DR. GARRETT JUSTIN MCKINNEY (Orcid ID : 0000-0002-6267-2203)

Article type : Resource Article

Resolving allele dosage in duplicated loci using genotyping by sequencing data: a path forward for population genetic analysis

Garrett J. McKinney[1*], Ryan K. Waples[1,2], Carita E. Pascal[1], Lisa W. Seeb[1], James E. Seeb[1]

[1] School of Aquatic and Fishery Sciences, University of Washington, 1122 NE Boat Street, Box 355020, Seattle WA 98195-5020, USA.

[*]**Corresponding author:** Garrett J. McKinney (email: gjmckinn@uw.edu, phone: 1-765-430-3272)

[2] Current address: Department of Biology, The Bioinformatics Centre, University of Copenhagen, 2200 Copenhagen, Denmark.

Running head: Genotyping paralogs in GBS data

Key words: Genome duplication, paralog, genotyping, amplicon sequencing, RADSeq

Abstract

Whole genome duplications have occurred in the recent ancestors of many plants, fish, and amphibians. Signals of these whole genome duplications still exist in the form of paralogous loci. Recent advances have allowed reliable identification of paralogs in genotyping by sequencing (GBS) data such as that generated from restriction-site associated DNA sequencing (RADSeq); however, excluding paralogs from analyses is still routine due to difficulties in genotyping. This exclusion of paralogs may filter a large fraction of loci, including loci that may be adaptively important or informative for population genetic analyses. We present a maximum-likelihood method for inferring allele dosage in paralogs and assess its accuracy using simulated GBS, empirical RADSeq, and amplicon sequencing data from Chinook salmon. We accurately infer allele dosage for some paralogs from a RADSeq dataset and show how accuracy is dependent upon both read depth and allele frequency. The amplicon sequencing dataset, using RADSeq-derived markers, achieved sufficient depth to infer allele dosage for all paralogs. This study demonstrates that RADSeq locus discovery combined with amplicon sequencing of targeted loci is an effective method for incorporating paralogs into population genetic analyses.

Introduction

Gene and genome duplication events provide raw material for evolution through release of duplicate gene copies (Ohno 1970). Gene duplication can result in adaptation when one of the descendent copies gains a new function or speciation when alternative silencing of genes leads to reproductive incompatibilities between populations (Lynch & Conery 2000). When duplicate genomic regions are retained intact, the efficiency of selection to drive evolution is enhanced due to the increase in effective population size that results from polyploidy (Allendorf *et al.* 2015).

There are multiple lines of evidence suggesting the general importance of duplication in evolution. Genome duplication events have coincided with the origin of vertebrates (Holland *et al.* 1994) and teleosts (Crow *et al.* 2006). Gene duplication has been shown to facilitate adaptation to harsh environments (Kondrashov 2012) and polyploidy is thought to have enabled survival of flowering plants during the Cretaceious-Tertiary extinction event (Fawcett *et al.* 2009). Duplication has also been linked to speciation in multiple taxa: (1) elevated diversification often follows whole genome duplications in angiosperms (Tank *et al.* 2015) , (2) a high rate of gene duplication has been implicated in the species radiation of African cichlids (Brawand *et al.* 2014), and (3) divergent evolution of duplicate genes leads to loss of fitness in hybrid *A. thaliana* (Bikard *et al.* 2009).

Duplication of individual genes and gene families has also facilitated adaptation in many species (Kondrashov 2012); example adaptations include immune function (Zhang *et al.* 2015; Sackton *et al.*

2017), heavy metal tolerance (Chow *et al.* 2012), pesticide resistance (Lenormand *et al.* 1998), and domestication related traits (Liu *et al.* 2009). While there has been considerable study on the impact of individual duplicated genes and general patterns of evolution following genome duplication, population genetics studies using genotyping by sequencing (GBS) methodologies typically exclude paralogs from analysis (discussed below). This has the effect of excluding potentially important loci as well as entire genomic regions in species with ancestral genome duplications.

Paralogous loci can arise through whole-genome duplication, autopolyploidy or allopolyploidy, or duplication of chromosomal regions through segmental or tandem duplication. Ancestral whole-genome duplications have occurred in many species of plants (Fawcett *et al.* 2009; Wang *et al.* 2013; Clevenger & Ozias-Akins 2015) but have also taken place in some species of fish (Ohno *et al.* 1968; Ferris & Whitt 1980; Allendorf & Thorgaard 1984) and amphibians (Mable *et al.* 2011; Schmid *et al.* 2015). While less common, some taxa exhibit extensive paralogy as a result of segmental or tandem duplications such as salamanders (Sun *et al.* 2012) and lungfish (Biscotti *et al.* 2016). These various mechanisms of elevating ploidy can lead to complicated patterns where different ploidy levels can exist within species (Gompert & Mock 2017) and within individuals (Allendorf & Thorgaard 1984), copy number variation can occur within genes (Lighten *et al.* 2014), and paralogs can exhibit disomic or polysomic inheritance within the same chromosome (Allendorf & Thorgaard 1984). For this study we will focus on tetraploid paralogs that are undifferentiated and genotyped as a single locus; these loci are prevalent in organisms with ancestral whole-genome duplications and are likely the most common type of paralog encountered in genomic analyses.

In GBS data, paralogs are frequently collapsed into a single locus due to sequence similarity and the short sequence reads generated with current sequencing technologies. This presents two distinct difficulties in the analysis of paralogs: identification and genotyping. In the past, paralogs that have been collapsed into a single locus could only be reliably identified through alignments to a reference genome or by genotyping haploid individuals; however, methods have recently been developed that leverage populations-level analysis of GBS data to distinguish paralogous from non-paralogous loci (Verdu *et al.* 2016; McKinney *et al.* 2017; Willis *et al.* 2017) as well as identify individuals with elevated ploidy levels (Gompert & Mock 2017). Once identified, paralogous loci are often excluded from population genetic analysis because allele dosage (copy number of each allele) is difficult to quantify for heterozygous individuals (reviewed in Dufresne *et al.* 2014).

Diploid and tetraploid loci differ in the allele dosages, and resulting allele ratios, that are possible within heterozygous individuals. Diploid loci have a single heterozygous genotype (AB) with an allele ratio of 1:1. Tetraploid paralogs (duplicate loci) have up to three heterozygous genotypes with allele ratios of:

AAAB (3:1); AABB (2:2); ABBB (1:3). A special case where tetraploid paralogs are inherited disomically (diverged duplicate loci) and one of the loci is fixed for an allele results in only two heterozygous genotypes, for example AAAB and AABB if the A allele is fixed in one copy of the paralog. Theoretically, these different allele dosages can be identified based on observed reads; however, random sampling of alleles during sequencing causes the observed read ratios for heterozygotes to deviate from the expected values (read ratios for homozygotes can only deviate due to sequencing error). This problem is exacerbated with low sequencing coverage due to the stochastic variation in the number of sequence reads generated for each allele. Uncertainty in estimating allele dosage has led to the common practice of filtering paralogs from GBS data, without any attempt at genotyping or inclusion in population genetic analyses (e.g. Hecht *et al.* 2013; Dufresne 2016; Verdu *et al.* 2016; Tarpey *et al.* 2017).

