SCIENTIFIC REPORTS

OPEN

Received: 14 July 2017
Accepted: 15 November 2017
Published online: 14 December 2017

Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering

Richard G. J. Hodel[1,2], Shichao Chen[2,3], Adam C. Payton[1], Stuart F. McDaniel[1,2,4], Pamela Soltis[2,4] & Douglas E. Soltis[1,2,4]

The widespread adoption of RAD-Seq data in phylogeography means genealogical relationships previously evaluated using relatively few genetic markers can now be addressed with thousands of loci. One challenge, however, is that RAD-Seq generates complete genotypes for only a small subset of loci or individuals. Simulations indicate that loci with missing data can produce biased estimates of key population genetic parameters, although the influence of such biases in empirical studies is not well understood. Here we compare microsatellite data (8 loci) and RAD-Seq data (six datasets ranging from 239 to 25,198 loci) from red mangroves (*Rhizophora mangle*) in Florida to evaluate how different levels of data filtering influence phylogeographic inferences. For all datasets, we calculated population genetic statistics and evaluated population structure, and for RAD-Seq datasets, we additionally examined population structure using coalescence. We found higher F_{ST} using microsatellites, but that RAD-Seq-based estimates approached those based on microsatellites as more loci with more missing data were included. Analyses of RAD-Seq datasets resolved the classic Gulf-Atlantic coastal phylogeographic break, which was not significant in the microsatellite analyses. Applying multiple levels of filtering to RAD-Seq datasets can provide a more complete picture of potential biases in the data and elucidate subtle phylogeographic patterns.

Choice of molecular markers remains a critically important consideration when designing a phylogeographic, phylogenetic, or population genetic study, as researchers must optimize the amount of informative genetic data they can obtain for a fixed and typically modest cost. In phylogeographic studies, theoretical considerations impact decisions regarding whether to include more individuals or more loci. Microsatellites (or simple sequence repeats, SSRs) have been one of the workhorses of phylogeographic studies for over two decades—their high variability made them popular for distinguishing between closely related conspecific or congeneric individuals[1–3]. Microsatellite markers are now being gradually replaced by RAD-Seq data for phylogeographic inference[4].

There are advantages and disadvantages to using microsatellites in phylogeographic studies[2,3,5]. Microsatellites are a known quantity; hundreds of thousands of studies that use SSRs are in the literature—primers are already available for many groups. In addition, many user-friendly software packages are available for all aspects of microsatellite analysis, from loci development to population genetic inference[3]. If primers are already developed for the taxa of interest, microsatellites can be inexpensive to implement. Additionally, if initial results necessitate adding a few additional individuals and/or loci, project costs will increase linearly with microsatellites. However,

[1]Department of Biology University of Florida, Gainesville, FL, 32611, USA. [2]Florida Museum of Natural History University of Florida, Gainesville, FL, 32611, USA. [3]College of Life Sciences and Technology Tongji University, Shanghai, 200092, China. [4]The Genetics Institute University of Florida, Gainesville, FL, 32610, USA. Correspondence and requests for materials should be addressed to R.G.J.H. (email: richiehodel@gmail.com)

there are caveats to using SSRs. Perhaps most importantly, a limited number of loci (usually < 25) can feasibly be employed in a typical microsatellite study. Also, the mutational properties of SSRs are unusually high and almost certainly do not reflect those of the genome as a whole. Thus, the property that makes microsatellites excellent for distinguishing different individuals may inflate statistics such as F_{ST} and heterozygosity relative to the rest of the genome. Furthermore, microsatellites can be just as expensive to implement as newer high-throughput sequencing (HTS) techniques if there are no existing genetic resources (e.g., no primers already developed, or no available transcriptomic or genomic resources)[6].

The use of RAD-Seq data has increased greatly over the past decade, largely because thousands of loci can be generated simultaneously for hundreds of individuals for a fixed, known cost[7]. RAD-Seq uses restriction enzymes (REs) to create a reduced representation library of the genome; single-nucleotide polymorphisms (SNPs) in regions of DNA between restriction sites are used to distinguish between individuals[8]. Barcoding to allow efficient multiplexing during sequencing keeps costs down, which can be as little as $40 per individual for thousands of loci, assuming judicious sharing of reagents, and a well-designed plan for multiplexing individuals[9–11]. Microsatellite genotyping has a similar cost per individual, assuming primers are not developed, but many fewer loci are obtained[3]. SNPs have several advantages over microsatellites, as they are less likely to exhibit homoplasy than SSRs[12].

Despite advantages, there are also several caveats to using RAD-Seq. Unless there is a reference genome, loci obtained using RAD-Seq are anonymous, and some loci may be non-neutral[7]. Additionally, biases may be introduced at several stages in a RAD-Seq protocol: (1) digestion with REs samples a non-random portion of the genome due to biases in base composition; this is potentially worse if methylation sensitive enzymes are used; (2) polymorphisms in restriction sites that can lead to segregating presence/absence polymorphisms that are very difficult to detect without very deep sequencing and negating the cost-savings of using RAD-Seq in the first place[7,13]; (3) preferential PCR amplification of some loci during library construction necessarily reduces coverage of other loci[13]; (4) sequencing errors and/or low sequencing depth leads to incorrect genotype calling[7]; and (5) false loci are constructed due to the misassembly of paralogous reads[14,15]. Many potential problems are resolved by multiple PCR steps to even out loci coverage and by improvements in software when processing loci, but concerns remain that RE-based reduced representation methods do not capture a representative snapshot of the genome[16]. One other concern with RAD-Seq loci is that manual data curation is impossible, and errors may go undetected even by the most careful researchers[14,17,18]. Finally, the biggest potential problem when using RAD-Seq is that low coverage and high proportions of missing data can make it difficult to infer heterozygotes accurately.

Previous studies have compared results from SNPs and SSRs, revealing that microsatellites provide much more information—up to an order of magnitude more—on a per-marker basis than SNPs[19,20]. However, SNP studies typically use several orders of magnitude more markers than an average SSR study. Evidence has shown that the large number of loci in SNP studies can effectively allow for more powerful inferences, even though the information at each locus is less than that in microsatellite markers[21]. Because of the low number of loci used in SSR studies, the standard practice is to aim to minimize missing data. However, the nature of current library preparation and sequencing means that higher percentages of missing data are an unavoidable part of RAD-Seq studies. Simulation studies have shown that the large amounts of missing data in RAD-Seq studies can inflate F_{ST} estimates due to allelic dropout[13,18]. As more loci were included in these simulations, F_{ST} appeared to increase because many loci had genotype data for only one or a few individuals. In many such loci $F_{ST} = 1$ because by chance the few individuals sampled were homogeneous within populations but different between populations, leading to high average F_{ST}. Heterozygosity can be similarly inflated if the more frequent allele is likely to be absent (e.g., because mutations in the restriction site, which lead to allelic dropout, are often in ancestral alleles that occur at a high frequency)[18]. Arnold, *et al.*[13] confirmed results from Gautier, *et al.*[18] and also concluded that other summary statistics, including Θ and π, could be inaccurately estimated from loci with missing data. In spite of these problems, more recent simulation studies have indicated that missing data in RAD-Seq studies may not lead to incorrect inference, and in fact including loci with missing data can be advantageous for identifying shallow divergences[22].

Convention in phylogeographic studies often is to require 75 or 80% of individuals to have data for a given locus—otherwise that locus is discarded from the analyses (e.g., refs[23–28]). Presumably, requiring a locus to be present in a certain number of individuals will eliminate loci with high missing data that may be the cause of misestimated parameters[13,18]. However, the choice of a cutoff is arbitrary and is typically not justified in phylogeographic studies— the number of SNPs is virtually always reported as a single fixed value (e.g., "we identified a total of 4,234 SNPs," Jackson, *et al.*[24]). In reality, the various parameter values that determine how many loci are constructed and retained in SNP alignment methods means that there is a range of loci that could conceivably be included in a study[27,29].

To date, no phylogeographic study has investigated the effect of varying amounts of missing data in an empirical RAD-Seq dataset, even those explicitly comparing RAD-Seq-generated SNPs and microsatellites[30,31]. To remedy this knowledge gap, we investigate the phylogeography of red mangroves (*Rhizophora mangle* L., Rhizophoraceae) in Florida, using both an existing microsatellite dataset[32], and new RAD-Seq SNP datasets that vary in number of loci and the percentage of missing data. We filtered RAD-Seq loci to generate a dataset that would approximate the number of loci and amount of missing data typically used in RAD-Seq phylogeography studies, and we also generated datasets with more or less stringent filtering to test the effects of increasing or decreasing the number of loci and percentage of missing data. Specifically, we address the following questions: (1) In RAD-Seq datasets, how are phylogeographic inferences affected by the number of loci used? (2) In RAD-Seq datasets, how are phylogeographic inferences affected by the percentage of missing data? (3) What are the important differences in performance between microsatellites and RAD-Seq data in population genetic and phylogeographic inference? (4) Do RAD-Seq data reveal any novel phylogeographic inferences not already recovered by microsatellites for red mangroves in Florida?

