MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2017) **17**, 955–965 doi: 10.1111/1755-0998.12647

Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage

STUART C. WILLIS,* CHRISTOPHER M. HOLLENBECK,* JONATHAN B. PURITZ,*,† JOHN R. GOLD* and DAVID S. PORTNOY*

Marine Genomics Laboratory, Department of Life Sciences, Texas A&M University-Corpus Christi, 6300 Ocean Drive, Corpus Christi, TX 78412, USA, †Marine Science Center, Northeastern University, 430 Nahant RD, Nahant MA 01908, USA

Abstract

Next-generation sequencing of reduced-representation genomic libraries provides a powerful methodology for genotyping thousands of single-nucleotide polymorphisms (SNPs) among individuals of nonmodel species. Utilizing genotype data in the absence of a reference genome, however, presents a number of challenges. One major challenge is the trade-off between splitting alleles at a single locus into separate clusters (loci), creating inflated homozygosity, and lumping multiple loci into a single contig (locus), creating artefacts and inflated heterozygosity. This issue has been addressed primarily through the use of similarity cut-offs in sequence clustering. Here, two commonly employed, postclustering filtering methods (read depth and excess heterozygosity) used to identify incorrectly assembled loci are compared with haplotyping, another postclustering filtering approach. Simulated and empirical data sets were used to demonstrate that each of the three methods separately identified incorrectly assembled loci; more optimal results were achieved when the three methods were applied in combination. The results confirmed that including incorrectly assembled loci in population-genetic data sets inflates estimates of heterozygosity and deflates estimates of population divergence. Additionally, at low levels of population divergence, physical linkage between SNPs within a locus created artificial clustering in analyses that assume markers are independent. Haplotyping SNPs within a locus effectively neutralized the physical linkage issue without having to thin data to a single SNP per locus. We introduce a Perl script that haplotypes polymorphisms, using data from single or paired-end reads, and identifies potentially problematic loci.

Keywords: nonmodel species, population genomics, single-nucleotide polymorphisms

Received 16 September 2016; revision received 1 December 2016; accepted 14 December 2016

Introduction

The field of population genetics, empowered by high-throughput DNA sequencing, is rapidly expanding the potential for high-resolution demographic, genomic and evolutionary analyses of nonmodel organisms (Mardis 2008). The technology has not yet reached the point where sequencing the full genome of many samples is cost or labour-efficient, so most studies rely on reduced-representation libraries to provide a manageable number of single-nucleotide polymorphisms (SNPs) to survey across individuals (Altshuler *et al.* 2000). Currently, there are several library-preparation approaches and bioinformatics procedures used to identify and genotype hundreds to thousands of SNPs in a panel of individuals (e.g. Okou *et al.* 2007; Van Tassell *et al.* 2008). One form of library preparation (restriction-site-associated DNA or RAD) takes advantage of the relative frequency of restriction endonuclease sites to tailor the number of fragments sequenced (Puritz *et al.* 2014b). The major challenge for most RAD sequencing projects applied to nonmodel organisms is to assemble a high-quality set of homologous sequences with minimal missing data across the greatest number of individuals, without use of a reference genome (Davey *et al.* 2011). This challenge has been met with many solutions and mixed degrees of success (Puritz *et al.* 2014a,b).

Assembling a RAD data set requires separation of reads into clusters corresponding to a single location on a haploid set of chromosomes (hereafter, single-copy locus). The challenge, therefore, is to identify highly similar sequences that occupy different chromosomal locations (hereafter, multicopy loci). These multicopy loci include paralogs, transposons and other, nonallelic similar sequences (Hohenlohe *et al.* 2011; Peterson *et al.* 2012) that may artificially cluster together during assembly. There are several approaches to detect multicopy loci

Correspondence: Stuart C. Willis; E-mail: swillis4@gmail.com

such as quantitative PCR (e.g. D'haene *et al.* 2010) or phylogenetic analysis of homologous sequences (e.g. Cannon & Young 2003), but none of these are cost-effective for the volume of data typical of a RAD population genetics data set. The problem is especially challenging for taxa with recent whole-genome duplications followed by partial 'diploidization', such as salmonids (Christensen *et al.* 2013).

Identification and elimination of multicopy loci in SNP data sets begins during bioinformatics assembly and filtering. An initial step in clustering reads is to select a cut-off for the number of base differences allowed among reads that are assembled into a contiguous sequence alignment (contig), what will thereafter be considered as corresponding to a single-copy locus (e.g. Catchen *et al.* 2011). A stringent cut-off can be applied at this step to restrict the number of multicopy loci; however, divergent alleles within a single locus may be split into different contigs (oversplitting) and this can inflate observed homozygosity, compromising downstream analyses that depend on unbiased estimates of heterozygosity (Catchen *et al.* 2011; Ilut *et al.* 2014; Harvey *et al.* 2015). Alternatively, a lower sequence similarity threshold can be used to avoid oversplitting and postassembly approaches can be employed to filter the data set and identify potential multicopy loci (Ilut *et al.* 2014).

One postassembly filtering approach is based on the observation that read depths derived from single-locus clusters theoretically form a distribution around a mean read depth (Emerson *et al.* 2010). Contigs with abnormally high read depth often signal the presence of multicopy loci (Emerson *et al.* 2010), meaning that secondary peaks or outliers in a frequency distribution of read depth per contig may indicate suspect alignments and can be used to choose thresholds for single- vs. multicopy loci. A second filtering approach (Hohenlohe *et al.* 2011) relies on the occurrence of fixed or near-fixed differences between nonallelic loci, which causes an excess of heterozygotes above the expected 50% for bi-allelic SNPs. Filters that employ this approach tend to eliminate SNP loci with proportions above this level or that deviate significantly from either Hardy–Weinberg or binomial expectations (Hohenlohe *et al.* 2011; Parchman *et al.* 2012). A third filtering approach, haplotyping, relies on the fact that closely linked SNPs can constitute haplotypes of which a diploid individual can have no more than two (Peterson *et al.* 2012; Ilut *et al.* 2014). Consequently, contigs that contain reads with three or more haplotypes within an individual can be flagged for inspection or removed (Parchman *et al.* 2012; Peterson *et al.* 2012). Unlike a filter for excess heterozygosity, which relies on significant divergence between alleles at multicopy loci, identifying excess haplotypes within individuals only requires that there are two or more variable SNP sites within a contig. The number of individuals exhibiting reads with more than two haplotypes can then be used as a cut-off to eliminate possible multicopy loci. These filters are designed to eliminate multicopy or artefactual contigs from population genomic data sets, and although many researchers may wish to identify true paralogous loci (potential sites of evolutionary innovation) in their data, the loci identified by these filters will often result from a variety of assembly and scoring errors also.

