Title
Theory, practice, and conservation in the age of genomics: The Galápagos giant tortoise as a case study

Permalink
https://escholarship.org/uc/item/6kz1f93w

Journal
Evolutionary Applications, 11(7)

ISSN
1752-4563

Authors
Gaughran, Stephen J
Quinzin, Maud C
Miller, Joshua M
et al.

Publication Date
2018-08-01

DOI
10.1111/eva.12551

Peer reviewed

ORIGINAL ARTICLE

WILEY Evolutionary Applications Open Access

Theory, practice, and conservation in the age of genomics: The Galápagos giant tortoise as a case study

Stephen J. Gaughran[1] (ID) | Maud C. Quinzin[1] | Joshua M. Miller[1] (ID) | Ryan C. Garrick[2] | Danielle L. Edwards[3] | Michael A. Russello[4] | Nikos Poulakakis[5,6] | Claudio Ciofi[7] | Luciano B. Beheregaray[8] | Adalgisa Caccone[1]

[1]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

[2]Department of Biology, University of Mississippi, Oxford, MS, USA

[3]Life and Environmental Sciences, University of California, Merced, Merced, CA, USA

[4]Department of Biology, University of British Columbia, Okanagan Campus, Kelowna, BC, Canada

[5]Department of Biology, School of Sciences and Engineering, University of Crete, Heraklion, Crete, Greece

[6]Natural History Museum of Crete, School of Sciences and Engineering, University of Crete, Heraklion, Crete, Greece

[7]Department of Biology, University of Florence, Sesto Fiorentino (FI), Italy

[8]Molecular Ecology Lab, School of Biological Sciences, Flinders University, Adelaide, SA, Australia

Correspondence
Stephen J. Gaughran and Adalgisa Caccone, Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA.
Emails: stephen.gaughran@yale.edu, adalgisa.caccone@yale.edu

Abstract

High-throughput DNA sequencing allows efficient discovery of thousands of single nucleotide polymorphisms (SNPs) in nonmodel species. Population genetic theory predicts that this large number of independent markers should provide detailed insights into population structure, even when only a few individuals are sampled. Still, sampling design can have a strong impact on such inferences. Here, we use simulations and empirical SNP data to investigate the impacts of sampling design on estimating genetic differentiation among populations that represent three species of Galápagos giant tortoises (*Chelonoidis* spp.). Though microsatellite and mitochondrial DNA analyses have supported the distinctiveness of these species, a recent study called into question how well these markers matched with data from genomic SNPs, thereby questioning decades of studies in nonmodel organisms. Using >20,000 genomewide SNPs from 30 individuals from three Galápagos giant tortoise species, we find distinct structure that matches the relationships described by the traditional genetic markers. Furthermore, we confirm that accurate estimates of genetic differentiation in highly structured natural populations can be obtained using thousands of SNPs and 2–5 individuals, or hundreds of SNPs and 10 individuals, but only if the units of analysis are delineated in a way that is consistent with evolutionary history. We show that the lack of structure in the recent SNP-based study was likely due to unnatural grouping of individuals and erroneous genotype filtering. Our study demonstrates that genomic data enable patterns of genetic differentiation among populations to be elucidated even with few samples per population, and underscores the importance of sampling design. These results have specific implications for studies of population structure in endangered species and subsequent management decisions.

KEYWORDS
Chelonoidis, conservation, genomics, population structure, sampling design, single nucleotide polymorphism

Modern molecular techniques provide unprecedented power to understand genetic variation in natural populations. Nevertheless, application of this information requires sound understanding of population genetics theory.

- Fred Allendorf (2017, p. 420)

1 | INTRODUCTION

The advent of high-throughput DNA sequencing has enabled the characterization of the genomes of model and nonmodel organisms alike. Genomewide data can improve the precision and accuracy of estimates of population parameters, enhancing our understanding of present-day structure, gene flow, and local adaptation (Funk, McKay, Hohenlohe, & Allendorf, 2012). These data have also facilitated more detailed reconstructions of historical events that impacted evolutionary trajectories within species (e.g., Emerson et al., 2010), and among closely related species (e.g., Chaves et al., 2016).

While whole-genome sequencing is still beyond the budget of many research programs, methods based on reduced-representation genomic libraries (e.g., double-digest restriction-site associated DNA sequencing, ddRAD-seq (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012)) allow tens or hundreds of thousands of single nucleotide polymorphisms (SNPs) to be discovered and reliably genotyped at a much reduced cost (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016). This is particularly beneficial for species of conservation concern, where limited resources and sampling constraints (i.e., few individuals are available) may be prevalent. No matter the application, though, well-designed population genetics studies aim to maximize their statistical power while minimizing costs.

Genomewide SNP data are currently being applied to a broad spectrum of conservation objectives. These range from informing captive breeding programs (e.g., Wright et al., 2015) and improving detection of hybridization and inbreeding depression (e.g., vonHoldt, Kays, Pollinger, & Wayne, 2016; Robinson et al., 2016), to delineating conservation units, assessing levels of adaptive genetic variation, and predicting viability in the face of anthropogenic impacts such as climate change (Brauer, Hammer, & Beheregaray, 2016; Henry & Russello, 2013; Rellstab, Gugerli, Eckert, Hancock, & Holderegger, 2015; Sork et al., 2016). The appeal of genomic approaches to conservation biology is heightened by indications that a large number of independent loci can alleviate issues associated with small sample sizes per population; when using thousands of loci, one can obtain reliable estimates of genetic diversity and population differentiation, so long as the true values of these parameters are sufficiently high (e.g., Li & Durbin, 2011; Willing, Dreyer, & van Oosterhout, 2012). Yet, as noted by Allendorf (2017), genomic datasets need to be analyzed within the context of a carefully considered sampling design. Shortcomings in sampling design can lead to erroneous conclusions (Meirmans, 2015), which can have profound consequences for any population-level study, but especially for those with direct management implications for threatened or endangered species.

