



A comparative analysis between SNPs and SSRs to investigate genetic variation in a juniper species (*Juniperus phoenicea* ssp. *turbinata*)

Cristina García^{1,2} · Erwan Guichoux³ · Arndt Hampe³

Received: 21 December 2017 / Revised: 31 October 2018 / Accepted: 1 November 2018
© The Author(s) 2018

Abstract

Genomic resources are a valuable research tool for understanding and forecasting the response of forest trees to global change and for developing science-based management strategies. Yet, many ecologically relevant tree species still lack such resources. The conifer genus *Juniperus* contains > 70 species that are widely distributed through the Northern Hemisphere, including several keystone species that form extensive forests in arid landscapes. To date, single-nucleotide polymorphism (SNP) markers have not been described for this ecologically important tree genus and the few described simple sequence repeat (SSR) markers result insufficient for performing reliable population demographic inference. Here, we report on the successful development of 19 new SSR and 147 SNP markers for Phoenician juniper (*Juniperus phoenicea* ssp. *turbinata*), a species widely distributed along the coasts of the Mediterranean Basin. We calculate a series of population genetic diversity estimates for each set of markers independently and for both sets combined. Our comparison shows that the higher per-locus information content of SSRs makes them the marker of choice for parentage and assignment studies, whereas SNPs provide more reliable demographic inferences (N_e and detection of a recent bottleneck). We also test and confirm the transferability of the new set of SNP markers to the closely related tetraploid species *J. thurifera*. Finally, we perform an orthology analysis with two gymnosperm model species to search for SNPs linked with functional genes.

Keywords Conifers · ddRADseq · Genetic diversity · Non-model tree species · Orthology analysis

Introduction

Genomic tools are now available for many organisms from all over the tree of life and enable ecologists and evolutionary biologists to perform fundamental tasks such as inferring

Communicated by F. Gugerli

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11295-018-1301-x>) contains supplementary material, which is available to authorized users.

✉ Cristina García
Cristina.Garcia-Perez@liverpool.ac.uk

¹ Institute of Integrative Biology, Department of Evolution, Ecology, and Behaviour, University of Liverpool, Crown Street, L69 7ZB Liverpool, UK

² Plant Biology, CIBIO/InBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Laboratório Associado Universidade do Porto, Campus Agrário de Vairão, Vairão 4485-661, Portugal

³ BIOGECO INRA, Université Bordeaux, 33610 Cestas, France

recent and historical demographic trends of populations in response to environmental changes (Aitken et al. 2008; Neale and Kremer 2011). However, the large size and the complexity of many tree genomes render the development of species-specific genomic resources for this group particularly challenging. For example, many gymnosperms have huge genomes with a high proportion (up to 60%) of repetitive content, and polyploidy is common among angiosperms (Ellegren 2014; Soltis et al. 2015). Although the advent of increasingly fast and affordable next-generation sequencing (NGS) technologies has partly solved these obstacles, genomic research on forest trees still remains largely focused on a few species of economic importance, whereas most ecologically relevant tree species still lack basic genomic resources. In addition, a strong geographic bias towards temperate tree species constrains our ability to decipher the long-lasting evolutionary impact of climate change across forested lands worldwide, a task of utmost importance to forecast the response of forests to emerging mega-disturbances (Millar and Stephenson 2015).

Forests and woodlands have been deforested for centuries in favor of agriculture and farming worldwide, but rural

abandonment has also prompted the expansion of native forest species from isolated remnant patches in the past decades, particularly in the Northern Hemisphere (Trumbore et al. 2015). Tracking the demographic expansion of formerly isolated forests and tailing the advance of invasive species are two major challenges for ecologists interested in maintaining the biodiversity of managed landscapes and the ecosystem services they provide. Accomplishing those challenges requires obtaining reliable demographic estimates to design scientifically informed management plans. Estimates such as the effective population size (N_e) can be inferred from genetic data of extensively genotyped populations, as already proven in endangered animal populations with fluctuating demographic trends (McCoy et al. 2013). Yet, unlike forestry and agronomic studies, ecological studies typically rely on a limited number (~10–20) of simple sequence repeat (SSR) markers that often prove insufficient for gaining reliable demographic inferences from genetic data. These studies could enormously benefit from the incorporation of single-nucleotide polymorphisms (SNPs) because of the following: (i) these provide less biased estimates of gene flow (Singh et al. 2013) and perform better in inferring the timing of recent and past demographic bottlenecks (McCoy et al. 2013); (ii) SNP calling is less prone to ambiguous visual scoring compared to SSR calling, although it is also strongly dependent on the parameters set in the bioinformatic pipelines; and (iii) SNPs allow not only demographic inferences but can potentially also characterize functional genetic variation (Eckert et al. 2010). Therefore, the development of SNP markers and the combination of different molecular marker types will improve our understanding of the evolutionary trends of tree species inhabiting managed landscapes, including species with large and complex genomes.

Several *Juniperus* species have increased their population sizes and geographical distribution during the past decades driven by land-use changes, namely rural abandonment, and increasingly arid conditions (Van Auken 2008). For example, fragmented populations of *Juniperus phoenicea* have rapidly expanded across a matrix of surrounding Mediterranean shrublands (García et al. 2014); former remnant patches of *J. thurifera* now have developed into extensive savanna-like forests in former agricultural landscapes of the central Iberian Peninsula (Escribano-Ávila et al. 2012; García-Cervigon et al. 2016), and *J. monosperma* has become the dominant species of the piñon-juniper ecotone in northern New Mexico following an extreme drought event in the 1950s (Allen and Breshears 1998). At a very different temporal scale, phylogeographical studies have retraced the postglacial expansions of juniper taxa that represent today the major tree species of many central Asian highlands and mountain ranges (Opgenoorth et al. 2010). All these junipers act as foundation species structuring their community and creating stable local conditions for other plant species (Whitham et al. 2006), and therefore, they are fundamental to forecast the response of

plant communities to recent climatic trends. The genus *Juniperus* is moreover central to evolutionary studies examining past tree distributional ranges driven by long-term geological and climate changes (Mao et al. 2010, 2012). Its complex evolutionary history still puzzles evolutionary biologists whose studies on deciphering the evolutionary history of this complex group would benefit from a greater availability of transferable genetic markers among closely related species (Li et al. 2012).

