adding time and expense, and in some cases precluding the use of commonly archived tissue types and preservation methods.

Here, we demonstrate a set of methodological and bioinformatic techniques which address many of these challenges in developing genetic resources for a ploidy-variable, polysomic species, the white sturgeon (*Acipenser transmontanus*). The sturgeons (Acipenseriformes) are a classic example of polysomic polyploidy in vertebrates. All extant sturgeons, which exhibit between ~120 and ~360 chromosomes, are hypothesized to be polyploid relative to an extinct diploid ancestor which had 60 chromosomes (Rajkov, Shao, & Berrebi, 2014). The sterlet (*Acipenser ruthenus*), a Eurasian sturgeon with ~120 chromosomes, should by this ratio be tetraploid (4N), but in exploring gene content and homology of a draft genome, Du et al. (2020) discovered extensive, though incomplete, diploidization resulting from a "segmental deduplication" process, while others have inferred both disomic and tetrasomic inheritance of microsatellite markers in this species (Rajkov et al., 2014). By similar logic, the white sturgeon (*A. transmontanus* ; ~240 chromosomes) should be ancestrally octoploid (8N), though microsatellite inheritance patterns have suggested both tetrasomic and octosomic segregation (Drauch Schreier, Gille, Mahardja, & May, 2011). Intriguingly, white sturgeon occasionally exhibit spontaneous autopolyploidy generally resulting in increases of chromatin content by ~1.5 (12N, dodecaploid) (Drauch Schreier et al., 2011; A. D. Schreier, May, & Gille, 2013). And though of unknown fertility, backcrossed offspring (10N, decaploid) are often viable (J. P. Van Eenennaam et al., 2019), creating a wide range of ploidies within a single species.

White sturgeon are the largest freshwater fish in North America, reaching lengths up to 6.1m, though lengths of 2m are more common (Scott & Crossman, 1973). As euryhaline fish, white sturgeon may be found along the Pacific coast as far north as the Aleutian Islands and as far south as northern Baja, though their current strongholds include the Sacramento-San Joaquin, Columbia, and Fraser River Basins (Hildebrand et al., 2016). Although the Columbia basin hosts the largest total aggregation of white sturgeon, their distribution in this system is broken into a number of *de facto* population segments by dams and other river modifications that prevent almost all demographic exchange (Hildebrand et al., 2016). Several of these river sections contain populations classified by the US or Canada as threatened or endangered and even more population segments are in decline due to recruitment limitation resulting from habitat degradation (Hildebrand et al., 2016). While conservation management plans have been developed for most white sturgeon population segments, a lack of robust information about historical and contemporary movement, population structure, and recruitment patterns have hindered effective solutions for these fish, which can take a decade or more to mature (Hildebrand et al., 2016). Obtaining and utilizing genetic data, in particular, has seen challenge not unlike many polyploid species (Anders et al., 2011). While microsatellite markers for white sturgeon have been available for some time (Rodzen, Famula, & May, 2004), the unclear or mixed segregation patterns of these markers has made inferring robust genetic data difficult (Clark & Schreier, 2017; Drauch Schreier et al., 2011). In addition, although some researchers have achieved moderate success by coding the polysomic data as pseudo-dominant di-somic markers or by using ploidy-agnostic analysis methods (Rodzen et al., 2004; A. Drauch Schreier, Rodzen, Ireland, & May, 2012), this has nonetheless limited the types of analyses available.

To remedy these limitations, we developed a set of single-nucleotide polymorphism (SNP) markers using reduced representation genomic libraries and tested the reliability of polysomic segregation patterns by examining inheritance in known cross families and allele ratios in a large sample of individuals. The SNP markers were developed for survey with the 'genotyping-by-thousands' or 'GT-seq' method (Campbell, Harmon, & Narum, 2015), a multiplex amplicon-based method utilizing massively-parallel sequencing to cost-effectively survey hundreds of individuals simultaneously and providing read data approximately proportional to allelic dosage. We provide updated scripts to efficiently genotype polysomic individuals in a ploidy-aware manner by incorporating the *funkyPloid* function from the R package tripsAndDipR v0.2.0 (Delomas et al. submitted), which fits beta-binomial mixture models to the sets of allele read counts and compares the likelihoods of candidate ploidies. This permits each individual to be genotyped in accordance with its inferred ploidy. We demonstrate the utility of these SNPs to infer parentage/relatedness and estimate population and individual-level genetic parameters using a computer package specifically designed for polyploids, Polygene (Kang Huang et al., 2020).

Methods

To address the challenges identified above, our combined process consists of three areas: 1) discovery of single nucleotide polymorphisms in segregating in white sturgeon using reduced representation genomic library sequencing, including filtering steps to improve the likelihood that candidate markers represent single-locus variants and the lowest functional ploidy (somy) of the target species, 2) verification that the markers conform to the expected polysomic segregation and inheritance by examining allele count ratios in a large set of samples and the incidence of Mendelian incompatibilities in a set of known parent-offspring trios, and 3) evaluating the utility of these markers to resolve population genetic parameters via the accuracy of inferred relationships and relatedness and the precision of population statistic estimates.

RAD Library sequencing and SNP Ascertainment

Single nucleotide polymorphic (SNP) loci were identified from a restriction site-associated DNA (RAD) library created using a single enzyme (sbfI) (Miller, Dunham, Amores, Cresko, & Johnson, 2007; Baird et al. 2008; Puritz et al., 2014). Twenty-eight barcoded samples of white sturgeon from a broad geographic range were prepared, pooled equimolarly, and sequenced on an Illumina HiSeq (100bp paired end, quality trimmed to 80bp). The ascertainment panel included 12 samples from the lower/middle Columbia River, 14 from the lower/middle Snake River, and 2 from the Sacramento River. Forward reads were processed using the Stacks (Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011) pipeline, using assembly parameters of M of 2 (maximum mismatch; ustacks), N of 4 (maximum secondary mismatch; ustacks), n of 2 (maximum sample mismatch; cstacks), and m of 4 to 16 (minimum stack depth; ustacks) depending on the depth of coverage of that individual, i.e. the nearest integer using $million\ raw\ reads * 2$. The relatively low mismatch thresholds (versus a default 3) ensured that only very similar reads were assembled, and assisted in preventing homeologs (homologs derived from polyploidization) from being clustered (Dufresne et al., 2014; but see Ilut, Nydam, & Hare, 2014; O'Leary, Puritz, Willis, Hollenbeck, & Portnoy, 2018; Willis, Hollenbeck, Puritz, Gold, & Portnoy, 2017). Stacks were filtered to retain variants with minor allele frequencies above 5%, genotyped in at least 80% of individuals, and for combined depth between 1,050 and 1,600 sequence reads (first mode of a multi-modal distribution; Supplemental Figure 1). Using read ratios as proxy for genotypes of unknown ploidy, a principal components analysis was performed with the candidate SNPs. SNPs with the highest loadings on the first 10 Eigenvectors were selected for further development. Forward reads containing candidate markers were concatenated with their reverse-complemented paired reads to increase target length, and primers were developed using Primer3. Primers were tested individually using standard PCR conditions, and in combination using standard GT-seq multiplex conditions (Supplemental File 1). Pooling thresholds, the number of individuals that can be simultaneously genotyped in a single run, were tailored to produce >90% genotyping success across individuals, as discussed later.

Marker Evaluation

To examine the functional ploidy ('somy', or meiotic segregation pattern) of the identified SNPs in white sturgeon, we initially evaluated read ratios in the ascertainment panel individuals, which exhibited patterns reflective of 5 genotype categories (AAAA, AAAB, AABB, ABBB, BBBB), i.e. tetrasomy. To confirm this, we estimated ploidy from a larger group (N=3,514; Table 1) of white sturgeon from the Columbia, Fraser, and Sacramento River Basins, using read counts for the 325 SNPs and the R function *funkyPloid* (Delomas et al. submitted), implementing the beta-binomial model with uniform noise for candidate ploidies (somies) of 4N, 5N, and 6N. From this group, putative tetrasomic (4N) individuals were retained that had a minimum of 50k reads and a minimum log likelihood ratio to the next most likely ploidy (hereafter, minimum alternate LLR) of 25, resulting in 2,378 individuals retained for further analyses. We then used the read counts for each locus across these individuals to assess the ploidy of the locus with*funkyPloid* comparing 4N and 8N models. This was achieved by transposing the read count matrices input to *funkyPloid* . The *funkyPloid* function assesses the number of amplified copies of SNPs in the genome, and does not directly assess the pattern of segregation. Thus, this test cannot discriminate between a true tetrasomic locus and two co-amplified, disomic loci. However, given the chromosome complement and previous observations (e.g. Drauch Schreier et al., 2011), disomic loci are likely to be rare, and so here we only consider tetrasomic and octosomic segregation for these loci. It should also be noted that because this LLR metric includes no penalty for overfitting, 'noisy' 4N loci may exhibit higher likelihood for 8N despite only being present in four copies in the genome (Delomas et al. submitted). Thus, rather than evaluate the raw LLR results, we ranked each locus based on fit to 4N or 8N models.