Here, we explore the power of using thousands of SNP markers to study population structure, and the impact of sampling design and small sample sizes on detecting and describing that structure. To do this, we use genomic data from Galápagos giant tortoises (*Chelonoidis* spp.) as a case study, given that a recent study has questioned the genomic distinctiveness of several species within this genus (Loire et al., 2013). The Galápagos Islands are home to a radiation of endemic giant tortoises that includes 11 endangered and four extinct species (Figure 1). Taxonomic designations are supported by differences in morphology, geographic isolation of most species, and evidence of
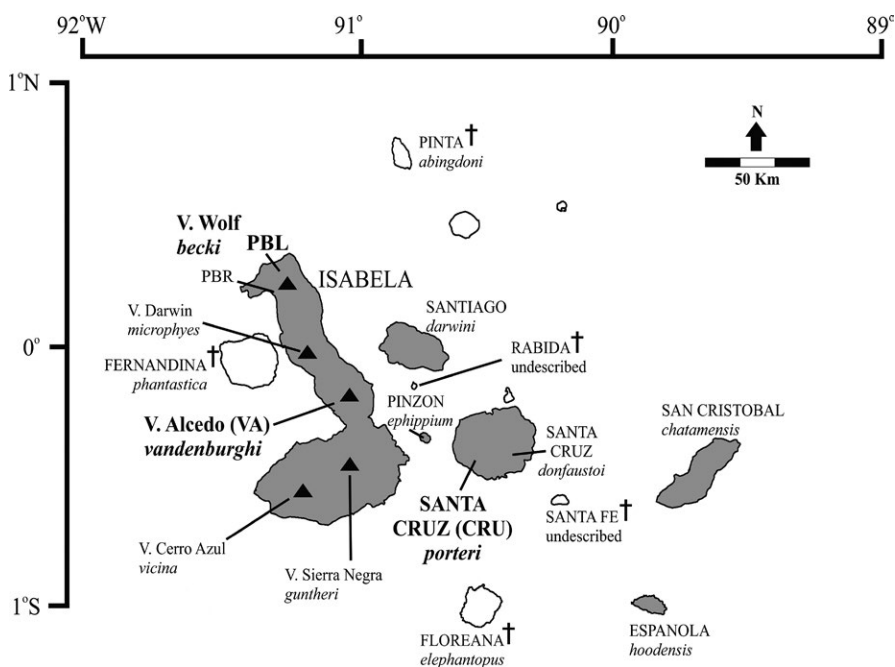
FIGURE 1 Distribution map of Galápagos giant tortoises throughout the archipelago. The islands with extant species are shown in gray, while the islands with extinct species are in white. Black triangles identify the location of the four volcanoes on Isabela Island, each with its own locally endemic tortoise species. Extinct species are identified by a cross symbol. Names of each species are in cursive with a black line pointing to the island or location within an island where they occur. The populations from the three species in this study are identified by two or three letter symbols in bold: CRU = *C. porteri*, Santa Cruz Island (La Caseta). VA = *C. vandenburghi*, Volcano Alcedo, central Isabela Island, and PBL = *C. becki*, Piedras Blancas, Volcano Wolf, northern Isabela Island

evolutionary divergence based on mitochondrial DNA (mtDNA) and nuclear microsatellite data (Ciofi, Milinkovitch, Gibbs, Caccone, & Powell, 2002; Beheregaray, Ciofi, Caccone, Gibbs, & Powell, 2003; Garrick et al., 2015; see Fig. S7a,b).

In contrast to previous studies (see section VIII in Appendix S1 for details; Ciofi et al., 2002; Beheregaray, Ciofi, Caccone, et al., 2003; Beheregaray et al., 2004; Russello et al., 2005; Poulakakis, Russello, Geist, & Caccone, 2012; Garrick et al., 2015; Poulakakis et al., 2015), Loire et al. (2013) challenged the genetic distinctiveness of three Galápagos giant tortoise species. Those authors collected transcriptome-derived genotypic data from ~1,000 synonymous SNPs from five captive individuals representing three species (*C. becki*, *C. porteri*, and *C. vandenburghi*). They did not detect significant differentiation, as measured by F_{ST}, when comparing two groups (one group of three *C. becki* individuals, and the second group consisting of two individuals, one *C. porteri*, and one *C. vandenburghi*). These two groups were constructed on what the authors identified as natural partitions, based on the observation that their samples fall into two different mtDNA clades (Fig. S6a; Poulakakis et al., 2012). Furthermore, Loire et al. (2013) did not detect homozygosity excess, as measured by F_{IT}, for which positive values would indicate population structure. Given that previous population genetic studies have largely relied upon data from mtDNA and microsatellites, such a discrepancy between these traditional markers and genomic SNPs could have wide-ranging implications, beyond the case of Galápagos giant tortoises, and therefore warrants further investigation.

In this study, we investigate the agreement of population structure analyses based on genomewide SNPs compared to those based on mtDNA sequences and microsatellite genotypes. To do this, we generated a dataset of tens of thousands of genomewide SNPs from 30 individuals representing the same three species (*C. becki*, *C. porteri*, and *C. vandenburghi*) considered by Loire et al. (2013). As these species form a recently diverged species complex, we treat each species as a population to compare against the null hypothesis that all Galápagos giant tortoises belong to a single species with one panmictic population. First, we address whether or not there is significant genomic differentiation among these three Galápagos giant tortoise species using newly generated SNPs. Then, we subsample our data to explore the effects of using only a few individuals per population and of pooling individuals from different populations on estimating genetic differentiation. From these subsampling simulations, we predict the range of F_{ST} estimates expected when using the sampling scheme of Loire et al. (2013). Finally, we re-analyze the raw RNA-seq data from Loire et al. (2013) to test our prediction.

2 | MATERIALS AND METHODS

2.1 | Sampling and sequencing

Samples were obtained during previously conducted collection expeditions (Beheregaray, Ciofi, Caccone, et al., 2003; Beheregaray, Ciofi, Geist, et al., 2003; Beheregaray et al., 2004; Caccone, Gibbs, Ketmaier, Suatoni, & Powell, 1999; Caccone et al., 2002; Ciofi et al., 2002, 2006;

Edwards, Garrick, Tapia, & Caccone, 2014; Edwards et al., 2013; Garrick et al., 2012, 2014; Poulakakis et al., 2008, 2012; Russello et al., 2005; Russello, Beheregaray, et al., 2007; Russello, Hyseni, et al., 2007). Approximately 10 samples per population for each extant species (*n* = 121 individuals in total) were selected for sequencing as part of a larger project on the phylogeography of Galápagos giant tortoises. These individuals were chosen as they displayed concordant and unambiguous genetic assignments between mitochondrial (control region, mt*CR*) and microsatellite (12 loci) ancestry based on a published database of 123 mitochondrial haplotypes (Poulakakis et al., 2012) and 305 genotyped individuals (Edwards et al., 2013) that include all the extant and extinct populations and species.

