DR. STUART C WILLIS (Orcid ID : 0000-0002-2274-1112)

Received Date : 16-Sep-2016
Revised Date : 01-Dec-2016
Accepted Date : 14-Dec-2016
Article type : Resource Article

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

Authors: Stuart C. Willis[1]*, Christopher M. Hollenbeck[1], Jonathan B. Puritz[1,2], John R. Gold[1], & David S. Portnoy[1]

*To whom correspondence should be addressed: swillis4@gmail.com

Authors' Institutions:
[1]Marine Genomics Laboratory, Department of Life Sciences, Texas A&M University-Corpus Christi, 6300 Ocean Drive, Corpus Christi, Texas 78412, USA.
[2]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 non-model species. Utilizing genotype data in the absence of a reference genome, however, presents a number of challenges. One major challenge is the tradeoff between splitting alleles at a single locus into separate clusters (loci), creating inflated homozygosity, and lumping multiple loci into a single contig (locus), creating artifacts and inflated heterozygosity. This issue has been addressed primarily through the use of similarity cutoffs in sequence clustering. Here, two commonly employed, post -clustering, filtering methods (read depth and excess heterozygosity) used to identify incorrectly assembled loci are compared with haplotyping, another post-filtering clustering 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 datasets 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: population genomics, non-model species, single nucleotide polymorphisms

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 non-model organisms (Mardis 2008). The technology has not yet reached the point where sequencing the full genome of many samples is cost or labor-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 non-model 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; Puritz *et al.* 2014b).

Assembling a RAD dataset 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, multi-copy loci). These multi-copy loci include paralogs, transposons, and other, non-allelic similar sequences (Hohenlohe *et al.* 2011; Peterson *et al.* 2012) that may artificially cluster together during assembly. There are several approaches to detect multi-copy loci 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 dataset. 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 multi-copy loci in SNP datasets begins during bioinformatics assembly and filtering. An initial step in clustering reads is to select a cutoff 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 cutoff can be applied at this step to restrict the number of multi-copy loci; however, divergent alleles within a single locus may be split into different contigs (over-splitting) and this can inflate observed homozygosity, compromising downstream analyses that depend on unbiased estimates of heterozygosity (Catchen *et al.* 2011; Harvey *et al.* 2015; Ilut *et al.* 2014). Alternatively, a lower sequence similarity threshold can be used to avoid over-splitting and post-assembly approaches can be employed to filter the dataset and identify potential multi-copy loci (Ilut *et al*. 2014).

One post-assembly 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 multi-copy 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. multi-copy loci. A second filtering approach (Hohenlohe *et al.* 2011) relies on the occurrence of fixed or near-fixed differences between non-allelic 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 (Ilut *et al.* 2014; Peterson *et al.* 2012). 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 multi-copy 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 multi-copy loci. These filters are designed to eliminate multi-copy or artifactual contigs from population genomic datasets, and though 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 dataset to one SNP per locus reduces the total information content. Fortunately, the information content of all SNPs in a dataset can be preserved and physical linkage artifacts 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 multi-copy loci for elimination from a SNP dataset. We evaluated filter performance by using four simulated RAD datasets containing multi-copy 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 bi-directional gene exchange (m = 0.01) until 0.1N generations in the past (after this, m = 0) were simulated, and 1,000 loci from 40 individuals (20 per population) were sampled. To introduce multi-copy 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 dataset contained 950 single-copy and 50 multi-copy 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, θ = 20) with a mean of $k*\theta$ = 32, mode of (k-1)*θ = 12, and a 95% probability interval of 2.6 - 97.2. These simple, multi-copy datasets 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 (μ = 7x10^{-9}), thus creating low and high genetic diversity, simple datasets. Simulations for the larger population also included a low but positive rate of recombination within fragments (ρ = 4Nr= 10, sites = 100). Complex multi-copy datasets also were generated to explore the performance of filtering for older, more divergent multi-copy loci which may feature fixed-site or nearly-fixed differences. Both sequence datasets (low/high diversity) were duplicated, and for reads from the 50 multi-copy 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, Texas). 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 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 by 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 multi-copy 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 (E. Garrison Boston College) in addition to the filtering below (see Supplemental Information).

