MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2017) doi: 10.1111/1755-0998.12654

Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species

ALISON G. NAZARENO,* JORDAN B. BEMMELS,† CHRISTOPHER W. DICK† and LÚCIA G. LOHMANN*
*Departamento de Botânica, Universidade de São Paulo, Rua do Matão 277, Cidade Universitária, CEP 05508-900, São Paulo, São Paulo, Brazil, †Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor MI 48109, USA

Abstract

High-throughput DNA sequencing facilitates the analysis of large portions of the genome in nonmodel organisms, ensuring high accuracy of population genetic parameters. However, empirical studies evaluating the appropriate sample size for these kinds of studies are still scarce. In this study, we use double-digest restriction-associated DNA sequencing (ddRADseq) to recover thousands of single nucleotide polymorphisms (SNPs) for two physically isolated populations of *Amphirrhox longifolia* (Violaceae), a nonmodel plant species for which no reference genome is available. We used resampling techniques to construct simulated populations with a random subset of individuals and SNPs to determine how many individuals and biallelic markers should be sampled for accurate estimates of intra- and interpopulation genetic diversity. We identified 3646 and 4900 polymorphic SNPs for the two populations of *A. longifolia*, respectively. Our simulations show that, overall, a sample size greater than eight individuals has little impact on estimates of genetic diversity within *A. longifolia* populations, when 1000 SNPs or higher are used. Our results also show that even at a very small sample size (i.e. two individuals), accurate estimates of F_{ST} can be obtained with a large number of SNPs (≥1500). These results highlight the potential of high-throughput genomic sequencing approaches to address questions related to evolutionary biology in nonmodel organisms. Furthermore, our findings also provide insights into the optimization of sampling strategies in the era of population genomics.

Keywords: *Amphirrhox longifolia*, ddRADseq, genetic diversity, single nucleotide polymorphism, Violaceae

Received 8 September 2016; revision received 4 January 2017; accepted 9 January 2017

Introduction

Population genetic studies are generally based on statistical rules of thumb that guide sample size selection. Several studies have sampled as many as 20–30 individuals per population to estimate genetic parameters (Luikart & Cornuet 1998; Ward & Jasieniuk 2009; Hale *et al.* 2012), while others have sampled as many individuals as possible (Hobas *et al.* 2013). However, sample sizes are per se a critical issue in evolutionary studies, leading to ambiguous, inconclusive or negative results when sampling is limited (Swatdipong *et al.* 2010; Nazareno & Jump 2012; Hobas *et al.* 2013). The establishment of an adequate sampling scheme has been problematic for a variety of molecular markers (Luikart & Cornuet 1998; Koskinen *et al.* 2004; Kalinowski 2005; Ward & Jasieniuk 2009; Hale *et al.* 2012; Hobas *et al.* 2013; Jeffries *et al.* 2016). A simulation study has shown that accurate estimates of population differentiation can be obtained from relatively small sample sizes using large numbers of SNPs (Willing *et al.* 2012). A single empirical study to date has determined the power of the reduction in the number of samples when SNPs or microsatellite markers are used (Jeffries *et al.* 2016). However, no other empirical population genomic studies have been able to define optimal sampling strategies (i.e. number of individuals and molecular markers) when thousands of biallelic single nucleotide polymorphism (SNP) molecular markers are employed.

High-throughput sequencing technologies that employ restriction enzymes to produce reduced representations of genomes [e.g. complexity reduction of polymorphic sequences (CRoPS), restriction-site-associated DNA sequencing (RADseq), multiplexed shotgun genotyping (MSG) and genotyping by sequencing (GBS), see Davey *et al.* 2011 for a review] are enabling us to discover, sequence and genotype a high number of SNPs

Correspondence: Alison Gonçalves Nazareno, Fax: +55 11 30917547; E-mail: alison_nazareno@yahoo.com.br and Lúcia G. Lohmann, Fax: +55 11 30917547; E-mail: llohmann@usp.br

for model (e.g. Ramos *et al.* 2009; Mammadov *et al.* 2010; Davey *et al.* 2011; Uitdewiligen *et al.* 2013) and nonmodel organisms (e.g. Emerson *et al.* 2010; Helyar *et al.* 2011; Hohenlohe *et al.* 2011; Deagle *et al.* 2015; Andrews *et al.* 2016). These new techniques have allowed us to address a range of evolutionary questions (e.g. Emerson *et al.* 2010; Hohenlohe *et al.* 2010; Catchen *et al.* 2013; Lozier 2014; Deagle *et al.* 2015; Martin *et al.* 2016; Ozerov *et al.* 2016; Vera *et al.* 2016), including the establishment of adequate sampling schemes (Willing *et al.* 2012). In general, larger sample sizes are thought to be better (Ryman & Palm 2006). However, sequencing large numbers of individuals per population using high-density SNP-based genome can be overkill (Morin *et al.* 2004; Willing *et al.* 2012; Jeffries *et al.* 2016), inflating costs and analytical time.

Previous attempts to determine adequate sampling strategies employing a variety of molecular markers (Ryman *et al.* 2006; Hale *et al.* 2012; Willing *et al.* 2012; Gonzáles-Ramos *et al.* 2015; Jeffries *et al.* 2016) have analysed allele frequencies (Hale *et al.* 2012; Willing *et al.* 2012) and used regression models (Bashalkhanov *et al.* 2009), or resampling techniques (Koskinen *et al.* 2004; Gonzáles-Ramos *et al.* 2015). Willing *et al.* (2012) demonstrated that small sample sizes are enough to assess interpopulation divergence when thousands of biallelic SNPs markers are used. While Willing *et al.* (2012) deserve credit for being an essential study on population genomics, addressing sampling issues based on simulations may not always hold true in nature where model assumptions are often violated (Koskinen *et al.* 2004). Thus, empirical studies evaluating the effect of sample size (i.e. minimum number of individuals and loci) are greatly needed to estimate the levels of genetic differentiation and genetic diversity within natural populations.

Here, we empirically estimate for the first time the adequate sampling size for population genomics studies. More specifically, we determine how many individuals and how many biallelic SNPs are needed to accurately estimate intrapopulation genetic diversity parameters such as the effective number of alleles, observed and expected heterozygosity, and genetic differentiation measured by F_{ST} (Weir & Cockerham 1984). To address these questions, we used resampling techniques and constructed simulated populations with a random subset of individuals and SNPs from populations of the Amazonian plant species *Amphirrhox longifolia* (A. St.-Hil.) Spreng (Violaceae). We selected *A. longifolia* for this study because this species is very abundant and broadly distributed on the banks of lowland Amazonian rivers, making it an excellent candidate for future population genomic approaches investigating Wallace's riverine barrier hypothesis (Wallace 1852).

Material and methods

Focal taxon

Amphirrhox longifolia (Violaceae) is a small, shrubby tree that is broadly distributed through tropical lowland forests from Costa Rica to eastern Brazil (Braun *et al.* 2012). It is self-incompatible and pollinated by bees, with seeds that are dispersed over short distances by an explosive mechanism (Braun *et al.* 2012). No information is available on the evolutionary history of *A. longifolia*, nor about how genetic variation is partitioned within and among its populations. The size of the genome for the diploid *A. longifolia* and other Violaceae is also unknown.