Sampling Location	Code	Latitude (N)	Longitude (W)	% Loci Missing
Bahia Honda Key	BHKFl	24.55286	81.76776	73.5
Convoy Point	CvPFl	25.46347	80.33133	81.2
Cape Canaveral	CpCFl	28.82173	80.75594	83.0
Hollywood	HwdFl	26.03841	80.11780	79.4
Islamorada	IsmFl	24.90031	80.65690	81.0
Key Largo	KyLFl	25.09569	80.42957	88.9
Melbourne	MlbFl	28.07435	80.60526	79.8
New Port Richey	NPRFl	28.25432	82.75723	69.5
Seahorse Key	ShKFl	29.10040	83.06185	65.8
Terra Ceia Bay	TCBFl	27.59172	82.57524	81.7
Vaca Key	VKyFl	24.71154	81.06992	85.1
West Palm Beach	WPBFl	26.67505	80.04259	83.9

Table 1. The twelve sampling locations (each containing eight individuals), their codes, GPS coordinates, and the percentage of loci that have missing data for each sampling location before any filtering.

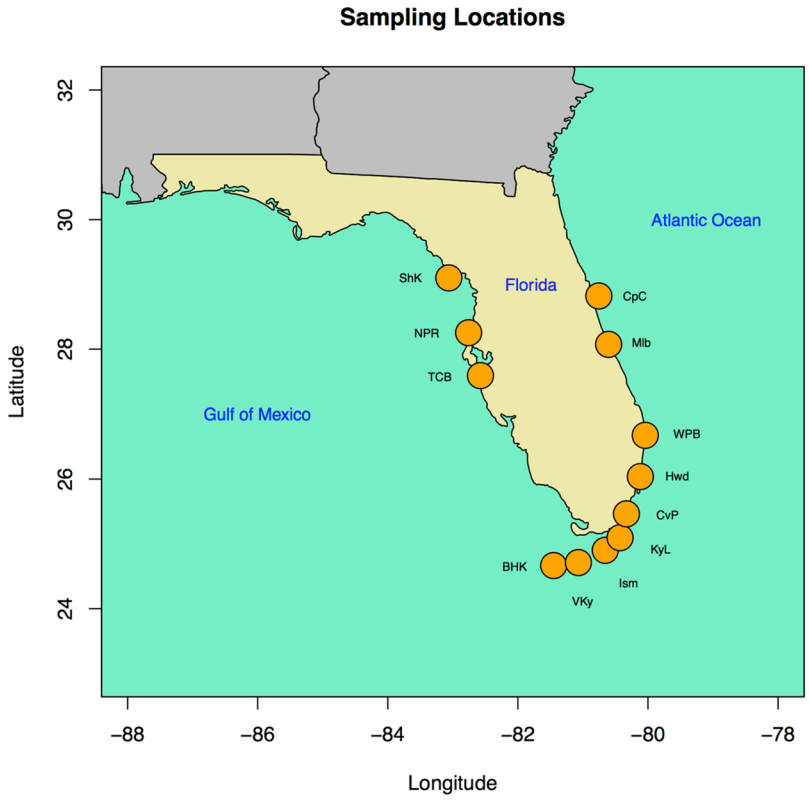

Figure 1. The 12 sampling locations (each with eight individuals) are indicated by orange circles. Sampling location codes are provided in Table 1. The map was generated using R (citation: R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/), and the R package 'maps' (citation: Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2017). maps: Draw Geographical Maps. R package version 3.2.0. https://CRAN.R-project.org/package=maps).

To address these questions, we used 96 red mangrove (*Rhizophora mangle*) individuals collected from 12 sampling locations on the coasts of Florida (Table 1, Fig. 1). Red mangroves are salt-tolerant trees that occur in coastal estuarine environments throughout the neotropics, experiencing high temperatures, frequent inundation, saline conditions, and periodic wave action associated with the coastal environment[33]. Red mangroves provide a variety of ecosystem services, including filtering water, providing habitat to animals, stabilizing shorelines, and protecting coastal environments from frequent wave action and occasional storm surges. Thus, red mangroves are important conservation targets—for which phylogeographic data can improve conservation strategies—making red mangroves a valuable study system.

Dataset	Individuals required to retain a locus	Number of loci	% individuals required to retain a locus
RAD_239	83	239	86.5
RAD_1180	75	1180	78.1
RAD_2317	65	2317	67.7
RAD_3831	50	3831	52.1
RAD_6255	30	6255	31.3
RAD_25198	1	25198	1.0

Table 2. The seven data sets used in this study; RAD-Seq data sets were generated by filtering loci from largest data set (RAD_25198). For all data sets (six RAD and one microsatellite), the total number of loci used is indicated.

Further analysis of phylogeographic patterns in red mangroves and other species occurring in the Florida peninsula is also warranted. Although previous studies of many coastal and marine taxa revealed a phylogeographic discontinuity at or near the southern tip of Florida[34,35], recent work on red mangroves using microsatellites failed to identify such a pattern[32,36]. Different types of molecular markers could reveal new phylogeographic insights, due to broader sampling of the genome, and provide a predictive framework for understanding how genetic variation in this iconic species will respond to climate change. Finally, red mangroves are an ideal system for comparing the performance of alternative genetic markers, given previous analyses of microsatellite loci[32,36,37] and the size of the genome (approximately 1 Gb; Hodel unpublished data, based on flow cytometry observations), enabling a rigorous test of the RAD-Seq method. Genome size is a necessary consideration with RAD-Seq; as genome size increases, the number of loci shared among many individuals for a given sequencing depth decreases.

Results

Datasets. Seven datasets, ranging from 8 loci (SSR_8) to 25,198 loci (RAD_25198; Table 2), were used to investigate in depth how variation in number of loci and percent missing data impacted phylogeographic inferences. These were selected from all possible datasets, for which basic statistics were calculated (Supplementary Figure 1). The name of each of the seven datasets contains information about locus type (RAD or SSR) and number of loci in the dataset. The smallest RAD dataset contained 239 loci (RAD_239), and the percentage of missing data for RAD datasets ranged from 11.7% to 78.1% (Table 3). The dataset RAD_1180, which required a locus to be present in 75 of 96 individuals (78.1%), most closely mimicked the amount of loci filtering typically used in a phylogeographic study. Therefore, in our analyses, we used this as a baseline dataset against which to compare other RAD datasets. Across sampling locations, the proportion of missing data was relatively uniform (Table 1); percentage of missing loci in the data matrix for a given sampling location ranged from 65.8% to 88.9%.

Population genetic analyses. Measures of heterozygosity were not significantly different between the microsatellite dataset and the RAD datasets; average H_O was 0.431 for the microsatellite dataset and 0.392 in RAD_1180, with a range from 0.354 to 0.477 across all RAD datasets (Table 3). Average H_E was 0.388 for the microsatellite dataset and 0.307 for RAD_1180 and ranged from 0.300 to 0.340 for all RAD-Seq datasets (Table 3). Average F_{ST} for microsatellites was 0.124, which was significantly greater than average F_{ST} for only one of the RAD datasets—the smallest (RAD_239; Table 3). Within the RAD datasets, average H_O was significantly greater in RAD_25198 than all others, and it was significantly lower in RAD_6255 than in all others; H_O did not predictably increase or decrease as the number of loci increased (Table 3). Additionally, within RAD datasets, average H_E was significantly greater in RAD_25198 than all others. F_{ST} ranged from 0.046 to 0.108 among the RAD datasets (Table 3). There was no significant difference in F_{ST} in the three smallest RAD datasets, but the three largest RAD datasets all had increased F_{ST} relative to the smaller datasets (Table 3). The dataset with the largest value of F_{ST} was RAD_6255 (Table 3). Average F_{IS} using microsatellites was not significantly different than F_{IS} calculated using RAD datasets; within RAD datasets, F_{IS} generally increased as more loci were added, although RAD_25198 had the lowest value of F_{IS} (Table 3). Many of the population genetic statistics were disproportionately affected by loci with very low or very high values of F_{ST}, F_{IS}, or heterozygosity (Fig. 2). The effect of extreme loci was particularly evident in the larger datasets (RAD_6255 and RAD_25198), in which there were large numbers of loci with extreme values (e.g., F_{ST} of 1.0; Fig. 2).