DNA was extracted from blood samples using a DNeasy Blood and Tissue kit (Qiagen) according to the manufacturer's instructions. We then prepared ddRAD libraries following Peterson et al. (2012). For each sample, 500 ng of genomic DNA was digested with the restriction enzymes *MluCI* and *NlaIII* (New England BioLabs) and ligated with Illumina-specific adaptors representing up to 18 unique barcodes and two index codes. Ligated fragments of samples were pooled into 13 libraries and size-selected to be ~310 bp (range 279–341 bp) with a BluePippin (Sage Science). Size-selected libraries included 12–24 individuals and were paired-end sequenced on 13 lanes of an Illumina HiSeq 2000 at the Yale Center for Genome Analysis.

2.2 | SNP calling

We used forward and reverse reads to generate a de novo assembly using the pyrad v.3.0.3 pipeline (Eaton, 2014). Reads were demultiplexed and assigned to each individual based on barcodes allowing for one mismatch. We replaced base calls of *Q* < 20 with an ambiguous base (*N*) and discarded sequences containing more than four ambiguities. We used 85% clustering similarity as a threshold to align the reads into loci. We set additional filtering parameters to allow for a maximum number of SNPs to be called: retaining clusters with a minimum depth of sequence coverage (Mindepth) >5 and a locus coverage (MinCov) >10, a maximum proportion of individuals with shared heterozygote sites of 20% (MaxSH = p. 20), and a maximum number of SNP per locus of 15 (maxSNP = 15). For subsequent analyses, we filtered this dataset using vcftools (Danecek et al., 2011) to generate a set of polymorphic loci (23,057 SNPs) with no missing data common to all three Galápagos giant tortoises populations of interest, abbreviated PBL, CRU, and VA and corresponding to the species *C. becki*, *C. porteri*, and *C. vandenburghi*, respectively (*n* = 10 individuals each).

2.3 | Analytical methods

F-statistics (F_{IT}, F_{IS}, global F_{ST}, and pairwise F_{ST}) were calculated using the diveRsity package in R (Keenan, McGinnity, Cross, Crozier, & Prodöhl, 2013), which uses a weighted Weir and Cockerham (1984) estimator. The same package was used to assess the statistical significance of these estimates by bootstrapping across loci. Through this method, we established 95% confidence intervals for each estimate, accepting as significant those that did not include 0. Pairwise

F_{ST} calculated from thousands of subsamples of the data (described below) were carried out in vcftools (Danecek et al., 2011) to streamline computation. We also used vcftools to calculate the number of loci out of Hardy–Weinberg equilibrium (HWE) for each population and for pooled populations.

As F_{ST} estimates rely on a priori assignment of individuals to groups that are typically based on geographic location, we used two methods that do not have this assumption to assess patterns of differentiation among our samples. To do this, we first carried out principal component analysis (PCA) on all 30 individuals, using the PLINK software (Chang et al., 2015). Principal components 1 and 2 were plotted against each other in R. To complement the multivariate analyses, we performed a Bayesian clustering analysis, implemented in the program STRUCTURE version 2.3.4 (Falush, Stephens, & Pritchard, 2003; Pritchard, Stephens, & Donnelly, 2000), also including all 30 individuals. STRUCTURE assumes a model with K unknown clusters representing genetic populations in HWE and then assigns individuals to each cluster based on allele frequencies. We ran 20 repetitions of STRUCTURE for $K = 1–5$, with a burn-in of 10,000 iterations and MCMC length of 50,000 iterations. These runs used the admixture model, correlated allele frequencies among populations, and did not assume prior population information. All other parameters were left at default values. Results were postprocessed and visualized using CLUMPAK (Kopelman, Mayzel, Jakobsson, Rosenberg, & Mayrose, 2015). We used mean log likelihood values (Pritchard et al., 2000) and the ΔK statistic (Evanno, Regnaut, & Goudet, 2005) to infer the best K (Fig. S5). Both analyses considered the 23,057 SNPs common to all individuals.

To further assess the power of our SNPs to detect population structure, we randomly subsampled individuals from each of the species and calculated pairwise F_{ST} for each species using these subsamples. We tested this for per-species sample sizes of $n = 2$, $n = 3$, and

$n = 5$. This process was repeated 1,000 times for each sample size. We also carried out a similar analysis maintaining all 10 individuals per population but randomly subsampling SNPs from our dataset. For these analyses, we used the following number of SNPs: 25, 50, 100, 200, 500, 1,000, 5,000, and 10,000. This was repeated 1,000 times for each sample size. Finally, we used a subsampling scenario that directly mimicked the one in Loire et al. (2013) to further evaluate the impact of limited sample sizes and pooling of samples from distinct species on F_{ST} estimates. As was done in Loire et al. (2013), we compared a set of three individuals from *C. becki* to a grouping that included one *C. porteri* plus one *C. vandenburghi* individual. To account for sample variation, we repeated this grouping process 1,000 times (described in full in section IV in Appendix S1).

3 | RESULTS

3.1 | Tortoise samples and ddRAD-seq dataset

Our sequencing generated a total of 3,094,399,092 retained reads (approximately 15—58 million reads per individual) after demultiplexing and filtering reads for quality and ambiguous barcodes and ddRAD-tags. de novo assembly of the data resulted in 48,004,056 ddRAD-tags (approximately 320,000–465,000 per individual). From these, we called SNPs and obtained 973,321 variable sites. We then narrowed those loci down to only loci with genotypes called in every individual in our three species dataset, for a total of 23,057 SNPs. For the three species of interest, the number of loci retained within populations and between population pairs is presented in Table 1. The average coverage per locus per individual was 12X (minimum 9; maximum 15).

3.2 | *F*-statistics using ddRAD-seq data

Calculation of *F*-statistics revealed values consistent with highly structured populations ($F_{IT} = 0.257$, 95% CI: 0.251–0.262; $F_{IS} = 0.079$, 95% CI: 0.073–0.084; and global $F_{ST} = 0.193$, 95% CI: 0.189–0.198). Using the SNPs in common to each population pair (Table 1), we found pairwise F_{ST} values of 0.169 (95% CI: 0.164–0.174) between PBL and CRU, 0.181 (95% CI: 0.175–0.187) between PBL and VA, and 0.233 (95% CI: 0.226–0.240) between CRU and VA (Table 2). These estimates were similar to, though higher than, F_{ST} estimates using 12 nuclear microsatellite markers (Garrick et al., 2015) for these species comparisons (Table 2 and Table S5).