Study area and field collections

The study area is located near the mouth of the Rio Negro (Novo Airão, Amazonas State, Brazil). The Rio Negro is the fifth largest river in the World and the largest tributary of the Amazon Basin (Latrubesse *et al.* 2005). The Rio Negro surroundings are covered by dense, tall, evergreen lowland and submontane forests, interspersed by other vegetation types such as open grasslands and scrubby vegetation (i.e. white sand *campinas*) (Macedo & Prance 1978). Deforestation along the Rio Negro has been minimal due to the infertile sandy soils that are unsuitable for agriculture. Samples for this study were collected in May 2015 from two populations of *A. longifolia*, that is populations A (02°01′36.4″S, 61°15′25.1″W) and B (02°07′15.5″S, 61°10′32.7″W), situated within 20 km from each other, both on the same side of the Rio Negro (Fig. 1). From each population, we sampled 35 reproductive individuals of *A. longifolia*. All individuals are separated by at least 50 m to prevent sampling from close relatives.

Library preparation and sequencing

We extracted genomic DNA from leaf samples of *A. longifolia* using the Macherey-Nagel kit (Macherey-Nagel GmbH & Co. KG), following the manufacturer's instructions. We created two genomic libraries (A and B) using a double-digest RADseq (ddRAD) protocol (Peterson *et al.* 2012), with modifications to minimize the risk of high variance in the number of reads per individual within a pool. Specifically, PCR (as detailed below) was performed on each individual samples and the amplicons were pooled for size selection, instead of pooling samples prior to size selection and PCR as recommended by Peterson *et al.* (2012). Before the digestion reactions, we assessed

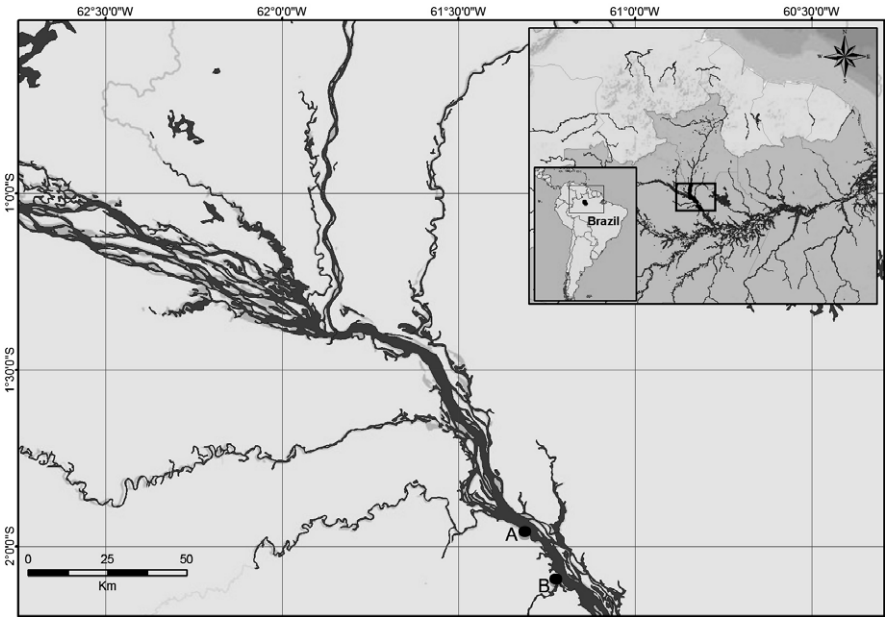

Fig. 1 *Amphirrhox longifolia* (A. St.-Hil.) Spreng populations (A and B) sampled in the wet season along the left bank of the Rio Negro, Amazon Basin, Brazil.

double-stranded DNA concentration for each sample using the Qubit dsDNA Assay Kit (Invitrogen) and made the necessary adjustments to bring each individual DNA in the pool to equal molar concentration. The initial DNA concentration for each sample varied from 350 to 450 ng μL^{-1}. Each sample was digested with the high-fidelity restriction enzymes EcoRI and MseI (New England Biolabs). Digestion reactions were carried out in a total volume of 20 μL, using 17 μL resuspended DNA, 5 units of EcoRI, 5 units of MseI and 1× CutSmart buffer (New England Biolabs) for 3 h at 37 °C, ending with a 20-min deactivation step at 65 °C. Reactions were then purified with the Agencourt AMPure XP system (Beckman Coulter), following the manufacturer's instructions, with elution in 40 μL TE buffer. To standardize the initial DNA mass to be added to the adapter ligation, we quantified the amount of cleaned digests for all samples using Qubit. Adapter ligations were carried out in a total volume of 30 μL, combining 42 ng DNA, 0.22 μM of a nonsample specific MseI adaptor (common for all samples), 0.33 μM of a sample specific EcoRI double-strand adaptor for each DNA sample, 1U of T4 DNA ligase (New England BioLabs), and 1.3× T4 ligase buffer which were incubated at 23 °C for 30 min. Reactions were then heat-killed at 65 °C for 10 min following a slow cooling to room temperature (23 °C). A total of 96 EcoRI double-stranded barcodes with unique 10-base pair sequences were created using python scripts; these barcodes can be found, together with the MseI oligo sequences, in Appendix S1 (Supporting information). Ligation products were cleaned with the Agencourt AMPure XP system and amplified in 20 μL PCRs that contained 13.5 μL of the ligation product, 0.2 μM of each primer (Appendix S1, Supporting information), 0.2 mM dNTPs, 1.0 mM $MgCl_2$, 0.5 U of iProof™ High-Fidelity DNA polymerase (Bio-Rad) and 2× of iProof buffer. The PCR protocol (98 °C for 30s, 20 cycles of 98 °C for 20 s, 60 °C for 30 s and 72 °C for 40 s, followed by a final extension at 72 °C for 10 min) was carried out in an Eppendorf PCR System. Before pooling samples at each library, DNA concentration of each sample ranged from 2.36 ng μL^{-1} (samples from library A) to 3.54 ng μL^{-1} (samples from library B). Multiplexed libraries were prepared with approximately equal amounts of DNA. We used an automated size-selection technology at 2% agarose cartridge (Pippin Prep; Sage Science, Beverly, MA) to select genomic fragments at a target range size of 375–475 bp. Size, quantity and quality of each individual library were measured on the Agilent 2100 Bioanalyzer (Agilent Technologies) using the Agilent DNA 1000 Kit. Each library was sequenced (100-bp single-end reads) on a half lane of an Illumina HiSeq 2000 flow cell (Illumina Inc., San Diego, CA, USA) at The Centre for Applied Genomics in Toronto, Canada (each half lane was pooled with 20 individuals from another study).

