Gene Duplication, Population Genomics, and Species-Level Differentiation within a Tropical Mountain Shrub

Alicia Mastretta-Yanes[1,*], Sergio Zamudio[2], Tove H. Jorgensen[3], Nils Arrigo[4], Nadir Alvarez[4], Daniel Piñero[5], and Brent C. Emerson[1,6]

[1]Centre for Ecology, Evolution and Conservation, School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom

[2]Centro Regional del Bajío, Instituto de Ecología A. C., Pátzcuaro, Michoacán, México

[3]Department of Bioscience, Aarhus University, Denmark

[4]Department of Ecology and Evolution, Biophore Building, University of Lausanne, Switzerland

[5]Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico

[6]Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), San Cristóbal de La Laguna, Santa Cruz de Tenerife, Spain

*Corresponding author: E-mail: a.yanes@uea.ac.uk.

Accepted: September 8, 2014

Data deposition: This project has been deposited at Dryad Repository under the accession doi:10.5061/dryad.n3jk5.

Abstract

Gene duplication leads to paralogy, which complicates the de novo assembly of genotyping-by-sequencing (GBS) data. The issue of paralogous genes is exacerbated in plants, because they are particularly prone to gene duplication events. Paralogs are normally filtered from GBS data before undertaking population genomics or phylogenetic analyses. However, gene duplication plays an important role in the functional diversification of genes and it can also lead to the formation of postzygotic barriers. Using populations and closely related species of a tropical mountain shrub, we examine 1) the genomic differentiation produced by putative orthologs, and 2) the distribution of recent gene duplication among lineages and geography. We find high differentiation among populations from isolated mountain peaks and species-level differentiation within what is morphologically described as a single species. The inferred distribution of paralogs among populations is congruent with taxonomy and shows that GBS could be used to examine recent gene duplication as a source of genomic differentiation of nonmodel species.

Key words: RAD-seq, de novo assembly, GBS, paralogy, Transmexican Volcanic Belt, *Berberis*.

Introduction

The development of genotyping-by-sequencing (GBS) methods (reviewed by Davey et al. 2011; Poland and Rife 2012) has accelerated the use of genomic data in population genetic studies of nonmodel organisms. This is particularly useful for plants, where population genetic studies have often struggled to obtain sufficient resolution from DNA sequence data with traditional Sanger sequencing approaches. For example, several plant phylogeographic studies (e.g., Tovar-Sánchez et al. 2008; Gugger et al. 2011; Mastretta-Yanes et al. 2011) have been substantially less informative than studies that have used comparable sequencing effort in animal taxa within the same geographic region (e.g., McCormack et al. 2008; Bryson et al. 2011, 2012; Ornelas et al. 2013). By applying GBS techniques sufficient nucleotide variation can be harnessed within plant species to address evolutionary questions, such as genetic association of adaptive traits (Parchman et al. 2012) and genomic divergence of hybridizing tree species (Stölting et al. 2013). However, applying GBS to plants poses a unique set of challenges, or exacerbates those common to other taxa (Deschamps et al. 2012; Morrell et al. 2012; Schatz et al. 2012). Plant genomes typically contain a large number of transposable elements (Feschotte et al. 2002), which causes GBS reads to map with equal probability to multiple positions within a reference genome. Polyploidy events have also occurred frequently throughout the evolutionary history of plant

species, as well as other types of gene duplication that can result in large multigene families (Lockton and Gaut 2005; Flagel and Wendel 2009), and thus a considerable amount of paralogous loci. Paralogous loci are typically treated as a nuisance variable and filtered from GBS data; however, the emergence of paralogous loci is a consequential process that contributes to genome evolution, and can thus be examined for the quantification of genomic differentiation among populations and species.

Paralogous loci arise by gene duplication, such that both copies evolve in parallel during the history of an organism (Fitch 1970; fig. 1*a*). Gene duplication can occur at the whole genome level (polyploidy event), but can also be limited to chromosome segments or single genes (Hurles 2004). Gene duplication can confound the assembly of genomic data because paralogs can be erroneously merged together as a single locus (fig. 1*c*), leading to difficulty in distinguishing allelic variation from differences among closely related gene family members (Dou et al. 2012; Hohenlohe et al. 2012). This issue is caused by relatively recent gene duplications (i.e., those origination within a genus or among closely related species), because more ancient duplication events occurring over much deeper time scales are expected to have accumulated enough differences to be assembled as different loci (fig. 1*d*). The confounding effect of gene duplication on the assembly of genomic data is particularly problematic for de novo assembly, but even if a reference genome is available, the short sequence reads that are typical of high-throughput sequencing may not map uniquely within a reference genome (Hohenlohe et al. 2012; Morrell et al. 2012).

Treating paralogs as a single locus generates spurious heterozygous genotype calls and can confound the estimation of genetic differentiation among individuals and populations. The magnitude of this effect will depend upon the characteristics of the focal genome, and the relatedness of the samples being analyzed. With regard to focal genome characteristics, plant and fish genomes contain more duplicated genes than mammals (Volff 2004; Lockton and Gaut 2005) and will thus, on average, provide a greater challenge for genome assembly because of paralogous loci. The evolutionary relatedness among samples is also important because paralogs are continuously arising within each evolutionary lineage (Lynch and Conery 2000; Hurles 2004; Langham et al. 2004). Thus, the more a focal group departs from a model of panmixia, the more paralogous loci one would expect to retrieve across all samples. In the extreme, one may expect different species, or sufficiently differentiated populations, to exhibit species-specific or population-specific paralogs.

Paralogous loci are typically entirely filtered from GBS data. This can be done at the stage of assembly and genotyping, for instance, by incorporating differences in coverage (Dou et al. 2012) or by testing the independence of biallelic single nucleotide polymorphisms (SNPs) for each pair of tags (Poland et al. 2012, but see also Gayral et al. 2013; Eaton 2014 for other approaches). Filtering can also be performed on the assembled data, for example, by retaining only those loci with the number of expected alleles and Hardy–Weinberg proportions (Hohenlohe et al. 2011; Catchen et al. 2013).