Multi-copy locus elimination by variant filtering and haplotyping

Three approaches for post-clustering, filtering of multi-copy loci (read depth, excess heterozygosity, and haplotyping) were investigated using both empirical and simulated data. Full details of filtering routines are described in Supplemental 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 dataset (Scheme 4) was generated with no filtering for multi-copy loci.

To filter multi-copy loci based on read depth (Schemes 1, 2a, and 3), SNPs were filtered by mean read depth across individuals, with cutoffs determined empirically for both simulated and empirical datasets (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 χ^2 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 multi-copy loci based on haplotyping (Schemes 1, 2c, and 3), a custom Perl script was employed (Supplemental 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 multi-copy loci that were eliminated at each step and in each filtering scheme were recorded (Table 1). For empirical data, where the true number of multi-copy 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 multi-copy loci and physical linkage on estimates of population-genetic parameters, the empirical dataset 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 dataset (MAF ≥2/2N alleles), and then the data were re-haplotyped (without further filtering). For schemes 1 and 3, filtered datasets 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 datasets, one of low and one of high genetic diversity, were generated for comparison with the empirical dataset. For both of these simulated datasets, SNP loci were filtered for ≤ 95% missing data for consistency with the empirical dataset and then filtered using a MAF of ≥2/2N. Analyses for each dataset were run with and without simulated multi-copy loci (removed manually), and thinned datasets were compared to unthinned datasets. After filtering with greater stringency for missing data (50% vs. 95%; Supplemental Information), these datasets consisted of ~5-10% of the original 1,000 contigs. Additionally, for datasets where multi-copy 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 datasets (one empirical, two simulated) and an estimate of unbiased population divergence (G''_{ST}) between pairs of samples within datasets. 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) were calculated using the program STRUCTURE, with the admixture model and correlated allele frequencies (Pritchard *et al.* 2000). No *a priori* population 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

Multi-copy loci filtering of simulated data

A total of 1,000 contigs from the low variability, simple and complex sequences were reconstructed using *dDocent*, as were 1,000 contigs from the high variability, simple sequences. In each case, the 50 multi-copy 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 multi-copy loci, were reconstructed from the high variability, complex dataset. Of the three remaining (expected) multi-copy loci, one contig contained only reads from the second population pair (in effect becoming a single-copy locus). The other two expected, multi-copy 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, multi-copy 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 multi-copy loci, making overall filtering less effective. The three filters applied separately (Scheme 2) were variously effective at eliminating multi-copy loci. The most effective filter alone, excess heterozygosity, did achieve 100% success eliminating multi-copy loci in the simulation involving the low-diversity, high-complexity dataset. When run separately, haplotyping was the least effective filter in terms of removal of multi-copy loci. However, haplotyping performed well in both high-complexity datasets and was more effective than depth filtering in the high-diversity, high-complexity dataset. This is due to the fact that multi-copy loci in high complexity datasets exhibited more divergent haplotypes, increasing the chance of recognizing extra haplotypes within individuals. Haplotyping also identified multi-copy loci not identified by the other two filters applied under Scheme 3, including all five of the anomalous, multi-copy loci from the high diversity, high complexity dataset.