TABLE 1 Number of polymorphic loci present in all individuals ($n = 10$ per species) used for analyses of each population (diagonal) and population pair (below diagonal)

	PBL (*C. becki*)	CRU (*C. porteri*)	VA (*C. vandenburghi*)
PBL (*C. becki*)	9,580		
CRU (*C. porteri*)	19,654	11,703	
VA (*C. vandenburghi*)	13,520	16,432	5,732

TABLE 2 Pairwise F_{ST} values between given species pairs. Above the diagonal, values calculated using our dataset of SNPs with no missing data and common to the population pair, along with 95% confidence intervals. Below the diagonal, values calculated using 12 microsatellite loci from Garrick et al. (2015) (see section VIII in Appendix S1). Data were obtained using 10 samples for each population (PBL, VA, CRU) for the three species

	PBL (*C. becki*)	CRU (*C. porteri*)	VA (*C. vandenburghi*)
PBL (*C. becki*)		0.169 (0.164–0.174)	0.181 (0.175–0.187)
CRU (*C. porteri*)	0.137		0.233 (0.226–0.240)
VA (*C. vandenburghi*)	0.163	0.202	

3.3 | PCA and STRUCTURE

The first two principal components of the PCA showed clear differentiation among individuals from the three species. PC1 accounted for approximately 12.0% of the variation among individuals, and PC2 accounted for approximately 9.3% of the variation among individuals (Figure 2). Similarly, both mean log likelihood values (Pritchard et al., 2000) and the ΔK statistic (Evanno et al., 2005) supported the existence of three distinct genetic units in the STRUCTURE analysis (Fig. S5). These groups correspond to the a priori geographic groupings used in F_{ST} estimates and to the three named species. Our separate analysis of loci out of HWE, the basis for the STRUCTURE algorithm, supported these findings as well. When each species was considered separately, out of 23,057 loci PBL showed 214 out of HWE, CRU showed 124 out of HWE, and VA showed 71 out of HWE. When the CRU and VA samples were pooled, the number of loci out of HWE rose to 1,326. When all three species were pooled and treated as one population, 2,422 loci were found to be out of HWE.

3.4 | Sample size, number of loci, and the effect of individual samples

In all population comparisons for the three sample sizes (n = 2, 3, or 5), the majority of estimates were within 0.03 of the F_{ST} value calculated using the complete dataset of 10 samples per population (Figure 3 and Figure S2). In every case, when the sample size was two, F_{ST} tended to be underestimated, though with a long tail of overestimated outliers.
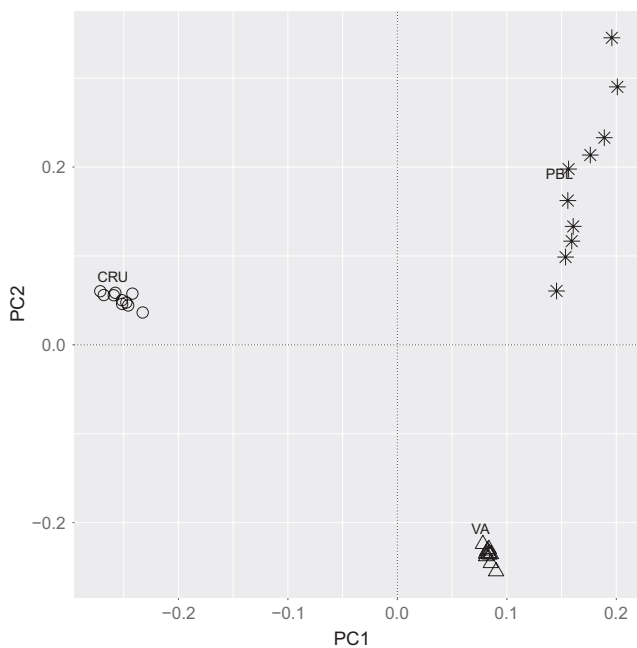
FIGURE 2 Principal component 1 (PC1) plotted against principal component 2 (PC2) for 30 individuals from three populations, resulting from PCA analysis on 23,057 SNPs. Stars, open circles, and open triangles identify individuals from the PBL (_C. becki_), CRU (_C. porteri_), and VA (_C. vandenburghi_) populations, respectively. The analysis was carried out using PLINK (Chang et al., 2015)

In all comparisons with sample sizes of three or five, this skew disappeared: We found that 95% of the estimates were within 0.05 of the estimate using 10 samples (Tables S1).

Our F_{ST} estimates from subsampled SNPs ranging from 25 to 10,000 SNPs appeared to have the statistical power to detect population structure between these population pairs when 10 individuals were used, with 95% of all estimates above 0 (Table S2A–C and Figure S3). However, as expected, with many fewer SNPs the range of 95% of the estimates was very wide (see section III in Appendix S1). For example, when only 100 SNPs were used to compare PBL and CRU, 95% of the F_{ST} estimates were between 0.1 and 0.255, while using 1,000 SNPs gave 95% of the F_{ST} estimates between 0.146 and 0.194 for the same comparison (Table S2A). We estimated nearly identical Fst values when using all SNPs common to a given population pair and when using the set of SNPs common to all three populations (Table S3).

3.5 | Effect of pooling samples

To test how pooling samples affected _F_-statistic estimates, we used the Loire et al. (2013) sampling design, pooling one individual from _C. porteri_ and one individual from _C. vandenburghi_ into one population and comparing this to three individuals from _C. becki_. When the set of common SNPs (n = 23,057) were included in the analysis, the F_{ST} estimates between 1000 pairs of these groups ranged from 0.045 to 0.136 (95%: 0.052–0.127, mean: 0.075). When only 1,000 SNP loci were used, as in Loire et al. (2013), the F_{ST} estimates ranged from 0.006 to 0.157 (95%: 0.031–0.134, mean: 0.076) (see Fig. S4A,B). This confirms that pooling samples from two populations, each representing different species, results in a strongly depressed F_{ST} estimate. However, these simulations highlight that the occurrence of genetic differentiation (i.e., positive F_{ST} values) should still be detectable even with this grouping scheme.