Recent studies that attempted to incorporate paralogs into population genetic analyses using GBS data revealed potential signals of selection (Limborg *et al.* 2017; Waples *et al.* 2017). Limborg *et al.* (2017) estimated population allele frequencies directly from read counts for each allele using *PolyFreqs* (Blischak *et al.* 2016) while Waples *et al.* (2017) scored only the presence/absence of each allele within each individual. These methods of incorporating paralogs into population genetic analysis can be useful, particularly when sequence depth is low, but are still limited because they do not provide accurate allele dosage. Resolving allele dosage is vital because individual genotypes are fundamental components of many population genetic analyses (Dufresne *et al.* 2014).

Allele dosage has been successfully inferred using both fluorescent-based microarrays (Gidskehaug *et al.* 2011) and with GBS when ultra-high depth sequencing was used (Lighten *et al.* 2014; Ferrandiz-Rovira *et al.* 2015; Biedrzycka *et al.* 2017); however, these methods are impractical for many studies for a variety of reasons. Microarrays require development resources that are outside the scope of many non-model organism studies. Ultra-high depth sequencing has been used for targeted studies that generally genotype one or a few genes and achieve sequencing depths of up to tens of thousands of reads. While useful for interrogation of individual genes or gene families, ultra-high depth sequencing is intractable for genome-wide population genetic analyses.

Our goal was to identify a practical read depth and analysis pipeline to enable the scoring of dosage in duplicated genes detected in GBS data. We introduce a maximum-likelihood method to genotype allele dosage in paralogs and evaluate accuracy in simulated tetraploid GBS data. We then apply our method to a restriction-site associated DNA sequencing (RADSeq) and amplicon sequencing dataset from Chinook salmon, a species that retains ~17% of the paralogs from the salmonid whole-genome duplication (McKinney *et al.* 2017).

We found that genotype rate (proportion of individuals assigned a genotype) per locus was influenced both by read depth and minor allele frequency with low read depth and high minor allele frequency associated with reduced genotype rate. Genotype rate relative to read depth varied by genotype with heterozygous genotypes requiring greater read depth to achieve 100% genotype rate than homozygous genotypes. Simulation results showed that a genotype rate of > 95% for heterozygous genotypes was achieved at a read depth between 76 and 100, and > 99% was achieved at a read depth between 126 and 150. The RADSeq dataset had sufficient read depth to reliably genotype only low minor allele frequency loci. The amplicon sequencing dataset had sufficient read depth to genotype all loci. Combining our method of genotyping with reliable methods of paralog identification will allow future studies to incorporate paralogs into population genetic analyses.

Materials and Methods

Scoring allele dosage

We constructed a polyploid genotyper (*PolyGen*) to consider all possible genotypes for a locus based on the number of alleles and ploidy of the locus. Our genotype calls are based on allele dosage and do not distinguish among all possible chromosomal arrangements (e.g., AAAB = ABAA and ABAB =AABB). Allele dosage is inferred using a maximum likelihood algorithm that performs equations in the following steps:

1) The relative dosage for each allele is calculated for each possible genotype. For a tetraploid locus with two alleles the relative dosage for each possible genotype is:

Genotype	Allele A relative dosage	Allele B relative dosage
AAAA	1	0
AAAB	0.75	0.25
AABB	0.50	0.50
ABBB	0.25	0.75
BBBB	0	1

2) The chance that a read will be sampled from a given allele, $p(a)$, given a particular underlying genotype is a function of the relative dosage of the allele in the genotype as well as the error rate:

$$p(a) = d_r(a) * (1 - \epsilon) + (1 - d_r(a)) * \epsilon$$

(1)

Where $d_r(a)$ is the relative dosage of allele a and epsilon (ϵ) is the sequencing error rate.

3) The overall loglikelihood of a genotype, $L(g)$, is obtained by summing the relative dosage loglikelihoods for each allele:

$$L(g) = \sum_{a}^{n} \ln\big(p(a)\big) * c_a$$

(2)

Where $p(a)$ is the chance that a read will be sampled from allele a, c_a is the count of observations of allele a, and there are *n* alleles.

4) The two most likely genotypes are compared using a likelihood ratio test with one degree of freedom (Hohenlohe *et al.* 2010). The most likely genotype is assigned if the likelihood ratio test is significant at $\alpha = 0.05$, otherwise no genotype is assigned.

We implemented this algorithm in an R script (File S1). This algorithm is capable of genotyping loci with any number of alleles and any ploidy level but we evaluate it with only tetraploid loci as that is the most common ploidy level for paralogs.

Simulated and Empirical Datasets

Simulated data was used to assess the ability of *PolyGen* to reconstruct known genotypes through a range of read depths and allele frequencies. An initial population of 2,000 individuals was constructed; for each paralogous locus, 2,000 genotypes were generated following Hardy-Weinberg expectations based on a uniform-random-assigned allele frequency between zero and one. A total of 250 individuals were then randomly sampled from the full population to represent the study sample. A total of 1,500 paralogous loci were generated; each locus was assigned an average read depth between 10 and 150 by randomly drawing from a discrete uniform distribution. For each individual, the total number of reads for a locus was obtained by sampling from a Poisson distribution with the mean equal to the average depth for that locus. The number of reads for the A allele was obtained by drawing from a binomial distribution where the number of trials equals the total read depth for the locus and the probability equals the proportion of that allele in the underlying genotype. A sequencing error rate of 1% was simulated by modifying the allele probability. The reads for the B allele were obtained by subtracting the reads for the A allele from the total reads. Simulation was conducted in R (see File S2).

The allele frequency and average sequence depth per locus was varied to assess how these parameters influence paralog genotyping. Genotype accuracy was assessed by comparing the genotypes inferred by *PolyGen* to the true simulated genotypes.

We also used empirical data to evaluate the reliability of genotyping paralogs with GBS data given the read depth expectations that were determined in the simulations. Empirical data were derived from both RADSeq and amplicon sequencing from three populations; the RADSeq dataset from these populations previously was used to choose a subset of markers for development into an amplicon sequencing panel (data not shown).

RADSeq data were generated for three populations of Chinook salmon inhabiting the large Kuskokwim River drainage in Western Alaska, USA (Goodnews, George, and Necons rivers). A total of 48 individuals per population was sampled; DNA was extracted and sequencing libraries prepared with the *SbfI* enzyme following the methods of Baird *et al.* (2008) and Everett *et al.* (2012). Samples were sequenced on a HiSeq 4000 with single-end 100bp reads; 96 samples were sequenced per lane. Two rounds of sequencing were conducted and the volume of DNA for each individual adjusted in the second round of sequencing to reduce variation in sequence reads per individual (Prince *et al.* 2017). Sequence data was processed with *STACKS* v1.31 (Catchen *et al.* 2011) using default settings with the following exceptions: process_radtags (-c -r -q -t 94), ustacks (-r --model_type bounded --bound_low 0 --bound_high 0.05), cstacks (-n 2). The *STACKS* catalog of variation was created using five individuals from each population and was combined with the catalog of variation from McKinney *et al.* (2017) to ensure consistent locus names between the studies. Loci genotyped in at least 80% of the samples and with a minor allele frequency of 0.05 in one or more populations were output from *STACKS* as a .vcf file. Allele-specific read counts from the vcf were used as input to *HDplot* (McKinney *et al.* 2017) to identify paralogs and as input to *PolyGen* to genotype paralogs.