Filtering in general was least effective in high-diversity datasets. This resulted from less effective mapping of higher variability reads onto contigs, thus reducing clarity of patterns needed to identify multi-copy loci. For example, mean depth for SNPs from multi-copy loci was 48.2 (range 13.5-69.0) and 47.3 (10.1-69.0) for the simple and complex, high-diversity datasets, respectively, versus 53.8 (18.3-72.6) and 53.2 (10.9-72.6) for the simple and complex, low-diversity datasets, 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 (Figure 1a). SNPs from multi-copy loci are shifted to the left in high-diversity datasets relative to low-diversity datasets and into depth bins constituting the first mode of the bi-modal distribution. Because of this shift more SNPs from multi-copy loci fell below the selected depth cutoff (maximum mean of 45 reads/individual). Similarly, values of and deviations between observed and expected heterozygosity were smaller in high-diversity datasets (0.237/0.241 and 0.244/0.244 mean observed/expected heterozygosity in simple and complex datasets, respectively) than low-diversity datasets (0.272/0.255 and 0.287/0.262 mean observed/expected heterozygosity in simple and complex datasets, respectively). Consequently, fewer loci exhibited excess heterozygosity when tested for deviations from HWE. Finally, a higher proportion of multi-copy loci with >2 haplotypes failed to be mapped within a single individual in high-diversity datasets, resulting in decreased efficiency of the haplotyping filter (Table 1). More permissive mapping parameters were not explored here, but it is possible that for datasets from populations with high genetic diversity (i.e., with a wide and overlapping range of sequence divergence between and within multi-copy and single-copy loci, respectively), less stringent initial mapping values would render these filters more effective.

Multi-copy 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 datasets (4.1 variants/contig) rather than simulated, high-diversity datasets (7.2 variants/contig). While the actual number of multi-copy loci in the empirical dataset was unknown, it likely is comparable to other non-polyploid, 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 (Figure 1b), with some contigs exhibiting obvious depth excesses (e.g., mean 4,918 reads/individual, versus an overall mode of 20). These contigs BLAST to known multi-copy loci such as ribosomal RNA genes. However, the observation of a single mode made it difficult to choose an effective read-depth threshold for discriminating multi-copy 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 cutoffs meant to approximate an upper confidence limit associated with the mode were examined: 2X the mode, the mode plus the difference between the mode and the minimum mean depth (mode+mode-min), and the 3rd quartile. The first (2X the mode) proved to be the least stringent for this dataset (read depth 40, approximately the 80th percentile) and was chosen as the experimental cutoff to potentially allow more multi-copy 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 1,555 (of 5,912 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 cutoff 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 (Figure 1b), 58.3% of SNPs the were below the selected experimental cutoff (40). One strategy would be to remove only contigs flagged by multiple filters, with the caveat that some multi-copy loci will remain (Table 1). The advantage of this strategy, however, depends on the effect of retaining multi-copy loci on downstream analyses.

Linkage, haplotypes, and population parameters

For the empirical dataset 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 datasets which had multi-copy 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 genome-wide 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 dataset 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 dataset of all SNPs than for haplotyped or thinned SNPs (Figure 2). This does not appear to result from the dataset of all SNPs being more informative, as the thinned and all-SNPs datasets had similar $\mathrm{G}''_{\mathrm{ST}}$ values (0.0014 ±0.0499 vs. 0.0012 ±0.0484, mean ± 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 datasets 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 datasets than in the empirical dataset, and suggests a greater opportunity for artifacts when the level of population divergence is small.

Discussion

Haplotyping SNPs within a contig provides a method to remove additional multi-copy loci or otherwise artifact-prone contigs from RAD datasets when used in combination with depth and excess heterozygosity filters. Both simulated and empirical datasets filtered with all three methods exhibited less heterozygosity than unfiltered datasets, and without the added burden of splitting single-copy loci resulting from using high similarity cutoffs for clustering sequences into contigs. When robust filtering, like that demonstrated here, is not applied to RAD datasets without a full reference genome, multi-copy loci (i.e. paralogs, transposons, and other, non-allelic similar sequences) will often be retained in the final dataset and this can lead to biased results in population genetic analyses. For example, there was higher heterozygosity (lower G_{IS} values) in datasets with no filtering of multi-copy loci as compared to those where multi-copy loci had been filtered (Table 3) or manually removed (Table 4); this is likely due to SNPs segregating independently in separate copies of multi-copy loci but being clustered into a single contig. This artifactual heterozygosity deflated measures of overall population divergence (G''_{ST}), although not substantially in the empirical datasets. This finding may reflect a higher proportion of multi-copy 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 multi-copy loci. However, the percentage of multi-copy loci falling below a given similarity cutoff, and therefore likely to be assembled incorrectly, will generally be difficult to predict *a priori* for non-model species.