We compared this to two measures of congruence with tetrasomic inheritance. First, we evaluated the percent of comparisons for each locus reflecting Mendelian incompatibilities in several parent-offspring genotypes of known crosses (4 dams, 2 sires, and 128 offspring genotyped at 90% completeness) of white sturgeon spawned by the Yakama Nation using a custom R script (Supplemental File 2). Mendelian incompatibilities were identified as any offspring genotype that was absent from the set of possible offspring genotypes from a pair of adults formed from all possible combinations of all potential diploid gametes assuming the absence of double reduction, which has not been identified in white sturgeon (Drauch Schreier et al., 2011; A. L. Van Eenennaam, Murray, & Medrano, 1998). Genotypes of each individual were obtained by modification of the GT-seq pipeline that accommodates different ploidies by integrating *funkyPloid* and assuming a normal distribution of read ratios for each allele ratio-genotype category with standard deviation starting at 0.05 for 4N and progressively reduced for each higher ploidy following s.d.=0.05*(4/ploidy) to avoid overlap of allele ratio categories (https://github.com/stuartwillis/gt-seq-ploidy). Allele ratios that fall outside the 95% confidence bounds for each genotype category are scored as missing. As such, the genotype thresholds are more stringent for increasing ploidies, and higher sequencing effort may be required to precisely estimate read ratios and genotype higher ploidy samples at all loci. Second, we visually inspected the allele ratio plots of the 2,378 4N individuals and rated each locus on a scale of 1-4 for conformance to the expected allele ratios for tetrasomy (95% confidence interval of normal distribution). The ratings were 1: almost all (~<5%) ratios fall within expected 95% bounds of 5 genotype classes; 2: ratio medians fall within expected 95% bounds of 5 genotype classes but with minor shifts towards the reference or alternate allele (allele bias), and ~<25% of ratios out of bounds; 3: ~5 genotype classes, but ratio medians often fall outside bounds (medians strongly skewed), and/or many (˜>25%) allele ratios fall outside confidence intervals; 4: 6+ genotype classes, homozygote allele ratio medians fall off x,y axes, and/or no distinct genotype classes. Example plots are provided with bounds reflecting 8N genotype categories (Supplemental Figure 2).

Marker Performance

We evaluated the accuracy of these markers for predicting parentage using a dataset of 326 offspring from a partial cross of 5 dams and 6 sires from the Yakama hatchery genotyped at a minimum threshold of 80% completeness (29 full sibling families, ranging from 3 to 23 offspring). Ploidy-accurate genotypes were generated by the updated GT-seq pipeline for polyploids. From all potential sire-dam-offspring trios, we estimated the percent of Mendelian incompatibilities between involving both vs. one or neither true parent using a custom R script (Supplemental File 2). For comparisons in which sex is unknown or both parents may not be included, we used the “Paternity” estimation routine of Polygene (Kang Huang et al., 2020), which includes several population genetic routines adapted for polyploids, though some of these are not applicable across samples of different ploidy. To evaluate performance of single-parent assignment, we included only 4 dams and 2 sires for 326 of the offspring, and examined the LOD score when both, one, or neither parent was present in the candidate set. For circumstances where candidate parents cannot be identified *a priori* , or where sibship relationships are of greater interest, we evaluated the performance of the Huang et al. (2015) maximum likelihood estimator of relationship against known relationships among the 326 offspring in this Yakama set. Although the presence or degree of meiotic double-reduction, resulting in gametes that carry both of a pair of sister chromatids, is not well known in white sturgeon, we applied all Polygene analyses under the “pure random chromatid segregation” (PRCS) model, which provides for some amount of double-reduction. Similarly, we also estimated sibship and full sibling families using Colony2 (Jones & Wang, 2010). Colony2 is designed for diploids, but accepts dominant data, so each SNP locus was recoded as two pseudo-dominant loci (Rodzen et al., 2004; Wang & Scribner, 2014). We ran analyses (full likelihood, high precision, 3 medium length runs) using different estimates of genotyping error (0.001 to 0.05), with parents absent or with all 11 parents present in different arrangements. Arrangement of parents included: separated by sex, together in a single set but ordered, and together but unordered, in each case with probability of inclusion of 0.9, and with all or score 1+2 loci only. Other parameters were left as default.

We also evaluated the utility of these SNPs in estimation of population genetic parameters useful in understanding the differences among and relationships between white sturgeon in different population segments. Using the 3,514 sequenced sturgeon, we evaluated the relationship between sequencing depth, genotype completeness, heterozygosity, and confidence in ploidy estimate represented as minimum alternate LLR. We also compared the genotypes of 142 of these fish which had been sequenced more than once to estimate genotyping error. We then filtered to retain individuals with 4N ploidy from minimum alternate LLR >10 and genotyped at a minimum of 80% completeness. We utilized only 4N individuals because PolyGene is currently limited to a single ploidy per population. We removed known stocked and hatchery individuals from the filtered set, and, to filter unknown stocked individuals, we excluded all but one individual from any set of individuals within a reach with relatedness estimates greater than 0.2 (K. Huang et al., 2015), resulting in 1,203 individuals. We reasoned that unknown stocked fish would exhibit high levels of relatedness due to the generally limited number of breeders available to hatchery operations. Based on estimates from known relationships (the Yakama fish), the applied value should eliminate all full-siblings and most half-siblings while minimizing the unintentional removal of unrelated individuals, although the actual results will be population-specific (Table 1). There has been considerable discussion of whether siblings should be filtered from population genetic datasets (Wang, 2018; Waples & Anderson, 2017), and we acknowledge that statistics calculated from this dataset will have been affected by this filtering but correct for potentially large bias that can be introduced from non-random sampling (Wang, 2018).

Using a custom R script (Supplemental File X), we calculated the minor allele frequency (MAF) of each locus for each reach with N>7 individuals to examine the information content of these loci for identifying distinct population segments. For each population with N>50 individuals, we calculated linkage disequilibrium among loci using Fisher’s G test and conformance to expectations of Hardy-Weinberg (HW) equilibrium using

Raymond & Rousset's (1995) estimator from the Markov chain (5k burn-in, 100 batches of 5k iterations), both applied in PolyGene and with correction for multiple tests (false discovery rate, FDR; Benjamini & Hochberg, 1995). Similarly, we examined differences across sub-populations in estimated inbreeding values using Weir's (1997) estimator in PolyGene. Finally, we estimated genetic divergence, as Nei's (1973) $\mathrm{F_{ST}}$analog, among a selection of sub-populations for which sample size was sufficient, and estimated the precision of this statistic by 100 50% jackknife replicates across loci and calculation of the 95% confidence interval from these assuming a normal distribution.

Results and Discussion

SNP Ascertainment

Read clustering and marker filtering produced 8,781 candidate SNPs. Primer design and testing from SNPs with strong loadings on PCA axes ultimately produced 325 primer combinations with consistent amplification, probe-matching patterns, and low cross-locus interference. These were retained for multiplex in the "Atr325" GT-seq panel (primers and probes for genotyping in Supplemental Table 1).

Marker Evaluation

Across all sequenced samples, individuals had on-target reads, those reads corresponding to amplified target fragments, ranging from 1 to 1.71 million, with a mean of 221k reads, corresponding to OT read percentages (OT reads/total reads) of mean 55% (~0 to 86%). The allele ratio plots revealed that most loci conformed to expectation of tetrasomic segregation (Number of loci with scores of 1, 2, 3, and 4: 97, 133, 56, and 39, respectively). Moreover, score 4 loci, which often did not exhibit five clear genotype categories, did not clearly conform to expectations of octosomy either (Supplemental Figure 2). Rather, there was a moderate and significant correlation ($p<0.001$) between these allele ratio scores and both percent Mendelian incompatibilities in parent-offspring comparisons ($R^2 = 0.59$) and ranked likelihood of octosomic inheritance ($R^2 = 0.36$)(Figure 1). This suggests that most or all of these SNP loci exhibit tetrasomic inheritance, though some may be affected by allelic bias or co-amplification of non-target loci, the latter of which may reflect homeologs in some cases. In addition, even the score 4 loci had a mean percent Mendelian incompatibility of 0.08, suggesting that despite some noise or skew, most of these loci still exhibit relatively reliable genetic signal.

While white sturgeon are ancestral octoploids (Drauch Schreier et al., 2011), it remains unclear what the contemporary meiotic segregation patterns are for this and other sturgeon. Indeed, even studies using similar data sources, such as microsatellite genotype and inheritance patterns, have come to conflicting conclusions (Drauch Schreier et al., 2011; Ludwig, Belfiore, Pitra, Svirsky, & Jenneckens, 2001). Perhaps not surprisingly, Du et al. (2020) found the sterlet genome exhibited both diploid and tetraploid characteristics, with entire segments of some chromosomes diploidized (homeologs lost or diverged), and others still syntenous across large stretches. Unfortunately, microsatellite markers may be equally subject to prolonged homeology, and are thus only helpful to resolve the question of segregation in so far as primers are exclusive to a single segregating locus rather than amplifying multiple separately segregating homeologs (Clark & Schreier, 2017). Moreover, this ambiguity as to how many homeologs are being amplified by any particular marker has complicated efforts to utilize genetic data from white sturgeon to aid conservation. Results from the present study suggest that a significant part of the white sturgeon genome is functionally tetraploid, although our bioinformatic process, which excluded variants with higher divergence and/or read depth, could have biased our survey to parts of the genome further along the process of tetraploidization (but see A. L. Van Eenennaam et al., 1998). Nonetheless, the observation that almost all of the molecular markers presented herein are clearly tetrasomic in 4N/8N individuals makes their application to conservation genetic analyses more straightforward, as demonstrated below.

Marker Performance

Counts of Mendelian incompatibilities (MI) between true parent-offspring trios, while generally not zero as a result of genotyping error, were distinctly less than those in comparisons including one or two non-parents.

Similarly, the distributions of LOD for correct and incorrect single parent assignments were distinct regardless of whether both, one, or neither of the parents were included in the candidate set (Figure 2). Estimates of relatedness among offspring from these crosses were very near theoretical expectations, although slightly downwardly biased for full- and half-sibling relationships (Figure 3). However, the ranges of both sibling types and unrelated individuals overlapped, making relationship estimation from this relatedness measure informative but imprecise.