Previous studies have reported the development of a handful of SSR markers for closely related juniper species that comprises the section Sabina (Mao et al. 2010), such as *J. sabina* (Geng et al. 2017), *J. thurifera* (Teixeira et al. 2014), and *J. phoenicea* (Molecular Ecology Resources Primer Development Consortium et al. 2013) but they resulted insufficient to perform demographic inference and parentage analyses. Here, we aim to contrast the performance of two types of molecular markers on reliably inferring population demographic parameters in long-lived tree species that inhabit managed landscapes. To attain that goal, we designed 19 SSR and 147 SNP markers for Phoenician juniper (*Juniperus phoenicea* ssp. *turbinata*), a woodland-forming conifer species that is widely distributed across the Mediterranean Basin and the Macaronesian archipelagos (Adams 2011). Phoenician juniper woodlands have been severely damaged for centuries by rural activities but are currently expanding in protected areas (García et al. 2014) and on abandoned agricultural lands (Bello-Rodríguez et al. 2016). We derive a series of population genetic diversity estimates for a well-known expanding population (García et al. 2014) using the two marker sets either independently or in combination to compare their respective performance. Then, we test the transferability of the new SNP markers to the closely related tetraploid species *J. thurifera*. Finally, we search for functionally equivalent genes based on the extensive genome sequences available for two gymnosperm model species, *Pinus pinaster* and *Pinus taeda* in order to facilitate prospective landscape genomic studies with juniper species.

Material and methods

Plant material and DNA extraction

Juniperus phoenicea ssp. *turbinata* grows primarily on stabilized coastal dunes and rocky seashores (Adams 2011). The species is monoecious, wind-pollinated and produces berry-like cones (arcestides) that are dispersed by frugivorous vertebrates. Populations greatly differ in their conservation state ranging from well-preserved stands (mainly in protected or inaccessible areas) to chronically fragmented and jeopardized sites (Bello-Rodríguez et al. 2016). We used plant material from a 1.2-ha permanent study plot previously established in

an expanding juniper population that occupies coastal stabilized dunes in the Reserva Biológica de Doñana (lat. 37.017085, long. -6.554601; Huelva province, Spain) (García et al. 2014). For the present study, we randomly chose 83 individuals from this long-term study plot to be genotyped. We also included 11 samples of *Juniperus thurifera* collected in the Parque Natural del Alto Tajo (lat. 41.063114, long. 2.198333; Guadalajara province, Spain) to test for the cross-species transferability of the newly developed SNP markers.

Nuclear DNA was extracted from 5 to 10 mg of dry leaf tissue per individual. This material was placed into a 2 ml screwed-top tube with two steel beads (4 mm in diameter). Tubes were frozen in liquid nitrogen before grinding at 30 Hz for 2 min and 30 s with a Mixer Mill MM300 (Retsch, Germany). When tissue failed to grind into powder, we applied a second grinding cycle. For DNA isolation, we used a genomic DNA plant tissue kit NucleoSpin Plant II (Macherey-Nagel, Duren) and followed the manufacturer's instructions.

SSR marker development and genotyping

Marker development An equimolar mixture of genomic DNA from 15 individuals sampled from the same focal population with a final concentration of 16.5 ng/μL measured with a fluorometer (Invitrogen Qubit 4, Life Technologies, Singapore) and an average A_{260}/A_{280} ratio ca. 1.8 was sent to Ecogenics (www.ecogenics.ch) to proceed with high-throughput microsatellite isolation following Malausa et al. (2011). Size-selected fragments from genomic DNA were enriched for SSR content by using magnetic streptavidin beads and biotin-labeled CT and GT repeat oligonucleotides. The SSR-enriched library was analyzed on an Illumina MiSeq platform using the Nano 2 × 250 v2 format. After the assembly using ABySS software (Simpson et al. 2009), SSRs were detected using MISA Perl script (Thiel et al. 2003) that yielded 9313 contigs or singlets containing a microsatellite insert with a tetra- or a trinucleotide of at least 6 repeat units or a dinucleotide of at least 10 repeat units. Suitable primer design was possible in 4451 SSR candidates. Primers were designed with Primer3 using standard settings (Rozen and Skaletsky 1999). Initially, we selected 96 loci with lengths between 7 and 21 repeats following van Asch et al. (2010) and with amplicon sizes ranging from 100 to 400 bp to facilitate multiplex setting. Among them, 21 loci amplified successfully and 19 loci showed polymorphism with a subset of 15 individuals randomly sampled in our study area. Previous studies have shown a similar percentage of success in developing SSRs for non-model tree species (De Bellis et al. 2016; Schoebel et al. 2013).

Genotyping We genotyped all 83 *J. phoenicea* ssp. *turbinata* individuals at the 21 selected SSR loci. For this purpose, we optimized multiplex PCR conditions for the set of primers following a three-primer approach (Schuelke 2000). Multiplexed PCRs were performed in a 20-μL final volume

containing 1× buffer [67 mM Tris-HCl pH 8.8, 16 mM $(\text{NH}_4)_2\text{SO}_4$, 0.01% Tween-20], 3.5 mM MgCl_2 , 0.01% BSA (Roche Diagnostics, Germany), 0.25 mM dNTP, 0.5 U *Taq* DNA polymerase (Biolone, UK), 2 μL of the primer premix, and 5 μL (50 ng) of genomic DNA. The primer premix contained 0.5 μM of the primer M13 and a final concentration of the forward and reverse primers according to Table SM1. We used Multiplex Manager 1.2 (Holleley and Geerts 2009) to compute the temperature of annealing (Table SM1) and the complementary threshold (the maximum number of AT or CG matches for any two primers within a multiplex reaction), which was set to seven. Amplified fragments were analyzed on an ABI 3130xl Genetic Analyzer and sized using GeneMapper 4.0 (Applied Biosystems, USA) and LIZ 500 size standard.