3.6 | Re-analysis of Loire et al. transcriptome data

Given that our analyses of ddRAD-seq data showed clear genetic structure among the populations from the three species, and our subsampling simulations (Fig. S4) predicted that positive F_{ST} values should still be detectable using the grouping scheme adopted by Loire et al. (2013), we re-analyzed the original RNA-seq data generated for that publication to further assess the source of the discrepancy. We downloaded the publically available RNA sequencing data generated by Loire et al. (2013) from the NCBI's Sequence Read Archive and recalled SNPs after aligning these reads to a draft genome assembly of a closely related species of Galápagos giant tortoise, _C. abingdonii_ (unpublished data; see methods in section VII in the Appendix S1). With these transcriptome-derived SNP data, we estimated an F_{ST} of 0.054 (95% CI: 0.049–0.058) when comparing the three _C. becki_ samples (PBL) to the combined two _C. porteri_ and _C. vandenburghi_ samples (CRU and VA). Notably, this F_{ST} value falls within our predicted range of F_{ST} estimates generated by subsampling the ddRAD-seq data. Our F_{IT} estimate for this dataset was −0.121 (95% CI: −0.129 to −0.113), with F_{IS} estimated to be −0.185 (95% CI: −0.192 to 0.177).
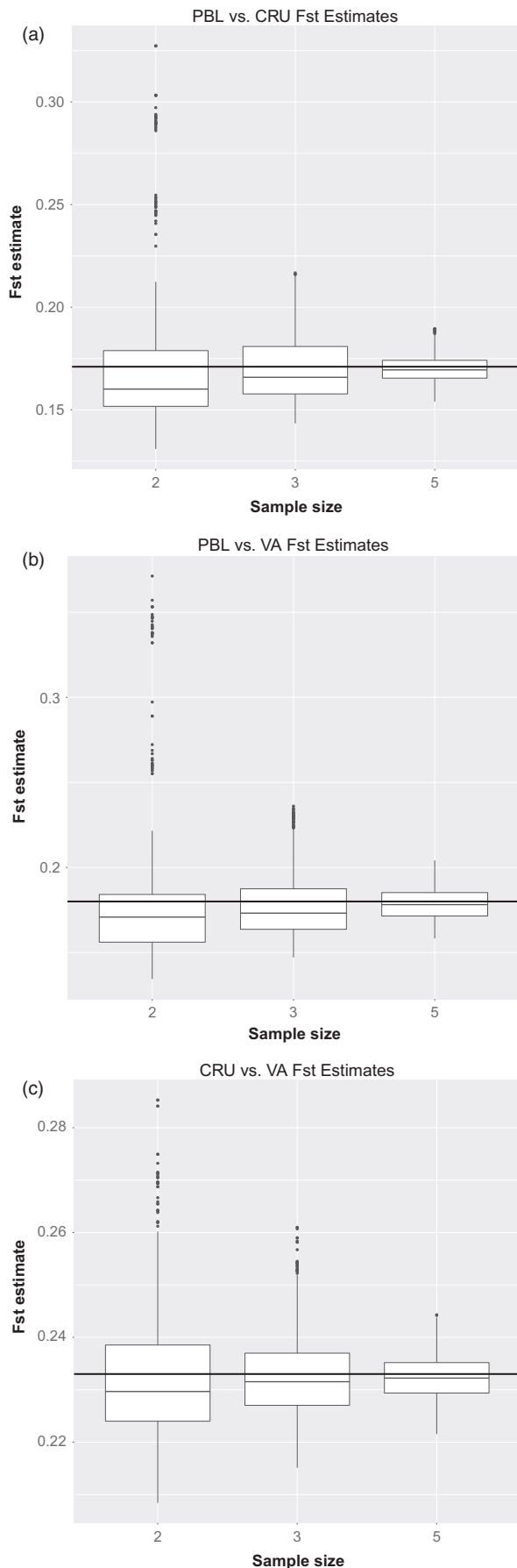
FIGURE 3 Boxplots of pairwise F_{ST} estimates using 1,000 randomly drawn subsamples of individuals for each sample size (n = 2, 3, or 5) from each population. PBL, CRU and VA correspond to population samples from *C. becki*, *C. porteri*, and *C. vandenburghi*, respectively. (a) is the pairwise comparison of PBL and CRU, (b) is the pairwise comparison of PBL and VA, and (c) is the pairwise comparison of CRU and VA. The horizontal black line in each boxplot marks the F_{ST} value calculated using all 10 individuals from each population in the pairwise comparison (see Table S1). Lower hinge corresponds to first quartile (25th percentile); upper hinge corresponds to third quartile (75th percentile). Whiskers indicate points within 1.5 times the interquartile range (IQR), with outliers indicated as points beyond that range

Plotting the first two principal components of a PCA of these five samples showed clear clustering of the conspecific samples from *C. becki*, while the single samples from *C. vandenburghi* (VA) and *C. porteri* (CRU) are distinct from each other and from the *C. becki* samples (Fig. S6).

4 | DISCUSSION

4.1 | Strong evidence of population structure

Using genomewide SNP data we found evidence for significant differentiation among the three species considered (*C. becki*, *C. porteri*, and *C. vandenburghi*), consistent with the findings of decades of research in this system (Beheregaray, Ciofi, Caccone, et al., 2003; Beheregaray, Ciofi, Geist, et al., 2003; Beheregaray et al., 2004; Ciofi et al., 2002; Garrick et al., 2015; Poulakakis et al., 2008, 2012, 2015; Russello et al., 2005; Russello, Beheregaray, et al., 2007; Russello, Hyseni, et al., 2007). Our estimate of F_{IT} (0.257), which was a focal metric used in the previous study (Loire et al., 2013), was positive and significantly different from zero. Positive values of F_{IT} indicate an excess of homozygous loci in the sample set. This could suggest the existence of population structure in the total sample set. This possibility is reinforced by the finding of very high and significantly different from zero F_{ST} estimates for the same comparisons (between 0.17 and 0.24; Table 2, Fig. S1). Interpreting significantly positive F_{IS} values, such as the one calculated from our ddRAD-seq dataset, can be difficult (Allendorf & Luikart, 2009). This could be due to substructure within one or more populations, sampling stochasticity, and/or recent demographic changes in relatively small populations. It could also be that such small populations are not necessarily expected to be in HWE due to the increased influence of genetic drift (Allendorf & Luikart, 2009).

To assess whether there was additional genetic structure outside of our a priori assignment of individuals based on their geographic location, we also analyzed the 30 samples in our ddRAD-seq dataset using two methods without prior assignment of each sample to a group. Both principal component (Figure 2) and Bayesian clustering analyses (Fig. S5) clearly discerned three genetically distinct clusters that corresponded to the samples from the three species tested in our pairwise F_{ST} estimates. This echoed our per-locus analysis of HWE, which showed that treating all 30 individuals from the three named species as a single population dramatically increased the number of loci out of HWE.