Relationships estimated by Colony2 were accurate to a degree, but not always comprehensive. When predicted genotyping error was relatively high (0.05), the number of full sibling families and dyad full sibships was estimated accurately; dyad half sibships, albeit incomplete (99.9%), was also at its highest, and the number of contributing parents (N_s) was estimated correctly. However, the probability of sibship was undesirably low for both full (mean 0.51; range 0.224-0.511) and half (0.31; 0.001-0.489) sibships, which may nonetheless result from utilizing a pseudo-dominant data format. As allowed genotyping error decreased (0.01 to 0.001), the number of full sibling families increased (29 to 36), with commensurate decreases in exclusion probability, and some full siblings were assigned as half siblings (1-3%), though their dyad probability of sibship was still more similar to full than half siblings (0.43; 0.214-0.489). With increased stringency in genotyping error, completeness of dyad half sibships also declined from 99% to 89%. Not surprisingly, the offspring segregated into new families with parents inferred to be absent from the input set tended to have moderately higher rates of MI with true parents. For example, the mean percent MI of accurately assigned offspring to an included male was 1.4% MI (range 0 to 3.3%), while the mean of his offspring assigned to a male inferred by the program was 2.4% (range 0.04 to 5.5%) (Supplemental Figure 3). Reducing the loci utilized to only those 230 loci with read ratio 1 or 2 scores did not improve results; in fact, the completeness of half-sibships allowing 5% genotyping error declined slightly (from 99.9% to 99.3%), and one sample pair, inadvertent replicates (clones), was identified as a separate full-sibling family even after being identified as clones by the program, suggesting that despite some noise, these loci provide important discriminatory power for identifying relationships.

Providing parents as separate sexes or as an ordered list (sires then dams) did not affect outcome from Colony2. Interestingly, however, when a single, unordered set of parents was provided as both potential sires and dams, results became erratic between single runs, with additional inferred (hypothetical) parents, reversed genders, and extra full sibling families, but no inaccurate assignments. Across 3+ combined runs, though, parents in an unordered list were assigned correctly with the exception that inferred gender of dams and sires as a group was occasionally reversed, and notwithstanding the aforementioned effects of genotyping error rates on full sibling families and sibship. We thus recommend multiple combined runs be made to ensure accurate results. Moreover, providing parents in any form appeared to result in improved identification, with increased sibship completeness for concomitant rates of genotyping error (e.g. 0.01 error: 99.3% with parents mixed; 92.8% with no parents). Importantly, however, in none of these analyses were unrelated individuals ever identified as siblings, and the estimated number of contributing parents (N_s) was, excluding the clone family, always correct (11). In addition, it is worth noting that these data consist of many related individuals with only a minimum of 80% genotype completeness, making the estimation of global allele frequencies, and thus relationship probability, more challenging than datasets with fewer offspring per family (Colony2 manual). Although we did not explore it, increasing the prior for sibship size (default of 1), specifying allele frequencies estimated from a less related set, and achieving greater genotype completeness, may improve the precision of relationships estimated for polyploid organisms with pseudodominant data in this program.

The inference of parentage, sibship, and relatedness are active areas in sturgeon conservation because many of the conservation management plans of the most recruitment limited populations call for supplementation through hatchery spawning and/or rearing (Hildebrand et al., 2016). While these plans exhibit great potential, they must be done with care because of the potential for genetic swamping of the wild population by alleles from just a few breeding individuals (Thorstensen, Bates, Lepla, & Schreier, 2019). This has been recognized for some time, however, and most *ex situ* spawning programs address this where possible by making factorial crosses of wild parents that are only spawned in a single brood year (Jager, 2005). Variance in survival across families can undermine these factorial and normalized supplementation designs, decreasing the genetic diversity reintroduced by hatchery offspring. For example, using parentage alongside PIT tag recordings, Schreier et al. (A. Schreier, Stephenson, Rust, & Young, 2015) found that several year classes of offspring that were surviving after 3 years did not reconstitute the genetic diversity of the brood stock, not to mention the adult population at large.

These observations have reinforced the push to monitor both the variability in recruitment success and long-term genetic effects of hatchery supplementation, objectives that depend on determining the relationship or number of contributing spawners in supplemented and/or wild fish. Because the number of broodstock in most locations will not themselves be an adequate representation of overall population genetic variation, programs that collect naturally produced eggs and larvae for hatchery rearing followed by repatriation as juveniles may capture the offspring of more spawning adults and therefore better represent standing genetic diversity (Thorstensen et al., 2019). While promising, repatriation techniques are only effective in situations where recruitment limitation results from survivorship in life stages promoted by tenure in the hatchery, spawning sites and times can be identified effectively and the number of adults spawning there exceeds broodstock constraints of nearby hatcheries, and survivorship variation among families is stochastic or reflects natural patterns (e.g. maternal health). For example, using sibship to estimate the number of spawners, Jay et al. (2014) identified strong variation in the number of spawners among spawning locations and dates, meaning repatriation programs would be well served to collect in multiple sites and times. Nonetheless, these authors estimated numbers of contributing spawners that would be difficult to reproduce with practical limitations on hatchery broodstock numbers (see also Blankenship, Schumer, Van Eenennaam, & Jackson, 2017). In any event, supplementation and repatriation programs operate on the assumption that relatedness in stocked offspring does not diminish genetic diversity or promote inbreeding depression in small populations, a conjecture that is more easily tested using the markers and techniques we have demonstrated here.

Supplementation and repatriation programs also presume that fitness of stocked offspring (i.e. fecundity and survival of their progeny) is similar to *in situ* individuals, although it is as yet unclear how variation in ploidy in white sturgeon, which may be exacerbated by human intervention, affects this parameter (J. P. Van Eenennaam et al., 2019). Our pipeline, thanks to integration of *funkyPloid* , allows simultaneous ploidy estimation and ploidy-aware genotyping. However, a minimum coverage of at least 100k reads is recommended to accurately score heterozygous genotypes and inform ploidy estimation. Confidence in ploidy estimates (minimum alternate LLR) is correlated with sequencing depth ($R^2 = 0.64$, p<0.001, Figure 4a). Minimum alternate LLR appears to be more closely tied to genotype completeness ($R^2 = 0.73$, p<0.001) than to heterozygosity ($R^2 = 0.56$, p<0.001), although both factors are influential (Figure 4a, Supplemental Figure 4). Similarly, genotype completeness appears to be more closely correlated to sequence coverage ($R^2 = 0.51$, p<0.001) than is heterozygosity ($R^2 = 0.39$, p<0.001), and together these observations indicate that multiplexing protocols should be optimized for genotyping completeness, which indirectly provides more accurate estimates of heterozygosity, in order to bolster confidence in ploidy estimates. For our samples, genotyping at 90% completeness generally required a minimum of ~100k on-target reads per sample (Figure 4b), or on average ~300 reads per marker for each individual (5 and 10 percentiles of samples >90% complete were 96.4k and 119.2k reads, respectively). In addition to more accurate estimates of ploidy, by comparing 142 samples genotyped twice or more we observed that after achieving [?]90% completeness, mean genotyping error (incorrect number of genotypes/number of typed loci, excluding missing data) was no more than 1.1%.

For several years, it has been known that individuals with nuclear DNA content indicative of dodecaploidy (12N), or sometimes 16N, were present in white sturgeon hatchery populations (Drauch Schreier et al., 2011; A. D. Schreier et al., 2013). Similar variants were also reported in Siberian sturgeon culture (*Acipsenser baerii;* Havelka, Bytyutskyy, Symonová, Ráb, & Flajšhans, 2016). It appears that this process likely results from retention of the second polar body after fertilization (Gille, Famula, May, & Schreier, 2015), and may be promoted by handling for hatchery spawning or rearing (J. P. Van Eenennaam et al., 2019). Autopolyploid white sturgeon have not been observed to show diminished survivorship or fertility, and their backcross offspring, which most often show the expected ploidy (10N), are often viable (Drauch Schreier et al., 2011;

Gille et al., 2015; Leal, Clark, Van Eenennaam, Schreier, & Todgham, 2018; Leal, Van Eenennaam, Schreier, & Todgham, 2020). This has raised several additional important questions regarding the fertility of these backcross fish and the ploidy of gametes and offspring produced. Although 12N fish appear to suffer no immediate fitness loss, and indeed in some autopolyploid sturgeon exhibit increased vigor (Beyea, Benfey, & Kieffer, 2005), it seems likely that these 10N (and likely pentasomic) suffer reduced fertility, and their aneuploid offspring, if viable, reduced vigor and fertility (J. P. Van Eenennaam et al., 2019). If so, this presents a problem for conservation aquaculture programs, if hatchery spawning techniques increase the incidence of autopolyploids, and there is no indication that repatriation programs are immune to this phenomenon either. The rate of natural ploidy variation in white sturgeon, which would be the standard to which to compare, is unclear and an active area of research. One constraint to address this, however, is that current methods for ploidy estimation rely on fresh tissue samples (Fiske et al., 2019), generally precluding the use of archived tissues. The pipeline presented here, however, provides for simultaneous genotyping and ploidy estimation using any form of DNA-bearing tissue. Although beyond the scope of this study, we did find and exclude several putative autopolyploid and backcross individuals among the Yakama offspring and *in situ* samples using this technique (results not shown).

In addition to their utility in determining relationships and estimating ploidy, we expect these SNP markers to be useful for identifying population structure and dispersal between population segments (Ogden et al., 2013; Roques, Chancerel, Boury, Pierre, & Acolas, 2019). Although these SNPs were initially identified as those with a minor allele frequency (MAF) >0.05 in the ascertainment panel, several of them exhibited a mean MAF in our filtered *in situ* samples below this value (Figure 5). However, most of these low-mean-MAF SNPs exhibited variation in MAF among localities that should make them useful for discriminating different populations. Notably, many loci exhibited as much variation in MAF among reaches within the Columbia basin as between sites in the Columbia and those in the Fraser and Sacramento River basins.