SNP development and genotyping

Double-digest restriction-site-associated DNA library preparation and sequencing Nine of the 83 individuals were used for double-digest restriction-site-associated DNA (ddRAD) library preparation (Peterson et al. 2012) and one sample was duplicate. ddRAD library preparation protocol followed the methods described by Pukk et al. (2015) with minor modifications, including purification by solid-phase reversible immobilization (SPRI) bead solution (CleanNA, Netherlands) after each step. In short, 650 ng of DNA was digested for 3 h at 37 °C with two rare-cutting restriction enzymes, 10 U of *AseI* (restriction site 5' ATTAAT 3') and *PstI* (restriction site 5' CTGCAG 3') (New England Biolabs, USA). The ligation consisted of 0.04 μM of P1-*AseI* and A-*PstI* adapters, which were added to 8 μL of restriction reaction together with 0.5 mM of ATP, 1× of T4 DNA Ligase Buffer, and 800 U of T4 DNA Ligase (New England Biolabs). A barcode (Ion Xpress 1-9) was added to the A-*PstI* adapter to identify each sample a posteriori. The 20-μL ligation reaction was carried out at 22 °C for 2 h and heat-inactivated for 11 min at 65 °C before cooling (1 °C per minute). Libraries were then quantified with the Ion Library TaqMan Quantitation Kit (Thermo Fisher Scientific) before equimolar library pooling. The pool was loaded on an automated size-selection system (Pippin Prep—Sage Science, USA) with a 2% agarose cartridge to extract DNA fragments of 290–310 bp. The sized pool (30 μL) was amplified in a 100-μL reaction containing 1× Q5 High Fidelity PCR Master mix and 0.6 μM of Ion Torrent primers A and P1 (New England Biolabs). PCR consisted of 98 °C for 30 s followed by 12 cycles of 98 °C for 10 s, 58 °C for 30 s, and 65 °C for 30 s. The quality and quantity of the pool were measured on a Bioanalyzer 2100 using High Sensitivity DNA kit (Agilent Technologies, USA) and the final library was diluted to 10 pM before sequencing on an Ion Torrent PROTON (Thermo Fisher Scientific).

Quality control and SNP discovery Raw sequences were demultiplexed based on their barcodes and quality filtered using the default settings of the Ion Torrent BaseCaller (> Q16 with a window size of 30 bases). Filtered reads were used for de novo identification of putative SNPs using three programs implemented in STACKS (Catchen et al. 2013). First, we applied *process_radtags* to trim all reads to 200 bp. Then, we used *denovo_map* to build the catalog of loci and call SNPs with the following parameters: minimum number of identical reads $m = 15$, number of mismatches allowed between loci when processing a single individual $M = 2$, number of mismatches allowed between loci when building the catalog $n = 3$. Finally, *populations* generated the FASTA and VCF files used for downstream analysis. Putative SNPs were filtered out using different criteria: (i) SNP called for all nine samples; (ii) unique SNP within each stack; (iii) an identical genotype for the two technical replicates; and (iv) three genotypic classes identified and SNP not present in the first or last 20 bases of the sequence.

SNP selection for genotyping, SNP calling, and genotyping of individuals A total of 187 candidate SNPs were submitted for assay design using the MassARRAY® Assay Designer version 4.0.0.2 (Agena Biosciences). Four multiplexes of 156 SNPs (three 40-plex and one 36-plex) were designed for the SNP genotyping, which was performed using the iPLEX Gold chemistry following Gabriel et al. (2009) on a MassArray System (Agena Biosciences, USA). Data analysis was completed using MassARRAY Typer Analyzer 4.0.26.75 (Agena Biosciences). We filtered out all monomorphic SNPs, loci with weak or ambiguous signal (i.e., displaying more than three clusters of genotypes or unclear cluster delimitation) and loci with > 6% missing data.

Functional annotation of SNP markers

An orthology analysis was performed to search for functional annotation of the 187 sequences of the candidate SNPs using Blastall 2.2.15 (Zhang et al. 2000). The BlastN algorithm was used to align the flanking sequence of the SNPs with the transcriptome of *Pinus pinaster* (Canales et al. 2014, SustainPine v3.0, <http://www.scbi.uma.es/sustainpinedb/>). For *Pinus taeda*, the BlastX algorithm was used with the ConGenIE database (<http://congenie.org/>). The E value cut-off was set at 10^{-10} . We considered only the best blast hit for biological interpretation.

Evaluation of SSR and SNP markers to perform population genetic analyses

We first examined the quality of the 21 SSR markers based on the analysis of the multilocus genotypes of 15 individuals run on an ABI3730 (Applied Biosystems, California) (Table SM1).

Specifically, we evaluated the genetic correlation among SSR markers by estimating the Hardy-Weinberg equilibrium (HWE, Table SM2) and by testing for genotypic linkage disequilibrium as implemented in GENEPOP (Rousset 2008) (Table SM3). Additionally, PopGenReport version 3.0 served to infer the frequency of null alleles based on the Brookfield estimator (Brookfield 1996) (Table SM1). Finally, we gauged the polymorphism of each marker by assessing (Table SM1): (i) the number of alleles (A); (ii) the effective number of alleles (A_e), an estimate of the number of equally frequent alleles in an ideal population, A_e is of interest for comparison of allelic diversity across loci with diverse allele frequency distributions; (iii) unbiased expected and observed heterozygosity (uH_e and H_o , respectively); (iv) the polymorphic information content (PIC), a measure of the informativeness of a genetic marker (Botstein et al. 1980); (v) the probability of identity (PI, i.e., the average probability that two unrelated individuals drawn from the same randomly mating population will have the same multilocus genotype by chance); and (vi) the probability of exclusion (PE, the probability of excluding a putative parent pair). We used PICcalc (Nagy et al. 2012) to obtain the PIC per locus and we implemented GenAIEx 6.502 (Peakall and Smouse 2012) to obtain all other estimates.