Results of our analyses of population structure using tens of thousands of genomewide SNPs are concordant with earlier studies using mtDNA haplotypes and microsatellite genotypes (Beheregaray, Ciofi, Caccone, et al., 2003; Beheregaray, Ciofi, Geist, et al., 2003; Beheregaray et al., 2004; Ciofi et al., 2002; Garrick et al., 2015; Poulakakis et al., 2008, 2012, 2015; Russello et al., 2005; Russello, Beheregaray, et al., 2007; Russello, Hyseni, et al., 2007). These findings definitively resolve concerns raised by Loire et al. (2013) regarding whether these traditional markers were accurately reflecting the genetic distinctiveness of Galápagos giant tortoise species. Importantly, our results not only revealed the same genetic clustering as earlier studies, but also showed the same patterns of genetic distance. As in the microsatellite studies, we found slightly greater genetic differentiation between *C. becki* and *C. vandenburghi* (PBL and VA: F_{ST} = 0.181) than between *C. becki* and *C. porteri* (PBL and CRU: F_{ST} = 0.169), and the greatest differentiation between *C. porteri* and *C. vandenburghi* (CRU and VA: F_{ST} = 0.233) (Table 2). While qualitatively the same, our F_{ST} estimates are notably higher than those calculated using microsatellites (Table 2), a finding predicted by the mathematics of using biallelic vs. multiallelic loci (Putman & Carbone, 2014), which has also been found in other systems (e.g., Payseur & Jing, 2009).

4.2 | Impact of sample size and number of loci on detecting population structure

Population genetic theory (Nei, 1978), simulations (Willing et al., 2012), and empirical work (Reich, Thangaraj, Patterson, Price, & Singh, 2009) support the idea that a dataset of thousands of loci should have the power to detect population structure with high precision, even when only a few individuals per population are analyzed. We tested this idea with our Galápagos giant tortoise ddRAD-seq SNP data by estimating F_{ST} from subsamples of two, three, and five individuals from each population and comparing them to the same estimates obtained from 10 individuals per population. All tested sample sizes were able to detect significant F_{ST} values, though using three or five samples yielded more precise estimates than using only two (Figure 3; Tables S1). These analyses are consistent with the idea that accurate F_{ST} values can be estimated using as few as two or three samples per population if thousands of SNPs are analyzed. Likewise, we found that for highly differentiated populations such as those studied here, hundreds of SNPs were sufficient to accurately describe population structure when ten individuals per population were used. This empirical evidence should be helpful in the design of future conservation genetics studies that aim to describe population structure, in which case additional samples may lead to diminishing returns for improving statistical power. This will be especially useful for endangered or elusive species for which sampling may present a severe limitation.

4.3 | Sampling design matters

Our genomewide SNP data detected high and significant differentiation among these three species, even when only two or three

individuals from each were used in the analysis (Figure 3). While these results were strongly supported, they failed to explain the discrepancy described by Loire et al. (2013), who used over 1,000 synonymous SNPs from transcriptome sequencing data and found no differentiation between the same three species. Their sample size of five captive individuals does not by itself account for the discrepancy between the two studies, because, as we show above (Fig. S4), using thousands of SNPs should give sufficient power to detect population structure in Galápagos giant tortoises, even when sample size is that small.

Instead, sampling design, and specifically grouping of individuals into inappropriate population units, rather than sample size likely biased the statistical power of Loire et al.'s (2013) study. Their sampling scheme divided the five individuals into two groups, which did not reflect the population divergence of the three species. Specifically, this mixed group included two individuals, each from different species (CRU, *C. porteri* from Santa Cruz Island; VA, *C. vandenburghi* from central Isabela Island), and another group of three individuals from the other species (PBL, *C. becki* from northern Isabela Island). The justification for this grouping was based on the closer phylogenetic relationship of mtDNA haplotypes from *C. porteri* and *C. vandenburghi* (Caccone et al., 1999; Russello, Beheregaray, et al., 2007; Russello, Hyseni, et al., 2007) compared to haplotypes found in the PBL *C. becki* population. This choice is problematic for several reasons (detailed in the Appendix S1 section VIII). Most importantly, *F*-statistics are a reflection of population differentiation, not of phylogenetic relatedness. Treating the individuals from *C. porteri* and *C. vandenburghi* as belonging to the same population biased the *F*-statistics estimates by leading to an increase in within-group variation, and therefore depressed F_{ST} values. This within-group structure, which distorts *F*-statistics, is known as Wahlund effect (Wahlund, 1928).

The problem outlined above is clear in our pairwise analysis using >20,000 SNPs, which shows that while the *C. becki* population sample is about equally differentiated from the *C. porteri* and *C. vandenburghi* ones, the ones from *C. porteri* and *C. vandenburghi* are more differentiated from each other than from the *C. becki* population sample (Table 2). To empirically test for the Wahlund effect under this sampling scheme, we simulated a scenario in which three samples from *C. becki* were compared to a population consisting of one *C. porteri* and one *C. vandenburghi* sample. Repeating this sampling scenario 1,000 times, we found significantly depressed mean F_{ST} estimates, as low as 0.075, with 95% of comparisons ranging from 0.052 to 0.127 (Fig. S4a). Even more strikingly, when we limited the analysis to a similar number of markers as Loire et al. (2013) and used 1,000 randomly drawn SNPs, the range of 95% of the estimates increased to 0.031–0.134.

4.4 | RNA-seq data support population structure

While our subsampling simulations showed a clear Wahlund effect when samples from two different species (*C. porteri* and *C. vandenburghi*) were combined into one grouping, these F_{ST} estimates were still positive (mean F_{ST} = 0.075). We therefore would have expected Loire et al. (2013) to find a similar estimate in their analysis of RNA-seq

data, but they reported no significantly positive F_{ST} value. To investigate this discrepancy, we re-analyzed their raw sequencing data by aligning it to a Galápagos giant tortoise reference genome. Using the SNPs from this re-analysis, we estimated an F_{ST} of 0.054, which is similar to our expected F_{ST} under their sampling design (Fig. S4). Our estimates of F_{IS} and F_{IT} for the RNA-seq dataset were negative, a surprising result that may be related to the sampling design, the specific individuals included in that study, or the deviations from HWE that can occur in small populations (Kimura & Crow, 1963). This last point is due to the assumption of large numbers in HWE, which is violated in small populations (Allendorf & Luikart, 2009).