Pairwise F_{ST}. The values of pairwise F_{ST} for each sampling location relative to other sampling locations were remarkably consistent across datasets (Table 4). For most sampling locations pairwise F_{ST} estimated by SSRs was approximately twice as large as RAD dataset estimates. For every dataset, pairwise F_{ST} between Seahorse Key and all other sampling locations was the highest. For every RAD dataset, Islamorada had the lowest value for pairwise F_{ST}, but the SSR dataset identified West Palm Beach as the sampling location with the lowest estimate of pairwise F_{ST}. Even as the amount of missing data increased, the pairwise F_{ST} estimates remained consistent; RAD_25198 produced similar values to smaller RAD datasets (Table 4).

F_{IS} by sampling location. Cape Canaveral was the location with the highest F_{IS} using the microsatellite data (SSR_8), followed by Key Largo and Seahorse Key (Table 5). Meanwhile, for all RAD datasets, Seahorse Key had one of the lowest (i.e., most negative) F_{IS} values among all populations. Within the RAD datasets, the number of loci and/or amount of missing data affected F_{IS}. For example, in Key Largo, the largest dataset yielded a value

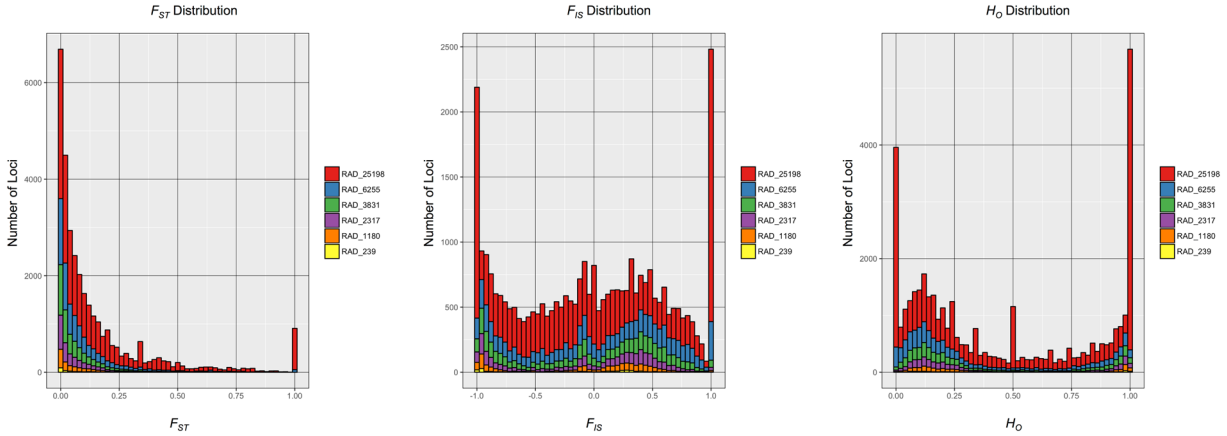

Figure 2. Stacked histograms of per locus estimates of F_{ST}, F_{IS}, and H_O for each of the RAD datasets. Datasets with more loci are stacked on top of datasets with fewer loci.

of 0.015, while the smallest dataset had a value of −0.194. This was not a large absolute change in F_{IS}, but the interpretation of this statistic changed based on whether it is positive or negative (higher values indicate a greater level of inbreeding). In general, within RAD datasets, F_{IS} increased as loci were added, although this trend was not universal, especially in the largest RAD dataset. For instance, in Bahia Honda Key, F_{IS} was lowest in the largest dataset RAD_25198 (25,198 loci, 78.1% missing data). Conversely, in Islamorada, F_{IS} was lowest in the smallest dataset (RAD_239, 11.7% missing data).

Heterozygosity by sampling location. Observed heterozygosity for each sampling location ranged from 0.320 (Seahorse Key) to 0.451 (Hollywood) when averaged across all datasets (Table 6). For most datasets, Seahorse Key was the sampling location with the lowest H_O, although notably RAD_25198 identified six other sampling locations with lower H_O than Seahorse Key (Table 6). Similarly, most datasets reported Hollywood as the sampling location with the highest H_O, but SSR_8 found Convoy Point and Islamorada had higher H_O than Hollywood, and RAD_25198 identified five other sampling locations with greater H_O, with Key Largo having the highest H_O (Table 6). For most sampling locations, measures of H_O, when ranked relative to other sampling locations, remained similar across all RAD datasets except RAD_25198. Interestingly, the values of H_O ranked relative to other sampling locations were more similar between SSR_8 and the five smallest RAD datasets (RAD_239-RAD_6255) than any of the five smallest RAD datasets were to RAD_25198 (Table 6).

PCA and SVDQuartets. The PCA analysis revealed that microsatellite data did not identify clear groupings of individuals based on sampling location or other geographical divisions (Fig. 3). Similarly, RAD_239 did not differentiate the samples into discrete clusters. However, RAD_1180, RAD_2317, RAD_3831, RAD_6255, and RAD_25198 all divided the samples into two groups with minimal overlap in the PCA visualization: one group was Gulf Coast samples, and the other group was Atlantic Coast samples (Fig. 3). Closer inspection of the PCAs revealed that most of the Cape Canaveral individuals formed a discrete cluster intermediate between the two other clusters (Gulf and Atlantic). Most RAD datasets had sufficient resolution to place Cape Canaveral between the Gulf and Atlantic clusters, but the use of a small number of loci (i.e., RAD_239) was unable to show this relationship. Furthermore, the two largest datasets, RAD_6255 and RAD_25198, showed Cape Canaveral individuals clustering more closely to the Atlantic than the Gulf cluster.

The 50% majority-rule consensus bootstrap trees generated with SVDQuartets showed substantial variation between datasets when inferring genealogical relationships between individuals and/or sampling locations (Fig. 4). In many cases, dataset RAD_239 did not identify genealogical relationships that were recovered with other datasets with more loci. However, certain key relationships among individuals were consistently shown in multiple datasets with thousands of loci. In every dataset except RAD_239 (i.e., every dataset with at least 1180 loci), all Seahorse Key (ShKFl) samples formed a clade (Fig. 4). In four datasets (RAD_2317, RAD_3831, RAD_6255, RAD_25198), all Gulf Coast (NPRFl, ShKFl, TCBFl) samples (except one individual: NPRFl_R8), together with all Cape Canaveral (CpCFl) samples, formed a clade that is sister to all remaining Atlantic Coast samples plus NPRFl_R8. Interestingly, this relationship was not recovered in RAD_1180, the dataset with 'ideal' filtering of loci—but all datasets with more loci (and therefore more missing data) did recover the relationship. Each RAD dataset had nodes with varying levels of bootstrap support (Fig. 4). Datasets with fewer loci showed few nodes with bootstrap support >70%; RAD_239 had three such nodes. More loci resulted in more nodes with bootstrap support >70%, up to a point: RAD_1180 had six highly supported nodes, RAD_2317 had eight, and RAD_3831 had the most with nine. Then the number of highly supported nodes slightly declined as more loci were added: RAD_6255 and RAD_25198 each had eight nodes with bootstrap support >70% (Fig. 4).

Sampling loci. Analyzing differently sized samples of loci from RAD_25198 and SSR_8 provided several crucial insights. A microsatellite dataset with seven loci sampled from SSR_8 performed better in estimating F_{ST} than a dataset with six loci, although in each case, all 100 sampled replicates fell within the 95% confidence interval of