Closely linked SNPs can also pose complications in data analysis when associations due to linkage are treated as a statistical association among loci resulting from consanguinity, selection or population structure (Kaeuffer *et al.* 2007). Over time scales of population-level processes, SNPs within a fragment of a few hundred base pairs in length are expected to exhibit background linkage disequilibrium (LD) and thus should not be considered independent markers (Falush *et al.* 2003; Kaeuffer *et al.* 2007). This presents a dilemma for researchers who wish to glean as much information as possible from their data as the total observed SNPs will be greater than the number of segregating loci. In addition, considering that SNPs contain less information per locus than multi-allelic markers such as microsatellite loci (Morin *et al.* 2009), thinning the data set to one SNP per locus reduces the total information content. Fortunately, the information content of all SNPs in a data set can be preserved and physical linkage artefacts removed by haplotyping SNPs within segregating loci.

Here, we explore the efficacy of using read depth, excess heterozygosity and haplotyping, sequentially, separately and in combination to identify multicopy loci for elimination from a SNP data set. We evaluated filter performance using four simulated RAD data sets containing multicopy loci, generated with a combination of either high or low mutation rate and either simple or complex evolutionary history. We also evaluated an empirical data set generated from a marine fish with low population structure and high genetic diversity. Finally, we examined bias and precision in estimating population-genetic parameters by retaining and considering all SNPs as independent loci, thinning to a single SNP per contig, or haplotyping SNPs within contigs.

Methods

Simulated RAD data

Sequence reads from a double-digest RAD library (i.e. paired reads of fixed-length, allelic sequences) were simulated using the *simrrls* Python script (D. Eaton, Yale),

creating reads of a user-specified library type. The EGGLIB library (De Mita & Siol 2012) was used to specify demographic parameters that affect allelic coalescence and simulate sequences under those conditions. Two large, randomly mating populations that diverged from a common, homogenous population $4N$ generations in the past, followed by bidirectional gene exchange ($m = 0.01$) until $0.1N$ generations in the past (after this, $m = 0$), were simulated, and 1000 loci from 40 individuals (20 per population) were sampled. To introduce multicopy loci (in this case double-copy), another pair of populations, with the same demographic history but which had diverged from the first pair of populations $20N$ generations in the past, followed by zero gene exchange, were simulated. From this second pair of populations, sequences from 50 of the 1000 loci (5%) were sampled and combined with reads from the first pair of populations. The resulting data set contained 950 single-copy and 50 multicopy loci. Simulated sequences consisted of paired 100-base pair (bp) forward and reverse reads, with the number of reads per locus per individual specified with a gamma distribution ($k = 1.6$, $\theta = 20$) with a mean of $k*\theta = 32$, mode of $(k-1)*\theta = 12$ and a 95% probability interval of 2.6–97.2. These simple, multicopy data sets also included sequencing errors and insertion–deletion mutations introduced at default rates ($P = 0.001$ per site). Data were simulated at lower ($N = 35\,000$) and higher ($N = 70\,000$) population sizes, with a constant mutation rate ($\mu = 7 \times 10^{-9}$), thus creating low and high genetic diversity, simple data sets. Simulations for the larger population also included a low but positive rate of recombination within fragments ($\rho = 4Nr = 10$, sites = 100). Complex multicopy data sets also were generated to explore the performance of filtering for older, more divergent multicopy loci, which may feature fixed-site or nearly fixed differences. Both sequence data sets (low/high diversity) were duplicated, and for reads from the 50 multicopy loci derived from the second pair of populations, the 5th G of every odd read was changed to an A and the fourth G of every even read was changed to a T. While this procedure did not create fixed differences between locus copies from each population pair, it increased the likelihood of divergent haplotypes over in situ mutation alone.

Empirical data

Empirical data consisted of a reduced-representation, genomic library of red drum, *Sciaenops ocellatus*, created using a modified version of the double-digest, restriction-associated DNA sequencing (ddRAD) protocol of Peterson *et al.* (2012). The data set was composed of 100-bp paired-end reads for 40 individuals sampled from two localities (Lower Laguna Madre and Sabine Lake, TX). These localities, while demographically independent over a single generation, are part of the same western 'regional population' of red drum (Hollenbeck 2016) and could thus be considered to consist of one or two clusters of individuals. Details of library construction can be found in Puritz *et al.* (2014a) and data be obtained from NCBI's Short Read Archive (SRA) under Accession No. SRP041032.

Reference construction, read mapping, variant calling and preliminary filtering

Both simulated and empirical data were processed using the DDOCENT pipeline *v*.2 (Puritz *et al.* 2014a), which facilitates efficient construction of a reference genome (catalog of putatively orthologous sequences), quality trimming of sequence reads, alignment-based mapping of trimmed reads to the reference and calling of polymorphic positions using a probabilistic model and considering a priori sampling units. For both simulated and empirical data, the reference set was created from unique, untrimmed sequences that were present at least twice within individuals (K1 = 2) and at least twice among individuals (K2 = 2), and then clustered at no less than 80% sequence similarity (c = 0.8), from which a consensus sequence was derived. These parameters are expected to bypass the majority of sequencing errors, which are expected to occur in only a single sequence, and provide effective clustering of even divergent alleles within loci, with the possibility of clustering reads from multicopy loci with similar sequences (Ilut *et al.* 2014). Quality-trimmed reads were mapped by alignment to the reference consensus sequences, using mapping parameter values of 1, 3 and 5 for match score, mismatch cost and gap-opening penalty, respectively. Variant calling was performed with FREEBAYES (Garrison & Marth 2012) on BAM files of aligned reads. Polymorphisms (which initially included complex, insertion–deletion, multi-allelic and bi-allelic variants) were filtered for quality and missing data with a combination of VCFTOOLS (Danacek *et al.* 2011) and VCFLIB (Garrison 2014; Boston College) in addition to the filtering below (see Appendix S1, Supporting information).