Identifying and genotyping SNPs

Files containing all raw sequence reads for all *A. longifolia* individuals were analysed in STACKS v. 1.35 (Catchen *et al.* 2011; Catchen *et al.* 2013b) for de novo assembly. Initially, we used the process_radtags program in STACKS to assign reads to individuals and eliminate poor quality reads as well as reads devoid of the expected EcoRI cut site (options—barcode_dist 4 -q -e ecoRI). All sequences were processed in ustacks to produce consensus sequences of RAD tags. The program ustacks takes a set of short-read sequences from a single sample as input and aligns them into exactly matching stacks. A maximum-likelihood framework (Hohenlohe *et al.* 2010) was then applied to estimate the diploid genotype for each individual of *A. longifolia* at each nucleotide position. In our analysis, the optimum minimum depth of coverage to create a stack was set to three sequences, the maximum distance allowed between stacks was set to two nucleotides, and the maximum number of stacks allowed per de novo locus was set to three. We enabled the stacks assembly deleveraging algorithm (−d), which resolves overmerged tags, and the removal algorithm (−*r*), which drops highly repetitive stacks from the algorithm. The alpha value for the SNP model was set to 0.05. Cstacks was used to build a catalog of consensus loci containing all the loci from all the individuals and merging all alleles together. Then, each individual genotype was compared against the catalog using sstacks. We subsequently used rxstacks to exclude problematic loci with a log-likelihood less than −100 and loci that matched a single catalog locus (conf_limit = 0.25) or any nonbiological haplotypes (−prune_haplo) in more than 25% of the individuals. We then ran the POPULATIONS software within Stacks to identify the loci found in at least 90% of all samples at each population ($P = 1$, $r = 0.9$), with sequencing depth of 12×. We included only the first SNP per locus in the final analysis because the use of multiple SNPs within loci strongly affects statistical power (Morin *et al.* 2009). All raw sequence reads are available from the National Center for Biotechnology Information Short Read Archive (Accession no. PRJNA362221).

Population characterization

We characterized the populations of *A. longifolia* in terms of the number of raw reads sequenced and number of unlinked SNPs identified. We used GENALEX 6.5 (Peakall & Smouse 2006) to remove SNP markers showing deviation from Hardy–Weinberg equilibrium (HW) for each population of *A. longifolia*. We then used the software BAYESCAN v. 2.1 with 20 pilot runs of 10 000 iterations, a burn-in of 50 000 iterations and a final run of 100 000 iterations to remove SNPs potentially under balancing and divergent selection. To minimize false positives, prior odds of the neutral model were set to 10 000 (i.e. the neutral model is 10 000 times more likely than the model with selection; Foll & Gaggiotti 2008).

After filtering SNP markers (i.e. SNPs deviating from HW equilibrium or under selection) for each population, we estimated the number of effective alleles (i.e. a measure of the maximum possible diversity if all alleles had the same frequency, A_e) as $1/\Sigma p_i^2$, where p_i is the frequency of the ith allele (Kimura & Crow 1964). We estimated the unbiased expected genetic diversity (i.e. unbiased expected heterozygosity, uH_e) by applying the formula described by Nei & Roychoudhury (1974) and the observed heterozygosity (H_o) by directly counting the individuals that were heterozygous at each locus. Population genetic statistics were averaged across loci using GENALEX 6.5 (Peakall & Smouse 2006). We estimated the inbreeding coefficient for each population using Wright's Fixation Index F (Nei & Chesser 1983). We also estimated pairwise genetic differentiation (F_{ST}) for *A. longifolia* populations using an ANOVA approach following Weir & Cockerham (1984). We used the SPAGEDI program (Hardy & Vekemans 2002) to compute F and F_{ST}. We determined the significance of the deviation of F and F_{ST} values from 0 through jackknife, using the same software. We also assessed the impact of missing data (ranging from 0% to 50%) on the number of SNPs and on the estimates of genetic diversity parameters of *A. longifolia* population.

Evaluating the effects of sample size

We used resampling techniques to investigate the effect of sample size (i.e. number of individuals) on estimates of genetic diversity and differentiation. Prior to evaluating the effects of sample size directly, we initially performed a power analysis to determine the minimum number of resampling replicates and SNPs that would be needed to ensure accurate estimation of genetic parameters. For all within-population genetic diversity estimates, a single resampling scheme was used to generate resampled data sets differing in number of replicates (x), SNPs (k) and individuals (n). Specifically, for each *A. longifolia* population, we constructed simulated data sets consisting of different numbers of resampling replicates (x = 100, 200, 400, 600 and 750), each represented by all combinations of different sample sizes (number of individuals per population, n = 2, 4, 6, 8, 10, 15 and 20) and number of SNPs (k = 50, 100, 200, 500, 1000, 1500, 3000, and 3500). To construct each resampling replicate, we selected a random subset of individuals from the empirical data set (n = 35) using a custom script in R (R Core Team 2014; Appendix S2, Supporting information). The resulting simulated 'populations' were

considered independent because sampling was carried out without replacement (i.e. no *A. longifolia* individual was included more than once in the same replicate). However, the same individual could be included in more than one replicate of the simulated data set for each sample. For each simulated population, we further randomly resampled $k = 50$ to 3500 SNPs. The R output files for the simulated populations were then converted into Genepop infiles using the software FORMATOMATIC v. 0.8.1 (Manoukis 2007). Finally, we used GENALEX 6.5 (Peakall & Smouse 2006) to estimate intrapopulation genetic diversity parameters (i.e. A_E, H_O, and unbiased H_E) for each replicate at each sample size (i.e. for each simulated population). Because the reduction in population sizes will become increasingly more common due to habitat fragmentation and degradation, we also performed all analyses using a not filtered data set (i.e. SNPs deviating from HW equilibrium or under selection were included because these tests are not possible when small sample sizes are used).

Although it has previously been noted that genetic diversity estimates are often not directly comparable across populations with different numbers of individuals due to ascertainment bias in small populations (Petit *et al.* 1998; Leberg 2002; Kalinowski 2004; Pruett & Winker 2008), these biases may be much less prominent for low-diversity markers such as SNPs (Pruett & Winker 2008). Furthermore, our resampling technique ensured that all loci were resampled and used to calculate genetic diversity estimates without any filtering (i.e. regardless of whether or not loci were polymorphic in the resampled populations). This approach eliminated any possible effect of ascertainment bias, making our estimates across different sample sizes directly comparable. As such, our resampling methods and results are directly applicable to situations in which the same set of loci is used to estimate genetic diversity among populations with different sample sizes.

To estimate the degree of genetic differentiation among populations, a slightly different subsampling strategy was used to resample the 1620 loci shared between populations. First, we resampled SNPs from populations A and B individually rather than combining all individuals into a single pool. Therefore, the population size (n) used to calculate population differentiation (F_{ST}) refers to the number of individuals per population (for a total of $2n$ individuals across both populations). Second, because F_{ST} is estimated specifically over polymorphic loci, we believe that it is most appropriate to compare F_{ST} values calculated over resampled polymorphic loci only. Thus, while resampling the data for a given sample size, we continued to resample SNPs (without replacement) until the desired number of loci (polymorphic in at least one simulated population) was obtained. Using this resampling strategy, simulated data sets were generated for sample sizes that varied from two to 20 individuals per population ($n = 2$, 4, 6, 8, 10, 15 and 20) and 50 to 1500 SNPs ($k = 50$, 100, 200, 500, 1000 and 1500), for $x = 100$ replicates each. Resampling was conducted using a custom script in R (Appendix S3, Supporting information). We then estimated population genetic differentiation by calculating F_{ST} (Weir & Cockerham 1984) using the R package DIVERSITY (Keenan *et al.* 2013). By analogue with Wright's F-statistics, the method of Weir & Cockerham (1984) uses an ANOVA approach to estimate the intra- and interpopulation variance components that are used to estimate F_{ST} (Weir & Cockerham 1984). This analysis does not require standardization for markers with few allelic states (e.g. SNPs; Meirmans & Hedrick 2011); therefore, we do not report results for any 'corrected' F_{ST} analogues such as G'_{ST} or Jost's D (Keenan *et al.* 2013).