A total of 59 paralogs from the RADSeq dataset were chosen to develop an amplicon sequencing panel. Amplicon sequencing (GTseq, Campbell *et al.* 2015) was conducted on an additional 48 individuals from each population. Sequencing libraries were prepared following the methods of Campbell *et al.* (2015). These samples were sequenced in combination with other Chinook salmon samples and loci sequenced for another amplicon sequencing study (data not shown). The total sequencing effort included 300 individuals and 1,200 loci on a single lane of an Illumina HiSeq4000 with single-end 100bp reads. A custom perl script was used to obtain read counts for each allele at each locus; these read counts were then used as input to *PolyGen*.

Allele Frequency Estimation

We estimated allele frequencies for the simulated paralogs and the empirical paralogs by counting the observed occurrences of each allele in the genotypes output by *PolyGen*. For the simulated data, we genotyped loci under tetraploid and diploid models to determine how genotyping unidentified paralogs as diploid loci affects allele frequency estimates. We then compared the estimated allele frequency to the true allele frequency for the simulated paralogs to assess accuracy in allele frequency estimation.

Results

Genotyping

We simulated a dataset of 1,500 paralogs with varying average read depths from 10 to 150 reads per locus to examine: 1) genotype rate and 2) genotype accuracy and discrepancies. Genotype rate per locus was influenced both by average read depth per locus as well as the locus minor allele frequency (Figure 1A). Reduced read depths resulted in a lower genotype rate, and for any given read depth, minor allele frequency was negatively correlated with genotype rate. Genotype rate for a given read depth was strongly dependent on the underlying true genotype (Table 1). Genotype rate for homozygous genotypes reached 99% at a read depth of only 26 but heterozygous genotypes required much greater read depths to reach similar genotype rates. Heterozygous genotypes reached >95% genotype rate between 76 and 100 reads. The genotype rate for all genotypes was >99% between 126 and 150 reads. Genotype accuracy was generally high with few miscalled genotypes at any read depth (Table 1). A maximum miscall rate of ~5% was seen for AABB heterozygotes with less than 25x coverage; this dropped to ~1.3% between 51x and 75x coverage. There were essentially no miscalled genotypes for read depths above 76 with a maximum 0.5% incorrect calls for true AABB genotypes. Read depths >100 resulted in genotype rates >95% and genotype accuracy >99% for all genotype classes.

Plotting the allele ratio and read depth for each simulated genotype showed complete separation in allele ratio distributions for heterozygous genotypes after 150 reads; below 150 reads the amount of overlap increased with decreasing read depth (Figure 2 A). Plotting assigned genotypes revealed that the pattern of uncalled genotypes coincided with regions of overlap in allele ratios between genotypes (Figure 2 B).

A total of 17,810 RADSeq loci passed genotype rate and minor allele frequency filters; 2,806 (16%) of these loci were identified as paralogous by *HDplot* (Figure S1). The average read depth per locus was 43 and ranged from 5 to 112 (Figure S2A). The genotype rate per locus was influenced both by average read depth per locus and minor allele frequency (Figure 1B) which was in concordance with results from the simulated data; however, the spread of genotype rate relative to average read depth was broader than in the simulated datasets.

Fifty-nine paralogs were developed into assays for amplicon sequencing. Average depth was 817 reads with a range of 56 to 2164 (Figure S2B); average genotype rate was >99.9% with a range of 98.7% to 100%.

Histograms of allele ratios revealed clear peaks associated with each genotype class (Figure 3). The three categories for disomically inherited paralogs (AABB diverged duplicates, Figure 3A) were easily distinguished from the five category plots for tetrasomically inherited paralogs (Figure 3B).

Allele Frequency Estimation

Allele frequency estimates were systematically biased towards 0.5 when the simulated paralogous loci were genotyped under diploid assumptions (Figure 4A). Treating tetraploid loci as diploid led to elevated estimates of allele frequency for true frequencies less than 0.5 and decreased estimates for true frequencies greater than 0.5. When simulated paralogous loci were genotyped under tetraploid assumptions, the estimated allele frequencies closely tracked the true allele frequencies but loci with low read depth showed a slight downward bias in estimated relative to true frequency for true frequencies less than 0.5 and a slight upward bias in estimated frequencies for true frequencies greater than 0.5 (Figure 4B).

Discussion

Genotyping

Accuracy of genotyping was high regardless of read depth but both read depth and minor allele frequency strongly influenced the call rate of paralogs (Figure 1A, Table 1). This pattern was the result of genotype-specific relationships between read depth and genotype rate. Genotyping accuracy was >99.9% for homozygous genotypes at read depths of 26 whereas heterozygous genotypes required read depths of >100 to achieve similar accuracy (Table 1). Loci with low minor allele frequency tend to have high genotype rates regardless of read depth because there are few heterozygous genotypes. Loci with high minor allele frequency are more strongly influenced by read depth because of the greater proportion of heterozygous genotypes.

The relationship among genotype rate, read depth, and minor allele frequency has important implications for downstream analyses. Genotype rate filters are commonly used to identify high-quality loci in RADSeq datasets. This could lead to systemic biases in the retained loci when applied to paralogous loci. A total of 1,471 paralogs in our RADSeq dataset (52% of total paralogs) were retained with a genotype rate filter of 80%. The average minor allele frequency of the retained loci was much lower than the discarded loci (0.11 vs. 0.37). For some analyses, such as site frequency spectrum, the bias introduced by

locus filtering alone would in turn lead to biased interpretations. For loci that are retained, heterozygous individuals are more likely to have uncalled genotypes than homozygous individuals; this has the potential to skew results in studies with insufficient read depth to reliably genotype paralogs.

Our simulations suggest that tetraploid paralogs can reach near-perfect genotyping at average read depths of 100. While this depth is greater than found in most RADSeq studies, it is achievable with appropriate consideration of genome size, number of loci generated by RADSeq method, number of individuals sequenced, and sequencer output. RADSeq methods that further reduce the number of sequenced loci, such as ddRAD (Peterson *et al.* 2012), could also be used to achieve greater sequence depth for genotyping paralogs.

Amplicon sequencing, targeting a fixed locus set, is a more tractable solution to obtain sufficient read depth to genotype paralogs. Amplicon sequencing methods such as GT-seq (Campbell *et al.* 2015) or RAD-capture (Ali *et al.* 2016) generally achieve greater read depth than RADSeq, and the number of individuals and loci sequenced per lane can be manipulated to achieve desired read depths. We achieved a 19-fold increase in read depth for paralogs in the amplicon sequencing dataset relative to the RADSeq dataset. We report the results for amplicon sequencing of 144 individuals and 59 loci, but note that the samples and loci in this study shared a sequencing lane with another project. The average read depth of 817 reads per paralogous locus was achieved with a total of 300 individuals and 1,200 loci sequenced on a single lane of an Illumina HiSeq 4000. The results from our simulation as well as our amplicon data suggest that a read depth between 100 and 150 is adequate to reliably genotype paralogs. Under this assumption we have room to increase our loci or samples 6 to 8 fold and still achieved sufficient read depth.

Variation in reads per allele

Multiple factors will contribute to the successful genotyping of paralogs. We simulated variation in average read depth per locus to demonstrate the effect of read depth, but equally important is the variation in reads sequenced per allele at a locus. In the simulated dataset, variation in reads per allele was modeled by assuming that each allele has an equal probability of being sequenced. The RADSeq dataset showed more uncalled genotypes for a given depth than the simulated dataset, suggesting other sources of variation in reads per allele. The contributing factors to variation in reads per allele are unclear; possible contributors include initial DNA quality, methods of library preparation, sequencing technology, and PCR duplicates. Of these possibilities, the effect of PCR duplicates is easiest to ascertain; however, the Chinook salmon dataset was obtained using single-end sequencing, so PCR duplicates could not be identified.