Nevertheless, the consequences of downward biases in estimates of inbreeding and population divergence caused by retaining multi-copy loci are not easy to predict, and depend on the intended purpose of the data. In situations of very low but non-zero 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 multi-copy 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 dataset, although pipelines that reconstruct fewer multi-copy loci and less often over-split 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 datasets 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 datasets retain more information than thinned SNP datasets, population statistics in this study from filtered datasets were quite similar between thinned SNP and haplotype datasets. 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**.

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**.

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

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**.

Emerson KJ, Merz CR, Catchen JM, *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the Natinal 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.

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

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, Duffie Judy C, Seeholzer GF, *et al.* (2015) Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. *PeerJ* **895**.

Hedrick PW (2005) A standardized genetic differerntiation 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.

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, Réale1 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: Gst 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 inlodgepole pine. *Molecular Ecology* **21**, 2991-3005.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HH (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLoS ONE* **7**, e37135.

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 (2014a) dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organism. *PeerJ* **2**.

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.

Data Accessibility

Empirical Illumina sequences data for red drum be obtained from NCBI's Short Read Archive (SRA) under Accession SRP041032. Scripts for generating the simulated sequence data as well as some automated filtering have been posted to github (https://github.com/jpuritz/).

Tables

Table 1. Results of filtering of simulated ddRAD datasets. 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, multi-copy loci filtered divided by the total simulated, multi-copy loci available. The number of multi-copy loci available to filter at each step may not necessarily match the number remaining in a previous step because some number of multi-copy loci were eliminated in

intermediate filtering steps not directed towards multi-copy loci. The third through fifth columns list the total number of contigs reconstructed by the *dDocent* pipeline, the number of multi-copy loci clusters recovered, and the number of SNPs scored across all clusters. The last columns are the number of simulated multi-copy loci remaining after filtering and the number of those multi-copy loci observed to possess more than two haplotypes.

Table 2. Results of filtering of the empirical ddRAD dataset. 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 multi-copy loci specific filtering also are shown (Scheme 4). For each scheme, the final remaining number of contigs and SNPs with ≤ 5% missing data are listed.

Table 3. Dataset characteristics and population statistics for red drum from Lower Laguna Madre and Sabine Lake, TX, USA. Data were filtered for minor allele frequency (MAF >1/2N alleles). Results are shown from three multi-copy loci filtering schemes: SNPs filtered by each method sequentially (Scheme 1), all SNPs from contigs identified in combination (Scheme 3), or no multi-copy loci filtering (Scheme 4). Number of remaining contigs (#contigs) and SNPs (#SNPS) for each filtering scheme are shown for datasets 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. Dataset characteristics and population statistics for simulated data with simple haplotypes. Data from two simulations (low and high variability) are shown with and without multi-copy loci removed from final datasets. Data were filtered for minor allele frequency (MAF > 1/2N alleles). The number of remaining contigs (#contigs) and SNPs (#SNPs) are shown for datasets 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}").

Figure Legends

Figure 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 cutoff above which contigs are flagged as multi-copy loci.

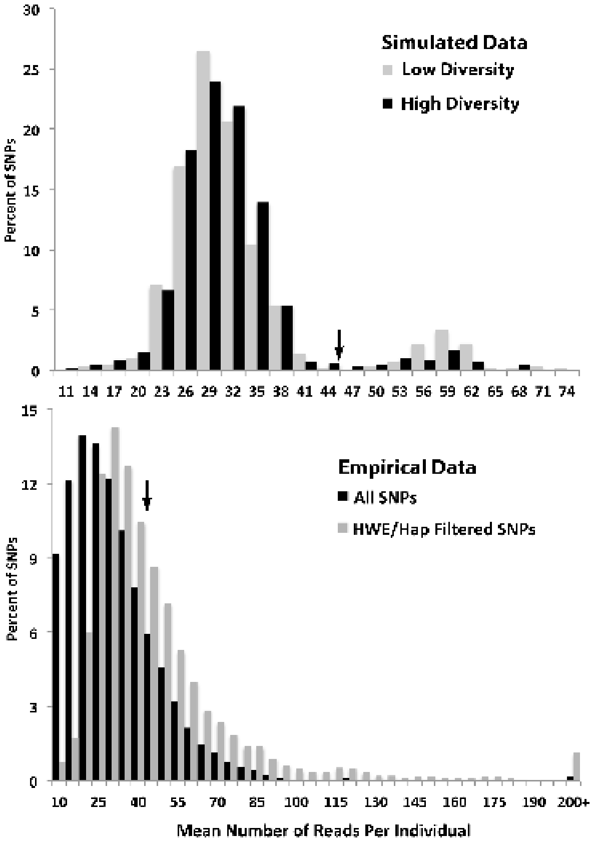
Figure 2. Bar plots of posterior probability of individual assignment for 39 red drum to K=2 clusters, using the program STRUCTURE for three versions of the ddRAD dataset.