Despite gene duplication representing an analytical challenge for GBS, it is also a major source of evolutionary novelty

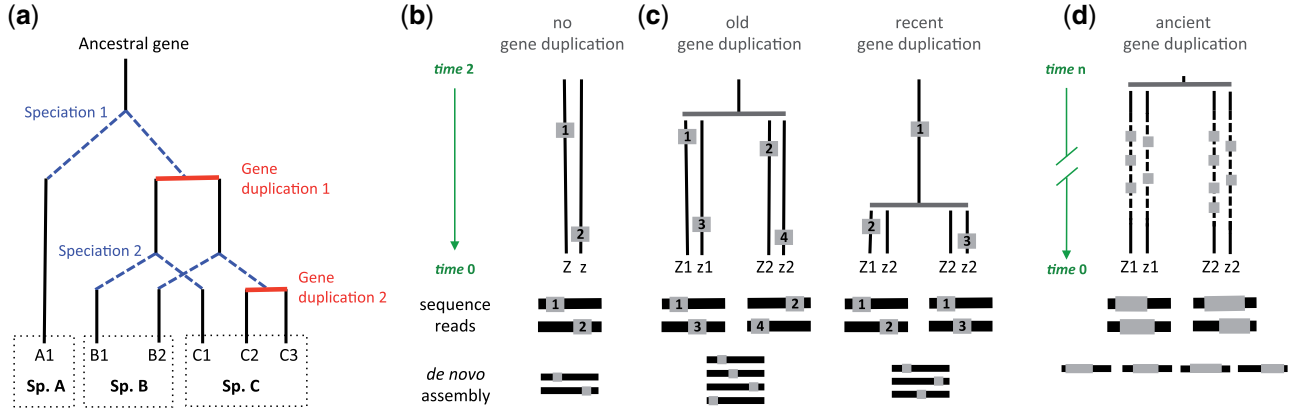

Fig. 1.—(*a*) Paralogy and orthology relationships among six contemporary genes (A1–C3) in three species (A–C), adapted from Jensen (2001). Paralogous genes are produced by duplication events (red horizontal line) and orthologous by speciation (blue dashed inverted "Y"). A given gene in one species may have more than one ortholog in another species (e.g., B1 and B2 in species B are orthologs of A1 in species A) and paralogs are not necessarily restricted to the same species (e.g., B1 and C2 are paralogs). (*b*) On a locus Z (with the alleles Z and z), mutation events (gray boxes) lead to the formation of two possible sequence reads (coverage not shown) that are correctly assembled as two alleles of the same locus. (*c*) Loci that are the product of gene duplication (Z1 and Z2) produce reads that cannot be distinguished from allelic variation and are assembled as a single locus with several alleles, generating erroneous SNP calls. Loci produced by relatively old duplication events would accumulate more nucleotide differences than recently duplicated loci. Therefore, if paralogs are merged as single locus, the products of old duplication events will generate more (spurious) alleles than paralogs from more recent duplication events. (*d*) Loci produced by more ancient duplication events would accumulate enough differences to be assembled as different loci.

(Ohno 1970; Lewis 1951). Therefore, by treating paralogs as a nuisance parameter and discarding them, potential signatures of evolution and adaptation are also being discarded. A duplicated gene copy may acquire a new function (Ohno 1970), specialize for a subset of the functions originally performed by the ancestral single-copy gene (Lynch and Force 2000), or contribute to protein dosage effects in response to environmental variables (Kondrashov et al. 2002). These processes are particularly relevant for plant evolution, as most plant diversity seems to have arisen following the duplication and adaptive specialization of preexisting genes (Lockton and Gaut 2005; Moore and Purugganan 2005; Flagel and Wendel 2009). For example, many plant genes involved in pathogen recognition and herbivory defense arose through gene duplication (Moore and Purugganan 2005). However, there are also several examples of adaptive gene duplications in bacteria, yeast, fish, insect, and mammal species (Kondrashov 2012). In addition to functional diversification, gene duplication can also promote speciation through the passive accumulation of genomic divergence (Lynch and Conery 2000). For example, following the duplication of an essential gene in *Arabidopsis thaliana*, populations varied with respect to the copy that retained functionality, which acts as a postzygotic barrier among populations (Bikard et al. 2009).

Here, rather than seeking to remove paralogous loci, we use GBS data for the explicit purpose of investigating the distribution of putative recent gene duplication events among plant populations. We use double-digest restriction-site associated DNA sequencing (ddRAD) data sampled from the nonmodel plant species *Berberis alpina* (Berberidaceae) and its close relatives to characterize both 1) genomic relationships among individuals based on putative orthologs and 2) the distribution of paralogous loci of recent origin among sampling localities and species. The inferred distribution of paralogous loci among sampling locations and species is congruent with genomic differentiation estimated from presumed orthologous loci, and reveals species-level differentiation within what is morphologically described as a single species. More broadly, our study shows that GBS can be used to study, without a reference genome, gene duplication as a source of population divergence and evolutionary novelty in nonmodel species.

Materials and Methods

Study System and Sampling

Berberis alpina is a shrub that grows from 3,200 to 4,200 m above sea level (masl) on alpine grasslands of the Transmexican Volcanic Belt (TMVB), a system of isolated high-altitude mountains in tropical Mexico (fig. 2). The TMVB is a biodiversity hotspot (Myers et al. 2000) where temperate-to-cold adapted plant species are thought to have either survived through, or diversified in situ during, the Pleistocene climate fluctuations (Toledo 1982; Graham 1999). *Berberis moranensis* grows at lower altitudes in the TMVB (1,800–3,150 masl; Zamudio 2009a) and is expected to be closely related to *B. alpina*.

Mountain peaks from 3,300 to 4,200 masl within the TMVB and nearby areas of the Altiplano Sur (AS) and of the Sierra Madre Oriental (SMOr) were surveyed for *B. alpina* (sensu Zamudio 2009b) during September–October 2010 and April–May 2011 (fig. 2). The species was found in a total of seven locations, which represents its known distribution within the TMVB and the AS (fig. 2). It was not found in the surveyed mountains of the SMOr. Samples of *B. moranensis,* a closely related species that grows up to 3,150 masl, were collected in Cerro San Andrés (fig. 2), where *B. alpina* is absent. Samples of the outgroups *Berberis trifolia* and *Berberis pallida* were collected at lower elevations (~2,000–2,300 masl) of the TMVB (fig. 2) in October 2012. Sampling was performed with SEMARNAT permission No. SGPA/DGGFS/712/2896/10. Herbarium specimens of *B. alpina* and *B. moranensis* were prepared and deposited within the Herbario Nacional in Mexico City (MEXU).

Molecular Methods

Based on data from related species, the sampled *Berberis* species are likely diploid with a genome size of between 0.50 and 1.83 Gb (Rounsaville and Ranney 2010). We used ddRAD data from Mastretta-Yanes et al. (2014a), which consist of 75 individually tagged specimens of *B. alpina* and *B. moranensis* (6–10 per population), 3 samples of each outgroup (*B. trifolia* and *B. pallida*), and 15 replicated samples, with at least one replicate per population or species. Briefly, the ddRAD libraries were prepared using the enzymes *Eco*RI-HF and *Mse*I using a modified version of Parchman et al. (2012) and Peterson et al. (2012) protocols. Samples were divided into three groups, each sequenced using single-end reads (100 bp long) in a separate lane of an Illumina HiSeq2000.