Allele Frequency Estimation

Accurate allele frequency estimates are important for assessing population genetic parameters such as F_{ST}. GBS datasets are typically filtered so that retained loci conform to HWE expectations for diploid loci. This filtering method failed to identify approximately two-thirds of the paralogs in a previous Chinook salmon study (McKinney et al. 2017); it is likely that other studies using HWE expectations likely failed to identify a large proportion of paralogs, and these paralogs were subsequently genotyped as diploid loci. Allele frequency estimates were systemically biased when paralogs were genotyped as diploid loci. This problem is likely both common and unrecognized in organisms with mixed ploidy. Until recently it was difficult to reliably identify paralogs in GBS data, particularly in organisms with mixed ploidy, and many paralogs likely escaped detection. Accurate paralog identification, either using genomic resources or tools such as *HDplot*, is a critical first step for accurately estimating allele frequencies. If read depth is sufficient then individual genotypes could be obtained using *PolyGen*, allowing for allele frequency estimation. If read depth is insufficient for inferring individual genotypes, then allele frequencies could be estimated using programs such as *polyFreqs* (Blischak *et al.* 2016) that estimate frequencies directly from read counts or by using presence/absence methods of allele scoring.

Importance of Paralogs in Population Genetics

Paralogs are commonly excluded from population genetic studies due to genotyping difficulties. While exclusion has been a practical solution to a real problem, the impact of excluding paralogs is more than the loss of a few loci. Drift and selection act differently on paralogs than non-paralogs, with selection acting more efficiently on paralogs due to increased effective population size (Meirmans & Van Tienderen 2013). In addition, paralogs are often distributed non-randomly throughout the genome (Linardopoulou *et al.* 2005). In duplicated salmonids, a large fraction of loci are retained as paralogs even though the majority of the genome has rediploidized. These retained duplicates are concentrated in the distal ends of eight pairs of chromosome arms that are conserved across the salmonid genus *Oncorhynchus* (Brieuc *et al.* 2014; Kodama *et al.* 2014; Larson *et al.* 2016; Waples *et al.* 2016). Presumably many retained duplicates in other autotetraploids are distally located as well (see Allendorf *et al.* 2015; Limborg *et al.* 2016). For these species, excluding paralogs in population genetics studies will result in the failure to interrogate entire regions of the genome. Finally, different genomic regions can reveal different information about species histories. For coalescent approaches, regions of genome duplication will have a deeper time to most recent common ancestor (TMRCA) than non-duplicated regions due to differences in effective population size (Allendorf *et al.* 2015). The deeper TMRCA of duplicated regions could allow researchers to look back further in the demographic or evolutionary history than they could with non-duplicated regions.

Here we offer a solution for genotyping allele dosage in paralogs and demonstrate GBS approaches to successfully genotype and incorporate paralogs into population genetic analysis. This method, in combination with recently developed methods for identifying paralogs in GBS datasets, will enable the incorporation of paralogs into population genetic analyses and unlock analysis of duplicated genomic regions.

Acknowledgements

We thank Wes Larson for providing constructive comments on the manuscript. This research was partially funded by the Alaska Sustainable Salmon Fund under study #44812 and #44913 from NOAA, U.S. Department of Commerce, administered by the Alaska Department of Fish and Game (ADFG). The statements, findings, conclusions and recommendations are those of the authors and do not necessarily reflect the views of the NOAA, the U.S. Department of Commerce, or the ADFG.

Data Accessibility

RADSeq data is available in NCBI SRA SRP129033.

Amplicon sequencing data is available in NCBI SRA SRP129894.

Supporting Information

File S1. R code for polygen algorithm.

File S2. R code to create simulated GBS read count data.

Figure S1. Results from HDplot for RADSeq dataset. Read-ratio deviation (*D*, y-axis) is plotted against heterozygosity (*H*, x-axis). Loci identified as singletons are in blue, loci identified as duplicates are in pink, and loci identified as diverged duplicates (disomically inherited duplicates) are in green.

Figure S2. Histogram of average read depth for paralogs identified in the datasets for: (A) RADSeq and (B) amplicon sequencing.

References

Ali OA, O'Rourke SM, Amish SJ*, et al.* (2016) RAD capture (Rapture): flexible and efficient sequence-based genotyping. *Genetics* **202**, 389-400.

Allendorf FW, Bassham S, Cresko WA*, et al.* (2015) Effects of crossovers between homeologs on inheritance and population genomics in polyploid-derived salmonid fishes. *J Hered* **106**, 217-227.

Allendorf FW, Thorgaard GH (1984) Tetraploidy and the evolution of salmonid fishes. In: *Evolutionary Genetics of Fishes* (ed. Turner B), pp. 1-53. Plenum Publishing Corporation.

Baird NA, Etter PD, Atwood TS*, et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376.

Biedrzycka A, Sebastian A, Migalska M, Westerdahl H, Radwan J (2017) Testing genotyping strategies for ultra-deep sequencing of a co-amplifying gene family: MHC class I in a passerine bird. *Molecular Ecology Resources* **17**, 642-655.

Bikard D, Patel D, Le Mette C*, et al.* (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* **323**, 623-626.

Biscotti MA, Gerdol M, Canapa A*, et al.* (2016) The lungfish transcriptome: a glimpse into molecular evolution events at the transition from water to land. *Sci Rep* **6**, 21571.

Blischak PD, Kubatko LS, Wolfe AD (2016) Accounting for genotype uncertainty in the estimation of allele frequencies in autopolyploids. *Mol Ecol Resour* **16**, 742-754.

Brawand D, Wagner CE, Li YI*, et al.* (2014) The genomic substrate for adaptive radiation in African cichlid fish. *Nature* **513**, 375-381.

Brieuc MSO, Waters CD, Seeb JE, Naish KA (2014) A dense linkage map for Chinook salmon (*Oncorhynchus tshawytscha*) reveals variable chromosomal divergence following an ancestral whole genome duplication event. *Genes Genomes Genetics* **4**, 447-460.

Campbell NR, Harmon SA, Narum SR (2015) Genotyping-in-Thousands by sequencing (GT-seq): a cost effective SNP genotyping method based on custom amplicon sequencing. *Mol Ecol Resour* **15**, 855-867.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics* **1**, 171-182.

Chow EW, Morrow CA, Djordjevic JT, Wood IA, Fraser JA (2012) Microevolution of *Cryptococcus neoformans* driven by massive tandem gene amplification. *Mol Biol Evol* **29**, 1987-2000.

Clevenger JP, Ozias-Akins P (2015) SWEEP: a tool for filtering high-quality SNPs in polyploid crops. *Genes Genomes Genetics* **5**, 1797-1803.

Crow KD, Stadler PF, Lynch VJ, Amemiya C, Wagner GP (2006) The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. *Mol Biol Evol* **23**, 121-136.