Multicopy locus elimination by variant filtering and haplotyping

Three approaches for postclustering filtering of multicopy loci (read depth, excess heterozygosity and haplotyping) were investigated using both empirical and simulated data. Full details of filtering routines are

described in Appendix S1 (Supporting information). The first (Scheme 1) was applied to individual SNPs and employed the three filtering approaches sequentially in the order read depth (a), excess heterozygosity (b) and haplotyping (c). In this scheme, each filtering step received only data remaining after a previous filtering step. Schemes 2 and 3 were applied jointly to all the SNPs in a contig rather than to individual SNPs. Scheme 2 employed the three filtering approaches separately (a-c), while Scheme 3 employed the three approaches separately but then combined results from all three. For comparison, a fourth data set (Scheme 4) was generated with no filtering for multicopy loci.

To filter multicopy loci based on read depth (Schemes 1, 2a and 3), SNPs were filtered by mean read depth across individuals, with cut-offs determined empirically for both simulated and empirical data sets (see Results and Discussion). In Scheme 1, only high depth SNPs were removed; in Schemes 2a and 3, entire SNP-containing contigs were removed if any of the constituent SNPs failed to pass the filter. To filter paralogs based on excess heterozygosity (Schemes 1, 2b and 3), the proportion of heterozygotes at each SNP locus was estimated using VCFTOOLS. For SNPs with >50% heterozygotes, a chi-square test was used to assess whether each conformed to expectations of Hardy–Weinberg equilibrium (HWE) and a correction for multiple tests (Benjamini & Hochberg 1995) was applied. In Scheme 1, SNPs significantly in excess of 50% heterozygotes were removed; in Schemes 2b and 3, any contig with one or more SNPs in excess of 50% heterozygotes and not in HWE was removed. To filter multicopy loci based on haplotyping (Schemes 1, 2c and 3), a custom Perl script was employed (Appendix S1, Supporting information). The script identifies multi-SNP genotypes for each individual at each contig, compares this to a catalog of haplotypes (spanning both read pairs) for each individual at each contig and flags homozygotes errantly called heterozygotes, based on genotyping error, and true heterozygotes with more than two haplotypes. In addition, the script discards variants observed in only one or two reads as sequencing errors. The user is able to set a cut-off for the number of genotyping errors and for extra haplotype-containing individuals allowed per contig, and for missing data. In this study, cut-offs were set such that if one or more individuals had >2 haplotypes at a contig, that contig was removed.

For simulated data, the number of multicopy loci that were eliminated at each step and in each filtering scheme was recorded (Table 1). For empirical data, where the true number of multicopy loci was unknown, the total number of contigs eliminated with each filter was recorded (Table 2).

Population statistics and effects of physical linkage

To examine possible effects of filtering multicopy loci and physical linkage on estimates of population-genetic parameters, the empirical data set was filtered using Schemes 1, 3, and 4. For Schemes 1 and 3, the haplotyping filter was run on data with no minor allele frequency (MAF) cut-off because rare alleles, while not necessarily desirable for many population-genetic analyses, are quite useful in identifying excess haplotypes at a locus within individuals. After initial haplotype filtering, SNPs were filtered using a MAF cut-off where the least common allele had to be observed at least twice in a given data set (MAF ≥2/2N alleles), and then the data were rehaplotyped (without further filtering). For Schemes 1 and 3, filtered data sets were thinned to a single SNP per contig (the first SNP, by default) for comparison to data sets containing all filtered SNPs (unthinned) and haplotypes (Table 3). For Scheme 4, only thinned and unthinned data sets were compared.

Two simulated, simple data sets, one of low and one of high genetic diversity, were generated for comparison with the empirical data set. For both of these simulated data sets, SNP loci were filtered for ≤95% missing data for consistency with the empirical data set and then filtered using a MAF of ≥2/2N. Analyses for each data set were run with and without simulated multicopy loci (removed manually), and thinned data sets were compared to unthinned data sets. After filtering with greater stringency for missing data (50% vs. 95%; Appendix S1, Supporting information), these data sets consisted of ~5–10% of the original 1000 contigs. Additionally, for data sets where multicopy loci had been removed, data were haplotyped for comparison to thinned and unthinned data sets as above (Table 4).

GENODIVE (Meirmans & Van Tienderen 2004) was used to generate estimates of the effective number of alleles (A_E) and the inbreeding coefficient (G_{IS}) for each of the three data sets (one empirical, two simulated) and an estimate of unbiased population divergence (G_{ST}'') between pairs of samples within data sets. G_{ST}'' is a measure of divergence, calibrated to the maximum possible divergence given the number of alleles at a locus, and consequently permits a direct comparison between bi-allelic loci (i.e. SNPs) and multi-allelic loci (haplotyped contigs) (Hedrick 2005; Meirmans & Hedrick 2011). Confidence intervals for G_{IS} and G_{ST}'' were generated using 10 000 bootstrap replicates across loci. Population assignment probability to two clusters (K = 2) was calculated using the program STRUCTURE, with the admixture model and correlated allele frequencies (Pritchard *et al.* 2000). No a priori population