De Novo Assembly of RAD Data

After demultiplexing and quality trimming of raw reads, final sequences were 84 bp long. Data were de novo assembled using the software Stacks v. 1.02 (Catchen et al. 2011, 2013) with the parameter values $m = 3$, $M = 2$, $N = 4$, $n = 3$, *max_locus_stacks* = 3, and an SNP calling model with an upper bound of 0.05. These settings 1) optimize the recovery of a large number of loci while reducing the SNP and RAD allele error rates, and 2) filter a fraction of putative paralogous loci merged as a single locus (Mastretta-Yanes et al. 2014b). After de novo assembly, the data were filtered to keep only those samples having more than 50% of the mean number of loci per sample, and only those loci present in at least 80% of the barcoded samples. Replicates were used to estimate error rates as in Mastretta-Yanes et al. (2014b) for each of the subsets of samples described in the sections below. For the

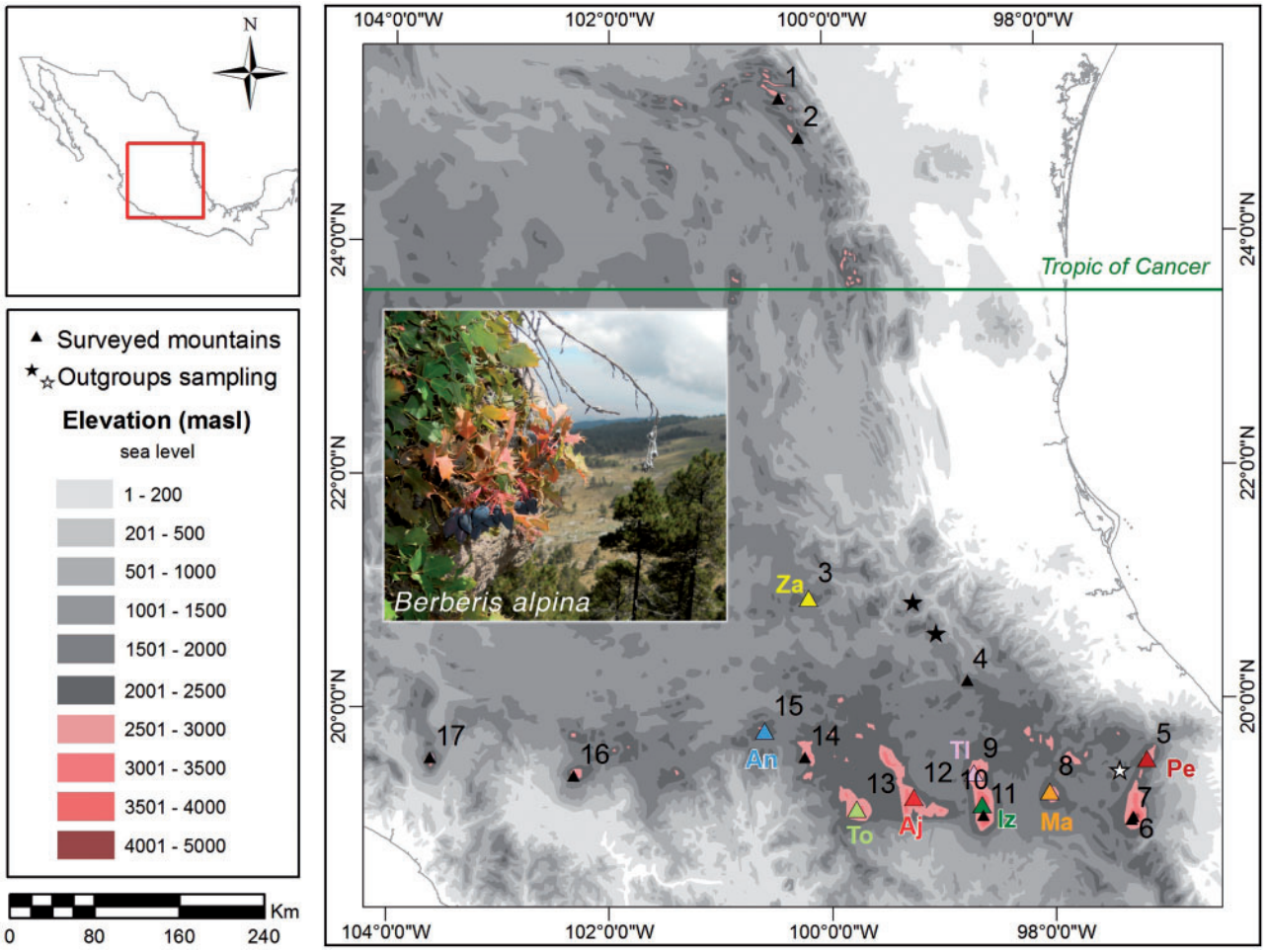

Fig. 2.—Surveyed mountains for *B. alpina* within the SMOr (1–2), AS (3), and the TMVB (4–17). Populations where *B. alpina* was found are Nevado de Toluca (To), Ajusco (Aj), Tlaloc (Tl), Iztaccihuatl (Iz), La Malinche (Ma) and Cofre de Perote (Pe) (To–Pe are referred as *B. alpina* ingroup), and Zamorano (Za). *Berberis moranensis* was collected from Cerro San Andrés (An, blue triangle). *Berberis pallida* (black stars) and *B. trifolia* (white star) were outgroups.

population genomic analyses, only one sample per replicate pair was used.

Considerably fewer loci were recovered in *B. pallida,* which is likely explained by mutations affecting restriction enzyme cutting sites and hence a distant evolutionary relationship with the other *Berberis* species in the study. This species was therefore excluded from further analyses.

Identifying Paralogs from Recent Gene Duplications

Here, we refer to a RAD-locus as a short DNA sequence produced by clustering together RAD-alleles; in turn, RAD-alleles differ from each other by a small number of SNPs in certain nucleotide positions (SNP-loci). During de novo assembly, two nucleotide mismatches ($M = 2$) were allowed among reads to form a putative RAD-locus. Among individuals, loci were merged as a single locus if they presented up to three mismatches ($n = 3$), which allows loci that are fixed differentially among different populations or species (thus represented independently across individuals) to be merged as a single locus (Stacks manual). During the formation of putative RAD-loci within individuals (determined by the *-M* parameter), and during the merging of monomorphic loci among individuals (determined by the *-n* parameter), it is expected that paralogous loci would be assembled as a single locus, leading to the formation of loci with three or more (spurious) alleles (fig. 1*c*). Thus, if more than two alleles per locus are allowed during de novo assembly, data will likely contain merged paralogs. Here, a maximum of three alleles per locus was allowed (*max_locus_stacks* = 3) to filter out paralogs of relatively old origin. This filter retains paralogs derived from more recent gene duplications events, because loci produced by recent gene duplications are expected to have accumulated fewer mutations than older duplicated loci (fig. 1*c*), and should thus produce less (spurious) alleles if merged as a single locus. Notice that "old origin" is a relative term, implying that loci are still similar enough to resemble allelic variation. Paralogs from more ancient duplications, such as those ones shared

across many genera and plant families (Lockton and Gaut 2005), are expected to have accumulated enough differences to be assembled as different loci (fig. 1d).