Convincingly, a PCA of Loire et al.'s (2013) SNP data revealed a tight cluster of the three PBL samples, whereas the CRU and VA samples were distinct both from the PBL cluster and from each other (Fig. S6). This pattern of principal components mirrors the one that we found with our 30 sample dataset for the same populations (Figure 2). These results, which match our expectations based on subsampling simulations (Fig. S4), suggest that the lack of significantly positive F_{ST} values found by Loire et al. (2013) is due not just to small sample size and inappropriate grouping of samples, but also the genotype filters employed in their initial analysis. The original Loire et al. (2013) methods describe a genotype filter that assigns posterior probability to genotypes based on HWE. We suspect that this may not be a reliable method when genotyping a pool of individuals from different species, as these samples will not meet the assumption of HWE. Our SNP calls of their data may have also been improved by mapping the RNA sequence reads to a draft Galápagos giant tortoise reference genome, as suggested by others (Shafer et al., 2016). However, our ddRAD-seq SNP data were called without mapping to a reference, so this methodological difference cannot completely explain the loss of signal.

5 | CONCLUSIONS

Reduced-representation sequencing offers practical ways to take advantage of the power of population genomics, even when samples and funds are limited (Narum, Buerkle, Davey, Miller, & Hohenlohe, 2013). Yet, thoughtful study design remains an essential component. Our analyses clearly showed that tortoises representing each of three named species exhibit high genetic differentiation at the genomic level, as demonstrated through high and significant F_{ST}, and positive F_{IT} estimates, as well as through principal component and Bayesian clustering analyses. Using thousands of SNPs gives high statistical power to detect population structure even when sample sizes of individuals are as few as two or three individuals. However, the heterogeneity of samples within a population can confound calculations using small sample sizes in unpredictable ways. Reduced sample size also limits the diversity of analyses that can be performed, especially limiting those that do not rely on a priori population designation, such as PCA and Bayesian clustering algorithms. Ultimately, we found that both our ddRAD-seq data and a re-analysis of RNA-seq data generated by Loire et al. (2013) were

consistent with the findings of earlier microsatellite and mtDNA studies. We therefore expect genomewide SNPs to support the conclusions of population genetic studies of Galápagos giant tortoises beyond the three species considered here.

Distinguishing populations and evolutionary lineages, such as the giant tortoise species analyzed here, is a vital role for population genetic analyses to play in conservation (Funk et al., 2012). Results from such analyses can assist in protected area designation (Larson et al., 2014), inform appropriate legal protections (vonHoldt, Cahill, et al., 2016), and guide captive breeding strategies (de Cara, Fernández, Toro, & Villanueva, 2011; Lew et al., 2015). We show that, as long as population genetics theory is carefully taken into account, the use of genomewide data enabled by high-throughput sequencing can be a powerful tool in these conservation efforts, even when sample sizes are limited.

ORCID

Stephen J. Gaughran http://orcid.org/0000-0002-9413-5074

Joshua M. Miller http://orcid.org/0000-0002-4019-7675

REFERENCES

Allendorf, F. W. (2017). Genetics and the conservation of natural populations: Allozymes to genomes. *Molecular Ecology*, 26, 420–430.

Allendorf, F. W., & Luikart, G. (2009). *Conservation and the genetics of populations*. Malden, MA: Blackwell Publishing Inc.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17, 81–92.

Beheregaray, L. B., Ciofi, C., Caccone, A., Gibbs, J. P., & Powell, J. R. (2003). Genetic divergence, phylogeography and conservation units of giant tortoises from Santa Cruz and Pinzon, Galapagos Islands. *Conservation Genetics*, 4, 31–46.

Beheregaray, L. B., Ciofi, C., Geist, D., Gibbs, J. P., Caccone, A., & Powell, J. R. (2003). Genes record a prehistoric volcano eruption in the Galapagos. *Science*, 302, 75.

Beheregaray, L. B., Gibbs, J. P., Havill, N., Fritts, T. H., Powell, J. R., & Caccone, A. (2004). Giant tortoises are not so slow: Rapid diversification and biogeographic consensus in the Galapagos. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 6514–6519.

Brauer, C. J., Hammer, M. P., & Beheregaray, L. B. (2016). Riverscape genomics of a threatened fish across a hydroclimatically heterogeneous river basin. *Molecular Ecology*, 25, 5093–5113.

Caccone, A., Gentile, G., Gibbs, J. P., Fritts, T. H., Snell, H. L., Betts, J., & Powell, J. R. (2002). Phylogeography and history of giant Galapagos tortoises. *Evolution*, 56, 2052–2066.

Caccone, A., Gibbs, J. P., Ketmaier, V., Suatoni, E., & Powell, J. R. (1999). Origin and evolutionary relationships of giant Galapagos tortoises. *Proceedings of the National Academy of Sciences of the United States of America*, 96, 13223–13228.

de Cara, M. Á. R., Fernández, J., Toro, M. A., & Villanueva, B. (2011). Using genome-wide information to minimize the loss of diversity in conservation programmes. *Journal of Animal Breeding and Genetics*, 4, 7. https://doi.org/10.1186/s13742-015-0047-8

Chang, C., Chow, C., Tellier, L., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4, 7. https://doi.org/10.1186/s13742-015-0047-8

Chaves, J. A., Cooper, E. A., Hendry, A. P., Podos, J., De León, L. F., Raeymaekers, J. A. M., ... Uy, J. A. C. (2016). Genomic variation at the tips of the adaptive radiation of Darwin's finches. *Molecular Ecology*, 25, 5282–5295.

Ciofi, C., Milinkovitch, M. C., Gibbs, J. P., Caccone, A., & Powell, J. R. (2002). Microsatellite analysis of genetic divergence among populations of giant Galapagos tortoises. *Molecular Ecology*, 11, 2265–2283.

Ciofi, C., Wilson, G. A., Beheregaray, L. B., Marquez, C., Gibbs, J. P., Tapia, W., ... Powell, J. R. (2006). Phylogeographic history and gene flow among giant Galapagos tortoises on southern Isabela Island. *Genetics*, 172, 1727–1744.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27, 2156–2158.

Eaton, D. A. R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, 30, 1844–1849.

Edwards, D. L., Benavides, E., Garrick, R. C., Gibbs, J. P., Russello, M. A., Dion, K. B., ... Caccone, A. (2013). The genetic legacy of Lonesome George survives: Giant tortoises with Pinta Island ancestry identified in Galapagos. *Biological Conservation*, 157, 225–228.

Edwards, D. L., Garrick, R. C., Tapia, W., & Caccone, A. (2014). Cryptic structure and niche divergence within threatened Galápagos giant tortoises from southern Isabela Island. *Conservation Genetics*, 15, 1357–1369.