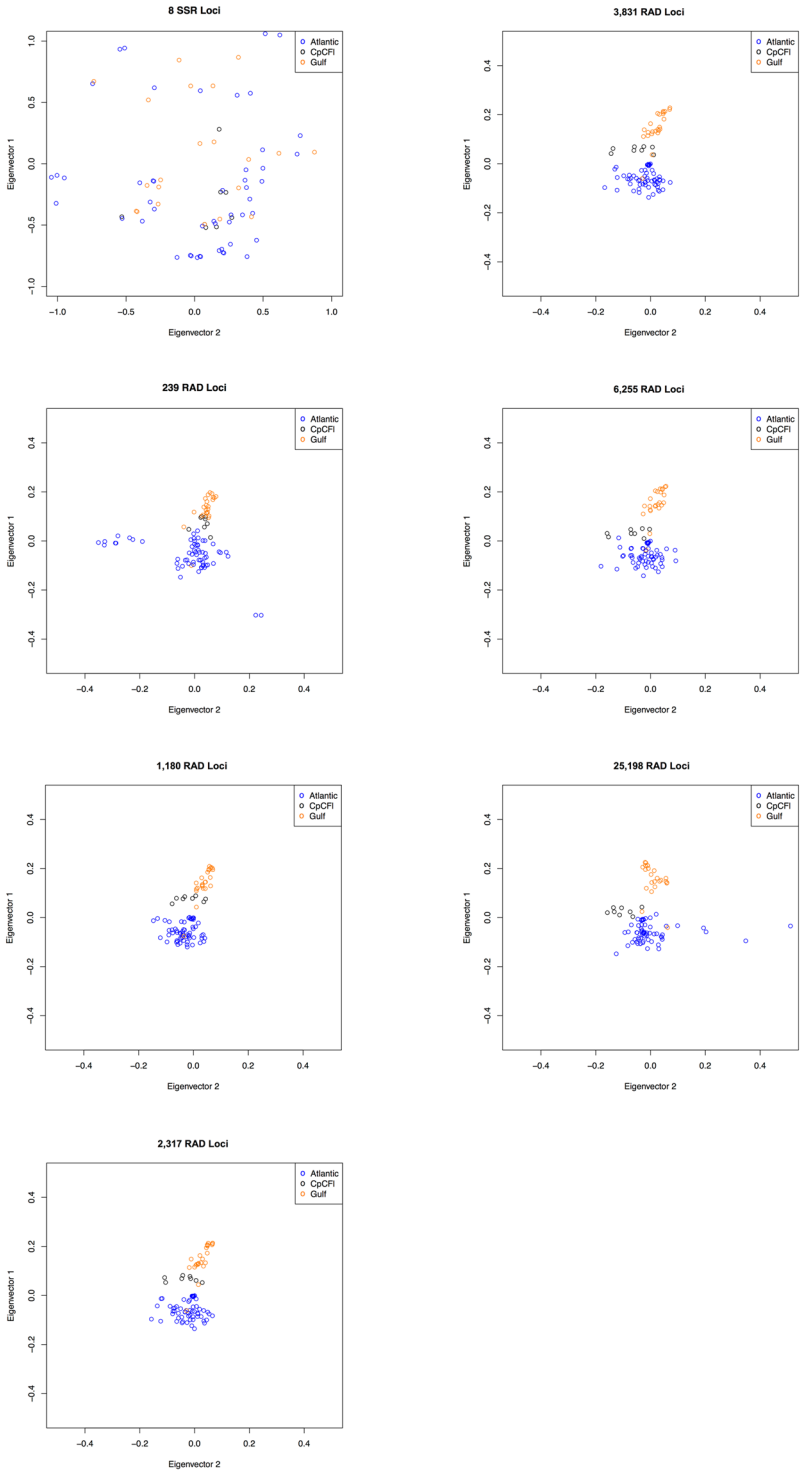

Figure 3. Principle component analysis (PCA) for all seven data sets. Note that the scales of the axes of the SSR_8 plot are different than the axes of all the RAD plots.

F_{ST} for the complete SSR_8 dataset (Fig. 5). For all RAD datasets, the value of F_{ST} estimated using only originally filtered data is different from all 100 permuted values of F_{ST} calculated from an equivalent number of loci sampled from the largest dataset (RAD_25198). For almost all datasets, F_{ST} based on sampled loci was less than F_{ST} using original loci, except for one dataset (RAD_6255), F_{ST} based on the sampled loci was greater. Strikingly, in none of the datasets did the confidence intervals from the sampled loci overlap with the confidence intervals of the estimated F_{ST} values from the original data (Fig. 5). The percentage of missing data in the largest dataset clearly had an immense impact. Even when very few loci (e.g., 239 loci) were sampled from the largest dataset, the

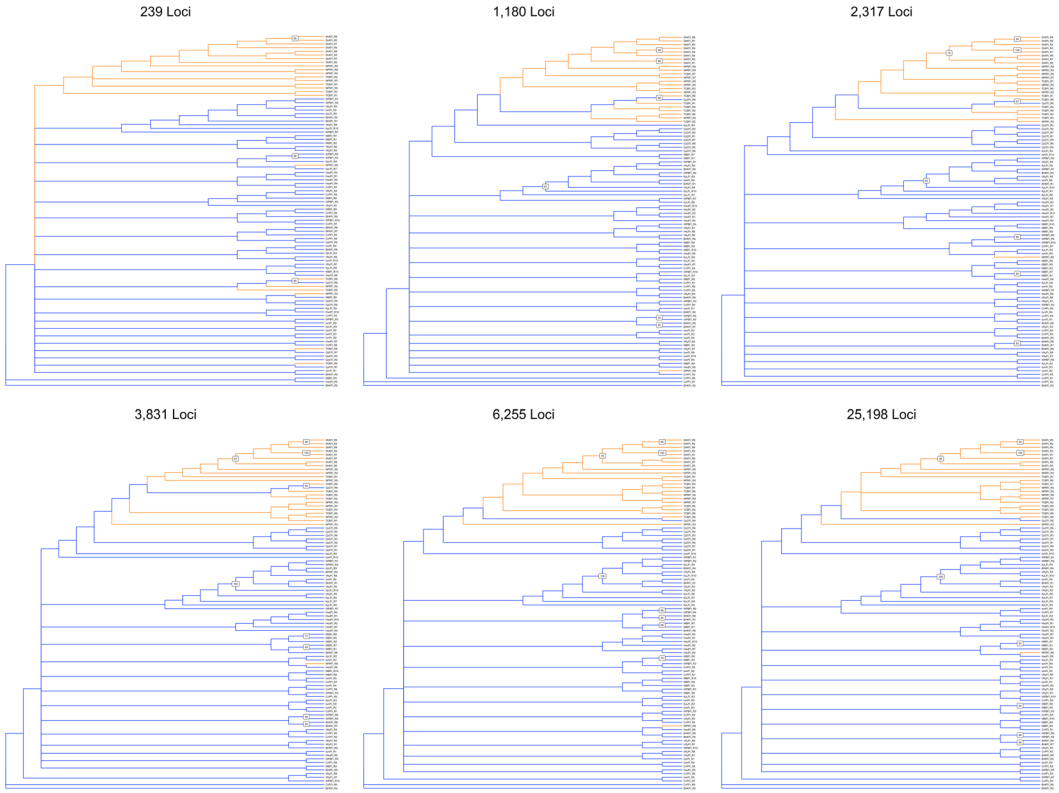

Figure 4. Trees estimated using every individual for each RAD dataset in SVDQuartets. Orange branches indicate individuals from sampling locations in the Gulf of Mexico, and blue branches represent individuals from Atlantic sampling locations.

Dataset	% Missing	F_{ST}		F_{IS}		H_O		H_E	
RAD_239	11.7	0.046	±0.009	-0.365	±0.107	0.410	±0.028	0.300	±0.013
RAD_1180	17.1	0.057	±0.004	-0.298	±0.061	0.398	±0.016	0.307	±0.007
RAD_2317	22.2	0.057	±0.002	-0.253	±0.033	0.390	±0.010	0.312	±0.004
RAD_3831	29.4	0.066	±0.002	-0.213	±0.027	0.382	±0.006	0.315	±0.003
RAD_6255	41.3	0.108	±0.002	-0.164	±0.022	0.356	±0.005	0.306	±0.002
RAD_25198	78.1	0.080	±0.003	-0.403	±0.016	0.477	±0.003	0.340	±0.002
SSR_8	0.0	0.124	±0.067	-0.110	±0.694	0.431	±0.187	0.388	±0.111

Table 3. Relevant population genetic statistics for each of the seven data sets used in this study. For each column, warmer colors indicate lower values and cooler colors show higher values. Immediately to the right of each of the four columns (F_{ST}, F_{IS}, H_O, H_E) is the 95% confidence interval for each statistic.

distribution of F_{ST} values clustered around the estimated F_{ST} using all 25,198 loci (Fig. 5), indicating that missing data, not number of loci, affected the differences in measured F_{ST}.

Discussion

Insights regarding choice of loci. Our results indicate that filtering loci using the standard cutoff (i.e., 75–80% of individuals must possess data for a given locus for that locus to be retained) should not be the gold standard in RAD-Seq studies—it is possible to retain many more loci without inflated statistics[23–28]. F_{ST} increased as missing data increased, as predicted by simulation studies, but the relationship is more nuanced than previously assumed. F_{ST} increases as the percentage of missing data increases—up to a point—and then decreases from RAD_6255 to RAD_25198, as the percentage of missing data nearly doubles, from 41.3% to 78.1% (Table 3). When more loci are included, the distribution of F_{ST} across the genome is more completely sampled. However, adding loci with more missing data may cause analyses to miss low-frequency alleles in the loci with extensive missing data, which would add error to average estimates of F_{ST}. Sampling analyses confirmed that F_{ST} generally