Table 1 Results of filtering of simulated ddRAD data sets

Data diversity	Multicopy haplotypes	Total contigs reconstructed	#Multicopy contigs	#SNPs	Filtering scheme	Filter by depth	Filter by H_o	Filter by #haplotypes	Multicopy loci left	Multicopy loci >2 haps
Low	Simple	1000	50	3641	1	30/50 (60%)	0/15 (0%)	2/15 (13%)	13	5
					2a	47/50 (94%)	–	–	3	–
					2b	–	49/50 (98%)	–	1	–
					2c	–	–	37/49 (76%)	12	2
					3			Combined filters:	0	
Low	Complex	1000	50	3714	1	28/50 (56%)	3/18 (17%)	1/15 (7%)	14	5
					2a	49/50 (98%)	–	–	1	–
					2b	–	50/50 (100%)	–	0	–
					2c	–	–	46/50 (92%)	4	1
					3			Combined filters:	0	
High	Simple	1000	50	7097	1	17/50 (34%)	0/32 (0%)	4/32 (13%)	28	16
					2a	42/50 (84%)	–	–	8	–
					2b	–	40/50 (80%)	–	10	–
					2c	–	–	35/50 (70%)	15	14
					3			Combined filters:	7	
High	Complex	1002*	52 (47)*	7187	1	16/52 (31%)	5/36 (14%)	7/31 (23%)	24	17
					2a	42/52 (81%)	–	–	10	–
					2b	–	47/52 (90%)	–	5	–
					2c	–	–	44/52 (85%)	8	6
					3			Combined filters:	2	

For each simulated condition (low/high diversity, simple/complex), contigs were filtered sequentially by depth, observed heterozygosity (H_O) and haplotyping (Scheme 1), filtered separately by depth, heterozygosity or haplotyping (Schemes 2a–c) or filtered in combination (Scheme 3). Values recorded in each filtering step are number of simulated, multicopy loci filtered divided by the total simulated, multicopy loci available. The number of multicopy loci available to filter at each step may not necessarily match the number remaining in a previous step because some number of multicopy loci were eliminated in intermediate filtering steps not directed towards multicopy loci. The third through fifth columns list the total number of contigs reconstructed by the DDOCENT pipeline, the number of multicopy loci clusters recovered and the number of SNPs scored across all clusters. The last columns are the number of simulated multicopy loci remaining after filtering and the number of those multicopy loci observed to possess more than two haplotypes. *See text for explanation.

Table 2 Results of filtering of the empirical ddRAD data set

Reference contigs	# Contigs ≥1 SNP	Total SNPs before filtering	Filtering scheme	Filter by depth (2× mode)	Filter by H_o	Filter by # haplotypes	Remaining contigs (≤5%)	Remaining SNPs (≤5%)
40 329	31 758	124 500	1	3727	30	1553	5677	13 280
			2a	4274	–	–	6826	20 182
			2b	–	353	–	10 621	32 160
			2c	–	–	2554	8332	20 647
			3	Combined filters: 5912			5271	12 664
			4	No paralog filtering			10 886	33 679

The number of reference contigs and contigs containing variants (≥1 SNP) from the DDOCENT pipeline, as well as the total SNPs before filtering, are shown. Rows list the number of contigs that were filtered sequentially by depth, observed heterozygosity (H_O) and haplotyping (Scheme 1), filtered separately by depth, heterozygosity, or haplotyping (Scheme 2a–c) or filtered in combination (Scheme 3). The number of contigs and SNPs retained with basic, but no multicopy loci specific filtering also are shown (Scheme 4). For each scheme, the final remaining number of contigs and SNPs with ≤5% missing data is listed.

Table 3 Data set characteristics and population statistics for red drum from Lower Laguna Madre and Sabine Lake, TX, USA

Multicopy filtering	#Contigs	Markers	# SNPs	# Alleles (A_E)	G_{IS} (95% CI)	G_{ST}'' (95% CI)
1. Sequential	4932	All SNPs	9964	19 928 (1.31)	−0.0103 (−0.0145: −0.0062)	0.0032 (0.0019:0.0045)
		Haplotypes	9964	14 691 (1.61)	−0.0108 (−0.0155: −0.0060)	0.0032 (0.0015:0.0049)
		Thin SNPs	4932	9864 (1.31)	−0.0102 (−0.0162: −0.0043)	0.0037 (0.0018:0.0057)
2. Combined	4590	All SNPs	9476	18 952 (1.30)	−0.0094 (−0.0136: −0.0052)	0.0030 (0.0017:0.0044)
		Haplotypes	9476	13 868 (1.62)	−0.0096 (−0.0142: −0.0049)	0.0029 (0.0011:0.0047)
		Thin SNPs	4590	9180 (1.31)	−0.0085 (−0.0145: −0.0025)	0.0034 (0.0014:0.0055)
3. None	9870	All SNPs	26 787	53 574 (1.36)	−0.0719 (−0.0764: −0.0675)	0.0027 (0.0020:0.0035)
		Thin SNPs	9870	19 740 (1.34)	−0.0441 (−0.0505: −0.0377)	0.0027 (0.0014:0.0039)

Data were filtered for minor allele frequency (MAF >1/2N alleles). Results are shown from three multicopy loci filtering schemes: SNPs filtered by each method sequentially (Scheme 1), all SNPs from contigs identified in combination (Scheme 3) or no multicopy loci filtering (Scheme 4). Number of remaining contigs (#contigs) and SNPs (#SNPS) for each filtering scheme are shown for data sets of all SNPs, haplotypes or thinned SNPs. Listed for each are number of alleles recovered, effective number of alleles (A_E) and estimates and 95% confidence intervals for the inbreeding coefficient (G_{IS}) and for population divergence (G_{ST}'').