Polymerase chain reaction (PCR) and sequencing error may also result in more than two alleles per locus within an individual (Hohenlohe et al. 2012; Catchen et al. 2013). However, the distribution of error-based alleles is stochastic, whereas merged paralogous loci should produce population-wide shared polymorphism. Thus, merged paralogs can be identified by their signature on the site frequency spectrum (SFS; Hohenlohe et al. 2012): Paralogous loci accumulate mutations independently, so assembling them as different alleles of the same locus produces spurious polymorphic positions at which all individuals would be heterozygous, with the exception of those that may have suffered allele dropout. This should bias the SFS toward heterozygosity with an excess of loci where the frequency of the major allele (p) is $p = 0.5$. Here, we consider any RAD-locus where $p = 0.5$ in at least one SNP-locus within a given population to be a potentially paralogous locus. Such loci were further examined among other populations and species, because some orthologous loci may by chance be at $p = 0.5$ in a given population, but it would be unlikely to observe this in two or more populations or within a related species. If an RAD-locus was identified as a potential paralog in two or more populations or species, it was considered to be shared among those taxa. However, if $p = 0.5$ in only one population or species, the RAD-locus was considered to be a private potential paralog (i.e., the locus was present in other populations, but with $p \neq 0.5$).

The data set was divided into the following three subsets of RAD-loci: 1) All loci, 2) putative orthologs—excluding all potential paralogs, and 3) putative orthologs within *B. alpina*—excluding potential paralogs shared between two or more sampling locations of *B. alpina*, or between two or more species, which generates a subset of loci that should be orthologous within *B. alpina*. The frequency of the major allele within each locus was estimated for each of the three data sets. Allele frequencies were estimated at each SNP-locus for each population and species by running the "populations" program of Stacks version 1.17 with the de novo assembled RAD-loci. The distribution of potential paralogous loci was examined and plotted with R version 2.15 (R Core Team 2012).

Structuring of Genetic Variation and Population Genomic Analyses

Preliminary analyses revealed the Cerro Zamorano population to be highly differentiated from other *B. alpina* populations (see Discussion), so it was treated as a different lineage from *B. alpina*. Hereafter, we use "*B. alpina* ingroup" to refer to the subset of *B. alpina* samples that excludes the Cerro Zamorano population.

Principal Coordinate Analysis (PCoA) was performed for all loci and for the putative orthologs. For each of these two data sets the PCoA was first performed with all samples, and then excluding the outgroup and the Cerro Zamorano population. Pairwise F_{ST} between populations were estimated using both subsets of loci. The percentage of polymorphic loci, heterozygosity, π, and F_{IS} at each nucleotide position were estimated for *B. alpina* ingroup using all loci and the subset of putative orthologs within *B. alpina*. All population genetic estimates were calculated using the "populations" program of Stacks.

Distribution of Potential Paralogous Loci among Populations and Species

The distribution of shared and private potential paralogs among populations and species was further examined by controlling for unequal sample sizes, by randomly sampling four individuals (the smallest sample size) per locality. Total, shared and private potential paralogous loci were identified as described above, with the exception that shared loci were defined as those shared with *B. alpina* ingroup populations.

A linear regression was used to test whether the proportion of private potential paralogous loci increases with population differentiation, the latter calculated as the mean F_{ST} per population using the putative orthologous loci subset (pairwise matrix from table 1, without the outgroup). The analysis was performed using R with and without the Cerro Zamorano population.

Table 1

Pairwise F_{ST} for the Putative Orthologs Subset (Filtering Out All Putative Paralogous Loci)

	Iz	*Ma*	*Pe*	*Tl*	*To*	Za	An	Out
Aj	0.0383	0.0663	0.0972	0.0248	0.0534	0.5387	0.0757	0.4649
Iz		0.0648	0.1042	0.0299	0.0643	0.5623	0.0973	0.4909
Ma			0.0954	0.0582	0.0903	0.5634	0.1377	0.4932
Pe				0.0848	0.1216	0.4991	0.1609	0.4050
Tl					0.0534	0.5861	0.0984	0.5074
To						0.6116	0.1276	0.5339
Za							0.7225	0.6976
An								0.6393

NOTE.—*Berberis alpina* ingroup populations are shown in italics in the first five columns. *Berberis moranensis* (An) and *Berberis trifolia* (Out) are in the last columns and are shown as a reference for the values found among different species. El Zamorano (Za) population shows F_{ST} values higher than those found for *B. moranensis* (An) and *B. trifolia* (Out).

Morphological Evaluation

We examined characters that have been informative for Mexican *Berberis* taxonomy (Zamudio 2009a, 2009b) to assess morphological and ecological differentiation among populations of *B. alpina*. Variables included leaf morphology (rachis length, number of leaflets, leaflet texture, shape of blades, number and length of teeth), growth habit and habitat preferences (vegetation type, substratum, and altitudinal distribution). Morphological characters were examined in specimens from extant herbarium material (MEXU, ENCB, IEB, XAL) including two *B. alpina* populations from SMOr (Sierra del Doctor and Cerro Pingüical; table 2) that was not possible to sample for the molecular analysis. Habitat characteristics and altitudinal distribution were recorded from field observations. Specimens of *B. moranensis* from throughout its distribution were also examined for comparison.

Results

RAD-Seq Data Yield and Error Rates

The number of samples recovered (excluding one sample per replicate pair) after de novo assembly and quality filtering was 2 for *B. trifolia*, 9 for *B. moranensis*, 4 for the Cerro Zamorano population, and 6–10 for each *B. alpina* population (supplementary table S1, Supplementary Material online). A total of 6,292 RAD-loci (84 bp long) and 6,105 SNP-loci were recovered after the de novo assembly and quality control steps. For the subset of putative orthologs (filtering all potential paralogs), a total of 4,030 RAD-loci and 3,843 SNP-loci were recovered. A total of 5,461 RAD-loci and 5,274 SNP-loci were recovered for the subset of putative orthologs within *B. alpina*. RAD-allele and SNP error rates, percentage of missing data, and mean coverage per locus per sample are reported in table 3. Broadly, for each data set the allele error rate ranges from 3.5% to 5.9% and the SNP error rate from 1.3% to 2.2%, with approximately 20% of missing data and a mean coverage of approximately 10.5 (table 3). Decreases from 5.9% to 4.1% for the RAD-allele error rate and from 2.2% to 1.5% for the SNP error rate represent significant differences ($P < 0.001$ and $P < 0.01$, respectively) between the data set of all loci and the data set excluding the 831 putative paralogs within *B. alpina* ingroup (see below for how these loci were defined). To confirm that this was not a chance effect, we randomly filtered 831 loci from the entire data set of samples and reestimated error rates. We repeated this 100 times, and across all repetitions the RAD-allele and SNP-locus error rates were 5.9–6.0% and 2.1–2.2%, respectively, which are not significantly different from the error rates found when no loci are removed ($P > 0.7$ for all repetitions for both types of error rate).