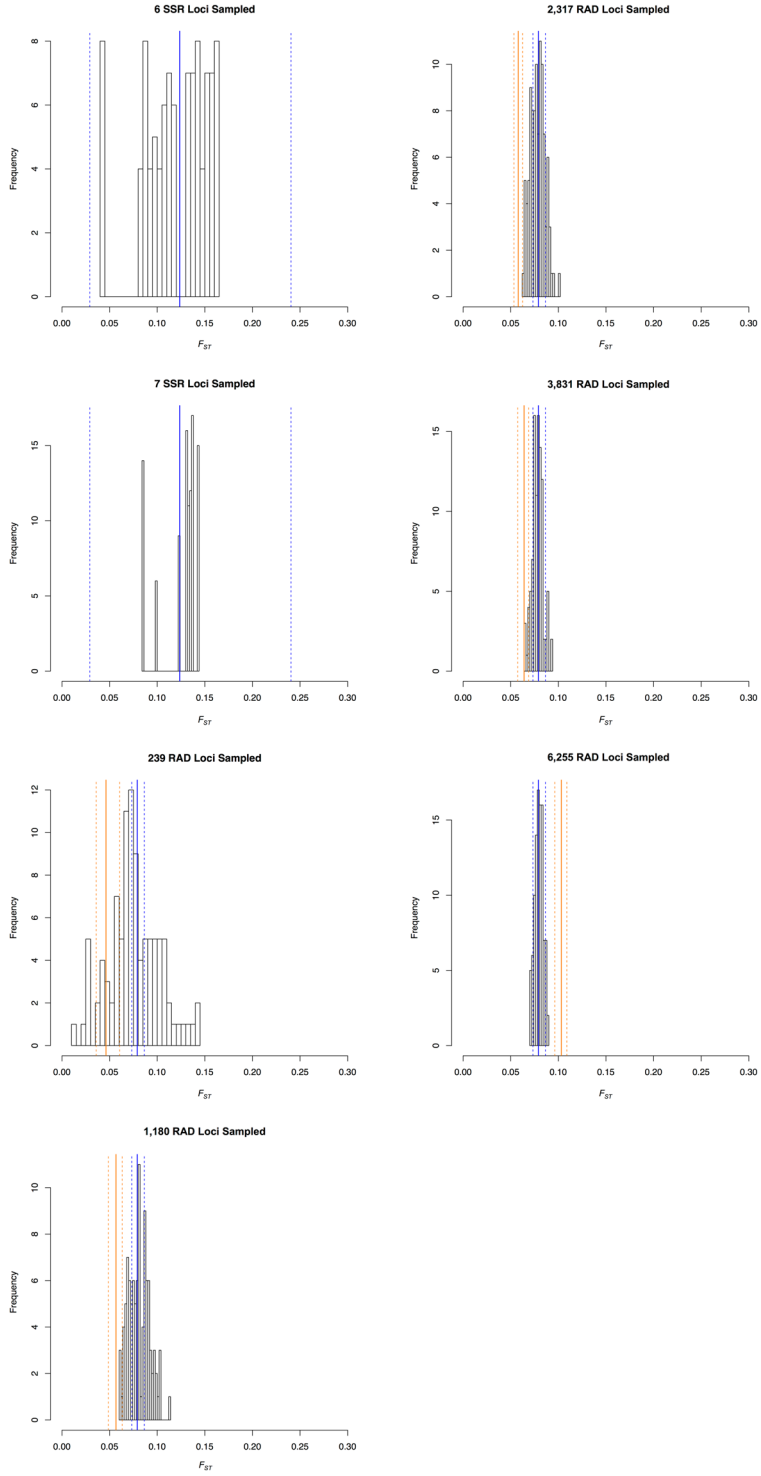

Figure 5. Histograms showing the distribution of the 100 samplings of loci from a larger data set. In the first two panels, six and seven SSR loci, respectively, were randomly sampled 100 times from the SSR_8 data set, and the distribution of the 100 calculations of F_{ST} are shown. The solid blue line indicates the parameter value estimated using all eight loci, and the dashed blue lines show the 95% confidence interval. In the remaining five plots, the histogram shows parameter estimates using the number of loci (239, 1,180, 2,317, 3,831, and 6,255, respectively) in the data set randomly sampled from RAD_25198 100 times. The solid blue lines indicate the F_{ST} value estimated using all 25,198 loci, and the dashed blue lines show the 95% confidence interval. The solid orange lines indicate F_{ST} estimated using the original data set (RAD_239, RAD_1180, RAD_2317, RAD_3831, and RAD_6255, respectively) and dashed orange lines show the 95% confidence interval for this estimate.

	SSR_8	RAD_239	RAD_1180	RAD_2317	RAD_3831	RAD_6255	RAD_25198	Average
BHKFl	0.101	0.039	0.055	0.049	0.052	0.068	0.063	0.061
CpCFl	0.155	0.055	0.069	0.066	0.069	0.075	0.078	0.081
CvPFl	0.085	0.040	0.048	0.046	0.050	0.061	0.062	0.056
HwdFl	0.163	0.052	0.056	0.059	0.063	0.073	0.076	0.078
IsmFl	0.106	0.030	0.040	0.039	0.044	0.051	0.050	0.052
KyLFl	0.108	0.063	0.071	0.073	0.081	0.100	0.097	0.085
MlbFl	0.134	0.043	0.052	0.052	0.055	0.068	0.074	0.068
NPRFl	0.130	0.045	0.060	0.062	0.064	0.083	0.081	0.075
ShKFl	0.279	0.117	0.134	0.138	0.143	0.158	0.152	0.160
TCBFl	0.100	0.056	0.077	0.076	0.083	0.101	0.098	0.084
VKyFl	0.135	0.045	0.050	0.052	0.056	0.067	0.069	0.068
WPBFl	0.083	0.046	0.058	0.056	0.053	0.070	0.073	0.063

Table 4. Pairwise F_{ST} for each sampling location (i.e., one sampling location versus all others) for each of the seven datasets. Within each data set, lower (warmer colors) and higher (cooler colors) values of F_{ST} are shown using color-coding.

increased as missing data increased (Fig. 5). Heterozygosity was less affected by missing data, as there was little or no change in either observed or expected heterozygosity when the percentage of missing data ranged between 11.7% (RAD_239) and 41.3% (RAD_6255). Only the largest dataset (RAD_25198, with 78.1% missing data) showed significantly higher heterozygosity than other datasets. Some simulation studies reported that missing data could inflate F_{ST}, and would likely inflate estimates of heterozygosity, leading to calls for removing loci with incomplete sampling[13]. However, more recent simulation studies showed that removing loci with higher mutation rates, which are more likely to have missing data, negatively impacted phylogenetic analyses[22]. Our study shows the importance of thoroughly exploring how loci are filtered in empirical systems. Extreme amounts of missing data yield higher estimates of F_{ST} and heterozygosity and lower estimates of F_{IS} (Table 3). A large number of loci in RAD_25198 had very high values of certain statistics (e.g., hundreds of loci with $F_{ST} > 0.975$ and thousands of loci with $H_O > 0.975$), which severely impacted average estimates of these statistics (Fig. 2). Notably, not all datasets have these extreme loci—dataset RAD_3831, which only requires 52.1% of individuals to have data for a given locus, and had 29.4% missing data, did not suffer from extreme loci, despite liberal filtering.

Although missing data caused some statistics to increase, it did not dramatically affect our conclusions. For many analyses, using datasets other than RAD_1180, especially RAD_2317 and RAD_3831, did not change the interpretation of the results. Regardless of which of the three datasets was used, F_{ST} was relatively low—between 0.057 and 0.066. Importantly, nearly doubling the amount of missing data from 17.1% (RAD_1180, the 'gold standard') to 29.4% (RAD_3831) resulted in a very small increase in F_{ST} and no significant change in other statistics (F_{IS}, H_O, H_E; Table 3). Furthermore, using very few loci (RAD_239) did not significantly change any of the statistics estimated using RAD_1180 (Table 3). Our data indicate that the often-used cutoff of 75–80% individuals with locus data is arbitrary, and different cutoffs should be considered and evaluated on a case-by-case basis to ensure an appropriate number of loci are used. The results suggest that in many cases, only minimal filtering of loci is needed, and many more loci can be retained than typically are. Researchers who wish to maximize the number of loci in their study could likely use very low cutoffs (e.g., require a locus to have data for >10% of individuals).

Typically, microsatellite datasets have lower F_{ST} values relative to SNPs due to a larger number of alleles, although simulation studies have shown evidence that F_{ST} can be elevated up to an order of magnitude in microsatellite datasets due to factors such as correlated allele frequencies[38]. Average F_{ST} ranged from 0.046 to 0.124 across all datasets—so there is not high differentiation detected in any dataset (Table 3). When using any RAD dataset except RAD_239, F_{ST} calculated using RAD loci was statistically indistinguishable from the microsatellite dataset (Table 3). In theory, a three-to-four-fold change in F_{ST} could alter biological conclusions—possibly with deleterious results (e.g., identifying populations or management units that would be prioritized for conservation)—but no matter how the loci were filtered, there was a relatively small range of estimated F_{ST}. RAD-Seq studies where larger values of F_{ST} were detected could exhibit larger absolute changes in F_{ST} when using different loci filtering cutoffs (e.g., refs[39,40]).