data diversity	multi-copy haplotypes	total contigs reconstructed	# multi-copy contigs	# SNPs	filtering scheme	filter by depth	filter by H_o	filter by # haplotypes	multi-copy loci left	multi-copy loci >2 haps
low	**simple**	1,000	50	3,641	**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**	1,000	50	3,714	**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**	1,000	50	7,097	**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**	1,002*	52 (47)*	7,187	**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	

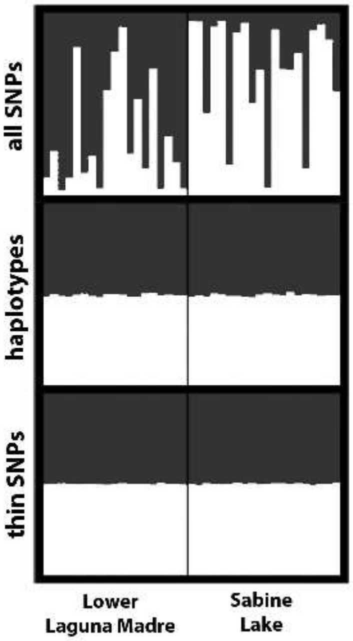
reference contigs	**# contigs ≥1 SNP**	**total SNPs before filtering**	**filtering scheme**	**filter by depth (2X mode)**	**filter by H_o**	**filter by # haplotypes**	**remaining contigs (≤5%)**	**remaining SNPs (≤5%)**
40,329	31,758	124,500	**1**	3,727	30	1,553	5,677	13,280
			2a	4,274	--	--	6,826	20,182
			2b	--	353	--	10,621	32,160
			2c	--	--	2,554	8,332	20,647
			3	combined filters: 5,912			5,271	12,664
			4	no paralog filtering			10,886	33,679

multi-copy filtering	# contigs	markers	# SNPs	# alleles (A_E)	G_{IS} (95%CI)	G_{ST}'' (95%CI)
1. sequential	4,932	all SNPs	9,964	19,928 (1.31)	-0.0103 (-0.0145:-0.0062)	0.0032 (0.0019:0.0045)
		haplotypes	9,964	14,691 (1.61)	-0.0108 (-0.0155:-0.0060)	0.0032 (0.0015:0.0049)
		thin SNPs	4,932	9,864 (1.31)	-0.0102 (-0.0162:-0.0043)	0.0037 (0.0018:0.0057)
3. combined	4,590	all SNPs	9,476	18,952 (1.30)	-0.0094 (-0.0136:-0.0052)	0.0030 (0.0017:0.0044)
		haplotypes	9,476	13,868 (1.62)	-0.0096 (-0.0142:-0.0049)	0.0029 (0.0011:0.0047)
		thin SNPs	4,590	9,180 (1.31)	-0.0085 (-0.0145:-0.0025)	0.0034 (0.0014:0.0055)
4. none	9,870	all SNPs	26,787	53,574 (1.36)	-0.0719 (-0.0764:-0.0675)	0.0027 (0.0020:0.0035)
		thin SNPs	9,870	19,740 (1.34)	-0.0441 (-0.0505:-0.0377)	0.0027 (0.0014:0.0039)

data diversity	multi-copy 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)



men_12647_f1.tif



men_12647_f2.tif