391 Dufresne F (2016) Don't throw the baby out with the bathwater: identifying and mapping paralogs in
392 salmonids. *Mol Ecol Resour* **16**, 7-9.
393 Dufresne F, Stift M, Vergilino R, Mable BK (2014) Recent progress and challenges in population genetics
394 of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools.
395 *Mol Ecol Resour* **23**, 40-69.
396 Everett MV, Miller MR, Seeb JE (2012) Meiotic maps of sockeye salmon derived from massively parallel
397 DNA sequencing. *BMC Genomics* **13**, 521.
398 Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance
399 to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A* **106**, 5737-5742.
400 Ferrandiz-Rovira M, Bigot T, Allaine D, Callait-Cardinal MP, Cohas A (2015) Large-scale genotyping of
401 highly polymorphic loci by next-generation sequencing: how to overcome the challenges to
402 reliably genotype individuals? *Heredity* **114**, 485-493.
403 Ferris SD, Whitt GS (1980) Genetic variability in species with extensive gene duplication: the tetraploid
404 Catastomid fishes. *American Naturalist* **115**, 650-666.
405 Gidskehaug L, Kent M, Hayes BJ, Lien S (2011) Genotype calling and mapping of multisite variants using
406 an Atlantic salmon iSelect SNP array. *Bioinformatics* **27**, 303-310.
407 Gompert Z, Mock KE (2017) Detection of individual ploidy levels with genotyping-by-sequencing (GBS)
408 analysis. *Mol Ecol Resour*.
409 Hecht BC, Campbell NR, Holecek DE, Narum SR (2013) Genome-wide association reveals genetic basis
410 for the propensity to migrate in wild populations of rainbow and steelhead trout. *Molecular*
411 *Ecology* **22**, 3061-3076.
412 Hohenlohe PA, Bassham S, Etter PD*, et al.* (2010) Population genomics of parallel adaptation in
413 threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**, e1000862.
414 Holland PW, Garcia-Fernandez J, Williams NA, Sidow A (1994) Gene duplications and the origins of
415 vertebrate development. *Dev Suppl*, 125-133.
416 Kodama M, Brieuc MS, Devlin RH, Hard JJ, Naish KA (2014) Comparative mapping between coho salmon
417 (*Oncorhynchus kisutch*) and three other Salmonids suggests a role for chromosomal
418 rearrangements in the retention of duplicated regions following a whole genome duplication
419 event. *Genes Genomes Genetics* **4**, 1717-1730.
420 Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing
421 environment. *Proc Biol Sci* **279**, 5048-5057.

Larson WA, McKinney GJ, Limborg MT*, et al.* (2016) Identification of multiple QTL hotspots in sockeye salmon (*Oncorhynchus nerka)* using Genotyping-by-Sequencing and a dense linkage map. *Journal of Heredity* **107**, 122-133.

Lenormand T, Guillemaud T, Bourguet D, Raymond M (1998) Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito *Culex pipiens*. *Evolution* **52**, 1705-1712.

Lighten J, van Oosterhout C, Paterson IG, McMullan M, Bentzen P (2014) Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Mol Ecol Resour* **14**, 753-767.

Limborg MT, Larson WA, Seeb LW, Seeb JE (2017) Screening of duplicated loci reveals hidden divergence patterns in a complex salmonid genome. *Mol Ecol*.

Limborg MT, Seeb LW, Seeb JE (2016) Sorting duplicated loci disentangles complexities of polyploid genomes masked by genotyping by sequencing. *Mol Ecol* **25**, 2117-2129.

Linardopoulou EV, Williams EM, Y. F*, et al.* (2005) Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437**, 94-100.

Liu GE, Ventura M, Cellamare A*, et al.* (2009) Analysis of recent segmental duplications in the bovine genome. *BMC Genomics* **10**, 571.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-1155.

Mable BK, Alexandrou MA, Taylor MI (2011) Genome duplication in amphibians and fish: an extended synthesis. *Journal of Zoology* **284**, 151-182.

McKinney GJ, Waples RK, Seeb LW, Seeb JE (2017) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Mol Ecol Resour* **17**, 656-669.

Meirmans PG, Van Tienderen PH (2013) The effects of inheritance in tetraploids on genetic diversity and population divergence. *Heredity (Edinb)* **110**, 131-137.

Ohno S (1970) *Evolution by gene duplication* Springer-Verlag, New York.

Ohno S, Wolf U, Atkin NB (1968) Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169-187.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7**.

Prince DJ, O'Rourke SM, Thompson TQ*, et al.* (2017) The evolutionary basis of premature migration in Pacific salmon highlights the utility of genomics for informing conservation. *Science Advances* **3: e1603198**.

Sackton TB, Lazzaro BP, Clark AG (2017) Rapid expansion of immune-related gene families in the house fly, *Musca domestica*. *Mol Biol Evol* **34**, 857-872.

Schmid M, Evans BJ, Bogart JP (2015) Polyploidy in Amphibia. *Cytogenet Genome Res* **145**, 315-330.

Sun C, Shepard DB, Chong RA*, et al.* (2012) LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol Evol* **4**, 168-183.

Tank DC, Eastman JM, Pennell MW*, et al.* (2015) Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol* **207**, 454-467.

Tarpey CM, Seeb JE, McKinney GJ*, et al.* (2017) SNP data describe contemporary population structure and diversity in allochronic lineages of pink salmon (*Oncorhynchus gorbuscha*). *Canadian Journal of Fisheries and Aquatic Sciences*.

Verdu CF, Guichoux E, Quevauvillers S*, et al.* (2016) Dealing with paralogy in RADseq data: in silico detection and single nucleotide polymorphism validation in Robinia pseudoacacia L. *Ecology and Evolution* **6**, 7323-7333.

Wang N, Thomson M, Bodles WJ*, et al.* (2013) Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Mol Ecol* **22**, 3098-3111.

Waples RK, Seeb JE, Seeb LW (2017) Congruent population structure across paralogous and nonparalogous loci in Salish Sea chum salmon (*Oncorhynchus keta*). *Mol Ecol*.

Waples RK, Seeb LW, Seeb JE (2016) Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (*Oncorhynchus keta*). *Mol Ecol Resour* **16**, 17-28.

Willis SC, Hollenbeck CM, Puritz JB, Gold JR, Portnoy DS (2017) Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Mol Ecol Resour* **17**, 955-965.

Zhang L, Li L, Guo X*, et al.* (2015) Massive expansion and functional divergence of innate immune genes in a protostome. *Sci Rep* **5**, 8693.

Table 1. Percent genotype accuracy for simulated data divided into average read depth intervals of 1-25 reads, 26-50, 51-75, 76-100, 101-25, and 126-150 reads. Within each read depth interval rows are the true genotype while columns are the genotype inferred by *PolyGen*.