Secondly, we performed population genetic analyses on the 83 genotyped *J. phoenicea* ssp. *turbinata* individuals for each marker type independently and for both types combined. We recorded the overall number of different alleles (allelic richness, A) and estimated the unbiased expected and observed heterozygosity (uH_e and H_o) as implemented in GenAIEx. We also calculated unbiased multilocus estimates of population inbreeding coefficient (F_{IS}) as implemented in INEST 2.2 (Chybicki et al. 2011) that has proved to be robust to the presence of null alleles (Campagne et al. 2012). As a measure of the usefulness of each set of markers for assignment studies, we measured average polymorphic information content (PIC), the overall parentage exclusion probability, the informativeness for inferring relationships (IR), and relatedness (Ir) as implemented in *Coancestry* (Wang 2011). The suitability of each set of markers to provide reliable demographic inferences was measured by estimating the contemporary effective population size (N_e) based on the linkage disequilibrium method as implemented in *NeEstimator* v2 (Do et al. 2014). We estimated N_e for the complete set of individuals successfully genotyped. We also evaluated the ability of each set of markers in estimating N_e of small-sized populations. To that end, we sampled genotypes randomly from our data set ($N = 83$) to create 100 populations of sizes $N = 10, 25, \text{ and } 50$, respectively. Then, we estimated N_e as above and recorded the average N_e and average confidence intervals over 100 populations for each size class. Finally, we tested the ability of each set of markers in detecting a recent bottleneck based on a sign test as implemented in *Bottleneck* (Cornuet and Luikart 1996).

Results

SSR and SNP development and selection

Two of the 21 new SSR markers (Junpho_017898 and Junpho_068482) were monomorphic and hence were discarded from all further analyses (Table SM1). Five of the remaining 19 loci did not meet HWE (Table SM2) and one pair of loci showed statistical evidence of linkage disequilibrium (LD) after applying the B - Y correction (Narum 2006) for multiple tests (Table SM3). A total of 102 million filtered reads (median size = 229 bp) were generated for the nine samples and 49,457 putative SNPs were identified. The SNP calling error rate, based on the replicated sample, was 5.5% for the 5661 loci for which all nine samples were available (19.6% for all SNPs for which at least the two technical replicates were called). The multi-criteria filtering generated a total of 187 SNPs retained for multiplex design. We discarded 40 SNPs because they did not properly amplify. We finally used a set of 147 SNPs for the population genetics analyses. Our final data set for these analyses contained ca. 2% (range 0–6%) of missing data per locus.

Estimates of genetic diversity

The overall number of SNP alleles more than tripled the number of SSR alleles (Table 1). Values of mean observed heterozygosity (H_o) across loci were similar for both sets of markers, whereas unbiased expected heterozygosity (uH_e) was higher for SSRs than for SNPs. The SSR-based estimate of the mean population inbreeding coefficient (F_{is}) exceeded the estimate obtained with SNPs. PIC values showed that SSR markers were on average more informative than SNP markers and, as a result, SSRs yielded higher values of IR, Ir, and parentage exclusion probability (Table 1). In turn, SSRs failed to detect a recent bottleneck that was successfully detected with SNPs either alone or in combination with SSRs. In addition, N_e estimates based on SNPs or the combination of SNPs and SSRs were more reliable both for the full data set ($N=83$) as well as for the smaller populations subsampled from the original data set ($N=10$, 25, and 50) (Table 1). For small population sizes ($N=10$ and $N=25$), SSRs frequently failed to provide a reliable N_e estimate. On the contrary, the SNPs, alone or in combination with SSRs, always successfully estimated N_e when we subsampled as low as $N=25$ individuals from the original data set.

Transferability and functional annotation of SNP markers

A total of 112 out of the 156 tested SNP loci amplified in the sister species *J. thurifera*. We detected only five polymorphic loci, possibly due to some extent to the relatively small

number of individuals tested. Only five (*P. pinaster*) and four (*P. taeda*) loci, respectively, could be annotated with strong E values ($6e-11$ to $6e-30$) and identities (75 to 88.7%) (Table 2). Two loci (1334_130 and 23712_40) emerged in both transcriptomes, whereas the remaining 182 loci (96.3%) did not match any sequence.

Discussion

The Aichi Biodiversity Targets include the preservation of the genetic diversity as one strategic goal, yet most non-model, but ecologically relevant, tree species lack genetic and genomic resources that allow researchers forecasting the fate of remnant populations in a changing world (Neale and Kremer 2011). As a result, the majority of recovery plans for endangered plant species overlook genetic factors as drivers of population extinction (Pierson et al. 2016), in spite of the ample existing evidence that underpins the role of genetic erosion and inbreeding in accelerating demographic decline (Frankham 2005). Therefore, by providing new genomic resources for a foundation tree species (*J. phoeniceas* spp. *turbinata*), our study can contribute to advancing both basic and applied ecological and evolutionary research (Mao et al. 2010).