Table 4 Data set characteristics and population statistics for simulated data with simple haplotypes

Data diversity	Multicopy loci	# Contigs	Markers	# SNPs	# alleles (A_E)	G_{IS} (95% CI)	G_{ST}'' (95% CI)
Low	No	55	All SNPs	151	302 (1.38)	−0.0142 (−0.0390:0.0107)	0.2107 (0.1626:0.2591)
			Haplotypes	151	167 (1.68)	−0.0067 (−0.0422:0.0271)	0.2677 (0.1972:0.3405)
			Thin SNPs	55	110 (1.35)	0.0039 (−0.0428:0.0515)	0.2656 (0.1729:0.3559)
Low	Yes	99	All SNPs	474	948 (1.69)	−0.4592 (−0.4874: −0.4300)	0.0782 (0.0595:0.0979)
			Thin SNPs	99	181 (1.58)	−0.3641 (−0.4361: −0.2891)	0.1426 (0.0871:0.2037)
High	No	80	All SNPs	359	718 (1.32)	0.0205 (−0.0014:0.0421)	0.2328 (0.2034:0.2617)
			Haplotypes	359	378 (2.16)	0.0099 (−0.0170:0.0383)	0.3272 (0.2736:0.3804)
			Thin SNPs	80	160 (1.34)	0.0105 (−0.0303:0.0509)	0.2148 (0.1652:0.2653)
High	Yes	123	All SNPs	753	1506 (1.59)	−0.3520 (−0.3755: −0.3275)	0.1089 (0.0926:0.1256)
			Thin SNPs	123	246 (1.51)	−0.2582 (−0.3237: −0.1908)	0.1360 (0.0978:0.1758)

Data from two simulations (low and high variability) are shown with and without multicopy loci removed from final data sets. Data were filtered for minor allele frequency (MAF > 1/2N alleles). The number of remaining contigs (#contigs) and SNPs (#SNPs) are shown for data sets of all SNPs, haplotypes or thinned SNPs. Listed for each are number of alleles recovered, effective number of alleles (A_E) and estimates and 95% confidence intervals for the inbreeding coefficient (G_{IS}) and for population divergence (G_{ST}'').

membership information was specified; runs consisted of 100 000 samples after 100 000 generations of burn-in. Because there were two simulated populations, and two localities (from a single regional population) from which empirical data were generated, assignment was estimated at K = 2.

Results

Multicopy loci filtering of simulated data

A total of 1000 contigs from the low variability, simple and complex sequences were reconstructed using DDOCENT, as were 1000 contigs from the high variability, simple sequences. In each case, the 50 multicopy loci (contigs with reads from both population pairs) were reconstructed into a single contig each, as expected (Table 1). However, a total of 1002 contigs, including 950 single-copy loci and 47 of the multicopy loci, were reconstructed from the high variability, complex data set. Of the three remaining (expected) multicopy loci, one contig contained only reads from the second population pair (in effect becoming a single-copy locus). The other two expected, multicopy loci were divided into two contigs each (total of four). One was split into two contigs, but each contig contained reads from each population pair, while the other split into two contigs where each contig contained only reads from the second population pair. Hereafter, these five are referred to as anomalous, multicopy loci.

Results of filtering by Schemes 1–3 are shown in Table 1. Overall, filtering by Scheme 3 (combined) was more effective than Scheme 1 (sequentially) and, in most cases, than Scheme 2 (separately). When applied sequentially to individual SNPs (Scheme 1), each filter removed data needed by the subsequent filter to identify multicopy loci, making overall filtering less effective. The three filters applied separately (Scheme 2) were variously effective at eliminating multicopy loci. The most effective filter alone, excess heterozygosity, did achieve 100% success eliminating multicopy loci in the simulation involving the low-diversity, high-complexity data set. When run separately, haplotyping was the least effective filter in terms of removal of multicopy loci. However, haplotyping performed well in both high-complexity data sets and was more effective than depth filtering in the high-diversity, high-complexity data set. This is due to the fact that multicopy loci in high-complexity data sets exhibited more divergent haplotypes, increasing the chance of recognizing extra haplotypes within individuals. Haplotyping also identified multicopy loci not identified by the other two filters applied under Scheme 3, including all five of the anomalous, multicopy loci from the high-diversity, high-complexity data set.

Filtering in general was least effective in high-diversity data sets. This resulted from less effective mapping of higher variability reads onto contigs, thus reducing clarity of patterns needed to identify multicopy loci. For example, mean depth for SNPs from multicopy loci was 48.2 (range: 13.5–69.0) and 47.3 (10.1–69.0) for the simple and complex, high-diversity data sets, respectively, versus 53.8 (18.3–72.6) and 53.2 (10.9–72.6) for the simple and complex, low-diversity data sets, respectively. No substantial difference was observed in depth for SNPs from single-copy loci (means: 28.4, 28.5, 28.3, 28.4). This pattern can be better understood by inspecting frequency distributions of mean depth across loci (Fig. 1a). SNPs from multicopy loci are shifted to the left in high-diversity data sets relative to low-diversity data sets and into depth bins constituting the first mode of the bimodal distribution. Because of this shift, more SNPs from multicopy loci fell below the selected depth cut-off (maximum mean of 45 reads/individual). Similarly, values of and deviations between observed and expected heterozygosity were smaller in high-diversity data sets (0.237/0.241 and 0.244/0.244 mean observed/expected heterozygosity in simple and complex data sets, respectively) than low-diversity data sets (0.272/0.255 and 0.287/0.262 mean observed/expected heterozygosity in simple and complex data sets, respectively). Consequently, fewer loci exhibited excess heterozygosity when tested for deviations from HWE. Finally, a higher proportion of multicopy loci with >2 haplotypes failed to be mapped within a single individual in high-diversity data sets, resulting in decreased efficiency of the haplotyping filter (Table 1). More permissive mapping parameters were not explored here, but it is possible that for data sets from populations with high genetic diversity (i.e. with a wide and overlapping range of sequence divergence between and within multicopy and single-copy loci, respectively), less stringent initial mapping values would render these filters more effective.