Identification and Distribution of Paralogous Loci *among* Populations and Species

A total of 2,262 RAD-loci were identified as potential paralogous loci. When examining the subset of all loci, the frequency

Table 2
Morphological Differences of *Berberis alpina* Populations and *Berberis moranensis*

	B. alpina			*B. moranensis*
Character	TMVB[a]	Pe[a]	Za and SMOr[b]	TMVB[c]
Growth habit	Low shrub 50–100 cm, or more	Low shrub 25–100 cm or more	Low shrub 10–60 cm	Shrub to tree 1–7(10) m
No. of leaflets	3–5(7)	3–5	3	(5)7–11(15)
Rachis length (terminal segment), cm	0.5–2(3)	1–2	Absent	(0.3)0.5–1.5(2)
Leaflets texture	Coriaceous	Coriaceous	Coriaceous and very rigid	Slightly coriaceous
Leaflets blades	Ovate to ample ovate	Ovate, oblong to elliptic	Oblong to elliptic	Lanceolate to ovate-lanceolate
No. of teeth by side	(3)4–7(9)	(3)4–7(12)	2–4(6)	(4)5–11(15)
Teeth length, mm	(1)2–5	2–5	5–10	1–2(5)
Substratum	Igneous rocks	Igneous rocks	Igneous or calcareous rocks	Igneous rocks
Vegetation type	Alpine grassland and upper limit of *Pinus hartwegii* and *Abies religiosa* forests	Alpine grassland and upper limit of *P. hartwegii* and *Ab. religiosa* forests	*Ab. religiosa*, *Pinus* spp. and *Quercus* spp. forests, never above timber line	*Ab. religiosa*, *Pinus* spp. and *Quercus* spp. forests and secondary vegetation after perturbation, never above timberline
Altitudinal distribution (masl)	3,200–4,200	3,300–4,180	2,800–3,250	(1,800)2,000–2,800(3,150)

[a]TMVB refers to sampled populations for *B. alpina* in the TMVB (*B. alpina* ingroup) as in figure 2, with the exception of Cofre de Perote (Pe) population.
[b]Cerro Zamorano (fig. 2) and SMOr populations: Sierra del Doctor (20°47′25″N, 99°33′53″W at 3,250 masl) and Cerro Pingüical (21°09′35″N, 99°42′02.4″W at 3,060 masl).
[c]Several localities within the TMVB at 1,800–3,150 masl.

of the major allele for each SNP-locus reveals that the majority of loci that are polymorphic across populations are fixed within each population (fig. 3*a*). The percentage of loci in the other categories decreases sharply and monotonically, but then increases abruptly within the category containing loci where $p = 0.5$. For *B. alpina* ingroup populations, the observed heterozygosity of 91% of these loci is $H_{obs} = 1$ and the F_{IS} value of 98% of the loci is negative, with $F_{IS} \leq -0.5$ in 77% of the cases. Out of the 2,262 potential paralogous loci, 831 have at least one SNP with $p = 0.5$ in two or more populations or species, and were considered putative paralogs within *B. alpina* ingroup. Around 99% of these SNP-loci show negative F_{IS} values for *B. alpina* ingroup populations, with $F_{IS} \leq -0.5$ in 69% of them and $H_{obs} = 1$ in 57%. Retaining only the presumable orthologs within *B. alpina* ingroup does not remove the overrepresentation of SNP-loci with both alleles at equal frequency within *B. moranensis* and the Cerro Zamorano population (fig. 3*b*), but it effectively removes the excess of loci where $p = 0.5$ within all *B. alpina* ingroup populations (fig. 3*b*).

The potential paralogs are not evenly distributed among sampling locations and species. In increasing order, the Cerro Zamorano population and *B. moranensis* exhibit proportionally more RAD-loci with at least one SNP where $p = 0.5$ (fig. 4 and supplementary fig. S2, Supplementary Material online), the majority of which are private (fig. 4 and supplementary fig. S2, Supplementary Material online). In contrast, within a given population of the *B. alpina* ingroup fewer loci were found to be at $p = 0.5$ (fig. 4 and supplementary fig. S2, Supplementary Material online). The number of private potential paralogs per population increases with their differentiation estimated from orthologous loci (fig. 6), both when the Cerro Zamorano population is included ($r^2 = 0.955$, $F_{1,6} = 128.3$, $P < 0.001$) and when it is excluded ($r^2 = 0.771$, $F_{1,5} = 16.85$, $P < 0.01$). The distribution of total, private, and shared potential paralogous loci is similar under unequal sample sizes ($n = 2$–10; supplementary fig. S2, Supplementary Material online) and equal sample sizes ($n = 4$; fig. 4).

Structuring of Genetic Variation

The PCoA from the subset of putative orthologous loci reveals that the Cerro Zamorano population explains as much of the variance as the outgroup, *B. trifolia*, whereas *B. moranensis* clusters closer to the remaining *B. alpina* populations (fig. 5*a*). Excluding the Cerro Zamorano population and *B. trifolia* (fig. 5*b*), results in separate clusters for *B. moranensis,* and for both the Cofre de Perote and Malinche populations of *B. alpina*, while Western populations (Aj, Tl, and Iz; fig. 2) form a single cluster.

For *B. alpina* ingroup populations, the pairwise F_{ST} matrix estimated with the putative orthologs ranges from 0.025 to 0.122 (mean = 0.070), with Cofre de Perote exhibiting the highest differentiation in all pairwise estimates (0.084–0.122, table 1). Pairwise F_{ST} values of the Cerro Zamorano population against *B. alpina* ingroup populations are larger (0.499–0.612) than values obtained by comparing any *B. alpina* ingroup population against the outgroup (0.405–0.534) or against *B. moranensis* (0.076–0.161).

Genetic Diversity within *B. alpina* Ingroup

When considering all nucleotide positions (i.e., including those not polymorphic) of the presumably orthologous loci within *B. alpina* ingroup, the percentage of polymorphic loci (notice that locus here refers to a nucleotide position within the RAD-loci) ranged from 0.304% to 0.482%; the average frequency of the major allele from 0.9990 to 0.9994; H_{obs} from 0.0011 to 0.0014; and π from 0.0010 to 0.0016 (supplementary table S1, Supplementary Material online). Cofre de Perote presented the highest genetic diversity (0.482% polymorphic loci, $H_{obs} = 0.0014$, and $\pi = 0.0016$); Nevado de Toluca presented the lowest levels of genetic diversity (0.304% polymorphic loci, $H_{obs} = 0.0011$, and $\pi = 0.0010$), with the remainder of the populations exhibiting intermediate levels. Cofre de Perote has substantially more private alleles (1,064) than both the remaining populations (293–485, supplementary table S1, Supplementary Material online) and *B. moranensis* (194, supplementary table S1, Supplementary Material online). When the same statistics are estimated including all potential paralogs (supplementary table S2, Supplementary Material online), the estimates of genetic diversity increase (e.g., H_{obs} increased from ≤ 0.0015 to ≥ 0.0026) and all F_{IS} values are negative.