Many juniper species are foundation species of semi-arid and arid ecosystems across the Northern Hemisphere, both in warm low-elevation and cold high-mountain habitats (Mao et al. 2010). To date, only 11 SSRs have been described for juniper species (Molecular Ecology Resources Primer Development Consortium et al. 2013), but these were not polymorphic enough to perform demographic inference or parentage analyses. The new set of SSR and SNP markers proved fully suitable to estimate genetic diversity, perform assignment or parentage analyses, and obtain reliable demographic inference. The issue of whether a small number of SSRs would do better in performing population genetic analysis than a large set of genome-wide distributed SNPs remains contentious. Our results show that mean values and variance of SSR-based estimates of genetic diversity (uH_e) exceeded those based on SNPs, as found in the previous studies (Fischer et al. 2017; Glover et al. 2010; Hamblin et al. 2007). This result confirms that estimates of genetic diversity typically show a wide variation among SSR markers (in terms of uH_e), which make SNPs more reliable when it comes to infer genome-wide genetic diversity, particularly when at least a few thousands of random SNPs are available (Fischer et al. 2017). The average PIC values show that both sets of markers, taken alone, still have a relatively moderate information content. However, the higher values of IR, Ir, and parentage exclusion probability obtained for SSRs (Table 1) indicate that they would be the marker of choice for assignment techniques such as parentage, maternity, and paternity analyses. On the

Table 1 Genetic diversity estimates for $N=83$ juniper (*Juniperus phoenicea* ssp. *turbinata*) individuals based on 19 simple sequence repeat (SSR) markers (left), 147 SNP markers (center), and both types of markers (right). For each set of markers, we report the number of polymorphic markers attained, the total number of different alleles, the range of the number of alleles per locus, the observed heterozygosity (H_o), the unbiased expected heterozygosity (H_e), the posterior mean estimate of the population inbreeding that takes into account the presence of null alleles (F_{is}), the mean polymorphic information content (PIC), the mean informativeness for inferring relationships (IR) and relatedness (Ir) according to Wang (2002), and the probability associated to a sign test to identify a recent bottleneck. We also report estimates of the contemporary effective population size (N_e) and their

associated confidence intervals based on the linkage disequilibrium method for three different sample sizes: $N=83$ (the full data set); $N=10$ (subset of 10 individuals randomly sampled 100 times from the full data set); $N=25$ (subset of 25 individuals randomly sampled 100 times from the full data set); $N=50$ (subset of 50 individuals randomly sampled 100 times from the full data set). Estimated N_e values for $N=10$, $N=25$, and $N=50$ are averaged for 100 simulated populations. Percentages indicate the number of times that N_e was successfully estimated over 100 simulated populations. Note that for both sets of markers, the estimated N_e sometimes exceeded the simulated population size (N) because the samples are integrated within a larger remnant forest patch composed of few hundreds of trees ($N \approx 750$ individuals)

	SSRs	SNPs	SSRs + SNPs
Number of markers	19	147	166
Total number of different alleles	93	291	384
Number of alleles per marker (range)	2–15	2	2–15
Observed heterozygosity (H_o)	0.37	0.38	0.38
Unbiased expected heterozygosity (uH_e)	0.44	0.39	0.39
Mean population inbreeding (F_{is})	0.0035 [0; 0.018]	0.0020 [0; 0.01]	0.018 [0; 0.01]
Mean polymorphic information content (PIC)	0.39	0.30	0.31
Overall multilocus parentage exclusion probability	0.98	1	1
Mean informativeness pairwise relationships (IR)	0.020	0.010	0.013
Mean informativeness pairwise relatedness (Ir)	0.06	0.04	0.04
Bottleneck (sign test)	$P > 0.05$	$P < 0.001$	$P < 0.001$
Contemporary effective population size (N_e)			
$N=83$	46 [34; 65]	36 [34; 39]	40.5 [38; 43]
$N=10$	12 [4; 128], 2%	43 [24; 790], 47%	47 [29; 119], 45%
$N=25$	39 [20; 241], 72%	30 [33; 53], 100%	46 [38; 58], 100%
$N=50$	51 [32; 99], 98%	39 [35–45], 100%	41 [38; 45], 100%

other hand, the SNP marker set clearly outperformed the SSRs in obtaining reliable demographic inferences, both in gaining robust estimates of N_e and in identifying a recent bottleneck event. This finding concurs with a previous study showing that SSRs typically fail to identify recent genetic bottlenecks (Peery et al. 2012). SNPs yielded reliable N_e estimates even for small-sized samples: for $N=25$, SNPs provided robust N_e values in 100% of the simulated populations, whereas SSRs only yielded reliable N_e estimates in 47%. This result suggests that, given a comparable budget to develop each set of markers (Hodel et al. 2016), SNPs are the marker type of choice for performing demographic inferences, particularly in small populations of non-model species. Note that this study is based on a remnant age-structured forest patch that has undergone successive cycles of population contraction and expansion. Therefore, genetic drift, founder effects, and non-random mating patterns are expected to impede allele frequencies to meet the Hardy-Weinberg proportions (HWP), which is the case for five out of the 19 SSRs. Similarly, these natural processes increase the probability that different markers show linkage disequilibrium (LD), as we found for one of 171 pairwise comparisons. As pointed out

by Waples (2015), once different sources of genotyping errors have been ruled out and the frequency of null alleles remains low ($< 10\%$), the presence of markers that do not meet the HWP and LD in natural populations could be the result of biological processes rather than genotyping errors. Although our results regarding the suitability of SSRs and SNPs for different types of evolutionary analyses might be influenced by the particular characteristics of the sample population, our results concur with previous studies concluding that the relative utility of each set of markers depends on the main goals of the study (Emanuelli et al. 2013; Singh et al. 2013; Yang et al. 2011), the availability of previous genetic resources, and the number of individuals sampled (Hodel et al. 2016; Nazareno et al. 2017).