Multicopy loci filtering of empirical data

Reference construction for the 40 red drum individuals resulted in 40 329 contigs (Table 2). A total of 124 500 variants were scored from reads mapped to these reference sequences, but only 79% of contigs contained variants. The average number of variants per variable contig was 3.7, which made these data similar to the simulated, low-diversity data sets (4.1 variants/contig) rather than simulated, high-diversity data sets (7.2 variants/contig). While the actual number of multicopy loci in the empirical data set was unknown, it likely is comparable to other nonpolyploid, bony fishes (e.g. <5% in stickleback, Ilut *et al.* 2014), and some results are still salient without this context. For example, the distribution of read depth was unimodal and highly skewed (Fig. 1b), with some contigs exhibiting obvious depth excesses (e.g. mean 4918 reads/individual, versus an overall mode of 20). Many of these contigs with extreme excess depth BLAST to known multicopy loci such as ribosomal RNA genes. However, the observation of a single mode made it difficult to choose an effective read-depth threshold for

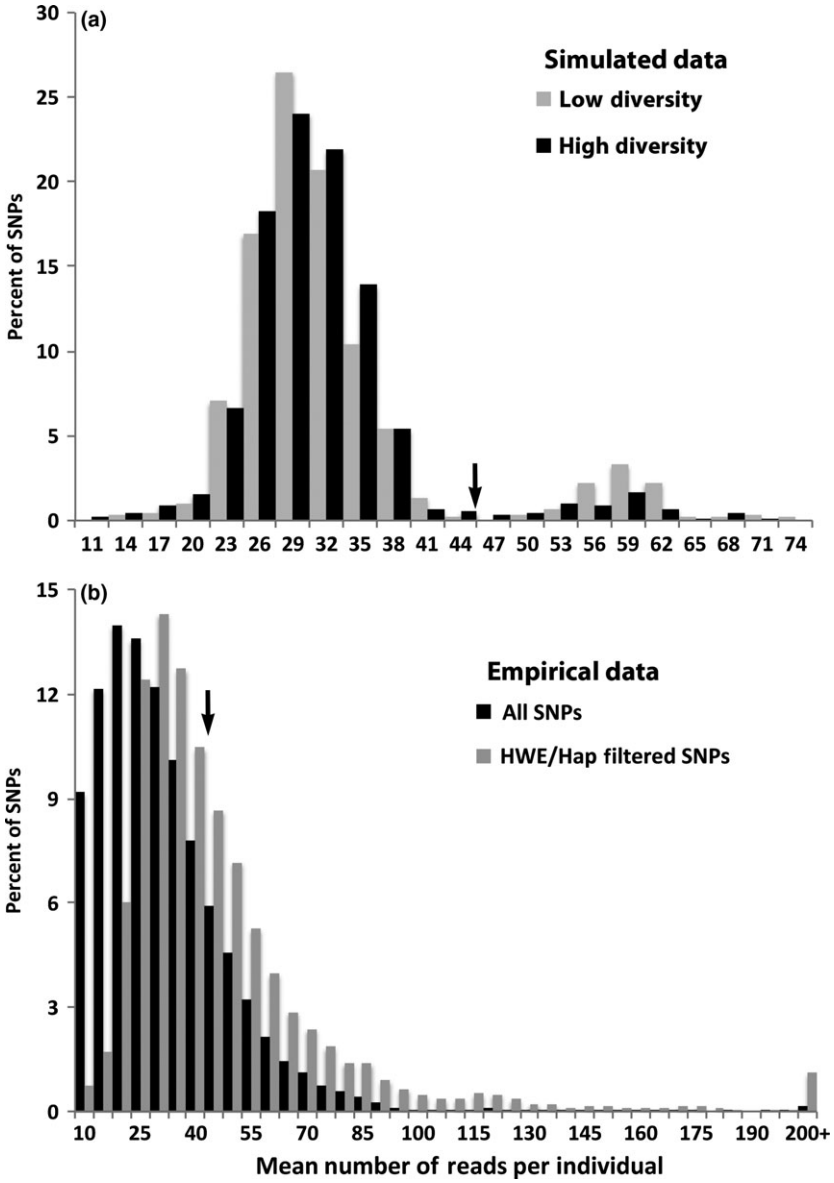

Fig. 1 Frequency distribution of mean number of reads per locus (depth/coverage): (a) simulated ddRAD data with 'simple' haplotypes; and (b) empirical ddRAD data from red drum. Arrows in each figure indicate the chosen read-depth cut-off above which contigs are flagged as multicopy loci.

discriminating multicopy loci. Working from the assumption that the majority of loci were single-copy, and that the observed peak corresponds to the mean depth for these loci, several cut-offs meant to approximate an upper confidence limit associated with the mode were examined: 2× the mode, the mode plus the difference between the mode and the minimum mean depth (mode + mode-min), and the 3rd quartile. The first (2× the mode) proved to be the least stringent for this data set (read depth 40, approximately the 80th percentile) and was chosen as the experimental cut-off to potentially allow more multicopy loci to remain in the data prior to excess heterozygosity and haplotype-based filtering. As with the simulated data, these filters removed fewer contigs than the depth filter, especially when applied sequentially and not strictly across entire contigs (Table 2); when applied in a combined manner, the heterozygosity and haplotype filters removed an additional 1555 (of 5912 total) contigs not flagged by the depth filter. Subsequently, the frequency distribution of depth for SNPs flagged by either excess heterozygosity or haplotyping was compared to the unfiltered distribution in an attempt to estimate an effective cut-off for read depth. While the depth distribution of flagged loci is shifted to the right as compared to the distribution of all loci, and most loci with high depth are flagged by excess heterozygosity and haplotyping filters (Fig. 1b), 58.3% of SNPs were below the selected experimental cut-off (40). One strategy would be to remove only contigs flagged by multiple filters, with the caveat that some multicopy

loci will remain (Table 1). The advantage of this strategy, however, depends on the effect of retaining multicopy loci on downstream analyses.

Linkage, haplotypes and population parameters

For the empirical data set, there was no clear difference among estimated population-genetic parameters based on all SNPs, haplotypes or thinned SNPs, despite haplotypes having a higher effective number of alleles (greater heterozygosity) per locus than SNPs (Table 3). Sequential versus combined filtering schemes also had little effect on estimated values. Estimates of inbreeding (G_{IS}) were negative and of similar magnitude with overlapping confidence intervals, reflecting high genetic diversity and effective population size in red drum (Gold *et al.* 2001; Turner *et al.* 2002). Estimates of population divergence (G_{ST}'') were similarly small, but confidence intervals did not include zero.