Similarly, the interpretation of F_{IS} and H_O could impact how data are considered in a biological context (e.g., identifying locations at risk for inbreeding depression). Positive values of F_{IS} and/or low values of H_O often indicate inbreeding, which means sampling locations are more vulnerable than other sampling locations. As noted earlier, SSR_8 identified five sampling locations with positive F_{IS} values (Table 5). Only one of these sampling locations (Key Largo) also had positive F_{IS} values in any of the RAD datasets. Clearly, marker choice (microsatellite versus RAD-Seq) affected the assessment of which populations are more vulnerable based on F_{IS} values. Agreement between these two markers types would facilitate identifying sampling locations vulnerable to inbreeding depression. However, it is understandable that different markers would lead to different results, as

	SSR_8	RAD_239	RAD_1180	RAD_2317	RAD_3831	RAD_6255	RAD_25198	Average
BHKFl	0.049	-0.207	-0.161	-0.132	-0.126	-0.118	-0.263	-0.140
CpCFl	0.095	-0.315	-0.208	-0.166	-0.159	-0.134	-0.113	-0.155
CvPFl	-0.300	-0.238	-0.201	-0.165	-0.158	-0.126	-0.109	-0.191
HwdFl	-0.206	-0.356	-0.274	-0.231	-0.206	-0.147	-0.105	-0.228
IsmFl	-0.130	-0.245	-0.163	-0.130	-0.111	-0.070	0.003	-0.128
KyLFl	0.081	-0.194	-0.168	-0.146	-0.133	-0.123	0.015	-0.109
MlbFl	-0.018	-0.169	-0.121	-0.096	-0.092	-0.075	-0.039	-0.092
NPRFl	-0.090	-0.317	-0.251	-0.225	-0.214	-0.209	-0.307	-0.234
ShKFl	0.067	-0.554	-0.506	-0.452	-0.426	-0.400	-0.427	-0.398
TCBFl	-0.041	-0.407	-0.357	-0.311	-0.300	-0.277	-0.335	-0.299
VKyFl	-0.109	-0.260	-0.214	-0.185	-0.158	-0.133	-0.086	-0.172
WPBFl	0.020	-0.233	-0.236	-0.218	-0.203	-0.180	-0.152	-0.177

Table 5. The variation in average inbreeding coefficient (F_{IS}) among data sets and populations. Within each data set, lower (warmer colors) and higher (cooler colors) values of F_{IS} are shown using color-coding. The average value of F_{IS} across all data sets for each population is shown in the last column of the table.

mutation rate can affect estimation of F_{IS}—in theory, as the mutation rate increases, F_{IS} should decrease. The data showed opposite result though, as F_{IS} was higher in the SSR_8 dataset for most sampling locations (Table 5). This is likely because estimates of H_O and H_E have larger variance when few loci are used, as in the SSR_8 dataset (Table 3). The relative values of H_O and H_E can dramatically affect the interpretation of F_{IS}, especially when H_O and H_E are similar (e.g., using the equation $F_{IS} = (H_E - H_O)/H_E$ would yield F_{IS} of 0.25 when $H_O = 0.4$ and $H_E = 0.5$, but $F_{IS} = -0.25$ if H_O and H_E are reversed). Within RAD datasets, estimates of F_{IS} for each sampling location were fairly consistent and F_{IS} increased as missing data increased, but this trend was not universal (Table 3). Identifying vulnerable sampling locations based on H_O revealed that RAD_25198 led to different conclusions than most other datasets. Across datasets, measures of H_O within each sampling location were consistent relative to other sampling locations, except for RAD_25198 (Table 6). Missing data impacted this analysis; a large number of loci in RAD_25198 had either very high or very low H_O, possibly leading to the pattern of H_O in RAD_25198 that contrasted with patterns in virtually every other dataset (Table 6, Fig. 2).

The PCA results show that as the number of loci increases, the definition of clusters improves, plateauing with RAD_2317 or RAD_3831. The clustering is similar in all RAD datasets with 1,180 or more loci, with Cape Canaveral individuals falling between the Gulf and Atlantic clusters. As more loci are added, the Cape Canaveral samples appear to be closer to the Atlantic cluster, especially in datasets RAD_6255 and RAD_25198 (Fig. 3). Taking into account the SVDQuartets results clarifies the clustering—all Cape Canaveral samples form a clade with all Gulf samples except one. However, this relationship is only present in datasets with 2,317 loci or greater—the putatively 'gold standard' dataset RAD_1180 does not show this relationship.

Phylogeographic patterns in red mangroves. Based on previous studies using microsatellite data[32,36], the relationship of Cape Canaveral samples to other sampling locations, as found here with both PCA and SVDQuartet analyses of RAD-Seq data, was surprising—previous studies did not find that any of the individuals in the Cape Canaveral population clustered with any of the Gulf samples. These new data could indicate an Atlantic-Gulf phylogeographic discontinuity, and that Cape Canaveral is an anomaly due to a lack of phylogeographic resolution, recent population founding, or human-mediated transplantation. The intermediate placement of Cape Canaveral in many of the PCAs suggests that it may actually cluster with the Atlantic samples, especially when considering datasets RAD_6255 and RAD_25198, indicating a phylogeographic break (Fig. 3). However, the SVDQuartets results place Cape Canaveral in a clade with the vast majority of Gulf samples, although this relationship is not highly supported in any datasets (i.e., bootstrap support is not >70% for this clade in any dataset) (Fig. 4). Assuming that Cape Canaveral is more closely related to Gulf samples, the age of the divergence between the two clades (Atlantic, Gulf + CpCFl) comes into question. Northern Florida represents the northern limit of the range of red mangroves[33]. Typically, populations of these trees in northern Florida are periodically extirpated due to freezing events, and these areas are re-colonized. The lower values of H_O in northern populations (CpCFl, MlbFl, ShKFl, NPRFl, TCBFl) relative to southern populations indicate that these populations were likely founded more recently from a small number of propagules. The Cape Canaveral population was likely founded by individuals from the Gulf Coast, suggesting that the divergence between the two clades (Atlantic, Gulf + CpCFl) is very recent.

Previous research indicates that gene flow is greater from the Gulf Coast to the Atlantic Coast in red mangroves; there may be ongoing gene flow from the Gulf to Cape Canaveral[32]. Alternatively, alleles from the Gulf Coast could have migrated into an existing Cape Canaveral population and proliferated due to other processes (e.g., drift). Another explanation for the sister relationship between the Gulf samples and Cape Canaveral is

	SSR_8	RAD_239	RAD_1180	RAD_2317	RAD_3831	RAD_6255	RAD_25198	Average
BHKFl	0.391	0.429	0.419	0.414	0.403	0.377	0.419	0.408
CpCFl	0.359	0.379	0.346	0.350	0.355	0.336	0.306	0.348
CvPFl	0.656	0.431	0.418	0.413	0.397	0.363	0.305	0.425
HwdFl	0.516	0.487	0.464	0.451	0.435	0.395	0.392	0.451
IsmFl	0.531	0.443	0.414	0.406	0.392	0.364	0.387	0.420
KyLFl	0.438	0.393	0.370	0.361	0.339	0.308	0.468	0.382
MlbFl	0.328	0.412	0.395	0.394	0.386	0.360	0.311	0.372
NPRFl	0.359	0.353	0.339	0.341	0.348	0.335	0.403	0.351
ShKFl	0.313	0.318	0.306	0.305	0.317	0.312	0.394	0.320
TCBFl	0.453	0.366	0.354	0.360	0.364	0.351	0.415	0.376
VKyFl	0.422	0.449	0.431	0.424	0.408	0.371	0.332	0.408
WPBFl	0.406	0.459	0.444	0.438	0.426	0.389	0.346	0.418

Table 6. The variation in observed heterozygosity (H_O) among data sets and populations. Within each data set, lower (warmer colors) and higher (cooler colors) values of H_O are shown using color-coding. The average value of H_O across all data sets for each population is shown on the bottom row of the table.

human-mediated transplantation of propagules or seedlings from the Gulf Coast to Cape Canaveral. However, all available publications and information from land managers who replied to requests for information confirm that any restoration that required importation of propagules used either local propagules or seedlings from the southern Atlantic Coast (ref.[41], personal communication with Rangers from Cape Canaveral National Seashore). Another possible explanation for this result is that red mangrove propagules were accidentally transported from the Gulf Coast of Florida to Cape Canaveral during construction of the Kennedy Space Center in the 1960s. Construction of the Space Center was a massive project. It is noteworthy that nearly 100,000 tons of steel was transported from the Gulf Coast to Cape Canaveral in numerous trucks; the transport of a few mangrove propagules during this process could easily have established a Gulf genotype in the Cape Canaveral area[42]. We conclude that, in contrast to microsatellites, RAD datasets recover a relationship between the Gulf Coast and the Atlantic Coast (excluding Cape Canaveral) that supports the presence of a maritime discontinuity in red mangroves. However, as red mangroves can disperse long distances, a population or populations that recently established in Cape Canaveral likely had a founder or founders that were predominantly of Gulf Coast origin. The fact that previous studies using SSRs did not elucidate this relationship is not surprising—both the PCA analysis and SVDQuartets analysis indicate that 1180 loci were barely sufficient to infer the placement of Cape Canaveral—datasets with many more loci were needed (Figs 3 and 4). The large number of loci required to resolve such relationships highlights why liberal filtering of RAD-Seq loci is advisable.