In conclusion, if the main goal of the study is to describe the distribution of genetic variation across the landscape (i.e., landscape genetics) or ascertaining parentage and pedigree relationships (for example in mating system studies), a handful of SSR markers represent a cost-efficient tool that can provide reasonably good results (Hodel et al. 2016). However, if the main interest focuses on obtaining reliable demographic inferences, then SNPs would be the marker of

Table 2 Identification of putative functional genes after applying an orthology analysis that compares the sequences of the new set of SNP markers developed for *Juniperus phoenicea* ssp. *turbinata* with thesequences obtained based on the transcriptome of *Pinus pinaster* (<http://www.scbi.uma.es/sustaininedb>) and the ConGenIE database for *Pinus taeda* (<http://congenie.org/>)

SNP ID	Unigen ID	% identity	E value	Function
<i>Pinus pinaster</i>				
1334_130	sp_v3.0_unigene34816	85.57	4 E-15	Nucleic acid binding protein, putative
20617_161	sp_v3.0_unigene106772	85.25	4 E-18	Transcription factor-like protein (<i>Arabidopsis thaliana</i>)
23712_40	sp_v3.0_unigene9536	87.5	6 E-29	Putative uncharacterized protein
17533_128	sp_v3.0_unigene184491	86.49	6 E-11	Leucine-rich repeat receptor-like protein kinase
13682_50	sp_v3.0_unigene104870	86.76	6 E-11	Putative uncharacterized protein At2g27790
<i>Pinus taeda</i>				
16940_143	PITA_000042173	80.00	2 E-27	Mannosyl-oligosaccharide 1,2-alpha-mannosidase MNS3 isoform X1
1334_130	PITA_000048852	71.74	1 E-16	Protein vip1-like
23712_40	PITA_000079720	73.33	4 E-15	Neutrophil cytosol factor
10725_145	PITA_000009562	75.00	1 E-30	Alpha-L-fucosidase 1-like

choice, particularly when the study entails small-sized populations such as in conservation and recovery plans Waters et al. (2013). Studies would of course benefit from applying both types of markers whenever possible (Graudal et al. 2014; Olsson et al. 2016). The transferability of the new set of SNPs to *J. thurifera* suggests that SNP markers could potentially be applied to address ecological and evolutionary studies entailing other closely related species of this key genus (Mao et al. 2010). Yet, we found a low level of polymorphisms in *J. thurifera*, which encourages the use of increasingly cost-effective de novo development of species-specific SNPs and sequence data to address evolutionary studies across taxa in non-model species.

The functional annotation of SNPs links population genetics and population genomics as it allows to dissect the contribution of neutral processes (such as gene flow) and selective processes (such as local adaptation) in determining current patterns of genetic variation across heterogeneous gradients (Eckert et al. 2010). Non-model species typically lack a reference genome, but an orthology analysis based on gymnosperm model species allowed us to determine whether any of our SNPs is located at a functionally relevant position in the genome. Given the large genome size of most gymnosperm species, finding SNPs associated to genes that code for ecologically relevant proteins is challenging and, as a result, we only identified genes with generic functions, such as nucleic acid-binding proteins. Moreover, the Sequenom sequences we used were very short (80–120 bp), a factor that increases the likelihood of failure when performing functional annotation. SNP development from transcriptomes would certainly increase the chances to find polymorphism linked to genes that mediate plant responses to environmental factors, such as prolonged droughts (Neale and Ingvarsson 2008). In a context of environmental change, identifying locally adapted genotypes is of utmost importance to pinpoint

source populations in breeding programs or to design assisted migration plans. By doing so, managers could potentially move best-adapted genotypes to locations where environmental conditions are expected to shift in the near future, for example by becoming more arid (Aitken et al. 2008; Jordan et al. 2017). Lastly, the possibility of transferring genomic resources among closely related species expands the application of molecular markers to address long-lasting evolutionary studies to identify hybridization events, depict diversification patterns, or compare genetic diversity among closely related species to infer the ecological and evolutionary factors that have shaped current patterns of genetic diversity among forest species with complex genomes (Petit and Hampe 2006).

Acknowledgments We are in debt to F. Valladares for collecting the *J. thurifera* samples and to J. Arias for lab work. We also thank Adline Delcamp and Isabelle Lesur for their help with the Mass Array genotyping experiments and the functional annotations, respectively. We acknowledge the two anonymous reviewers for their valuable input to improve the final version of this manuscript.

Data archiving statement Data sets with all multilocus genotypes will be available through the digital data repository www.dryad.org. The sequences and accession numbers provided by GenBank for SSR and SNP markers are listed in Table SMI.