Morphological Variation

Specimens from Cerro Zamorano, Sierra del Doctor, and Cerro Pingüical populations of *B. alpina* (Za–SMOr populations) are low rhizomatous shrubs (20–60 cm) that tend to have only

Table 3
RAD-Seq Data Yield and Error Rate for Each Subset of Loci

Data Subset	RAD-Loci	SNP-Loci	RAD-Locus Error Rate (%)	RAD-Allele Error Rate (%)	SNP Error Rate (%)	Missing Data (%)	Mean Coverage*
All loci	6,292	6,105	17.4 (10.3)	5.9 (1.3)	2.2 (0.06)	20	10.3 (4.2)
Putative orthologs	4,030	3,843	17.5 (10.4)	3.5 (1.1)	1.3 (0.04)	17	11 (4.3)
Putative orthologs within *Berberis alpina*	5,461	5,274	17.28 (10.3)	4.1 (1.2)	1.5 (0.04)	17	10.5 (4.3)

NOTE.—SD values are given in parenthesis.

three leaflets per leaf and a sessile terminal leaflet (not inserted on a conspicuous rachis' segment) (table 2). In contrast, populations from the TMVB are dense, appressed shrubs (20 cm to 1 m or more), flattened against rocks or cliffs and tend to have 3–5 (max. 7) leaflets with the terminal leaflet always inserted on a conspicuous segment of the rachis (table 2). A ubiquitous rachis and more than five leaflets are characteristic traits of *B. moranensis* (table 2). *Berberis alpina* populations of the TMVB inhabit the highest elevations, mostly on igneous rocks of alpine grasslands, whereas *B. alpina* populations of the Za–SMOr do not grow beyond the timberline and can grow on calcareous rocks. TMVB populations can co-occur with *B. moranensis* in the upper limit of conifer forests (supplementary table S1, Supplementary Material online).

Discussion

Paralogs Identification

A total of 2,262 RAD-loci were identified as potential paralogs, out of which 831 RAD-loci presented SNPs with $p = 0.5$ in more than one population or species and were identified as putative paralogs within *B. alpina*. Removing these loci produced a set of presumably orthologous RAD-loci for the *B. alpina* ingroup. This is similar to the approach taken by Hohenlohe et al. (2011) and Pujolar et al. (2014) to produce a data set of putative orthologs for population genetics analyses of fish species, by removing loci with high values of observed heterozygosity. Here, we explored the excess of heterozygosity by examining whether the loci where $p = 0.5$

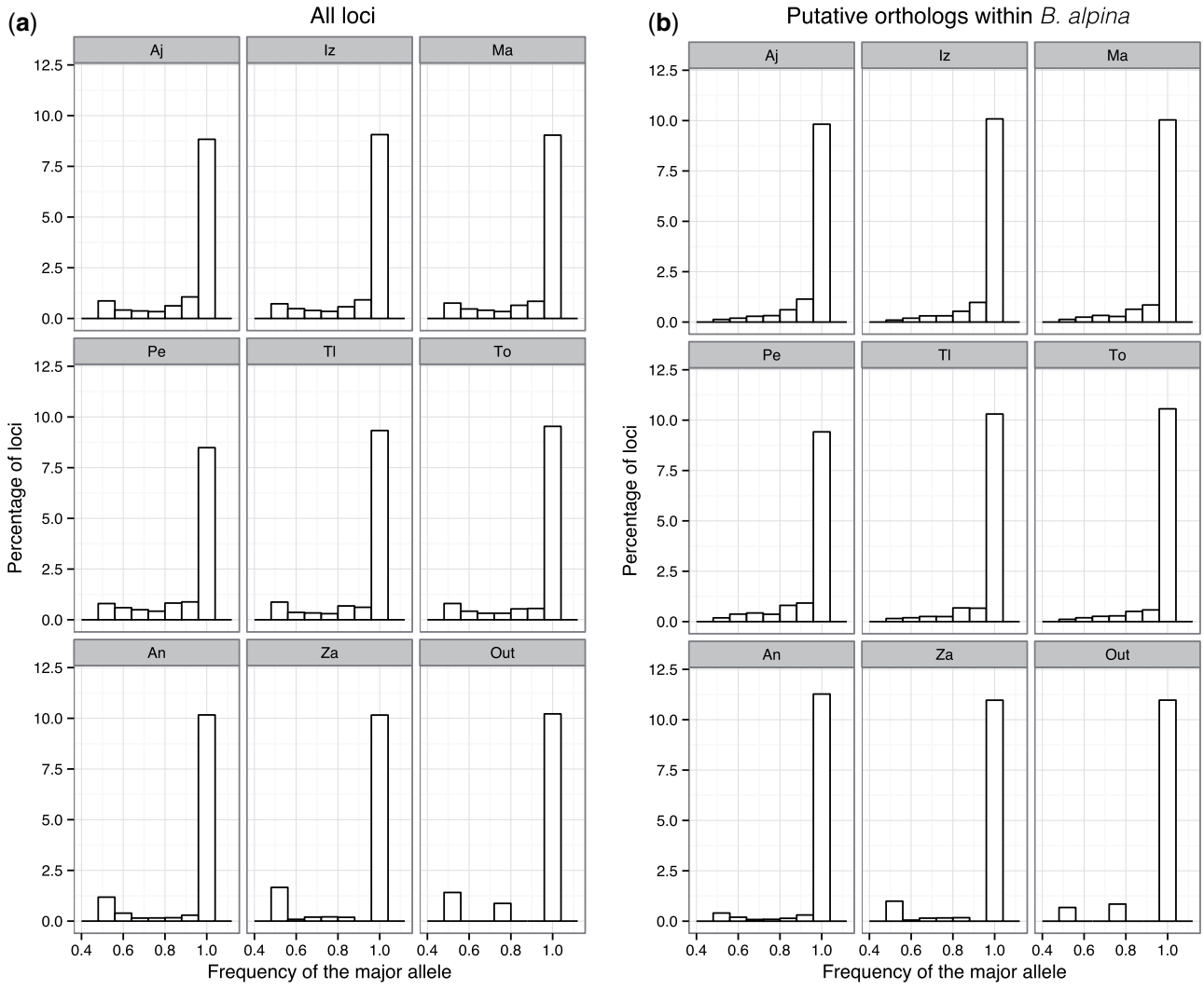

Fig. 3.—Distribution of the frequency of the major allele (*p*) for the SNP-loci for each *Berberis* spp. population. The plots on (*a*) correspond to all loci after de novo assembly and quality filtering. Notice that for every population a substantial percentage of loci is in the 0.5 category (left most bar). The plots on (*b*) show the distribution of the frequency of the major allele for the subset of loci presumably orthologous for *B. alpina* ingroup. This filtering removes the bias toward heterozygosity in *B. alpina* ingroup (top six panels), but not from *B. moranensis* (An) and the Cerro Zamorano population (Za). Notice that for Za and the outgroup (Out), small sampling sizes (4 and 2, respectively) affect the range of allele frequencies that can be recovered.

had high levels of H_{obs} and negative F_{IS}, as would be expected if these loci were the result of overmerging paralogous loci as a single locus. Then, we examined the effect of filtering the putative paralogs on the SFS and the estimation of population genetics statistics.