Box plots were used to assess the effects of sample size on intra- and interpopulation genetic diversity parameters because this approach is based on statistics that do not require assumptions about the shape of the data distribution (Krzywinski & Altman 2014). To assist in judging differences between means, the 95% confidence interval was obtained and inserted in the box plots using BOXPLOTR (Spitzer *et al.* 2014).

Results

Characterization of Amphirrhox longifolia populations

About 62 million (population A) and 80 million (population B) single-end raw reads of 101 bp were produced on one lane of HiSeq 2000 Illumina. Each read starts with a barcode sequence identifying a sample (up to 10 bp long) and the 6-bp restriction site followed by 85 bp of usable data. For population A, 98.18% of the reads (61 439 558) passed the default quality filters, including Phred quality scores >33, and contained an identifiable barcode. For population B, almost 99% of the reads (79 307 358) were retained for further analysis. Considering all samples from population A, the average number of valid reads was 1 919 986 ± 151 785 SE, but varied from 837 031 to 4 696 619. Although the average number of valid reads was 2 265 924 ± 125 951 SE (varying from 1 287 622 to 4 111 315) for population B, there is no significant difference between the means for both populations. Throughout the genome of *A. longifolia*, we identified 3646 and 4900 polymorphic SNPs for populations A and B, respectively, with maximum 10% missing data and minimum 12-fold coverage. For all significance levels (i.e. $\alpha = 0.05$, $\alpha = 0.01$, and $\alpha =$

0.001), a total of 1.81% and 2.94% of SNPs deviated from HWE for populations A and B, respectively. SNPs that significantly deviated from HWE (66 for population A and 144 for population B) were discarded before further analyses. We did not detect any loci that were under selection for any population of *A. longifolia*, with the false discovery rate (FDR) set to 0.05. As such, no other loci were removed from subsequent analyses.

The number of effective alleles (A_e) in *A. longifolia* populations varied from 1.070 ± 0.001 SE (population B) to 1.083 ± 0.002 SE (population A). The genetic diversity (H_e) in *A. longifolia* populations varied between 0.060 ± 0.001 SE (population B) and 0.069 ± 0.001 SE (population A). The fact that H_o was slightly higher than the H_e, varying from 0.062 ± 0.001 SE (population B) to 0.075 ± 0.002 SE (population A) indicates no inbreeding in *A. longifolia* populations ($F = -0.074 \pm 0.008$ SE, $P < 0.05$, for population A; $F = -0.025 \pm 0.006$ SE, $P < 0.05$, for population B). The F_{ST} estimated from 1620 neutral SNP markers shared between populations (i.e. all shared SNPs, as no loci under selection were detected using BAYESCAN) was significant and equal to 0.0785, with a 95% confidence interval from 0.0727 to 0.0881. No statistical differences were observed for the A_e, H_e, and H_o when SNPs that deviated from HWE were included. For instance, values ranged from 1.082 ± 0.002 SE (A_e), 0.069 ± 0.001 SE (H_e) and 0.073 ± 0.002 SE (H_o) for population A. In addition, when monomorphic loci (fixed in either population of *A. longifolia*) were included in the data set, no statistical differences were observed for the genetic diversity parameters A_e (1.081 ± 0.002 SE), H_e (0.067 ± 0.001 SE) and H_o (0.073 ± 0.002 SE).

When we increased the maximum percentage of missing data (r) from 0 to 50%, the number of SNPs increased significantly. For instance, for population A, the number of SNPs increased from 247 (0%, $r = 1$) to 90 740 (50%, $r = 0.5$). Similarly, the genetic diversity indices were higher with maximum 50% missing data ($H_o = 0.136 \pm 0.0005$ SE, $H_e = 0.123 \pm 0.0004$ SE, $F = -0.017 \pm 0.011$ SE) than with maximum 0% ($H_o = 0.063 \pm 0.004$ SE, $H_e = 0.060 \pm 0.003$ SE, $F = -0.013 \pm 0.000$ SE). However, when missing data were reduced to 0 or 10%, no significant differences among diversity parameters were found (Table S1, Supporting information).

Depicting sample sizes and number of loci for intra- and interpopulation genetic diversity

We assessed the impact of increasing sample sizes for intra- and interpopulation genetic diversity estimates by resampling 2, 4, 6, 8, 10, 15 and 20 individuals from empirical data sets obtained for two *A. longifolia* populations. Accurate estimates of population genetic parameters were recovered in our simulations with only $x = 100$ resampling replicates (Fig. 2). For instance, when we fixed the number of individuals (n) to two and the number of SNPs (k) to 100, no statistical difference was detected for the mean values of A_e, H_O and H_e even when the number of replicates was set to $x = 100$ [$A_e = 1.076$, 95% CI (1.071, 1.080); $H_o = 0.065$, 95% CI (0.061, 0.068); and $H_e = 0.046$, 95% CI (0.043, 0.048)] or 750 [$A_e = 1.073$, 95% CI (1.071, 1.074); $H_o = 0.064$, 95% CI (0.062, 0.065); and $H_e = 0.045$, 95% CI (0.044, 0.046)].

Our simulations allowed us to determine the minimum sample size of *A. longifolia* needed to ensure that the sample accurately reflects the genetic diversity of the empirical data sets. We only show the best scenario (i.e. minimum sample size required for a determined number of loci; Figs 3 and 4), as a large number of scenarios have resulted from the combination of the different sample sizes and number of loci (i.e. 56 scenarios for each intrapopulation estimates for each *A. longifolia* population and 42 scenarios for F_{ST} estimates). For population A, increasing sample sizes above eight individuals appears to have little impact on the mean H_e, when 1000 SNPs are considered, even when some loci in the data set were monomorphic or deviated from HWE. For instance, the mean values of unbiased H_e for $n = 8$ was 0.065 [95% CI (0.064, 0.066)] and for $n = 20$ was 0.067 [95% CI (0.065, 0.069)]. Considering all SNPs in the data set, the mean values of unbiased H_e for $n = 8$ was 0.063 [95% CI (0.062, 0.064)]. For A_e and H_o estimates, a small sample size ($n = 2$) with a moderate number of SNPs (i.e. 500 for A_e and 1000 for H_o) was sufficient to recover the genetic diversity found in *A. longifolia* populations [$A_e = 1.083$, 95% CI (1.079, 1.086); $H_o = 0.075$, 95% CI (0.071, 0.079); for $n = 2$, $A_e = 1.083$, 95% CI (1.081, 1.085) and $H_o = 0.074$, 95% CI (0.073, 0.076)] (Fig. 3). The same sample size ($n = 2$) was obtained when all SNPs were considered in the data set for the number of effective alleles and observed heterozygosity parameters [$A_e = 1.079$ with 500 SNPs, 95% CI (1.079, 1.086); $H_o = 0.069$ with 1000 SNPs, 95% CI (0.067, 0.071). For population B, sample sizes above six individuals appear to have little impact on the mean H_e, even when more than 1000 SNPs are considered. The mean values of unbiased H_e for $n = 6$ was 0.056 [95% CI (0.054, 0.058)] and for $n = 20$ was 0.059 [95% CI (0.058, 0.060)]. Using only 500 SNPs, a small sample size ($n = 2$) was enough to recover the A_e and H_o from this population of *A. longifolia* [$A_e = 1.070$, 95% CI (1.068, 1.072); $H_o = 0.062$, 95% CI (0.060, 0.064); for $n = 2$, $A_e = 1.067$, 95% CI (1.061, 1.074) and $H_o = 0.063$, 95% CI (0.062, 0.065)] (Fig. 3). Furthermore, increasing the number of SNPs above 1000 does not decrease the sample size needed to recover the overall genetic diversity of individual populations of *A.*