Funding information This work was funded by FCT research grant PTDC/BIA-BIC/5223/2014 and POCI-01-0145-FEDER-016817) granted to CG and the EU ERA-NET BiodivERsA project SPONFOREST (BiodivERsA3-2015-58). CG was also funded by the Fundação para a Ciência e a Tecnologia (FCT) through the Investigador Programme (IF/01375/2012). SNP development and genotyping were performed at the Genome Transcriptome Facility of Bordeaux (grants from the Conseil Régional d'Aquitaine no. 20030304002FA and 20040305003FA, from the European Union FEDER no. 2003227 and from Investissements d'Avenir ANR-10-EQPX-16-01).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Adams RP (2011) *Junipers of the world: the genus Juniperus*. Trafford, Bloomington
- Aitken S, Yeaman S, Holliday JA, Wang T, Curtis-McLane S (2008) Adaptation, migration or extirpation: climate change outcomes for tree populations. *Evol Appl* 1:95–111
- Allen CD, Breshears DD (1998) Drought-induced shift of a forest-woodland ecotone: rapid landscape response to climate variation. *Proc Natl Acad Sci U S A* 95:14839–14842
- Bello-Rodríguez V, García C, del Arco MJ, Hernández-Hernández R, González-Mancebo JM (2016) Spatial dynamics of expanding fragmented thermophilous forests on a Macaronesian island. *For Ecol Manag* 379:165–172
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331
- Brookfield JFY (1996) A simple new method for estimating null allele frequency from heterozygote deficiency. *Mol Ecol* 5:453–455
- Campagne P, Smouse PE, Varouchas G, Silvain JF, Leru B (2012) Comparing the van Oosterhout and Chybicki-Burczyk methods of estimating null allele frequencies for inbred populations. *Mol Ecol Res* 12:975–982. <https://doi.org/10.1111/1755-0998.12015>
- Canales J, Bautista R, Label P, Gómez-Maldonado J, Lesur I, Fernández-Pozo N, Rueda-López M, Guerrero-Fernández D, Castro-Rodríguez V, Benzekri H, Cañas RA, Guevara MA, Rodrigues A, Seoane P, Teyssier C, Morel A, Ehrenmann F, le Provost G, Lalanne C, Noirot C, Klopp C, Reymond I, García-Gutiérrez A, Trontin JF, Lelu-Walter MA, Miguel C, Cervera MT, Cantón FR, Plomion C, Harvenget L, Avila C, Gonzalo Claros M, Cánovas FM (2014) De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. *Plant Biotechnol J* 12:286–299. <https://doi.org/10.1111/pbi.12136>
- Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis toolset for population genomics. *Mol Ecol* 22:3124–3140. <https://doi.org/10.1111/mec.12354>
- Chybicki JJ, Oleksa A, Burczyk J (2011) Increased inbreeding and strong kinship structure in *Taxus baccata* estimated from both AFLP and SSR data. *Heredity* 107:589–600
- Cornuet JM, Luikart G (1996) Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144:2001–2014
- De Bellis F, Malapa R, Kagy V, Lebegin S, Billot C, Labouisse J-P (2016) New development and validation of 50 SSR markers in breadfruit (*Artocarpus altilis*, Moraceae) by next-generation sequencing. *Appl Plant Sci* 4:apps.1600021. <https://doi.org/10.3732/apps.1600021>
- Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR (2014) *NeEstimator v2*: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data. *Mol Ecol Res* 14:209–214. <https://doi.org/10.1111/1755-0998.12157>
- Eckert AJ, Bower AD, González-Martínez SC, Wegrzyn JL, Coop G, Neale DB (2010) Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Mol Ecol* 19:3789–3805
- Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol* 29:51–63
- Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troglio M, Myles S, Martínez-Zapater JM, Zyprian E, Moreira FM, Grandó MS (2013) Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. *BMC Plant Biol* 13:39. <https://doi.org/10.1186/1471-2229-13-39>
- Escribano-Ávila G, Sanz-Pérez V, Pías B, Virgós E, Escudero A, Valladares F (2012) Colonization of abandoned land by *Juniperus thurifera* is mediated by the interaction of a diverse dispersal assemblage and environmental heterogeneity. *PLoS One* 7:e46993
- Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, Holderegger R, Widmer A (2017) Estimating genomic diversity and population differentiation— an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. *BMC Genomics* 18:69
- Frankham R (2005) Genetics and extinction. *Biol Conserv* 126:131–140
- Gabriel S, Ziaugra L, Tabbaa D (2009) SNP genotyping using Sequenom MassARRAY iPLEX platform. *Curr Protoc Hum Genet*, Chapter 2: Unit 2.12. <https://doi.org/10.1002/0471142905.hg0212s60>
- García C, Moracho E, Diaz-Delgado R, Jordano P (2014) Long-term expansion of juniper populations in managed landscapes: patterns in space and time. *J Ecol* 102:1562–1571. <https://doi.org/10.1111/1365-2745.12297>
- García-Cervigon AI, Velázquez E, Wiegand T, Escudero A, Olano JM (2016) Colonization in Mediterranean old-fields: the role of dispersal and plant-plant interactions. *J Veg Sci* 28:627–638
- Geng Q, Qing H, Ling Z, Jeelani N, Yang J, Yoshikawa K, Miki NH, Wang Z, Lian C (2017) Characterization of polymorphic microsatellite markers for a coniferous shrub *Juniperus sabina* (Cupressaceae). *Plant Species Biol* 32:252–255. <https://doi.org/10.1111/1442-1984.12152>
- Glover KA, Hansen MM, Lien S, Als TD, Hoyhem B, Skaala O (2010) A comparison of SNP and STR loci for delineating population structure and performing individual genetic assignment. *BMC Genet* 11: 2–12
- Graudal L, Aravanopoulos F, Bennadji Z, Changtragoon S, Fady B, Kjær ED et al. (2014) Global to local genetic diversity indicators of evolutionary potential in tree species within and outside forests. *For Ecol Manag* 333:35–51
- Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. *PLoS One* 2:e1367
- Hodel RG, Segovia-Salcedo MC, Landis JB, et al. (2016) The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. *Appl Plant Sci* 4(6): apps.1600025. <https://doi.org/10.3732/apps.1600025>
- Holleley CE, Geerts PG (2009) Multiplex manager 1.0: a cross-platform computer program that plans and optimizes multiplex PCR. *BioTechniques* 46:511–517
- Jordan R, Hoffmann Ary A, Dillon Shannon K, Prober Suzanne M (2017) Evidence of genomic adaptation to climate in *Eucalyptus microcarpa*: implications for adaptive potential to projected climate change. *Mol Ecol* 26:6002–6020. <https://doi.org/10.1111/mec.14341>
- Li Z, Zou J, Mao K, Lin K, Li H, Liu J, Källman T, Lascoux M (2012) Genetic evidence for complex evolutionary histories of four high altitude juniper species in the Qinghai–Tibetan plateau. *Evolution* 66:831–845
- Malausa T et al (2011) High-throughput microsatellite isolation through 454 GS-FLX titanium pyrosequencing of enriched DNA libraries. *Mol Ecol Res* 11:644
- Mao K, Hao G, Liu J, Adams RP, Milne RI (2010) Diversification and biogeography of *Juniperus* (Cupressaceae): variable diversification rates and multiple intercontinental dispersal. *New Phytol* 188:254–272