Filtering out the putative paralogs for *B. alpina* ingroup removed the bias toward loci with $p = 0.5$ within these populations, but it remained noticeable for *B. moranensis* and the Cerro Zamorano population (fig. 3*b*). This is explained by a high number of private potential paralogs within both *B. moranensis* and the Cerro Zamorano population (267 and 617, respectively; fig. 4c). Under Hardy–Weinberg equilibrium (HWE), loci where most individuals are heterozygous are expected to be at the lowest frequency of the spectrum. Although it remains possible that some of the private potential paralogous loci detected here are actually true loci where $p = 0.5$, for *B. moranensis* and the Cerro Zamorano population they account for 18% and 37% of the nonfixed SNP-loci. Balancing selection could cause a bias toward heterozygosity but this should affect very few loci in the genome and it cannot explain all (or most, as some may not be due to allele drop out) individuals being heterozygous (as shown by $H_{obs} = 1$ in 91% and negative F_{IS} in 98% of the loci where $p = 0.5$). Biological explanations for such extreme heterozygosity within populations are lacking, and co-occurring PCR/sequencing error cannot have produced the bias, because samples were individually tagged and randomly sequenced in different lanes. The most parsimonious explanation is therefore that the inferred heterozygosity is an artifact of the assembly of independent loci as a single locus. Therefore, the $p = 0.5$ criterion used here for identifying potential paralogs among populations and species could be fine-tuned by formal tests of HWE deviations in data sets with sufficient sampling sizes per species. Finally, all things being equal, if the private

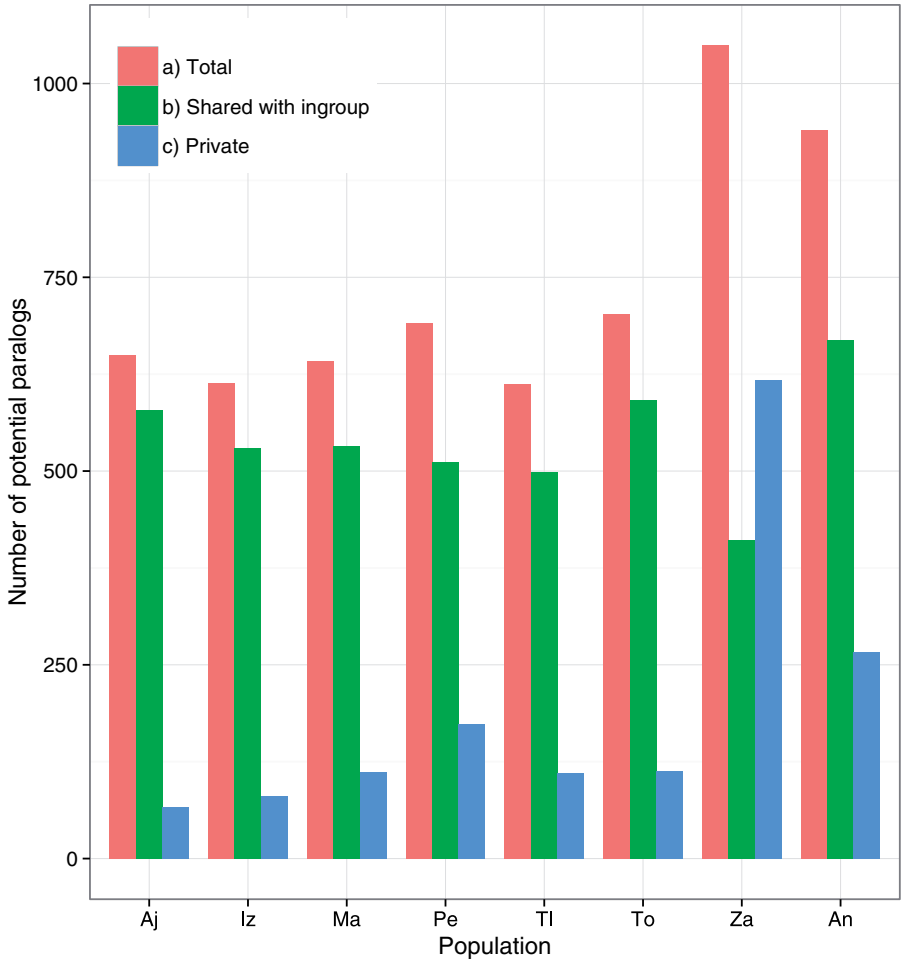

Fig. 4.—Distribution of RAD-loci with at least one SNP-locus where the frequency of the major allele equals 0.5 (potential paralogs). (*a*) There are more loci biased toward $p = 0.5$ in *B. moranensis* (An), the Cerro Zamorano population (Za), and *B. trifolia* (Out) than in *B. alpina* ingroup populations (Aj–To). (*b*) Most of the loci where $p = 0.5$ are the same loci in *B. alpina* ingroup and any given population or species, but (*c*) a substantial proportion of loci show $p = 0.5$ exclusively in *B. moranensis* or the Cerro Zamorano population. Sampling size ($n = 4$) is the same for every population.

potential paralogs were truly heterozygous loci their frequency within each population should be proportionally the same among populations. Interestingly, we found that the number of private potential paralogs increases with the differentiation estimated using only orthologs (fig. 6). This can be explained if the private potential paralogs were indeed the product of gene duplication, which is expected to occur independently within lineages and isolated populations.

Filtering deviations from HWE, such as bias toward heterozygosity caused by merged paralogs, is a necessary step for producing a set of putative orthologs, as evidenced by the following three observations. First, analyses including the putative paralogs yielded negative F_{IS} values for all populations of the *B. alpina* ingroup, and produced levels of polymorphic loci, H_{obs}, and π that were found to be erroneously higher (supplementary table S1, Supplementary Material online) than those obtained when these loci were excluded (table 1). Second, filtering out putative paralogs increased population differentiation estimates: After putative paralogous loci within *B. alpina* are filtered out, the first axis of the PCoA of all samples increases from 81% (supplementary fig. S1, Supplementary Material online) to 86% of the variance explained (fig. 5), and the mean of the F_{ST} pairwise values among the *B. alpina* ingroup populations increases from 0.060 (supplementary table S3, Supplementary Material online) to 0.077 (supplementary table S4, Supplementary Material online). This is to be expected from the erroneous assembly of paralogous loci as a single locus, as merged paralogs generate shared polymorphism among populations. Third, the removal of paralogous loci decreased both the RAD-allele and SNP error rates (from 5.9% to 4.1%, and from 2.2% to 1.5%, respectively), likely because paralogous loci have more "alleles," and are thus more prone to allele drop out, an important source of error for low coverage GBS data (Mastretta-Yanes et al. 2014b).