1-25 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	95.0	0.3	0.0	0.0	0.0	4.8
AAAB	1.8	49.9	1.2	0.0	0.0	47.2
AABB	0.0	2.4	20.6	2.4	0.1	74.5
ABBB	0.0	0.0	1.0	50.7	1.9	46.4
BBBB	0.0	0.0	0.0	0.2	94.9	4.8

26-50 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	99.6	0.0	0.0	0.0	0.0	0.4
AAAB	0.1	82.5	1.3	0.0	0.0	16.1
AABB	0.0	1.4	67.2	1.8	0.0	29.6
ABBB	0.0	0.0	1.1	84.3	0.2	14.4
BBBB	0.0	0.0	0.0	0.0	99.4	0.5

51-75 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	100.0	0.0	0.0	0.0	0.0	0.0
AAAB	0.0	93.6	0.6	0.0	0.0	5.8
AABB	0.0	0.5	88.4	0.8	0.0	10.3
ABBB	0.0	0.0	0.5	94.9	0.0	4.6
BBBB	0.0	0.0	0.0	0.0	100.0	0.0

76-100 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	100.0	0.0	0.0	0.0	0.0	0.0
AAAB	0.0	97.7	0.2	0.0	0.0	2.1
AABB	0.0	0.2	95.7	0.3	0.0	3.8
ABBB	0.0	0.0	0.2	98.3	0.0	1.5
BBBB	0.0	0.0	0.0	0.0	100.0	0.0

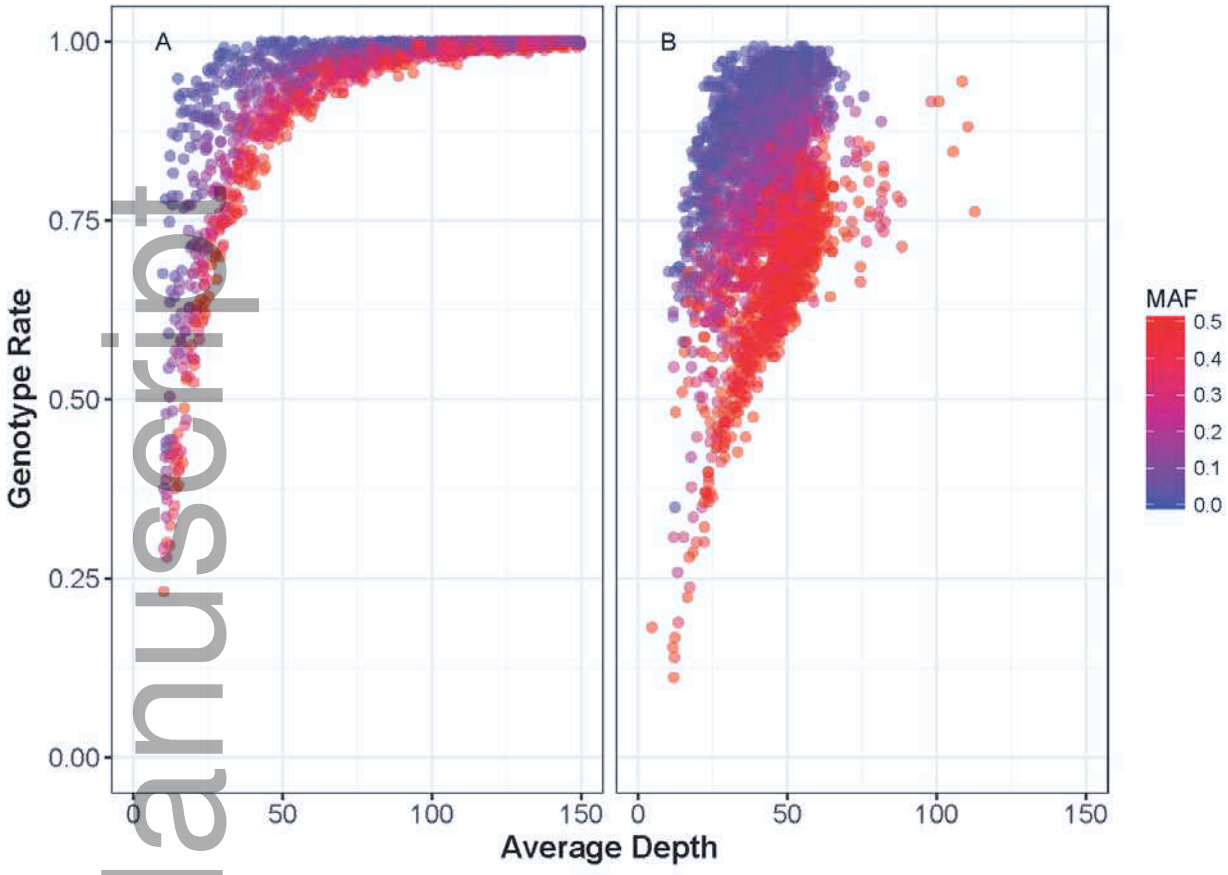
101-125 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	100.0	0.0	0.0	0.0	0.0	0.0
AAAB	0.0	99.1	0.1	0.0	0.0	0.8
AABB	0.0	0.1	98.3	0.1	0.0	1.5
ABBB	0.0	0.0	0.0	99.4	0.0	0.5
BBBB	0.0	0.0	0.0	0.0	100.0	0.0

126-150 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	100.0	0.0	0.0	0.0	0.0	0.0
AAAB	0.0	99.7	0.1	0.0	0.0	0.3
AABB	0.0	0.0	99.4	0.1	0.0	0.5
ABBB	0.0	0.0	0.0	99.8	0.0	0.2
BBBB	0.0	0.0	0.0	0.0	100.0	0.0

Figure 1. Genotype rate vs. average read depth per locus for A) simulated paralogs and B) RADSeq paralogs. Loci are displayed as dots and color coded by minor allele frequency (MAF). In both the simulated and the RADSeq data, genotype rate is dependent upon read depth and minor allele frequency. Reduced read depths result in lower genotype confidence and a decreased genotype rate. Loci with a lower minor allele frequency have an increased genotype rate relative to loci with a higher minor allele frequency for a given read depth.



492

493

494 **Figure 2.** Patterns of observed allele ratios by read depth for all loci in the simulated data. The true
495 genotypes are shown in A and the inferred genotypes are shown in B. The pattern of unassigned
496 genotypes mirrored regions of significant overlap in allele ratios among true genotypes.