Of all the comparisons of linkage disequilibrium between loci in populations with sufficient sample size, only 0.12% and 0.25% of comparisons were significant at FDR of 0.05 and 0.1, respectively. Among those significant comparisons, there did not appear to be a relationship between frequency of linkage association and allele ratio score (Figure 6a). The top five loci most involved in significant associations were Atr_72251-33 (1.19%), Atr_14917-56 (1.09%), Atr_36485-28 (0.99%), Atr_40343-66 (0.99%), and Atr_65359-46 (0.99%). In contrast, 34% of loci significantly deviated from HWE at an FDR of 0.05. However, there was a moderate and significant correlation ($R^2 = 0.47$; $p<0.001$) between the number of populations in which a locus was out of HW equilibrium and the allele ratio score, indicating that deviation from HWE was likely exacerbated by genotyping inaccuracy at a locus (Figure 6b). In addition, it is possible that filtering for related individuals in this dataset, either too strongly or too weakly, could also affect the incidence of significant HW tests.

All of the reaches with sufficient sample size exhibited median individual estimates of inbreeding that were above zero, although the median value and ranges of the estimates varied by reach, suggesting that these SNP loci will be useful for investigating trends of recruitment, population viability, and potential for inbreeding depression (Supplemental Figure 5). As observed previously (A. Drauch Schreier, Mahardja, & May, 2013), sub-populations in the Columbia basin appear to exhibit an isolation by distance pattern in which the fishes in the uppermost reaches (upper Snake, upper Columbia) exhibit the strongest divergence (Table 2). Notably, the precision on these estimates of F_{ST} was more strongly affected by sample size of individuals than by variation introduced by sampling different subsets of loci.

One of the challenges for supplementation programs in the most severely diminished population segments is obtaining enough unrelated brood stock from *in situ* populations so as to not further reduce standing genetic variation by swamping. In these cases, the use of translocated adults or hatchery reared young from other breeding stocks has been suggested but remains controversial because of the uncertainty of population structure among population segments (Hildebrand et al., 2016). While adult white sturgeon, before the creation of the hydropower barriers, likely would have been able to migrate throughout much of the Columbia Basin, fish in less impounded river systems (e.g. Fraser River) appear to show moderate site fidelity, an observation reinforced by population genetic structure (Andrea Drauch Schreier, Mahardja, &

May, 2012). Thus, even where movement patterns were not historically restricted for feeding or overwintering in Columbia River sturgeon, there may have been cryptic barriers or spawning site fidelity that reduced gene flow over longer distances (A. Drauch Schreier et al., 2013). It remains to be seen whether there is detectable population structure over shorter distances, or whether the detected entrainment of young fish has been sufficient to reduce population divergence following dam construction. Moreover, where translocation of adults or young is implemented, it will need to be closely monitored in studies for outbreeding depression, standing genetic variation, and potential local adaptation, which will be aided by the efficient genotyping of genetic markers described herein.

Improved Genetic Resources for Polyploids

Our technique of developing SNP markers using stringently filtered reduced representation genomic libraries for pairing with the ploidy-aware high throughput genotyping pipeline, as demonstrated here with white sturgeon, should make genetic data more accessible for a range of polyploid organisms. This process addresses several of the challenges inherent to polyploid organisms: discrimination of homeolog amplicons, dosage of polysomic alleles, inference of ploidy in each individual. Pairing these data with recent ploidy-flexible software will make investigations of polysomic organisms more efficient.

There are some caveats to this procedure, however. Perhaps the most idiosyncratic part of this combined technique is the SNP-discovery process. Our matching and filtering process was designed to retain only variants which segregated in otherwise highly similar sequence loci, which helps to discriminate sites, fixed or variants, in homeologs. This requires that a large number of candidate loci be surveyed initially so that a sufficient number of candidates are available for PCR testing after stringent filtering, thus requiring that the number of samples in the ascertainment panel be balanced with the sequencing effort to produce sufficient coverage across all the initial candidates and provide informative read-depth distributions. These parameters will have to be tailored to individual polyploid species, as will the exact filtering thresholds utilized to identify loci with the base segregation pattern (e.g. tetrasomy, in the case of white sturgeon).

One of the great advantages of our method is the ability to simultaneously infer ploidy and genotype individuals in ploidy variable species. While we provide a rough guideline of sequence depth for accurate ploidy inference and genotyping in white sturgeon, this value will need to be tailored for the number of markers surveyed and their on-target efficiency in individual species. It is worth re-iterating that the current genotyping function uses allele ratios predicted from a normal distribution with standard deviation that is inversely related to the number of genotype categories, i.e. ploidy. As ploidy increases, the width of the allowed distributions for each genotype category is reduced, and greater precision in allele ratios is required to genotype each marker. Thus, the required sequence coverage, which provides the sample size for each marker, will be higher with increasing ploidy. Similarly, the confidence level in ploidy inferences, or minimum alternate LLR, will also need to be tailored to individual species, their levels of heterozygosity, and ploidy range. While Delomas et a. (submitted) make some recommendations (e.g., a minimum LLR of 10 for the panel described here), individual researchers may find it useful to employ a higher or lower stringency threshold for genotyping, as the updated GT-seq genotyping pipeline currently only provides genotypes for individuals passing the user-specified minimum alternate LLR.

Two additional limitations to this ploidy estimation function worth noting. First, the function assumes that all loci within a single individual have the same ploidy. Accommodation of loci with multiple ploidies within an individual, e.g. tetraploid and octoploid loci within an ancestral octoploid, can be achieved by fitting models to each group of loci separately by ploidy. The likelihood across all loci could then be calculated as the product of the likelihoods for each group. Second, discrimination of ploidies that are exact multiples of one another may yield results that are less straightforward to interpret because of the current lack of a penalty function for overfitting (fitting noise with the higher ploidy model). For example, in the current dataset, for the same individuals from which the 4N allele plots were generated, a comparison of 4N and 8N models demonstrated that the 8N model had higher likelihood for 54% of samples, although only $<1\%$ of incorrect ploidy estimates had minimum alternate LLR higher than 10. As pointed out by Delomas et al. (submitted), the most likely 8N model will always have likelihood higher than or equal to the most likely 4N model, apart from deviations due to the threshold at which convergence of the EM algorithm is assumed, because 4N models are a subset of the space of all possible 8N models. However, ploidy can still be inferred in these situations: individuals of the lower ploidy will have LLR distributed close to zero and individuals of the higher ploidy will have LLR distributed further away from zero. Critical values for assigning ploidy can be chosen using LLRs from a set of known ploidy individuals. Alternatively, if individual ploidies are not known, but a group is presumed to have variable ploidy, the LLRs from individuals in this group are expected to have a bimodal distribution (one mode for each ploidy). If these modes are sufficiently distinct, critical values can be chosen to separate the two clusters. Nevertheless, we look forward to continued development of these functions to facilitate an even greater variety of tests.

Recent development of population genetic software that accommodates polysomic data has advanced the evolutionary analysis of contemporary polyploids, as demonstrated here for white sturgeon with Polygene. Polygene includes a number of functions that were previously inaccessible, including single-parent assignment and maximum likelihood estimates of relatedness. Sibship estimation remains unavailable, however, and for situations where candidate parents are not available, we demonstrated that SNP data could be coded as pseudo-dominant diploid data for use in Colony2 with reasonable, though time-consuming, efficacy. We note, however, that one limitation of Polygene is its treatment of ploidy as invariant within a "population", which prevents the estimation of parentage and other statistics across ploidy states. While other packages such as Genodive and adegenet do not appear to have this limitation, they nonetheless lack some of the functionality of Polygene. Though not demonstrated here, we envision scenarios in which simulations, relationships inferred from same-ploidy individuals, or known families, could be used to identify thresholds in ploidy-agnostic measures of relatedness to estimate relationships between individuals of different ploidy. As demonstrated here, these measures exhibit variance and potentially downward bias that suggest they need to be estimated for each population separately and treated conservatively. In any event, we expect that the greater accessibility of ploidy-correct genotype data, as provided by the techniques demonstrated here, will spur further development of software facilitating the genetic analysis of polyploids.

Conclusions

We described here the combination of techniques which will make genetic data for polyploids more accessible and facilitate population genetic and evolutionary studies of polysomic species. Together, these methods address some of the standing challenges to developing genetic resources for polyploids, including the identification of homeolog-specific genetic variants, inference of ploidy-aware genotypes, and estimation of parentage and other population genetic statistics. We demonstrate the efficacy of these techniques for white sturgeon, a polysomic and ploidy variable species of strong conservation concern in the Pacific Northwest. The molecular markers we present will greatly facilitate the conservation management of imperiled populations. These markers, along with our updated genotyping-by-thousands pipeline, allow the efficient generation of genetic data and estimation of ploidy for large numbers of samples from diverse tissue collections. These data will help researchers address a number of outstanding questions, including the efficiency and effects of alternative conservation schemes including hatchery spawning, repatriation, and translocation of white sturgeon in the Columbia River Basin. Although questions remain about the segregation of chromosomes and individual loci in white sturgeon, our examinations suggest the loci described here are reliably tetrasomic for putatively octoploid individuals.

Acknowledgements

The authors acknowledge the contributions of Nate Campbell and Andrew Matala in initial modification to the GT-seq pipeline to accommodate tetraploid data, analysis of the RAD data to identify variant sites, curation of the white sturgeon ascertainment samples and data, and filtering of candidate markers. Andrea Schreier provided insightful discussion on ploidy patterns in white sturgeon. Lori Maxwell and Stephanie Harmon assisted in the lab with preparation of samples and sequencing. The white sturgeon project was supported by Bonneville Power Administration (Grant no. 2008-907-00).