Conclusions

We cannot overemphasize the importance of thoroughly exploring RAD-Seq datasets when performing phylogeographic analyses—it is too easy to jump to conclusions when only using one arbitrary cutoff to filter loci. Our empirical data confirm that estimates of F_{ST} and/or heterozygosity may become inflated as missing data increase. However, this does not happen as quickly as implied in simulation studies as loci with missing data are added—liberal filtering of loci retains loci valuable for phylogeographic or phylogenetic inference, without inflating population genetic statistics. Thus, regardless of the cutoff value used to filter loci, researchers should investigate several other cutoffs with both increased and decreased amounts of missing data to appreciate fully the impact of missing data on parameters in their study. We found no evidence that the 75% or 80% cutoff commonly employed was optimal. In many analyses, other datasets with cutoffs ranging from 31.3% to 67.7% performed just as well as or better than RAD_1180. Many RAD-Seq studies aim to multiplex as many individuals as possible in a HTS run; our results show that retaining loci with more missing data is feasible and advantageous in empirical studies, and that researchers can include more samples in a single sequencing run. Our study confirmed that microsatellites were a valuable tool for inexpensively estimating population genetic statistics, such as F_{ST}, F_{IS}, and heterozygosity. However, this study revealed that the thousands of additional loci from across the genome provided by RAD-Seq increased phylogeographic resolution. We found that red mangroves likely have a phylogeographic discontinuity at the southern tip of Florida that was not detected in previous studies using SSRs[32,36] and that a single population from the Atlantic coast of Florida arose via recent colonization by propagules (either natural or human-mediated) from the Gulf coast.

Methods and Materials

Sample collection, DNA isolation. We collected leaf tissue from plants of *R. mangle* from 12 locations in Florida (Fig. 1). At each location, we collected one leaf from 10–20 individuals that were spaced at least 15 m apart to minimize collecting closely related individuals. For each sampling location, we randomly selected 8 individuals

to use in genetic analyses. GPS coordinates for each sampling location were recorded (Table 1). Each sampled leaf was placed in a labeled bag with silica gel and stored for 1–12 months; we then extracted DNA from the dried leaf tissue using a standard CTAB protocol[43].

Microsatellite amplification and analysis. We PCR-amplified eight nuclear microsatellite loci for *R. mangle* (RM 11, 19, 21, 36, 38, 41, 46, 47)[37]. An M13 protocol[44] was used to label amplicons with four fluorescent dyes (6-FAM, NED, PET, VIC). The PCR (25-μL reactions) contained: 5X buffer (5 μL), 2.5 mM $MgCl_2$ (2 μL), 2.5 mM dNTP (0.5 μL), 0.12 μM forward primer with M13 label attached (1.25 μL), 4.5 μM reverse primer (1.25 μL), 4.5 μM fluorescent dye (2.5 μL), H_2O (10 μL), Taq polymerase (0.5 μL), and 50 ng template DNA (2 μL). We carried out PCR in a Biometra T3 Thermocycler (Whatman Biometra, Goettingen, Germany) using the following cycles: initial denaturing at 94 °C for 3 minutes; 35 cycles of 94 °C (45 seconds), 52 °C (45 seconds), 72 °C (75 seconds); final elongation at 72 °C for 20 minutes. We used the Applied Biosystems 3730 DNA Analyzer (Applied Biosystems, Foster City, United States) at the University of Florida Interdisciplinary Center for Biotechnology Research to detect the fluorescent peaks. We determined microsatellite peaks in Geneious 6.5 (http://www.geneious.com/) using the GeneScan 600 size standard ladder for calibration[45].

RAD-Seq library preparation and data processing. We followed the double-digest RAD-Seq protocol developed by Peterson, *et al.*[46]. For each sample, we constructed 96 DNA libraries by digesting approximately 200 ng genomic DNA with *Eco*RI and *Mse*I. We then ligated Illumina adapters and unique 8–10-nucleotide barcodes to the DNA fragments. The DNA libraries were PCR-amplified in two separate reactions and pooled to minimize early PCR bias. We size selected 250–450-bp fragments using gel electrophoresis and sequenced the DNA fragments using the 1 × 100-bp setting on the Illumina HiSeq. 2500 platform. Raw sequence data were deposited in the NCBI Sequence Read Archive (accession numbers pending). We processed the raw Illumina reads using the FAST-X toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to filter sequences; we required 95% of bases to be above a quality score of 30 to retain a read. We then converted the sequences from FASTQ to FASTA, demultiplexed the reads, sorted them by barcodes, and trimmed the sequences by removing the final 2 bases to ensure that we were using only high-quality sequence data. We assembled the sequences into loci using the STACKS 1.24 pipeline[47] with the following parameter settings: -n 3 -m 3 -M 2 (parameters were optimized following Mastretta-Yanes, *et al.*[29]); all other parameters were left as the default. We selected seven datasets (one microsatellite and six RAD-Seq) and used a variety of analyses to compare the results produced by each dataset (Table 2). We used the 'populations' program in STACKS to produce an unfiltered dataset of RAD-Seq loci using the 'write single SNP' command and requiring a minor allele frequency >0.05. We then removed human, fungal, and microbial contamination from the loci and filtered loci by representation across individuals using an R script to create five smaller datasets (Data_aquisition.R; this script and all other scripts are available at https://github.com/richiehodel/red_mangrove_RAD_SSR). Filtered datasets were required to have locus data for a certain number of individuals for the given locus to be retained in the analysis; the number of individuals could range from 1–96 (Supplemental Fig. 1). The datasets were chosen such that they encompassed a wide range of loci and missing data.

Population genetic analyses. We used an R script (Basic_stats.R) and the R package 'hierfstat'[48] to calculate average F_{ST}, the inbreeding coefficient F_{IS}, H_O, and H_E for each of the seven datasets. To investigate how the number of loci affected comparisons of population genetic statistics among populations, we calculated pairwise F_{ST} (one sampling location versus all others combined) for each sampling location for each dataset using GenoDive[49] and an R script (Pairwise_Fst.R). Additionally, we calculated F_{IS} and H_O for each sampling location for each dataset to determine how measures that often inform conservation practices might be affected by the number of loci and amount of missing data. We measured how missing data were partitioned across sampling locations to verify that there were not any sampling locations with unusually high or low amounts of missing data (Table 1). Additionally, we investigated how several population genetic statistics were distributed across loci in each of the datasets (Stat_Distribution.R; Fig. 2).

Principal components and SVDQuartets. We used a principal component analysis (PCA) implemented in the R package 'SNPRelate'[50] to identify clusters of individuals in the RAD data with an R script (VCF_PCA.R) and GenoDive to run a PCA on the microsatellite data. After visualizing the initial results, we tested several ways of grouping sampling locations together based on geography. We used SVDQuartets[51] to determine genealogical relationships among individuals. This program selects the optimal topology for a quartet of taxa, and, after sampling millions of quartets, infers a phylogeny for all individuals based on choosing the quartets with the best scores and assembling them into a phylogenetic tree. We used an R script (Nexus_creation.R) to convert the output from the 'populations' program in STACKS into nexus files that could be read for the SVDQuartets analysis. For each RAD dataset, we evaluated all possible quartets and selected trees under the multispecies coalescent using QFM (Quartet Fiduccia Mattheyses) quartet assembly[52]. We used non-parametric bootstrapping (100 replicates for each dataset) to assess confidence in inferred genealogical relationships between individuals. The R script Tree_formatting.R was used to visualize and annotate the 50% majority-rule trees from SVDQuartets using the R packages 'ape'[53] and 'ggtree'[54].

Sampling loci. To test whether the number of loci or percentage of missing data for the loci used is the more important factor impacting measures of fixation, population differentiation, and heterozygosity, we randomly sampled from RAD_25198 (the RAD-Seq dataset comprising 25,198 loci) the equivalent number of loci contained in RAD_239, RAD_1180, RAD_2317, RAD_3831, and RAD_6255, respectively, and used these five sets of sampled loci in analyses. We used an R script (Subsample.R) to randomly sample loci without replacement from RAD_25198 and repeated the sampling 100 times for each dataset. We compared measures of F_{ST} calculated using the original datasets with results

calculated using the sampled loci from RAD_25198 (Fig. 5). We used bootstrapping to calculate 95% confidence intervals around F_{ST} for the original datasets and for the sets of loci sampled from RAD_25198 (Fig. 5).