Emerson, K. J., Merz, C. R., Catchen, J. M., Hohenlohe, P. A., Cresko, W. A., Bradshaw, W. E., Holzapfel, C. M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 16196–16200.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, 14, 2611–2620.

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

Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. *TREE*, 27, 489–496.

Garrick, R. C., Benavides, E., Russello, M. A., Gibbs, J. P., Poulakakis, N., Dion, K. B., ... Caccone, A. (2012). Genetic rediscovery of an "extinct" Galapagos giant tortoise species. *Current Biology*, 22, R10–R11.

Garrick, R. C., Benavides, E., Russello, M. A., Hyseni, C., Edwards, D. L., Gibbs, J. P., ... Caccone, A. (2014). Lineage fusion in Galapagos giant tortoises. *Molecular Ecology*, 23, 5276–5290.

Garrick, R. C., Kajdacsi, B., Russello, M. A., Benavides, E., Hyseni, C., Gibbs, J. P., Tapia, W., Caccone, A. (2015). Naturally rare versus newly rare: Demographic inferences on two timescales inform conservation of Galápagos giant tortoises. *Ecology and Evolution*, 5, 676–694.

Henry, P., & Russello, M. A. (2013). Adaptive divergence along environmental gradients in a climate-change-sensitive mammal. *Ecology and Evolution*, 3, 3906–3917.

vonHoldt, B. M., Cahill, J. A., Fan, Z., Gronau, I., Robinson, J., Pollinger, J. P., ... Wayne, R. K. (2016). Whole-genome sequence analysis shows that two endemic species of North American wolf are admixtures of the coyote and gray wolf. *Science Advances*, 2, e1501714.

vonHoldt, B. M., Kays, R., Pollinger, J. P., & Wayne, R. K. (2016). Admixture mapping identifies introgressed genomic regions in North American canids. *Molecular Ecology*, 25, 2443–2453.

Keenan, K., McGinnity, P., Cross, T. F., Crozier, W. W., & Prodöhl, P. A. (2013). diveRsity: An R package for the estimation and exploration of population genetics parameters and their associated errors. *Methods in Ecology and Evolution*, 4, 782–788.

Kimura, M., & Crow, J. F. (1963). The measurement of effective population number. *Evolution*, 17, 279–288.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, 15, 1179–1191.

Larson, W. A., Seeb, L. W., Everett, M. V., Waples, R. K., Templin, W. D., Seeb, J. E. (2014). Genotyping by sequencing resolves shallow population structure to inform conservation of Chinook salmon (*Oncorhynchus tshawytscha*). *Evolutionary Applications*, 7, 355–369.

Lew, R. M., Finger, A. J., Baerwald, M. R., Goodbla, A., May, B., & Meek, M. H. (2015). Using next-generation sequencing to assist a conservation hatchery: A single-nucleotide polymorphism panel for the genetic management of endangered delta smelt. *Transactions of the American Fisheries Society*, 144, 767–779.

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, 475, 493–496.

Loire, E., Chiari, Y., Bernard, A., Cahais, V., Romiguier, J., Nabholz, B., Lourenço, J. M., & Galtier, N. (2013). Population genomics of the endangered giant Galapagos tortoise. *Genome Biology*, 14, R136.

Meirmans, P. G. (2015). Seven common mistakes in population genetics and how to avoid them. *Molecular Ecology*, 24, 3223–3231.

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, 22, 2841–2847.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, 89, 583–590.

Payseur, B. A., & Jing, P. (2009). A genomewide comparison of population structure at STRPs and nearby SNPs in humans. *Molecular Biology and Evolution*, 26, 1369–1377.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One*, 7, e37135.

Poulakakis, N., Edwards, D. L., Chiari, Y., Garrick, R. C., Russello, M. A., Benavides, E., … Caccone, A. (2015). Description of a new Galapagos giant tortoise species (Chelonoidis; Testudines: Testudinidae) from Cerro Fatal on Santa Cruz Island. *PLoS One*, 10, e0138779.

Poulakakis, N., Glaberman, S., Russello, M., Beheregaray, L. B., Ciofi, C., Powell, J. R., & Caccone, A. (2008). Historical DNA analysis reveals living descendants of an extinct species of Galapagos tortoise. *Proceedings of the National Academy of Sciences of the United States of America*, 105, 15464–15469.

Poulakakis, N., Russello, M., Geist, D., & Caccone, A. (2012). Unravelling the peculiarities of island life: Vicariance, dispersal and the diversification of the extinct and extant giant Galapagos tortoises. *Molecular Ecology*, 21, 160–173.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155, 945–959.

Putman, A. I., & Carbone, I. (2014). Challenges in analysis and interpretation of microsatellite data for population genetic studies. *Ecology and Evolution*, 4, 4399–4428.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, 461, 489–494.

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, 24, 4348–4370.

Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., von Holdt, B. M., Marsden, C. D., Lohmueller, K. E., & Wayne, R. K. (2016). Genomic flatlining in the endangered island fox. *Current Biology*, 26, 1183–1189.

Russello, M. A., Beheregaray, L. B., Gibbs, J. P., Fritts, T., Havill, N., Powell, J. R., & Caccone, A. (2007). Lonesome George is not alone among Galápagos tortoises. *Current Biology*, 17, R317–R318.

Russello, M. A., Glaberman, S., Gibbs, J. P., Marquez, C., Powell, J. R., & Caccone, A. (2005). A cryptic taxon of Galapagos tortoise in conservation peril. *Biology Letters*, 1, 287–290.

Russello, M. A., Hyseni, C., Gibbs, J. P., Cruz, S., Marquez, C., Tapia, W., … Caccone, A. (2007). Lineage identification of Galapagos tortoises in captivity worldwide. *Animal Conservation*, 10, 304–311.

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2016). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, 8, 907–917.

Sork, V. L., Squire, K., Gugger, P. F., Steele, S. E., Levy, E. D., Eckert, A. J. (2016). Landscape genomic analysis of candidate genes for climate adaptation in a California endemic oak, *Quercus lobata*. *American Journal of Botany*, 103, 33–46.

Wahlund, S. (1928). Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. *Hereditas*, 11, 65–106.

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38, 1358–1370.

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.

Wright, B., Morris, K., Grueber, C. E., Willet, C. E., Gooley, R., Hogg, C. J., … Belov, K. (2015). Development of a SNP-based assay for measuring genetic diversity in the Tasmanian devil insurance population. *BMC Genomics*, 16, 791. https://doi.org/10.1186/s12864-015-2020-4

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.