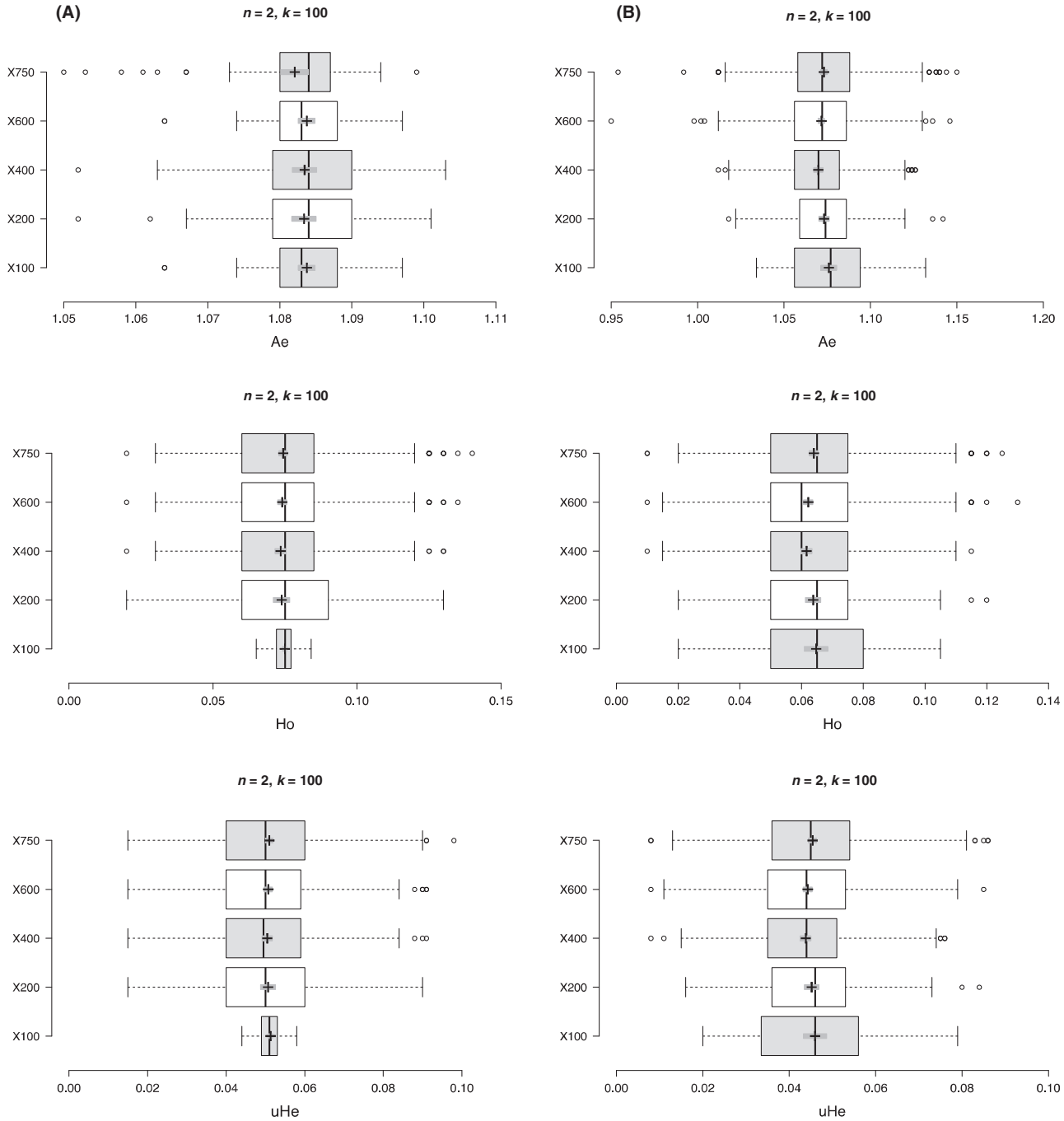

Fig. 2 Boxplots showing the minimum number of resampling replicates (x) needed to obtain accurate estimates of genetic diversity for populations A and B of *Amphirrhox longifolia* (Rio Negro, AM, Brazil). Centre lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles; outliers are represented by dots; crosses represent sample means; bars indicate 95% confidence intervals of the means. X = 100, 200, 400, 600, 750 resampling replicates are the sample points. A_E, number of effective alleles; H_O, observed heterozygosity; H_E, unbiased expected heterozygosity.

longifolia [for k = 1000, A_e = 1.083, 95% CI (1.082, 1.085), H_o = 0.074, 95% CI (0.073, 0.076) and H_e = 0.070, 95% CI (0.0696, 0.0713); for k = 3500, A_e = 1.082, 95% CI (1.080, 1.084), H_o = 0.075, 95% CI (0.074, 0.076) and H_e = 0.070, 95% CI (0.0698, 0.0705); Fig. S1, Supporting information].

As far as the degree of population genetic differentiation is concerned, increasing sample size above two