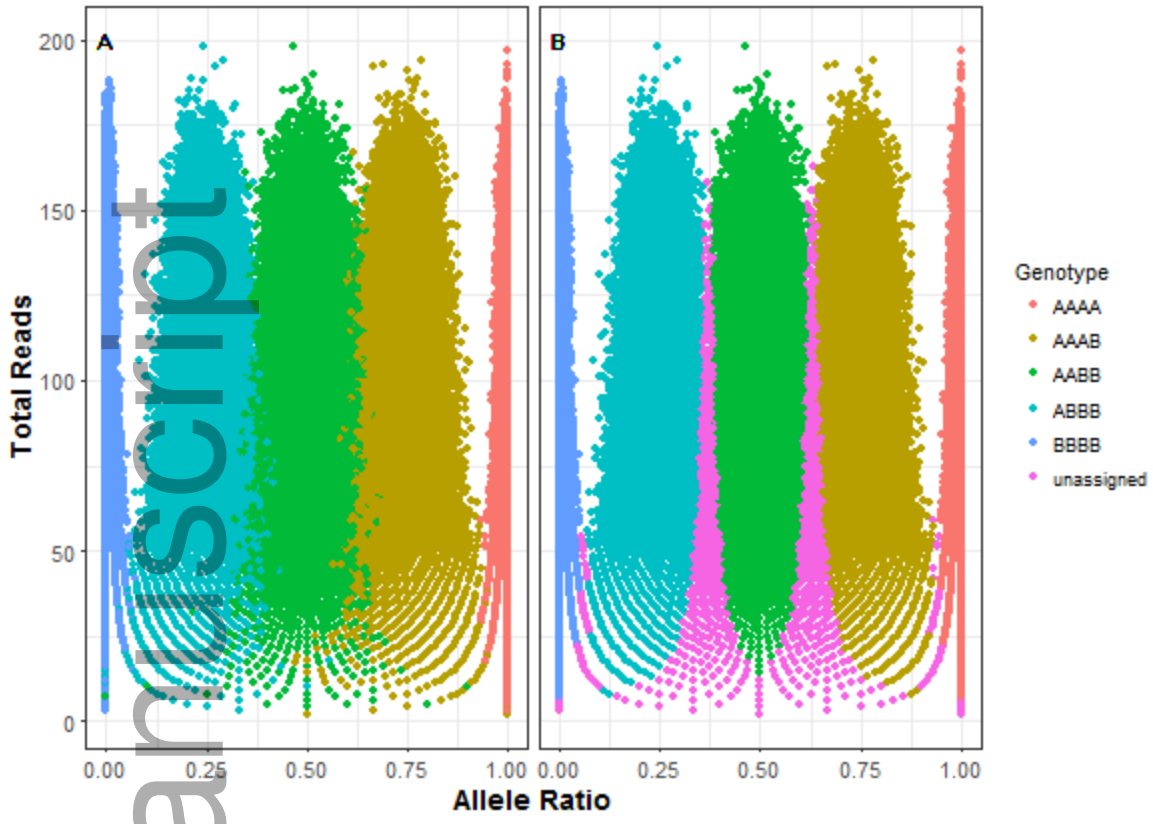


Figure 3. Histograms of observed allele ratios for A) locus RAD25055_38 and B) locus RAD48683_32 in the amplicon sequencing dataset. Allele ratio is given on the x-axis and number of individuals for each allele ratio is given on the y-axis. Locus RAD25055_38 had an average read depth of 505 and a 99% genotype rate. Only three genotype classes are seen in locus RAD25055_38 because the paralogs are inherited as independent disomic loci and one paralog has no allelic variation. Locus RAD48683_32 had an average read depth of 1,393 and exhibits all five genotype classes which is consistent with a tetrasomically inherited paralog.

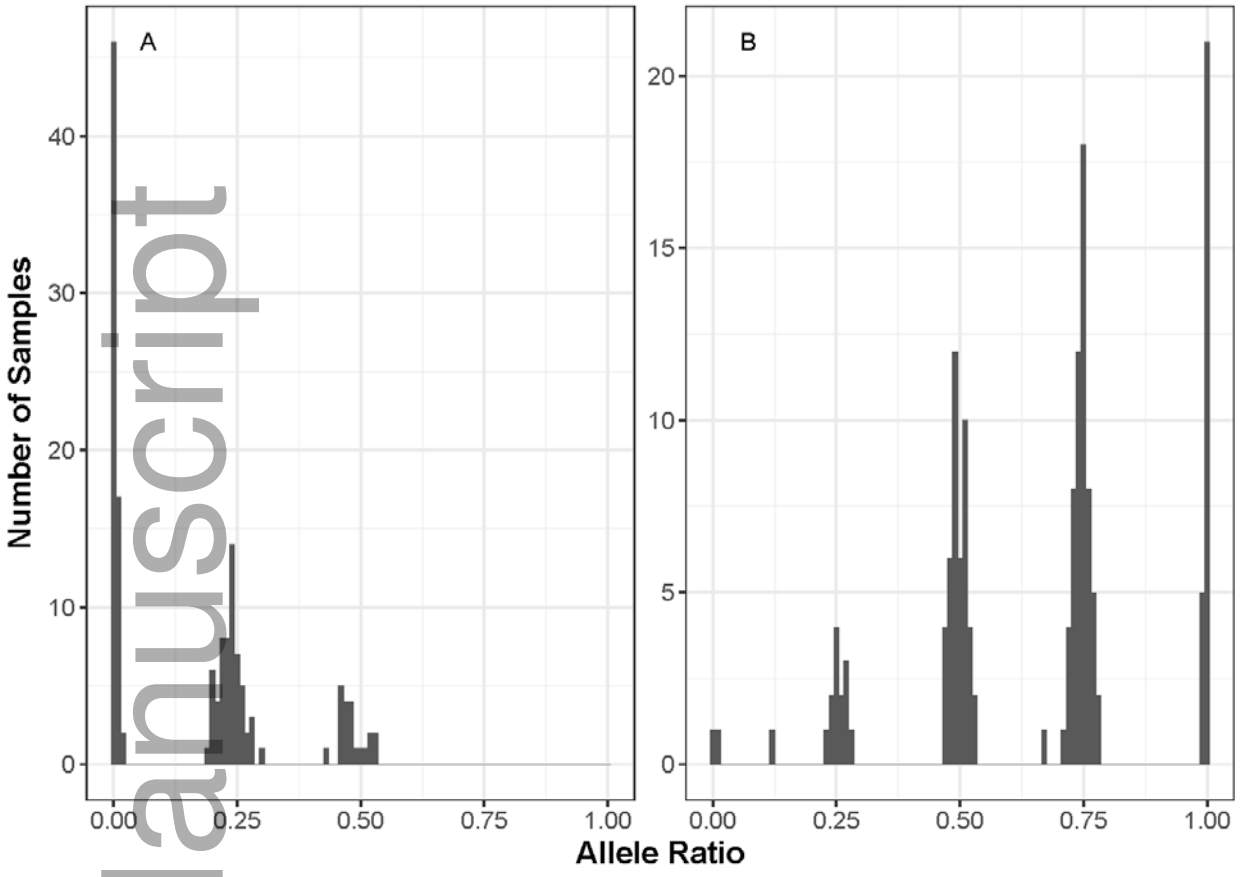
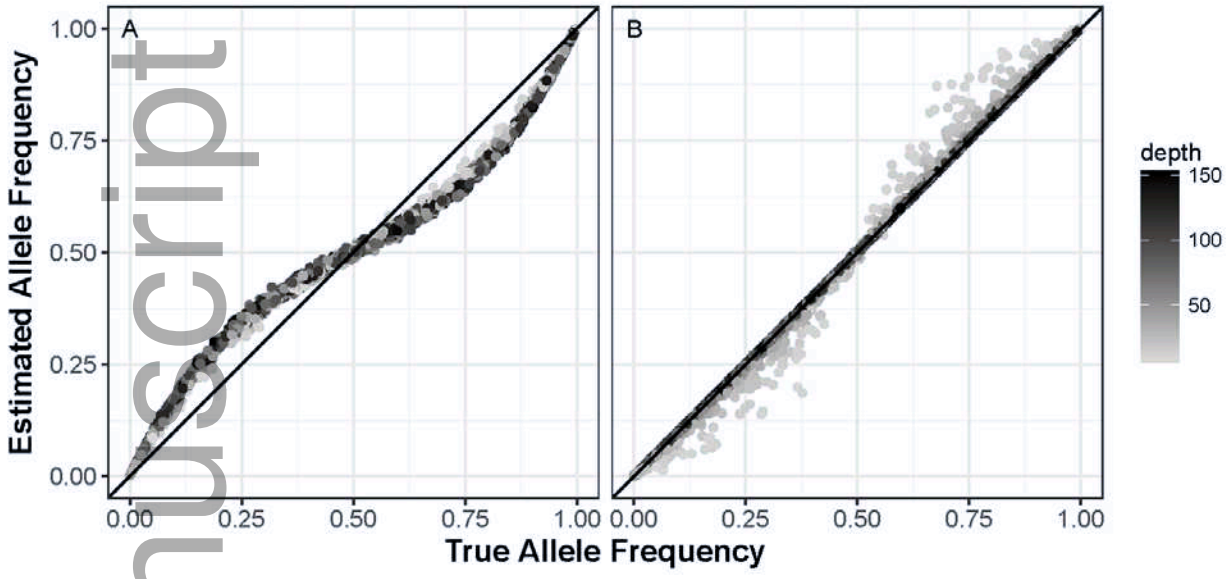


Figure 4. Allele frequency estimates for simulated duplicate loci when treated as A) diploid or B) tetraploid. Each dot is a locus, the true tetraploid allele frequency is given on the x-axis and the estimated allele frequency is given on the y-axis. The diagonal line shows the 1:1 relationship expected if estimated allele frequencies matched true allele frequencies. Allele frequency estimates are coded on a grayscale by the average read depth in B. Allele frequency estimates show a systemic bias when tetraploid loci are treated as diploid. Allele frequency estimates are accurate for read depths > 30 when tetraploid loci are genotyped with the correct ploidy but low read depth loci show a bias in allele frequency estimates.