- Mao K et al (2012) Distribution of living Cupressaceae reflects the break-up of Pangea. *Proc Natl Acad Sci U S A* 20:7793–7798
- McCoy RC, Garud NR, Kelley JL, Boggs CL, Petrov DA (2013) Genomic inference accurately predicts the timing and severity of a recent bottleneck in a non-model insect population. *Mol Ecol* 23:136–150
- Millar CI, Stephenson NL (2015) Temperate forest health in an era of emerging megadisturbance. *Science* 349:823–826
- Molecular Ecology Resources Primer Development Consortium et al (2013) Permanent genetic resources added to molecular ecology resources database 1 February 2013–31 March 2013. *Mol Ecol Res* 13:760–762. <https://doi.org/10.1111/1755-0998.12121>
- Nagy S, Poczai P, Cernak I, Tallér J (2012) PICcalc: an online program to calculate polymorphic information content for molecular genetic studies. *Biochem Genet* 50:670–672
- Narum SR (2006) Beyond Bonferroni: less conservative analyses for conservation genetics. *Conserv Genet* 7:811–811. <https://doi.org/10.1007/s10592-006-9189-7>
- Nazareno AG, Bemmels JB, Dick CW, Lohmann LG (2017) Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Mol Ecol Resour* 17:1136–1147. <https://doi.org/10.1111/1755-0998.12654>
- Neale DB, Ingvarsson PK (2008) Population, quantitative and comparative genomics of adaptation in forest trees. *Curr Opin Plant Biol* 11:149–155
- Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nat Rev Genet* 12:111–122
- Olsson S, Seoane-Zonjic P, Bautista R, Claros MG, González-Martínez SC, Scotti I, Scotti-Saintagne C, Hardy OJ, Heuertz M (2017) Development of genomic tools in a widespread tropical tree, *Symphonia globulifera* L.f.: a new low-coverage draft genome, SNP and SSR markers. *Mol Ecol Resour* 17:614–630. <https://doi.org/10.1111/1755-0998.12605>
- Opgenoorth L, Vendramin GG, Mao K, Miede G, Miede S, Liepelt S, Liu J, Ziegenhagen B (2010) Tree endurance on the Tibetan plateau marks the world's highest known tree line of the last glacial maximum. *New Phytol* 185:332–342. <https://doi.org/10.1111/j.1469-8137.2009.03007.x>
- Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28:2537–2539. <https://doi.org/10.1093/bioinformatics/bts460>
- Peery MZ et al (2012) Reliability of genetic bottleneck tests for detecting recent population declines. *Mol Ecol* 21:3403–3418. <https://doi.org/10.1111/j.1365-294X.2012.05635.x>
- 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. <https://doi.org/10.1371/journal.pone.0037135>
- Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. *Annu Rev Ecol Evol Syst* 37:187–214
- Pierson JC, Coates DJ, Oostermeijer JGB, Beissinger SR, Bragg JG, Sunnucks P, Schumaker NH, Young AG (2016) Genetic factors in threatened species recovery plans on three continents. *Front Ecol Environ* 14:433–440. <https://doi.org/10.1002/fee.1323>
- Pukk L, Ahmad F, Hasan SB, Kisand V, Gross R, Vasemägi A (2015) Less is more: extreme genome complexity reduction with ddRAD using ion torrent semiconductor technology. *Mol Ecol* 15:1145–1152
- Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Res* 8:103–106
- Rozen S, Skaletsky H (1999) Primer3 on the WWW for general users and for biologist programmers. In: Misener S, Krawetz SA (eds) *Bioinformatics methods and protocols*. Humana Press, Totowa, pp 365–386. <https://doi.org/10.1385/1-59259-192-2:365>
- Schoebel CN, Brodbeck S, Buehler D, Comejo C, Gajurel J, Hartikainen H, Keller D, Leys M, Ricanova S, Segelbacher G, Werth S, Csencsics D (2013) Lessons learned from microsatellite development for nonmodel organisms using 454 pyrosequencing. *J Evol Biol* 26:600–611
- Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nat Biotechnol* 18:233–234
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19(6):1117–1123
- Singh N, Choudhury DR, Singh AK, Kumar S, Srinivasan K, Tyagi RK, Singh NK, Singh R (2013) Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLoS One* 8:e84136. <https://doi.org/10.1371/journal.pone.0084136>
- Soltis PS, Marchant DB, Van der Peer Y, Soltis DE (2015) Polyploidy and genome evolution in plants. *Curr Opin Genet Dev* 35:119–125
- Teixeira H, Rodríguez-Echeverría S, Nabais C (2014) Genetic diversity and differentiation of *Juniperus thurifera* in Spain and Morocco as determined by SSR. *PLoS One* 9:e88996. <https://doi.org/10.1371/journal.pone.0088996>
- Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor Appl Genet* 106:411–422. <https://doi.org/10.1007/s00122-002-1031-0>
- Trumbore S, Brando P, Hartmann H (2015) Forest health and global change. *Science* 349:814–818. <https://doi.org/10.1126/science.aac6759>
- van Asch B, Pinheiro R, Pereira R, Alves C, Pereira V, Pereira F, Gusmão L, Amorim A (2010) A framework for the development of STR genotyping in domestic animal species: Characterization and population study of 12 canine X-chromosome loci. *Electrophoresis* 31:303–308
- Van Auken OW (ed) (2008) *Western North American Juniperus communities*, Ecological Studies. Springer, San Antonio
- Wang J (2002) An estimator for pairwise relatedness using molecular markers. *Genetics* 160:1203–1215
- Wang J (2011) Coancestry: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Mol Ecol Res* 11:141–145. <https://doi.org/10.1111/j.1755-0998.2010.02885.x>
- Waples RS (2015) Testing for Hardy-Weinberg proportions: have we lost the plot? *J Hered* 106:1–9
- Waters JM, Fraser CI, Hewitt GM (2013) The founder takes all: density-dependent processes structure biodiversity. *Trends Ecol Evol* 28:75–85
- Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, LeRoy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM, Fischer DG, Gehring CA, Lindroth RL, Marks JC, Hart SC, Wimp GM, Wooley SC (2006) A framework for community and ecosystem genetics: from genes to ecosystems. *Nat Rev Genet* 7:510–523. <https://doi.org/10.1038/nrg1877>
- Yang X, Xu Y, Shah T, Li H, Han Z, Li J, Yan J (2011) Comparison of SSRs and SNPs in assessment of genetic relatedness in maize. *Genetica* 139:1045–1054. <https://doi.org/10.1007/s10709-011-9606-9>
- Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *J Comput Biol* 7:203–214