Data availability. The datasets generated during the current study are available in the NCBI Genbank repository, https://www.ncbi.nlm.nih.gov/bioproject/PRJNA397667 (accession numbers SRR5918296-SRR5918355).

References

1. Guichoux, E. *et al*. Current trends in microsatellite genotyping. *Mol. Ecol. Resour.* **11**, 591–611 (2011).
2. Kalia, R. K., Rai, M. K., Kalia, S., Singh, R. & Dhawan, A. K. Microsatellite markers: an overview of the recent progress in plants. *Euphytica* **177**, 309–334 (2010).
3. Hodel, R. G. J. *et al*. The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. *Appl. Plant Sci*. **4**, (2016).
4. Seeb, J. E. *et al*. Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol. Ecol. Resour.* **11**, 1–8 (2011).
5. Gardner, M. G., Fitch, A. J., Bertozzi, T. & Lowe, A. J. Rise of the machines–recommendations for ecologists when using next generation sequencing for microsatellite development. *Mol. Ecol. Resour.* **11**, 1093–101 (2011).
6. Hodel, R. G. J. *et al*. A new resource for the development of SSR markers: Millions of loci from a thousand plant transcriptomes. *Appl. Plant Sci*. **4**, (2016).
7. Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G. & Hohenlohe, P. A. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* **17**, 81–92 (2016).
8. Baird, N. A. *et al*. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376 (2008).
9. Smith, A. M. *et al*. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res.* **38**, e142–e142 (2010).
10. Andolfatto, P. *et al*. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* **21**, 610–617 (2011).
11. Sonah, H. *et al*. An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. *PLoS One* **8**, e54603 (2013).
12. Rafalski, A. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* **5**, 94–100 (2002).
13. Arnold, B., Corbett-Detig, R. B., Hartl, D. & Bomblies, K. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* **22**, 3179–3190 (2013).
14. Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A. & Cresko, W. A. SNP discovery and genotyping for evolutionary genetics using RAD sequencing in *Methods in Molecular Biology* 157–178 (Clifton, 2012).
15. Xu, P. *et al*. Population genomic analyses from low-coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd. *Plant J.* **77**, 430–442 (2014).
16. Lowry, D. B. *et al*. Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Resour.* **17**, 142–152 (2017).
17. Davey, J. W. *et al*. Special features of RAD sequencing data: implications for genotyping. *Mol. Ecol.* **22**, 3151–64 (2013).
18. Gautier, M. *et al*. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* **22**, 3165–78 (2013).
19. Liu, N., Chen, L., Wang, S., Oh, C. & Zhao, H. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. *BMC Genet*. **6**(Suppl 1), S26 (2005).
20. Coates, B. S. *et al*. Comparative performance of single nucleotide polymorphism and microsatellite markers for population genetic analysis. *J. Hered.* **100**, 556–564 (2009).
21. Schopen, G. C. B., Bovenhuis, H., Visker, M. H. P. W. & van Arendonk, J. A. M. Comparison of information content for microsatellites and SNPs in poultry and cattle. *Anim. Genet.* **39**, 451–453 (2008).
22. Huang, H. & Knowles, L. L. Unforeseen consequences of excluding missing data from Next-Generation Sequences: Simulation study of RAD sequences. *Syst. Biol.* **65**, 357–65 (2016).
23. Catchen, J. *et al*. The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Mol. Ecol.* **22**, 2864–2883 (2013).
24. Jackson, A. M. *et al*. Population structure and phylogeography in Nassau grouper (*Epinephelus striatus*), a mass-aggregating marine fish. *PLoS One* **9**, e97508 (2014).
25. Bernardi, G., Azzurro, E., Golani, D. & Miller, M. R. Genomic signatures of rapid adaptive evolution in the bluespotted cornetfish, a Mediterranean Lessepsian invader. *Mol. Ecol.* **25**, 3384–3396 (2016).
26. Blanco-Bercial, L. & Bucklin, A. New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod *Centropages typicus*. *Mol. Ecol.* **25**, 1566–1580 (2016).
27. Rodríguez-Ezpeleta, N. *et al*. Population structure of Atlantic mackerel inferred from RAD-seq-derived SNP markers: effects of sequence clustering parameters and hierarchical SNP selection. *Mol. Ecol. Resour.* **16**, 991–1001 (2016).
28. Van Wyngaarden, M. *et al*. Identifying patterns of dispersal, connectivity and selection in the sea scallop, *Placopecten magellanicus*, using RADseq-derived SNPs. *Evol. Appl.* **10**, 102–117 (2017).
29. Mastretta-Yanes, A. *et al*. Restriction site-associated DNA sequencing, genotyping error estimation and *de novo* assembly optimization for population genetic inference. *Mol. Ecol. Resour.* **15**, 28–41 (2015).
30. Bradbury, I. R. *et al*. Transatlantic secondary contact in Atlantic Salmon, comparing microsatellites, a single nucleotide polymorphism array and restriction-site associated DNA sequencing for the resolution of complex spatial structure. *Mol. Ecol.* **24**, 5130–5144 (2015).
31. Jeffries, D. L. *et al*. Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Mol. Ecol.* **25**, 2997–3018 (2016).
32. Hodel, R. G. J., Cortez, M. B., de, S., Soltis, P. S. & Soltis, D. E. Comparative phylogeography of black mangroves (*Avicennia germinans*) and red mangroves (*Rhizophora mangle*) in Florida: Testing the maritime discontinuity in coastal plants. *Am. J. Bot.* **103**, 730–739 (2016).
33. Tomlinson, P. B. *The Botany of Mangroves* (Cambridge University Press, 2016).
34. Avise, J. C. Phylogeography: The History and Formation of Species (Harvard University Press, 2000).
35. Soltis, D., Morris, A., McLachlan, J. S., Manos, P. S. & Soltis, P. S. Comparative phylogeography of unglaciated eastern North America. *Mol. Ecol.* **15**, 4261–4293 (2006).
36. Kennedy, J. P. *et al*. Contrasting genetic effects of red mangrove (*Rhizophora mangle* L.) range expansion along West and East Florida. *J. Biogeogr.* **44**, 335–347 (2017).
37. Fu, R., Dey, D. K. & Holsinger, K. E. Bayesian models for the analysis of genetic structure when populations are correlated. *Bioinformatics* **21**, 1516–1529 (2005).
38. Rosero-Galindo, C., Gaitan-Solis, E., Cárdenas-Henao, H., Tohme, J. & Toro-Perea, N. Polymorphic microsatellites in a mangrove species, *Rhizophora mangle* L.(Rhizophoraceae). *Mol. Ecol. Notes* **2**, 281–283 (2002).
39. Kang, J., Ma, X. & He, S. Population genetics analysis of the Nujiang catfish *Creteuchiloglanis macropterus* through a genome-wide single nucleotide polymorphisms resource generated by RAD-Seq. *Sci. Rep.* **7**, 2813 (2017).

40. Manthey, J. D., Geiger, M. & Moyle, R. G. Relationships of morphological groups in the northern flicker superspecies complex (*Colaptes auratus* & *C. chrysoides*). *Syst. Biodivers.* **15**, 183–191 (2017).
41. Johnson, L.K. & Herren, L.W. Re-establishment of fringing mangrove habitat in the Indian River Lagoon 19–22 (Florida Department of Environmental Protection, 2008).
42. NASA Public Affairs. *The Kennedy Space Center Story* (Graphic House, 1991).
43. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15 (1987).
44. Schuelke, M. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* **18**, 233–234 (2000).
45. Kearse, M. *et al*. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–9 (2012).
46. Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S. & Hoekstra, H. E. Double Digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7**, (2012).
47. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: an analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–40 (2013).
48. Goudet, J. hierfstat, a package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* **5**, 184–186 (2005).
49. Meirmans, P. G. & Van Tienderen, P. H. Genotype and Genodive: two programs for the analysis of genetic diversity of asexual organisms. *Mol. Ecol. Notes* **4**, 792–794 (2004).
50. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
51. Chifman, J. & Kubatko, L. Quartet inference from SNP data under the coalescent model. *Bioinformatics* **30**, 3317–3324 (2014).
52. Reaz, R. *et al.* Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLoS One* **9**, e104008 (2014).
53. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, (289–290 (2004).
54. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**, 28–36 (2017).

Acknowledgements

We thank the following funding sources: Botanical Society of America, Florida SeaGrant, Sigma Xi, and the University of Florida Department of Biology. Publication of this article was funded in part by the University of Florida Open Access Publishing Fund.

Author Contributions

R.G.J.H., S.C. and A.C.P. collected the data, R.G.J.H., S.F.M., P.S.S. and D.E.S. contributed financially to the data collection, all authors designed analyses, R.G.J.H. wrote the manuscript, and all authors reviewed and edited the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-16810-7.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.