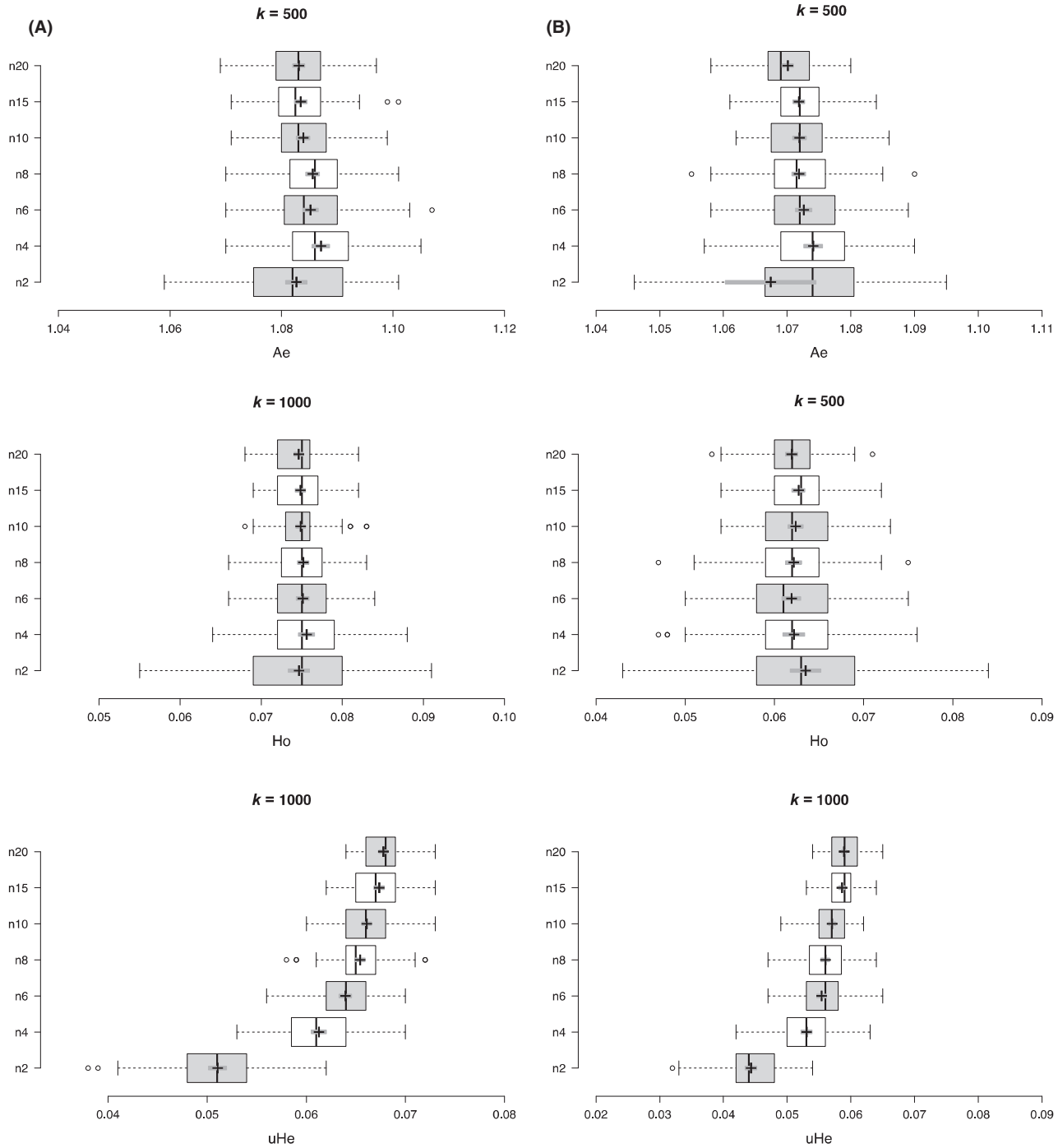

Fig. 3 Boxplots of genetic diversity indices based on 100 replicates for populations A and B of *Amphirrhox longifolia* (Rio Negro, AM, Brazil) depending on sample size (*n*, number of individuals). Centre lines show the medians; box limits indicate the 25th and 75th percentiles; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles; outliers are represented by dots; crosses represent sample means; bars indicate 95% confidence intervals of the means. *n*, sample sizes; *K*, number of SNPs; A_E, number of effective alleles; H_O, observed heterozygosity; uH_E, unbiased expected heterozygosity.

individuals appears to have little impact on the mean F_{ST} among resampling replicates, when 1500 polymorphic SNPs are considered (Fig. 4). For instance, the mean values of F_{ST} for $n = 2$ were 0.080 [95% CI (0.078, 0.083)] and for $n = 20$ were 0.079 [95% CI (0.076, 0.083)]. Furthermore, when loci are fixed in one or another *A. longifolia*

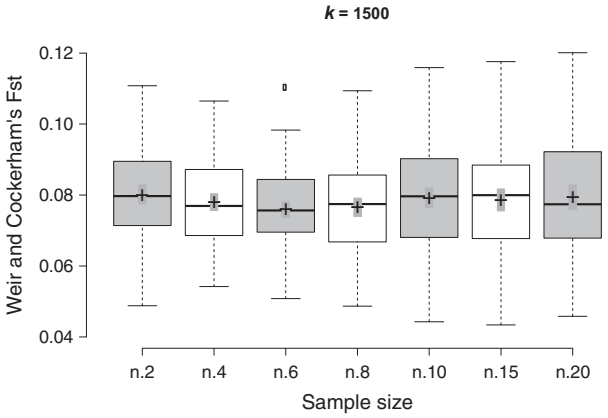

Fig. 4 Boxplot of Weir and Cockerham's F_{ST} (Weir & Cockerham 1984) for populations of *Amphirrhox longifolia* sampled along one bank of the Rio Negro (AM, Brazil). Centre lines show the medians; box limits indicate the 25th and 75th percentiles as determined by the R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles; outliers are represented by dots; crosses represent sample means; bars indicate 95% confidence intervals of the means. *n*, sample sizes; *K*, number of SNPs.

population, the mean values of F_{ST} were slightly small [0.076 [95% CI (0.059, 0.099)] for $n = 2$, and 0.077 [95% CI (0.069, 0.084) for $n = 20$]; however, these estimates were not significantly different from those estimated with polymorphic SNPs exclusively.

Discussion

Sampling design is a crucial aspect of population genetic studies (Luikart & Cornuet 1998; Manel *et al.* 2012; Yan & Zhang 2004; Cavers *et al.* 2005; Kalinowski 2005; Pruett & Winker 2008; Miyamoto *et al.* 2008; Morin *et al.* 2009; Nazareno & Jump 2012; Willing *et al.* 2012; Hale *et al.* 2012; Hobas *et al.* 2013). An ideal sampling design is one in which a sufficient number of individuals are sampled per population to obtain accurate estimates of genetic diversity and differentiation. In population genomics, where a large number of sequence variants (e.g. SNPs) are screened throughout the genome, no empirical study has ever evaluated the sample size needed to accurately estimate genetic diversity parameters in natural populations. Our study is the first to determine the sample sizes required for accurate estimates of genetic diversity and differentiation using empirical data. Our results suggest that, in general, relatively small sample sizes are likely to be sufficient.

Specifically, we found that six to eight individuals were sufficient to recover within-population genetic diversity estimates, even when monomorphic loci were included in the data set. In agreement with results derived from a theoretical study that used simulated SNPs data to assess population genetic structure (Willing *et al.* 2012), our study also showed that statistical power does not improve significantly when a large number of SNPs are used, indicating a reduction in interlocus sampling variance as more SNPs are sampled. However, the optimal sample size varied for all genetic diversity measures analysed (i.e. number of effective alleles, observed and expected heterozygosity). Confidence intervals overlapped for all sample sizes above two indicating that the two-individual sample was representative of the entire species (35-individual) sample in terms of the number of effective alleles (A_e) and observed heterozygosity (H_o) parameters in both populations studied (Fig. 3). For the H_o parameter, however, the number of SNPs varied from 500 (Population B) to 1000 (Population A). Overall, our results corroborate the idea that accurate genetic estimates can be obtained when large numbers of SNPs are employed (Morin *et al.* 2004; Willing *et al.* 2012).

Compared to traditional genotyping studies, generally lower sample sizes are required when high-throughput sequencing is used (Willing *et al.* 2012; Meirmans 2015; Jeffries *et al.* 2016). Here, we demonstrate how many individuals are required to estimate population genetic structure, a fundamental parameter for the understanding of evolutionary processes. Our results show that even when sample sizes are small (i.e. two individuals per population), accurate estimates of F_{ST} can be obtained when a large number of polymorphic SNPs are employed. This result was also recovered even when some loci were fixed in one of the *A. longifolia* populations. We assumed that the individuals sampled do not represent extreme variants or recent immigrants to the populations, and thus, these results should be approached with caution. Our results provide important insights for the population genomic era, which is revolutionizing the field of evolutionary genetics. In the light of our results, an ideal and less expensive sampling strategy for a plethora of (nonmodel) species is likely to be one in which a small number of individuals can be sampled for a large number of populations, even when widespread species are considered. This sampling scheme, where a small number of individuals will be sequenced, will greatly facilitate the use of modern genetic tools in population genetic studies, phylogeography and conservation biology, especially in developing countries where funding is generally scarce.