There were larger differences among population statistics estimated from all SNPs, haplotypes and thinned SNPs for simulated data sets, which had multicopy loci removed (Table 4). Population divergence estimated from haplotypes was larger than that from all or thinned SNPs. This may reflect increased power to resolve divergence with haplotypes or a sensitivity of G_{ST}'' to the number of alleles or heterozygosity (Kalinowski 2002; Meirmans & Hedrick 2011). G_{IS} values, alternatively, while different, had wide and overlapping confidence intervals, suggesting difficulty in accurately calculating a precise genomewide estimate for this parameter based on so few loci.

Another pattern appeared when assignment probabilities from STRUCTURE using all SNPS, haplotypes and thinned SNPs in the empirical data set were compared. While the mean level of assignment of samples into one of two clusters was small, reflecting low levels of population divergence, the variance in probability of individual assignment was much greater for the data set of all SNPs than for haplotyped or thinned SNPs (Fig. 2). This does not appear to result from the data set of all SNPs being more informative, as the thinned and all-SNPs data sets had similar G_{ST}'' values (0.0014 ± 0.0499 vs. 0.0012 ± 0.0484, mean $\pm$ standard deviation of thinned vs. all SNPs, respectively). Rather, when the analysis was run with SNPs in tight physical linkage, artificial clusters were formed on the mistaken interpretation that LD was the result of population structure. In contrast, the simulated, low-diversity data sets did not show this pattern. Instead, individuals were assigned back to their correct group with considerably higher posterior probability (mean: >0.97). This reflects the higher degree of population divergence in simulated data sets than in the empirical data set and suggests a greater opportunity for artefacts when the level of population divergence is small.

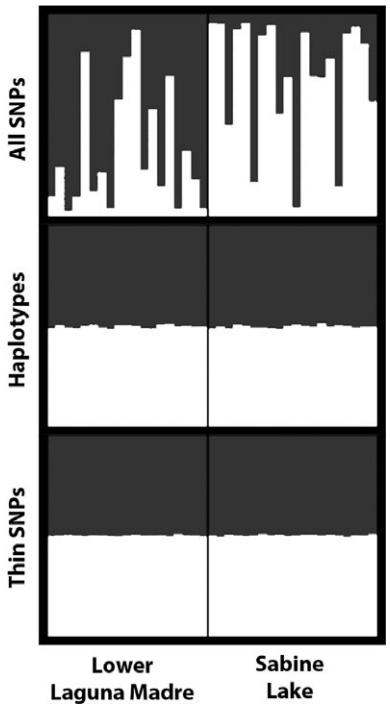

Fig. 2 Bar plots of posterior probability of individual assignment for 39 red drum to K = 2 clusters, using the program STRUCTURE for three versions of the ddRAD data set.

Discussion

Haplotyping SNPs within a contig provides a method to remove additional multicopy loci or otherwise artefact-prone contigs from RAD data sets when used in combination with depth and excess heterozygosity filters. Both simulated and empirical data sets filtered with all three methods exhibited less heterozygosity than unfiltered data sets, and without the added burden of splitting single-copy loci resulting from using high similarity cut-offs for clustering sequences into contigs. When robust filtering, like that demonstrated here, is not applied to RAD data sets without a full reference genome, multicopy loci (i.e. paralogs, transposons and other, nonallelic similar sequences) will often be retained in the final data set and this can lead to biased results in population-genetic analyses. For example, there was higher heterozygosity (lower G_{IS} values) in data sets with no filtering of multicopy loci as compared to those where multicopy loci had been filtered (Table 3) or manually removed (Table 4); this is likely due to SNPs segregating independently in separate copies of multicopy loci but being clustered into

a single contig. This artefactual heterozygosity deflated measures of overall population divergence (G_{ST}''), although not substantially in the empirical data sets. This finding may reflect a higher proportion of multicopy loci in simulated data relative to the empirical data, suggesting that artificially reduced heterozygosity is less of a problem for data derived from genomes with fewer multicopy loci. However, the percentage of multicopy loci falling below a given similarity cut-off, and therefore likely to be assembled incorrectly, will generally be difficult to predict a priori for nonmodel species.

Nevertheless, the consequences of downward biases in estimates of inbreeding and population divergence caused by retaining multicopy loci are not easy to predict, and depend on the intended purpose of the data. In situations of very low but nonzero population divergence, an increase in total heterozygosity could conceivably mask divergence and would provide biased estimates of gene flow and dispersal. For analyses that depend on unbiased and accurate estimates of heterozygosity or allele frequency spectra, the retention of paralogous loci may be more serious. For example, analyses such as genome scans depend on accurate estimates of neutral population divergence to identify outliers. Artificial downward bias in estimates of global levels of divergence might lead to more false positives for loci under directional selection, while multicopy loci might be identified as being under balancing selection (Foll & Gaggiotti 2008). This prediction should be true regardless of the bioinformatic pipeline used to produce the final marker data set, although pipelines that reconstruct fewer multicopy loci and less often oversplit alleles would naturally produce superior results in downstream analyses.

The results indicated that haplotyping is also a straightforward way to manage closely linked SNPs within a contig without loss of information content caused by thinning. Ignoring linkage can produce misleading results in analyses that assume observed LD is a result of demographic or evolutionary processes. This issue is potentially problematic for data sets that feature high diversity within and among populations and low divergence between populations, as was manifest in the clustering results from STRUCTURE. These results suggest that caution is warranted when using linked SNPs from populations with low expected genomic divergence to estimate assignment probabilities.

Finally, while it seems intuitive that haplotyped data sets retain more information than thinned SNP data sets, population statistics in this study from filtered data sets were quite similar between thinned SNP and haplotype data sets. In this case, this may reflect that the sheer number of SNPs recovered overcame any loss of signal associated with thinning (Kalinowski 2002; Willing *et al.* 2012). However, analyses that rely on locus-by-locus measures of divergence or linkage disequilibrium such as genetic mapping (e.g. Ball *et al.* 2010), estimates of identity, parentage or kinship (e.g. Lopéz Herráez *et al.* 2005), and LD-based estimates of effective population size (e.g. Waples & Do 2010) will find added benefit to haplotyping SNPs rather than thinning to a single SNP per contig because of the increased discriminatory power of additional alleles per locus.