515

Table 1. Percent genotype accuracy for simulated data divided into average read depth intervals of 1-25 reads, 26-50, 51-75, 76-100, 101-25, and 126-150 reads. Within each read depth interval rows are the true genotype while columns are the genotype inferred by PolyGen.

1-25 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	95.0	0.3	0.0	0.0	0.0	4.8
AAAB	1.8	49.9	1.2	0.0	0.0	47.2
AABB	0.0	2.4	20.6	2.4	0.1	74.5
ABBB	0.0	0.0	1.0	50.7	1.9	46.4
BBBB	0.0	0.0	0.0	0.2	94.9	4.8

26-50 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	99.6	0.0	0.0	0.0	0.0	0.4
AAAB	0.1	82.5	1.3	0.0	0.0	16.1
AABB	0.0	1.4	67.2	1.8	0.0	29.6
ABBB	0.0	0.0	1.1	84.3	0.2	14.4
BBBB	0.0	0.0	0.0	0.0	99.4	0.5

51-75 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	100.0	0.0	0.0	0.0	0.0	0.0
AAAB	0.0	93.6	0.6	0.0	0.0	5.8
AABB	0.0	0.5	88.4	0.8	0.0	10.3
ABBB	0.0	0.0	0.5	94.9	0.0	4.6
BBBB	0.0	0.0	0.0	0.0	100.0	0.0

76-100 Average Read Depth

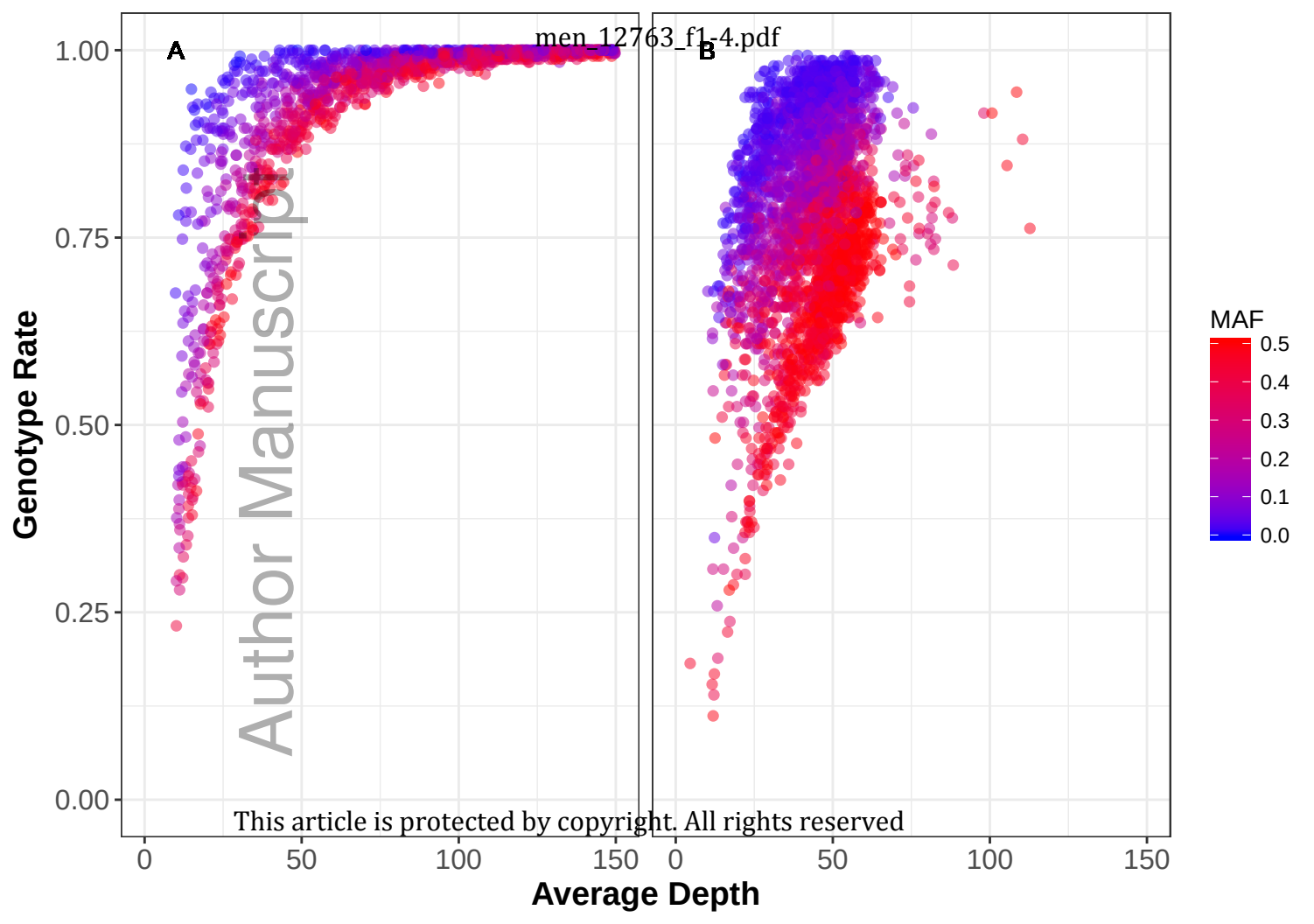
True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	100.0	0.0	0.0	0.0	0.0	0.0
AAAB	0.0	97.7	0.2	0.0	0.0	2.1
AABB	0.0	0.2	95.7	0.3	0.0	3.8
ABBB	0.0	0.0	0.2	98.3	0.0	1.5
BBBB	0.0	0.0	0.0	0.0	100.0	0.0

101-125 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	100.0	0.0	0.0	0.0	0.0	0.0
AAAB	0.0	99.1	0.1	0.0	0.0	0.8
AABB	0.0	0.1	98.3	0.1	0.0	1.5
ABBB	0.0	0.0	0.0	99.4	0.0	0.5
BBBB	0.0	0.0	0.0	0.0	100.0	0.0

126-150 Average Read Depth

True	AAAA	AAAB	AABB	ABBB	BBBB	unassigned
AAAA	100.0	0.0	0.0	0.0	0.0	0.0
AAAB	0.0	99.7	0.1	0.0	0.0	0.3
AABB	0.0	0.0	99.4	0.1	0.0	0.5
ABBB	0.0	0.0	0.0	99.8	0.0	0.2
BBBB	0.0	0.0	0.0	0.0	100.0	0.0


men_12763_f1-4.pdf
A
B
Genotype Rate
1.00
0.75
0.50
0.25
0.00
0
50
100
150
Average Depth
MAF
0.5
0.4
0.3
0.2
0.1
0.0
Author Manuscript
This article is protected by copyright. All rights reserved