Author contributions

Narum organized the development of the sturgeon SNP loci. Willis, Delomas, Parker, Miller, and Narum developed the idea for the study, and Willis and Delomas developed the methods employed. Willis, Anders, and Narum organized and chose the final samples utilized. Willis drafted the manuscript, and all authors reviewed and approved the drafts, as appropriate.

Conflicts of Interest Statement

The authors assert that they have no conflicts of interest regarding the present results or interpretations.

Data Accessibility

The genotyping and analysis scripts utilized in the study have been made available as supplemental data, or at https://github.com/stuartwillis/gt-seq-ploidy and https://github.com/delomast/tripsAndDipR.

Figure Legends

Figure 1. Conformance of white sturgeon SNP loci to expectations of tetrasomy. Top panel) Box plot of ranked conformance to a mixture model (Delomas et al. submitted) of tetraploidy (4N) vs. allele plot scores. Bottom Panel) Box plot of percent Mendelian incompatibilities by locus in known parent-offspring trios vs. allele plot scores.

Figure 2. Signal for parentage detection in white sturgeon SNPs. Top Panel) Percent Mendelian incompatibilities between parent-offspring and non-parent offspring trios from hatchery white sturgeon. Bottom Panel) Log odds scores for single parent assignment in Polygene when both, one, or neither parents are included for 326 offspring from the same hatchery cross.

Figure 3. Box plot of the Huang et al. (2015) maximum likelihood estimate of relatedness for 326 offspring from a partial 5x6 hatchery cross of white sturgeon. Full, full siblings; Half, half siblings.

Figure 4. Relationship between ploidy estimation confidence, heterozygosity, genotype completeness, and sequencing coverage. Top Panel) Ploidy estimation confidence (minimum alternative likelihood ratio) versus reads corresponding to target loci. Color scale represents percent genotype completeness. Inset graph shows lower left quadrant in detail. Bottom Panel) Percent genotype completeness versus reads corresponding to target loci. Color scale represents percent heterozygous genotypes. Dotted line corresponds to sequencing coverage generally achieving 90% genotype completeness (100k reads).

Figure 5. Minor allele frequency (MAF) of white sturgeon SNP loci in sampled populations, ordered by mean MAF (black line). “Minor allele” status was chosen based on mean allele frequency, though in some populations the minor allele is more common. Only populations with greater than seven individuals are shown. Orange, Columbia River Basin localities; green, outside Columbia Basin.

Figure 6. Statistics of genetic association for each white sturgeon SNP locus in populations with >50 individuals versus allele plot ratings. Top Panel) Number of statistically significant linkage associations. Bottom Panel) Number of statistically significant departures from Hardy-Weinberg equilibrium.

Supplemental Figure Legends

S.Figure 1. Frequency distribution plots of number of loci of a given depth for the RAD library before (blue) and after (orange) filtering by read depth.

S.Figure 2. Example allele plots for score 1-4 loci. Boundaries are indicated for each genotype category corresponding to octoploid genotypes (white: allele ratios shared with tetraploid genotypes; grey: allele ratios exclusive to octoploid genotypes; red: regions of indeterminate ratio recorded as missing).

S.Figure 3. Boxplot of percent Mendelian incompatibilities between true sire and offspring for those offspring correctly assigned to that sire (BLF-0021) by Colony2, or assigned to a missing father (*1).

S.Figure 4. Relationship between ploidy estimation confidence, heterozygosity, and sequencing coverage. Ploidy estimation confidence (minimum alternative likelihood ratio) versus reads corresponding to target loci. Color scale represents percent heterozygous genotypes. Inset graph shows lower left quadrant in detail.

S.Figure 5. Boxplot of individual inbreeding (Weir) by locality for select localities with N>7 individuals.

Table Captions

Table 1. Number of individual white sturgeon genotyped per locality, pre-filtering and post-filtering for genotype completeness, ploidy confidence, and relatedness.

Table 2. Mean and 95% confidence interval of genetic divergence (Nei's F_{ST} analog) from 50% jackknife replicates for select localities with N>7 individuals.

Supplemental Table Captions

Supplemental Table 1. GT-seq primers, probes, and variants for the 325 single nucleotide polymorphism loci from white sturgeon.

References Cited

Anders, P. J., Drauch-Schreier, A., Rodzen, J., Powell, M. S., Narum, S., & Crossman, J. A. (2011). A review of genetic evaluation tools for conservation and management of North American sturgeons: Roles, benefits, and limitations. *Journal of Applied Ichthyology* . doi: 10.1111/j.1439-0426.2011.01830.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing.*Journal of the Royal Statistical Society, Series B* , *57* (1), 289–300.

Beyea, M. M., Benfey, T. J., & Kieffer, J. D. (2005). Hematology and stress physiology of juvenile diploid and triploid shortnose sturgeon (Acipenser brevirostrum). *Fish Physiology and Biochemistry* ,*31* , 303–313. doi: 10.1007/s10695-005-1552-y

Blankenship, S. M., Schumer, G., Van Eenennaam, J. P., & Jackson, Z. J. (2017). Estimating number of spawning white sturgeon adults from embryo relatedness. *Fisheries Management and Ecology* . doi: 10.1111/fme.12217

Blischak, P. D., Kubatko, L. S., & Wolfe, A. D. (2018). SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics* . doi: 10.1093/bioinformatics/btx587

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

Catchen, J. M., Amores, A., Hohenlohe, P. A., Cresko, W. A., & Postlethwait, J. H. (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics‘* ,*1* , 3171–3182.

Clark, L. V., & Schreier, A. D. (2017). Resolving microsatellite genotype ambiguity in populations of allopolyploid and diploidized autopolyploid organisms using negative correlations between allelic variables. *Molecular Ecology Resources* . doi: 10.1111/1755-0998.12639

Comai, L. (2005). The advantages and disadvantages of being polyploid.*Nature Reviews Genetics* . doi: 10.1038/nrg1711

Crow, K. D., Stadler, P. F., Lynch, V. J., Amemiya, C., & Wagner, G. P. (2006). The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. *Molecular Biology and Evolution* . doi: 10.1093/molbev/msj020

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology* . doi: 10.1371/journal.pbio.0030314

Drauch Schreier, A., Gille, D., Mahardja, B., & May, B. (2011). Neutral markers confirm the octoploid origin and reveal spontaneous autopolyploidy in white sturgeon, Acipenser transmontanus. *Journal of Applied Ichthyology* . doi: 10.1111/j.1439-0426.2011.01873.x

Du, K., Stöck, M., Kneitz, S., Klopp, C., Woltering, J. M., Adolfi, M. C., . . . Schartl, M. (2020). The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nature Ecology and Evolution* . doi: 10.1038/s41559-020-1166-x

Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools.*Molecular Ecology* . doi: 10.1111/mec.12581

Fiske, J. A., Van Eenennaam, J. P., Todgham, A. E., Young, S. P., Holem-Bell, C. E., Goodbla, A. M., & Schreier, A. D. (2019). A comparison of methods for determining ploidy in white sturgeon (Acipenser transmontanus). *Aquaculture* . doi: 10.1016/j.aquaculture.2019.03.009

Gille, D. A., Famula, T. R., May, B. P., & Schreier, A. D. (2015). Evidence for a maternal origin of spontaneous autopolyploidy in cultured white sturgeon (Acipenser transmontanus). *Aquaculture* . doi: 10.1016/j.aquaculture.2014.10.002

Havelka, M., Bytyutskyy, D., Symonová, R., Ráb, P., & Flajšhans, M. (2016). The second highest chromosome count among vertebrates is observed in cultured sturgeon and is associated with genome plasticity.*Genetics Selection Evolution* . doi: 10.1186/s12711-016-0194-0

Hildebrand, L. R., Drauch Schreier, A., Lepla, K., McAdam, S. O., McLellan, J., Parsley, M. J., . . . Young, S. P. (2016). Status of White Sturgeon (Acipenser transmontanus Richardson, 1863) throughout the species range, threats to survival, and prognosis for the future.*Journal of Applied Ichthyology* . doi: 10.1111/jai.13243

Holland, P. W. H., Garcia-Fernandez, J., Williams, N. A., & Sidow, A. (1994). Gene duplications and the origins of vertebrate development.*Development* .

Huang, K., Guo, S. T., Shattuck, M. R., Chen, S. T., Qi, X. G., Zhang, P., & Li, B. G. (2015). A maximum-likelihood estimation of pairwise relatedness for autopolyploids. *Heredity* . doi: 10.1038/hdy.2014.88

Huang, Kang, Dunn, D. W., Ritland, K., & Li, B. (2020). polygene: Population genetics analyses for autopolyploids based on allelic phenotypes. *Methods in Ecology and Evolution* . doi: 10.1111/2041-210X.13338

Ilut, D. C., Nydam, M. L., & Hare, M. P. (2014). Defining loci in restriction-based reduced representation genomic data from nonmodel species: Sources of bias and diagnostics for optimal clustering.*BioMed Research International* . doi: 10.1155/2014/675158

Jager, H. I. (2005). Genetic and demographic implications of aquaculture in white sturgeon (Acipenser transmontanus) conservation. *Canadian Journal of Fisheries and Aquatic Sciences* . doi: 10.1139/f05-106

Jay, K., Crossman, J. A., & Scribner, K. T. (2014). Estimates of Effective Number of Breeding Adults and Reproductive Success for White Sturgeon. *Transactions of the American Fisheries Society* . doi: 10.1080/00028487.2014.931301

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* . doi: 10.1093/bioinformatics/btn129

Jones, O. R., & Wang, J. (2010). COLONY: A program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources* . doi: 10.1111/j.1755-0998.2009.02787.x

Leal, M. J., Clark, B. E., Van Eenennaam, J. P., Schreier, A. D., & Todgham, A. E. (2018). The effects of warm temperature acclimation on constitutive stress, immunity, and metabolism in white sturgeon (Acipenser transmontanus) of different ploidies. *Comparative Biochemistry and Physiology -Part A : Molecular and Integrative Physiology* . doi: 10.1016/j.cbpa.2018.05.021

Leal, M. J., Van Eenennaam, J. P., Schreier, A. D., & Todgham, A. E. (2020). Diploid and triploid white sturgeon (Acipenser transmontanus) differ in magnitude but not kinetics of physiological responses to exhaustive exercise at ambient and elevated temperatures. *Canadian Journal of Fisheries and Aquatic Sciences* . doi: 10.1139/cjfas-2019-0289

Ludwig, A., Belfiore, N. M., Pitra, C., Svirsky, V., & Jenneckens, I. (2001). Genome duplication events and functional reduction of ploidy levels in sturgeon (Acipenser, Huso and Scaphirhynchus). *Genetics* .