Acknowledgements

The authors thank members of the Marine Genomics Laboratory at Texas A&M University-Corpus Christi for fruitful discussions regarding molecular markers and RAD library preparation. Work was supported by an institutional grant (NA10OAR4170099) to the Texas Sea Grant College Program from the National Sea Grant Office, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, a Marine Fisheries Initiative Grant (NA12NMF4540082), National Marine Fisheries Service, U.S. Department of Commerce, a grant (#802) from the Texas Parks and Wildlife Department and by funding from the College of Science and Engineering and the Office of Research and Commercialization at Texas A&M University-Corpus Christi. This article is publication number 16 of the Marine Genomics Laboratory at Texas A&M University-Corpus Christi, and number 110 in the series Genetic Studies in Marine Fishes.

References

Altshuler DL, Pollara VJ, Cowles CR *et al.* (2000) A SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Ball AD, Stapley J, Dawson DA *et al.* (2010) A comparison of SNPs and microsatellites as linkage mapping markers: lessons from the zebra finch (*Taeniopygia guttata*). *BMC Genomics*, **11**, 218. doi: 10.1186/1471-2164-11-218.

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**, 289–300.

Cannon SB, Young ND (2003) ORTHOPARAMAP: distinguishing orthologs from paralogs by integrating comparative genome data and gene phylogenies. *BMC Bioinformatics*, **4**, 35.

Catchen JM, Amores A, Hohenlohe PA, Cresko WA, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1**, 3171–3182.

Christensen KA, Brunelli JP, Lamberg MJ *et al.* (2013) Identification of single nucleotide polymorphisms from the transcriptome of an organism with a whole genome duplication. *BMC Bioinformatics*, **14**, 325. doi: 10.1186/1471-2105-14-325.

Danacek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFTOOLS. *Bioinformatics*, **27**, 2156–2158.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Genetics*, **12**, 499–510.

De Mita S, Siol M (2012) EGGLIB: processing, analysis and simulation tools for population genetics and genomics. *BMC Genetics*, **13**, 27. doi: 10.1186/1471-2156-13-27.

D'haene B, Vandesompele J, Hellemans J (2010) Accurate and objective copy number profiling using real-time quantitative PCR. *Methods*, **50**, 262–270.

Emerson KJ, Merz CR, Catchen JM *et al*. (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences USA*, **107**, 16196–16200.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Garrison E (2014) *A Simple C++ Library for Parsing and Manipulating VCF files, + Many Command-Line Utilities*. Boston College, Boston, Massachusetts.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907.

Gold JR, Burridge CP, Turner TF (2001) A modified stepping-stone model of population structure in red drum, *Sciaenops ocellatus* (Sciaenidae), from the northern Gulf of Mexico. *Genetica*, **111**, 305–317.

Harvey MG, Judy CD, Seeholzer GF, Maley JM, Graves GR, Brumfield RT (2015) Similarity thresholds used in short read assembly reduce the comparability of population histories across species. *PeerJ*, **3**, e1066. doi:10.7287/peerj.preprints.864v1.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–1638.

Hohenlohe PA, Amish SJ, Catchen JM, Allendorf FW, Luikart G (2011) Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout. *Molecular Ecology Resources*, **11**, 117–122.

Hollenbeck CM (2016) *Genomic Studies of Red Drum (Sciaenops ocellatus) in US Waters*. Dissertation, Texas A&M University, College Station, Texas.

Ilut DC, Nydam ML, Hare MP (2014) Defining loci in restriction-based reduced representation genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering. *BioMed Research International*, **2014**, 1–9.

Kaeuffer R, Réale D, Coltman DW, Pontier D (2007) Detecting population structure using STRUCTURE software: effect of background linkage disequilibrium. *Heredity*, **99**, 374–380.

Kalinowski ST (2002) How many alleles per locus should be used to estimate genetic distances? *Heredity*, **88**, 62–65.

Lopéz Herráez D, Schäfer H, Mosner J, Fries HR, Wink M (2005) Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population. *Verlag der Zeitschrift für Naturforschung*, **60**, 637–643.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, **9**, 387–402.

Meirmans PG, Hedrick PW (2011) Measuring differentiation: G_{ST} and related statistics. *Molecular Ecology Resources*, **11**, 5–18.

Meirmans PG, Van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792–794.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66–73.

Okou DT, Steinberg KM, Middle C *et al*. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nature Methods*, **4**, 907–909.

Parchman TL, Gompert Z, Mudge J *et al*. (2012) Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, **21**, 2991–3005.

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**, e37135. doi:10.1371/journal.pone.0037135.

Pritchard JK, Stephens M, Donnelly PJ (2000) Inference of populations structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Puritz JB, Hollenbeck CM, Gold JR (2014) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, **2**, e431. doi:10.7717/peerj.431.

Puritz JB, Matz MV, Toonen RJ *et al*. (2014b) Demystifying the RAD fad. *Molecular Ecology*, **23**, 5937–5942.

Turner TF, Wares JP, Gold JR (2002) Genetic effective size is three orders of magnitude smaller than adult census size in an abundant, estuarine-dependent marine Fish (*Sciaenops ocellatus*). *Genetics*, **162**, 1329–1339.

Van Tassell CP, Smith TP, Matukumalli LK *et al*. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, **5**, 247–252.

Waples RS, Do C (2010) Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, **3**, 244–262.

Willing E-M, Dreyer C, van Oosterhout C (2012) Estimates of genetic differentiation measured by FST do not necessarily require large sample sizes when using many SNP markers. *PLoS ONE*, **7**, e42649. doi:10.1371/journal.pone.0042649.

All authors conceived of the study; S.C.W., C.M.H. and J.B.P. performed the data collection, simulation, and analysis; C.M.H. created the haplotyping script; all authors contributed to writing the manuscript.

Data accessibility

Empirical Illumina sequences data for red drum be obtained from NCBI's Short Read Archive (SRA) under Accession No. SRP041032. Scripts for generating the simulated sequence data as well as some automated filtering have been posted to github (https://github.com/jpuritz/).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1 Supplemental methods