Nevertheless, factors such as demographic history, intrinsic life-history traits (e.g. mating system, pollination and seed dispersal syndromes) and overall population characteristics (e.g. plant density, flowering phenology, demographic structure, spatial pattern and genetic structure) can also influence ideal sampling schemes. Little information is still available on how exactly these other factors should be taken into account

when designing sampling schemes for population genomics. However, it is reasonable to expect that plant species that are predominantly outcrossing, lack intrapopulation inbreeding and show low levels of genetic differentiation between populations, such as *A. longifolia* populations, would require very small sample sizes to accurately estimate population genetic parameters. In fact, a simulation study has shown that low levels of pairwise genetic differentiation as small as $F_{ST} = 0.01$ were detected by analysing only four individuals per population when a high number of SNPs were used (Willing *et al.* 2012).

The establishment of the most appropriate sample size and number of loci also depends on the study's objective (e.g. characterization of genetic diversity, hybridization, bottleneck detection, assignment test, parentage inference, genetic structure and connectivity) (Ryman *et al.* 2006; Morin *et al.* 2009; Hale *et al.* 2012; Willing *et al.* 2012; Hobas *et al.* 2013; Jeffries *et al.* 2016). However, only a few studies to date have examined the best combination of the number of loci and number of individuals to be sampled (Morin *et al.* 2009; Willing *et al.* 2012; Jeffries *et al.* 2016). While no population genomics studies have addressed such questions empirically to date, a few studies have investigated the impact of sampling empirically using microsatellite markers (Luikart & Cornuet 1998; Manel *et al.* 2002; Koskinen *et al.* 2004; Kalinowski 2005; Hale *et al.* 2012; Hobas *et al.* 2013; Gonzáles-Ramos *et al.* 2015). For instance, one population assignment test recommends 30–50 individuals when genotyping 10 loci in highly structured populations (Manel *et al.* 2012). On the other hand, 20–30 individuals and 5–20 polymorphic loci are recommended for detecting genetic bottlenecks (Luikart & Cornuet 1998). A previous study based on two polymorphic microsatellite markers has reported that 25 to 30 individuals are enough to accurately estimate expected heterozygosity, despite differences in taxon (Hale *et al.* 2012). When information about genetic patterns are available, some tools can help to optimize sampling for genetic studies focusing on genetic structure, hybridization, temporal sampling, bottlenecks, population connectivity and assignment tests (e.g. POWSIM by Ryman & Palm 2006; SPOTG by Hobas *et al.* 2013); however, mating patterns or overlapping generation are not taken into account in these programs. Specific recommendations may vary, but it is always important to consider the species' demographic history while establishing the most adequate sampling strategy (Hobas *et al.* 2013).

In this study, a combination of ddRADseq with high-throughput sequencing has allowed the discovery of thousands of SNPs for robust estimations of genetic diversity in populations of *A. longifolia*. No reference genomes are available for this plant species and the SNP markers developed here are the first of their kind for *A. longifolia*, providing a basis for future genome-wide population studies. This study has also demonstrated that SNP markers can accurately estimate the genetic diversity of *A. longifolia* populations even when small numbers of individuals are sampled. The next step will be to increase sampling to a wider range of populations of *A. longifolia* in order to gather in-depth information about the amount of genetic diversity found in their populations (A. G. Nazareno *et al.* unpublished). This information will also allow us to infer the evolutionary history of this abundant insect-pollinated Amazonian plant species.

Our results will also help to guide subsequent studies on threatened species and on species with reduced population sizes. These kinds of studies will be of particular importance in the future given that naturally small populations will become more and more common due to habitat loss and fragmentation. Our results are promising, given that they suggest that it may be possible to estimate genetic diversity and differentiation for such populations from very small sample sizes. Future genomic investigations are needed to determine whether these results will hold for taxa with contrasting life histories and when different goals are considered (e.g. characterization of genetic diversity, population assignment tests, bottleneck detection, and assessing genetic structure and genetic connectivity).

Acknowledgements

The authors thank the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for a postdoctoral fellowship to AGN (2013/12633-8; 2015/07141-4), and a collaborative BIOTA/Dimensions of Biodiversity grant cofunded by NSF, NASA & FAPESP to LGL (2012/50260-6). Additional funds were provided by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) through a Pq-1C grant to LGL (307781/2013-5) and the National Science Foundation (FESD 1338694) to CWD. We thank Verônica Thode, Maila Beyer, Beatriz Gomez, Annelise Frazão and Osmar Pereira for their great help during fieldwork. We also thank the Core Facility for Scientific Research (CEFAP) of the Universidade de São Paulo for computational support, as well as Nick Barton and two anonymous reviewers for comments and suggestions.

References

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17**, 81–92.

Bashalkhanov S, Pandey M, Rajora OP (2009) A simple method for estimating genetic diversity in large populations from finite sample sizes. *BMC Genetics*, **10**, 84.

Braun M, Dotter S, Schlindwein C, Gottsberger G (2012) Can nectar be a disadvantage? Contrasting pollination natural histories of two woody

Violaceae from the Neotropics. *International Journal of Plant Sciences*, **173**, 161–171.

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

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) STACKS: an analysis tool set for population genomics. *Molecular Ecology*, **22**, 3124–3140.

Cavers S, Degen B, Caron H *et al.* (2005) Optimal sampling strategy for estimation of spatial genetic structure in tree populations. *Heredity*, **95**, 281–289.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Deagle BE, Faux C, Kawaguchi S, Meyer B, Jarman SN (2015) Antarctic krill population genomics: apparent panmixia, but genome complexity and large population size muddy the water. *Molecular Ecology*, **24**, 4943–4959.

Emerson KJ, Merz CR, Catchen JM *et al.* (2010) Resolving postglacial phylogeography using high-throughput sequencing. *PNAS*, **107**, 16196–16200.

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

Gonzáles-Ramos J, Agell G, Uriz MJ (2015) Microsatellites from sponge genomes: the number necessary for detecting genetic structure in *Hemimycale columella* populations. *Aquatic Biology*, **24**, 25–34.

Hale ML, Burg TM, Steeves TE (2012) Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. *PLoS ONE*, **7**, e45170.

Hardy OJ, Vekemans X (2002) SPAGEDI: a versatile computer program to analyze spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Helyar SJ, Hemmer-Hansen J, Bekkevold D *et al.* (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. *Molecular Ecology Resources*, **11**, 123–136.

Hobas S, Gaggiotti O, Bertorelle G, ConGRESS Consortium (2013) Sample planning optimization tool for conservation and population genetics (SPOTG): a software for choosing the appropriate number of markers and samples. *Methods in Ecology and Evolution*, **4**, 299–303.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in three-spine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

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

Jeffries DL, Copp GH, Lawson Handley L, Olsén KH, Sayer CD, Hänfling B (2016) Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, **25**, 2997–3018.