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* . doi: 10.1126/science.290.5494.1151

Meirmans, P. G., Liu, S., & Van Tienderen, P. H. (2018). The Analysis of Polyploid Genetic Data. *Journal of Heredity* . doi: 10.1093/jhered/esy006

Meirmans, P. G., & Van Tienderen, P. H. (2013). The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity* . doi: 10.1038/hdy.2012.80

Meyer, A., & Van De Peer, Y. (2005). From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD). *BioEssays* . doi: 10.1002/bies.20293

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* . doi: 10.1101/gr.5681207

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America* . doi: 10.1073/pnas.70.12.3321

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology* , *0* (ja). doi: 10.1111/mec.14792

Ogden, R., Gharbi, K., Mugue, N., Martinsohn, J., Senn, H., Davey, J. W., . . . Congiu, L. (2013). Sturgeon conservation genomics: SNP discovery and validation using RAD sequencing. *Molecular Ecology* . doi: 10.1111/mec.12234

Ohno, S. (1971). Evolution by Gene Duplication. *Population (French Edition)* . doi: 10.2307/1530208

Puritz, J. B., Matz, M. V, Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology* , *23* (24), 5937–5942.

Rajkov, J., Shao, Z., & Berrebi, P. (2014). Evolution of polyploidy and functional diploidization in sturgeons: Microsatellite analysis in 10 sturgeon species. *Journal of Heredity* . doi: 10.1093/jhered/esu027

Raymond, M., & Rousset, F. (1995). An exact test for population differentiation. *Evolution* , *49* , 1280–1283. doi: 10.1111/j.1558-5646.1995.tb04456.x

Rodzen, J. A., Famula, T. R., & May, B. (2004). Estimation of parentage and relatedness in the polyploid white sturgeon (Acipenser transmontanus) using a dominant marker approach for duplicated microsatellite loci. *Aquaculture* . doi: 10.1016/S0044-8486(03)00450-2

Ronfort, J., Jenczewski, E., Bataillon, T., & Rousset, F. (1998). Analysis of population structure in autotetraploid species. *Genetics* .

Roques, S., Chancerel, E., Boury, C., Pierre, M., & Acolas, M. L. (2019). From microsatellites to single nucleotide polymorphisms for the genetic monitoring of a critically endangered sturgeon. *Ecology and Evolution* . doi: 10.1002/ece3.5268

Schreier, A. D., May, B., & Gille, D. A. (2013). Incidence of spontaneous autopolyploidy in cultured populations of white sturgeon, Acipenser transmontanus. *Aquaculture* . doi: 10.1016/j.aquaculture.2013.09.012

Schreier, A. Drauch, Mahardja, B., & May, B. (2013). Patterns of population structure vary across the range of the white sturgeon. *Transactions of the American Fisheries Society* . doi: 10.1080/00028487.2013.788554

Schreier, A. Drauch, Rodzen, J., Ireland, S., & May, B. (2012). Genetic techniques inform conservation aquaculture of the endangered Kootenai river white sturgeon Acipenser transmontanus. *Endangered Species*

Research . doi: 10.3354/esr00387

Schreier, A., Stephenson, S., Rust, P., & Young, S. (2015). The case of the endangered Kootenai River white sturgeon (Acipenser transmontanus) highlights the importance of post-release genetic monitoring in captive and supportive breeding programs. *Biological Conservation* . doi: 10.1016/j.biocon.2015.09.011

Schreier, Andrea Drauch, Mahardja, B., & May, B. (2012). Hierarchical patterns of population structure in the endangered fraser river white sturgeon (acipenser transmontanus) and implications for conservation.*Canadian Journal of Fisheries and Aquatic Sciences* , *69* , 1968–1980. doi: 10.1139/f2012-120

Scott, W., & Crossman, E. (1973). *Freshwater fishes of Canada, Bulletin 184* . Fisheries Research Board of Canada, Ottawa.

Soltis, D. E., Visger, C. J., Blaine Marchant, D., & Soltis, P. S. (2016). Polyploidy: Pitfalls and paths to a paradigm. *American Journal of Botany* . doi: 10.3732/ajb.1500501

Spoelhof, J. P., Soltis, P. S., & Soltis, D. E. (2017). Pure polyploidy: Closing the gaps in autopolyploid research. *Journal of Systematics and Evolution* . doi: 10.1111/jse.12253

Thorstensen, M., Bates, P., Lepla, K., & Schreier, A. (2019). To breed or not to breed? Maintaining genetic diversity in white sturgeon supplementation programs. *Conservation Genetics* . doi: 10.1007/s10592-019-01190-4

Van Eenennaam, A. L., Murray, J. D., & Medrano, J. F. (1998). Synaptonemal complex analysis in spermatocytes of white sturgeon, Acipenser transmontanus richardson (pisces, acipenseridae), a fish with a very high chromosome number. *Genome* , *41* , 51–61. doi: 10.1139/g97-101

Van Eenennaam, J. P., Fiske, A. J., Leal, M. J., Cooley-Rieders, C., Todgham, A. E., Conte, F. S., & Schreier, A. D. (2019). Mechanical shock during egg de-adhesion and post-ovulatory ageing contribute to spontaneous autopolyploidy in white sturgeon culture (Acipenser transmontanus). *Aquaculture* . doi: 10.1016/j.aquaculture.2019.734530

Wang, J. (2018). Effects of sampling close relatives on some elementary population genetics analyses. *Molecular Ecology Resources* . doi: 10.1111/1755-0998.12708

Wang, J., & Scribner, K. T. (2014). Parentage and sibship inference from markers in polyploids. *Molecular Ecology Resources* . doi: 10.1111/1755-0998.12210

Waples, R. S., & Anderson, E. C. (2017). Purging putative siblings from population genetic data sets: A cautionary view. *Molecular Ecology* . doi: 10.1111/mec.14022

Weir, B. S. (1997). Genetic Data Analysis II. *Biometrics* . doi: 10.2307/2533134

Wendel, J. F. (2000). Genome evolution in polyploids. *Plant Molecular Biology* . doi: 10.1023/A:1006392424384

Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., & Portnoy, D. S. (2017). Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Molecular Ecology Resources* , *17* (5), 955–965. doi: 10.1111/1755-0998.12647

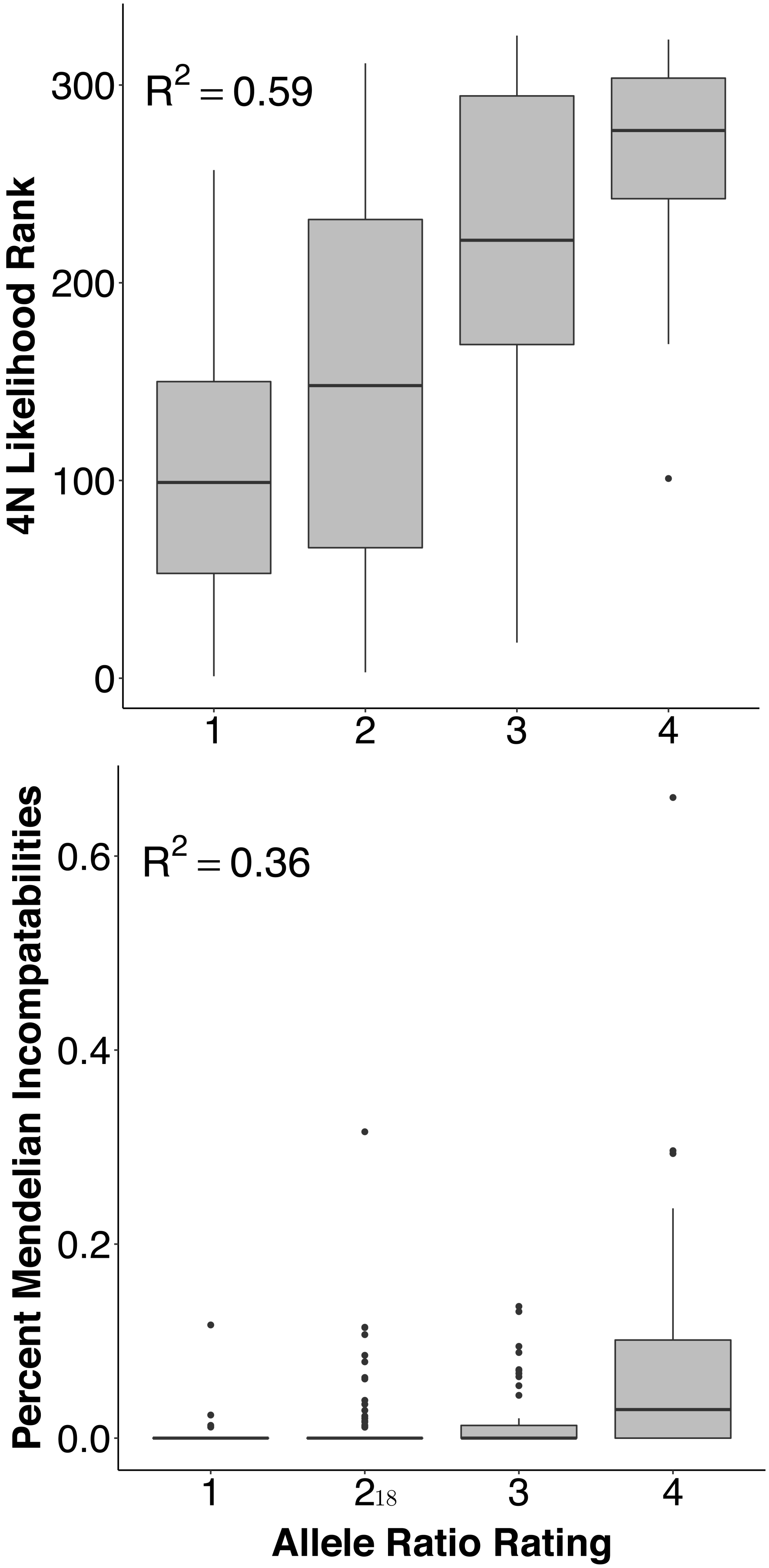
Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nature Reviews Genetics* . doi: 10.1038/35072009

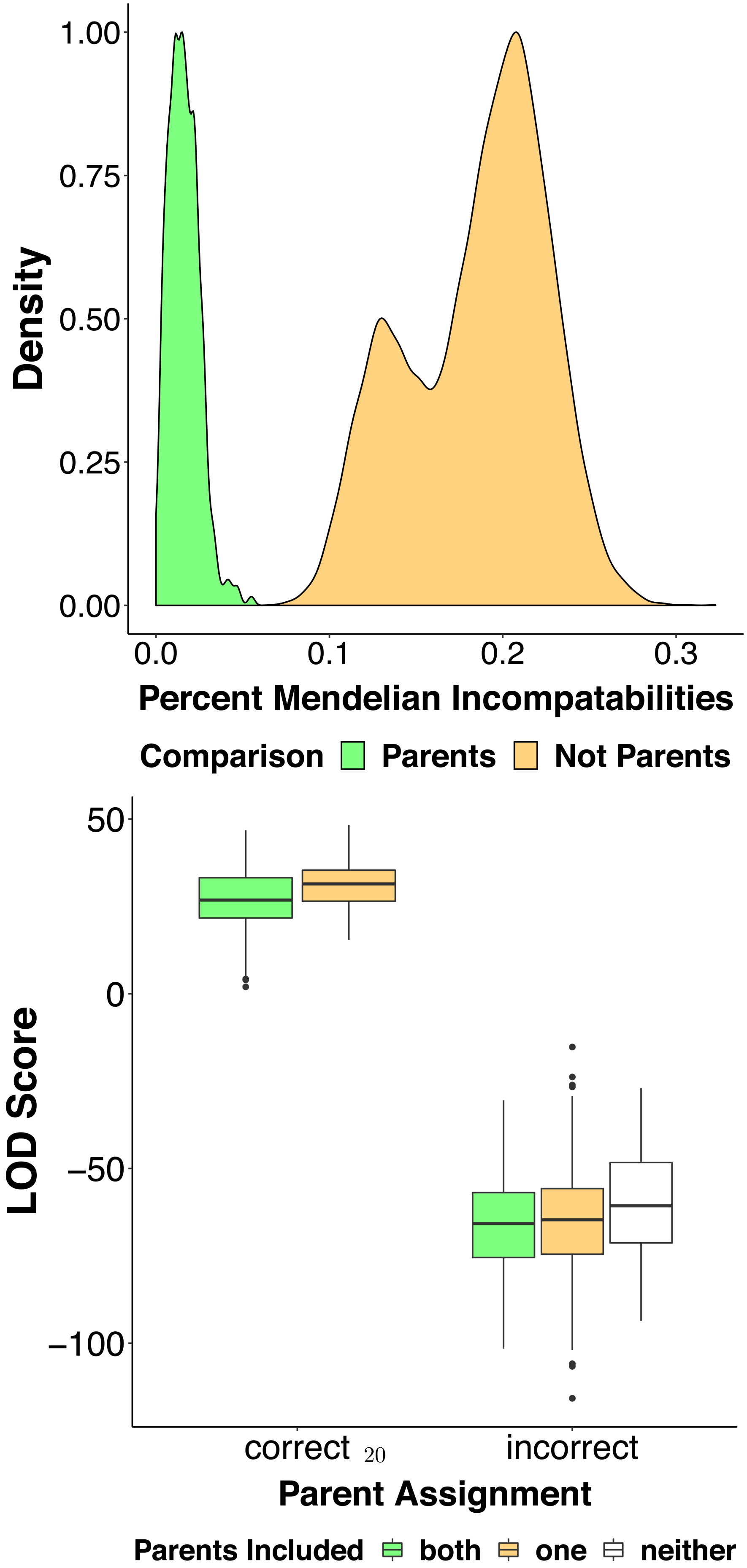
Hosted file

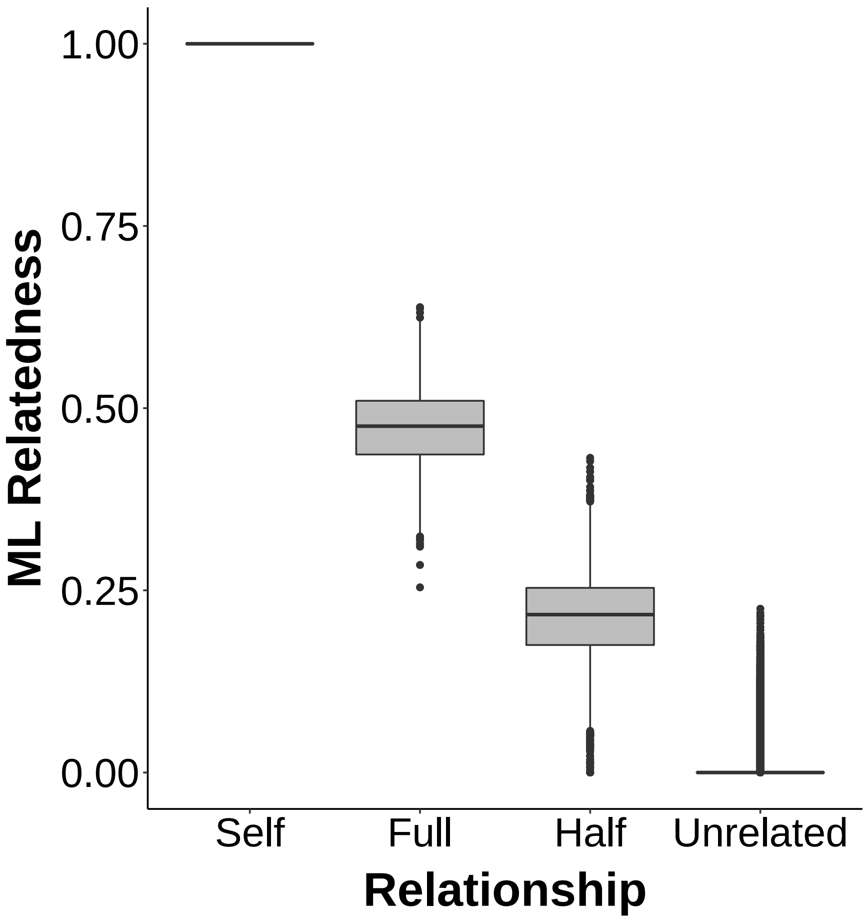
`Table1-sample-sizes.xlsx` available at https://authorea.com/users/353687/articles/477467-single-nucleotide-polymorphism-genotypes-and-ploidy-estimates-for-ploidy-variable-species-generated-with-massively-parallel-amplicon-sequencing