Kalinowski ST (2004) Counting alleles with rarefaction: private alleles and hierarchical sampling designs. *Conservation Genetics*, **5**, 539–543.

Kalinowski ST (2005) Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity*, **94**, 33–36.

Keenan K, McGinnity P, Cross TF, Crozier WW, Prodohl PA (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 (1964) The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.

Koskinen MT, Hirvonen H, Landry P-A, Primmer CR (2004) The benefits of increasing the number of microsatellites utilized in genetic populations studies: an empirical perspective. *Hereditas*, **141**, 61–67.

Krzywinski M, Altman N (2014) Visualizing samples with box plots. *Nature Methods*, **11**, 119–120.

Latrubesse EM, Stevaux JC, Sinha R (2005) Tropical rivers. *Geomorphology*, **70**, 187–206.

Leberg PL (2002) Estimating allelic richness: effects of sample size and bottlenecks. *Molecular Ecology*, **11**, 2445–2449.

Lozier JD (2014) Revisiting comparisons of genetic diversity in stable and declining species: assessing genome-wide polymorphism in North American bumble bees using RAD sequencing. *Molecular Ecology*, **23**, 788–801.

Luikart G, Cornuet J (1998) Empirical evaluation of a test for identifying recently bottlenecked populations from allele frequency data. *Conservation Biology*, **12**, 228–237.

Macedo M, Prance GT (1978) Notes on the vegetation of Amazonia II. The dispersal of plants in Amazonian white sand campinas: the campinas as functional islands. *Brittonia*, **30**, 203–215.

Mammadov JA, Chen W, Ren R *et al.* (2010) Development of highly polymorphic SNP markers from the complexity reduced portion of maize [*Zea mays*, L.] genome for use in marker-assisted breeding. *Theoretical Applied Genetics*, **121**, 577–588.

Manel S, Albert C, Yoccoz NG (2012) Sampling in landscape genomics. In: *Data Production and Analysis in Population Genomics* (eds Bonin A, Pompanon F). Humana Press, New York.

Manoukis NC (2007) FORMATOMATIC: a program for converting diploid allelic data between common formats for population genetic analysis. *Molecular Ecology Notes*, **7**, 592–593.

Martin CH, Crawford JE, Turner BJ, Simons LH (2016) Diabolical survival in Death Valley: recent pupfish colonization, gene flow and genetic assimilation in the smallest species range on earth. *Proceedings of the Royal Society B*, **283**, 20152334.

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

Meirmans PG, Hedrick PW (2011) Assessing population structure: F_{ST} and related measures. *Molecular Ecology Resources*, **11**, 5–18.

Miyamoto N, Fernández-Manjarrés JF, Morand-Prieur M-E, Bertolino P, Frascaria-Lacoste N (2008) What sampling is needed for reliable estimations of genetic diversity in *Fraxinus excelsior* L. (Oleaceae)?. *Annual Forest Science*, **65**, 403.

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

Nazareno AG, Jump AS (2012) Species-genetic diversity correlations in habitat fragmentation can be biased by small sample sizes. *Molecular Ecology*, **21**, 2847–2849.

Nei M, Chesser RK (1983) Estimation of fixation indices and gene diversities. *Annual Human Genetics*, **47**, 253–259.

Nei M, Roychoudhury AK (1974) Sampling variances of heterozygosity and genetic distance. *Genetics*, **76**, 379–390.

Ozerov MY, Gross R, Bruneaux M *et al.* (2016) Genomewide introgressive hybridization patterns in wild Atlantic salmon influenced by inadvertent gene flow from hatchery releases. *Molecular Ecology*, **25**, 1275–1293.

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetics software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conservation Biology*, **12**, 844–855.

Pruett CL, Winker K (2008) The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *Journal of Avian Biology*, **39**, 252–256.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/

Ramos AM, Croojimans RP, Affara NA *et al.* (2009) Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS ONE*, **4**, e6524.

Ryman N, Palm S (2006) POWSIM: a computer program for assessing statistical power when testing for genetic differentiation. *Molecular Ecology*, **6**, 600–602.

Ryman N, Palm S, André C *et al.* (2006) Power for detecting genetic divergence: differences between statistical methods and marker loci. *Molecular Ecology*, **15**, 2031–2045.

Spitzer M, Wildenhain J, Rappsilber J, Tyers M (2014) BOXPLOTR: a web tool for generation of box plots. *Nature Methods*, **11**, 121–122.

Swatdipong A, Primmer C, Vasemagi A (2010) Historical and recent genetic bottlenecks in European grayling, *Thymallus thymallus*. *Conservation Genetics*, **11**, 279–292.

Uitdewiligen JGAM, Wolters A-MA, D'hoop BB, Borm TJA, Visser RGF, van Eck HJ (2013) A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. *PLoS ONE*, **10**, e0141940.

Vera M, Díez-Delmolino D, García-Marín J-L (2016) Genomic survey provides insights into the evolutionary changes that occurred during European expansion of the invasive mosquitofish (*Gambusia holbrooki*). *Molecular Ecology*, **25**, 1089–1105.

Wallace AR (1852) On the monkeys of the Amazon. *Proceedings of the Zoological Society of London*, **20**, 107–110.

Ward SM, Jasieniuk M (2009) Review: sampling weedy and invasive plant populations for genetic diversity analysis. *Weed Science*, **57**, 593–602.

Weir BS, Cockerham CC (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.

Yan LN, Zhang DX (2004) Effects of sample size on various genetic diversity measures in population genetic study with microsatellite DNA markers. *Acta Zoologica Sinica*, **50**, 279–290.

A.G.N. and L.G.L. designed the study and coordinated sample collection. A.G.N. conducted molecular work, performed analyses and led the writing of the manuscript with input from all co-authors. J.B.B. and C.W.D. provided laboratory assistance, analytical input and troubleshooting.

Data accessibility

All input files are available from Dryad (doi:10.5061/dryad.bm98q). Raw sequence data files for all the *A. longifolia* individuals are available from the National Center for Biotechnology Information Short Read Archive (NCBI-SRA).

Supporting Information

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

Fig. S1 Boxplot showing the effect of number of SNPs on genetic diversity indices for *Amphirrhox longifolia* (Rio Negro, AM, Brazil). Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by the R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots; crosses represent sample means; bars indicate 95% confidence intervals of the means. *n*, sample sizes; X, number of SNPs; A_E, number of effective alleles; H_O, observed heterozygosity; uH_E, unbiased expected heterozygosity.

Table S1 Variation of genetic diversity parameters (H_O, observed heterozygosity; H_E, unbiased expected heterozygosity; F_{IS}, fixation index) depending on the percentage of missing data for one population of *Amphirrhox longifolia* (Rio Negro, Amazonas State, Brazil).

Appendix S1 The 96 EcoRI double-stranded barcodes with unique 10 base pair sequences and the MseI and Illumina PCR oligo sequences used in this study.

Appendix S2 The custom script in R used to select a random subset of individuals and loci, from the empirical dataset, in order to obtain genetic diversity estimates.

Appendix S3 The custom script in R used to select a random subset of individuals and loci, from the empirical dataset, in order to obtain genetic differentiation estimates.