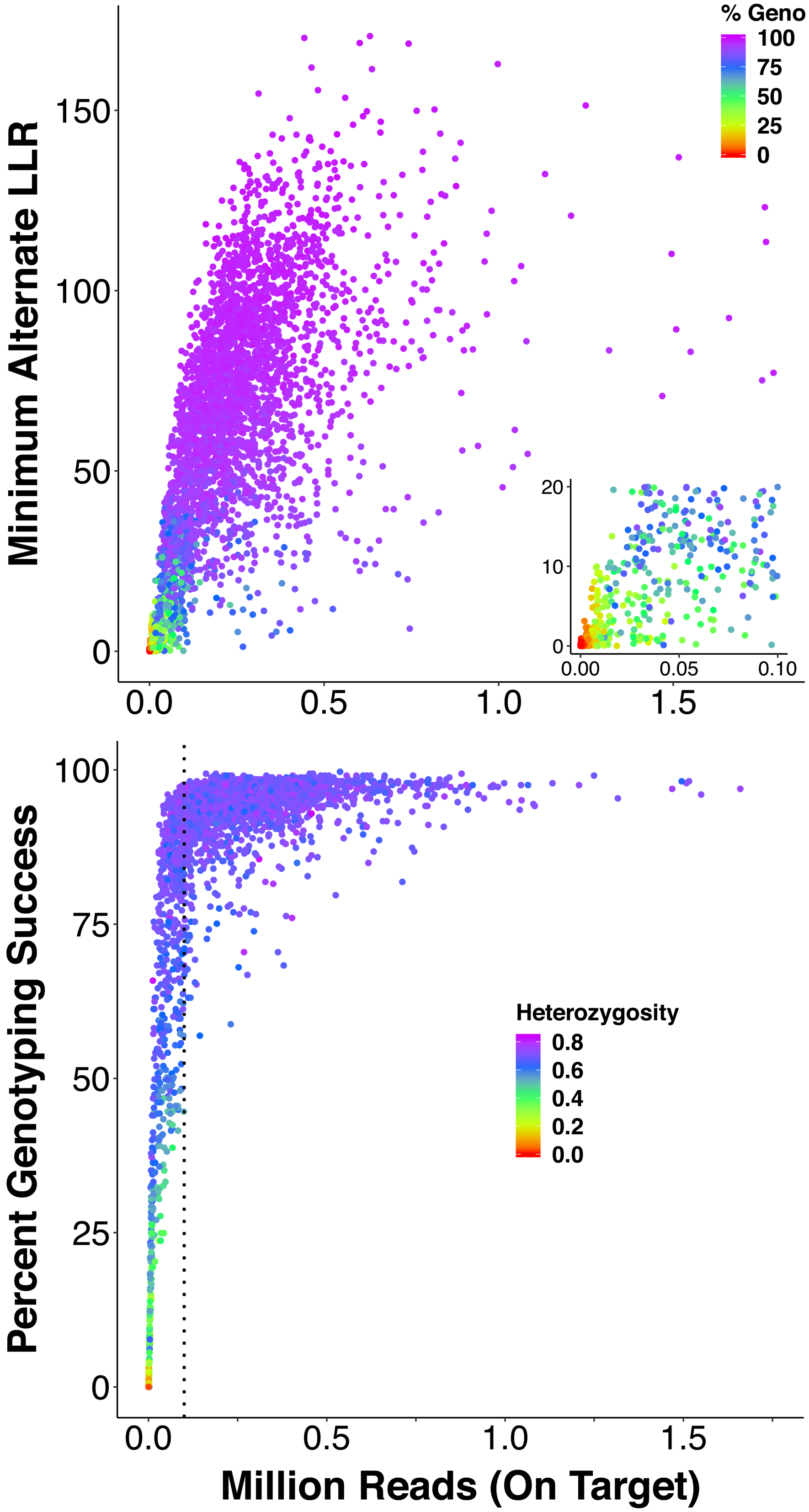
Hosted file

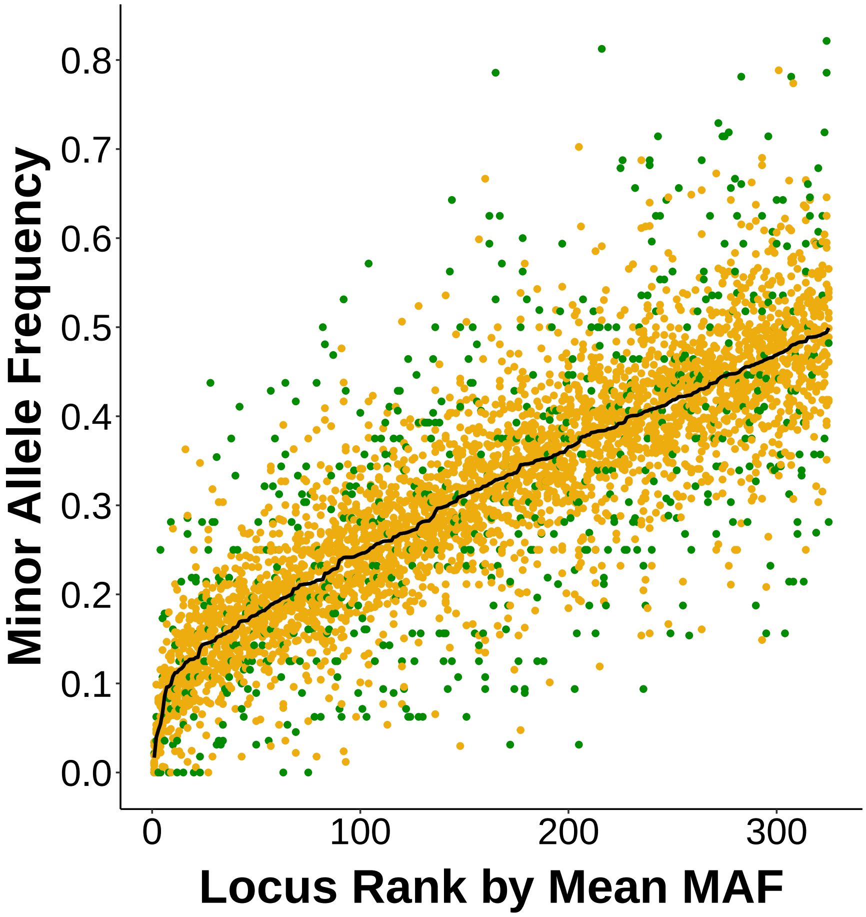
Table2_Fst_jk_intervals_formatted.xlsx available at https://authorea.com/users/353687/articles/477467-single-nucleotide-polymorphism-genotypes-and-ploidy-estimates-for-ploidy-variable-species-generated-with-massively-parallel-amplicon-sequencing

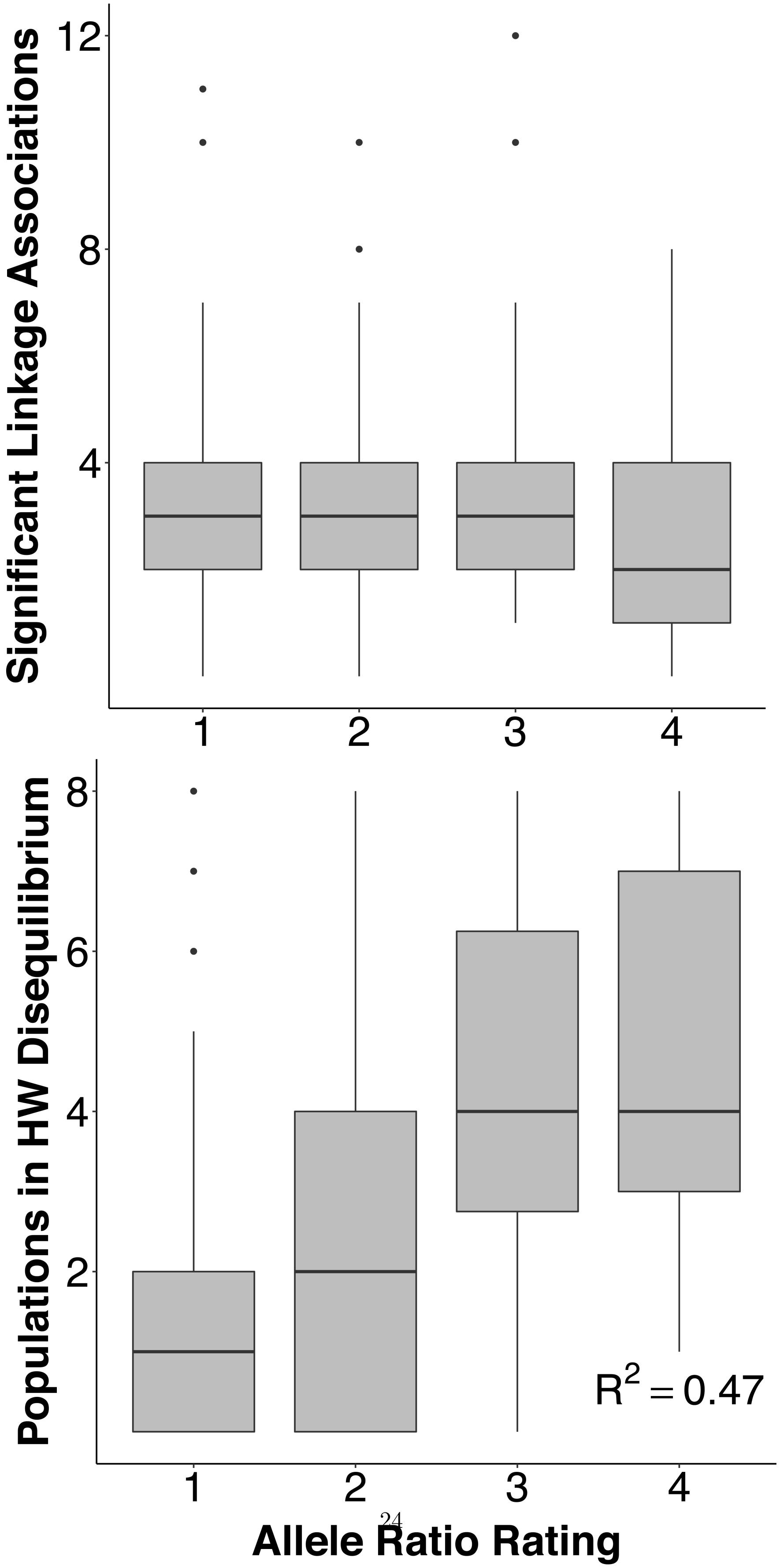

$R^2 = 0.59$
4N Likelihood Rank
300
200
100
0
1
2
3
4
$R^2 = 0.36$
Percent Mendelian Incompatabilities
0.6
0.4
0.2
0.0
1
2
3
4
Allele Ratio Rating


1.00
0.75
0.50
0.25
0.00
Density
0.0
0.1
0.2
0.3
Percent Mendelian Incompatabilities
Comparison 🟩 Parents 🟧 Not Parents
50
0
−50
−100
LOD Score
correct
incorrect
Parent Assignment
Parents Included 🟩 both 🟧 one □ neither


ML Relatedness
1.00
0.75
0.50
0.25
0.00
Self
Full
Half
Unrelated
Relationship


Minimum Alternate LLR
150
100
50
0
0.0
0.5
1.0
1.5
% Geno
100
75
50
25
0
20
10
0
0.00
0.05
0.10
Percent Genotyping Success
100
75
50
25
0
Heterozygosity
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
Million Reads (On Target)


Minor Allele Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
100
200
300
Locus Rank by Mean MAF


Significant Linkage Associations
12
8
4
1
2
3
4
Populations in HW Disequilibrium
8
6
4
2
$R^2 = 0.47$
1
2
3
4
Allele Ratio Rating