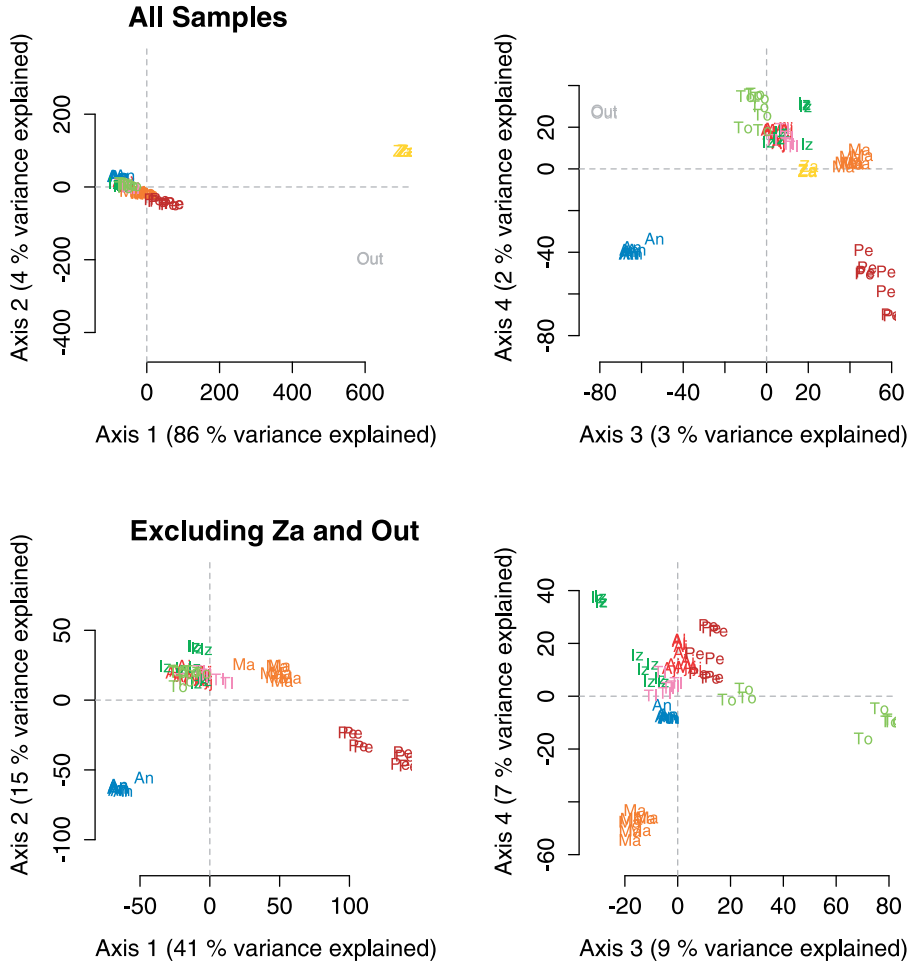

Fig. 5.—Principal coordinates analysis of the SNP-loci excluding all potential paralogs. (*a*) When all samples are analyzed axis 1 explains 86% of the variation and corresponds to the differences between El Cerro Zamorano and *B. trifolia* (Za and Out, respectively) to the rest of the populations. (b) If El Cerro Zamorano and *B. trifolia* are excluded, axes 1 and 2 separate *B. moranensis* (An) and the Cofre de Perote and Malinche (Pe and Ma) populations of *B. alpina*, explaining 41% and 15% of the variance, respectively. Populations ID and colors as in figure 2.

Origin of Paralogous Loci in *Berberis* Taxa and Populations

The older a gene duplication event is, the more nucleotide differences paralogous loci should accumulate, leading to an increased probability of recovering more than three "alleles" if they are merged as a single locus (fig. 1*c*). Eventually, the paralogs will accumulate enough differences to be assembled as different loci (fig. 1*d*). Thus, allowing a maximum of three alleles per locus, as done here, should retain paralogs of relatively recent origin. Because gene conversion causes paralogs to maintain sequence similarity (Lynch and Conery 2000), a fraction of the putative paralogs could be older. However, gene conversion occurs mostly within multigene families (Semple and Wolfe 1999), which in plants tend to have an ancient origin and be largely conserved among families (Flagel and Wendel 2009).

Regarding the duplication mechanism, ancient polyploidy events within *Berberis* cannot be fully discarded. However, given that 1) the potentially paralogous loci identified here are expected to have a recent origin, 2) that they represent only a fraction of the recovered RAD-loci (from 13% of the RAD-loci, for *B. alpina* ingroup to 17% for the Cerro Zamorano population), and 3) that they are not homogeneously distributed among populations and species, it is likely that they arose by gene duplication mechanisms other than whole genome duplication. These alternative duplication mechanisms (reviewed for plants by Freeling 2009) include segmental duplication events, transposable elements, and small-scale duplications (Lockton and Gaut 2005; Moore and Purugganan 2005; Flagel and Wendel 2009), and have been found to be responsible for the origin of recent paralogs within *Ar. thaliana* (Moore and Purugganan 2003).

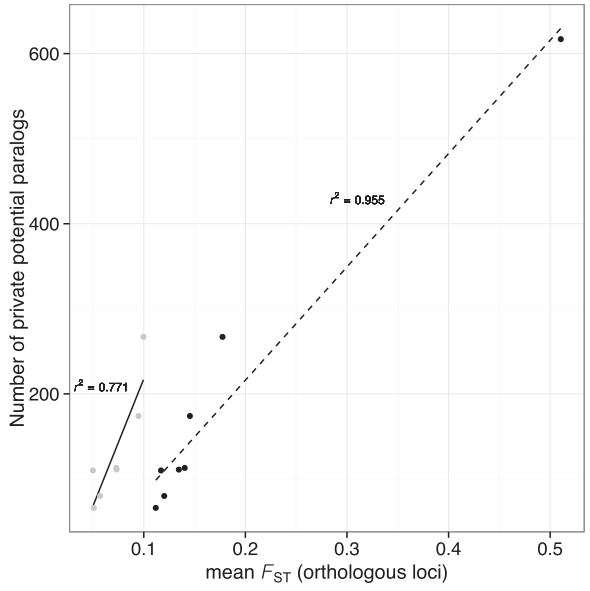

Fig. 6.—The number of private potential paralogs per population increases with their differentiation estimated from orthologous loci. The *x* axis corresponds to the mean F_{ST} per population from the pairwise matrix among populations and species estimated excluding all potential paralogs. The *y* axis corresponds to the number of private potential paralogs as in figure 4. Regression was performed with the Cerro Zamorano population (black dots, dashed line, $F_{1,6} = 128.3$, $P < 0.001$) and without it (gray dots, solid line, $F_{1,5} = 16.85$, $P < 0.01$).

Population Differentiation and a Cryptic *Berberis* Species

Berberis alpina sampled from Cerro Zamorano was found to be strongly genetically differentiated from all other *B. alpina* populations, forming a distinct cluster in the PCoA that explained as much of the variation as the outgroup (fig. 5). Additionally, F_{ST} values between Cerro Zamorano and the other *B. alpina* sampling locations are higher than those between the outgroup and the other *B. alpina* sampling locations (table 1 and supplementary table S4, Supplementary Material online). The Cerro Zamorano population also exhibits a high number of RAD-loci that are likely to comprise private paralogous loci (fig. 4*c*). Za–SMOr populations of *B. alpina* present habitat and leaf morphology differences from both TMVB populations of *B. alpina* and from *B. moranensis* (supplementary table S1, Supplementary Material online). Such morphological characters are not necessarily indicative of species level differentiation, but considered together with the genomic differentiation it would appear that Za–SMOr should be recognized as a different species from the *B. alpina* TMVB populations. Species level differentiation of the *Berberis* sp. from the Cerro Zamorano from *B. alpina* from the TMVB is also congruent with 1) analyses showing that the SMOr, the AS and the TMVB are different biogeographic units (Arriaga et al. 1997; Morrone et al. 2002), 2) the fact that Cerro Zamorano is an old (~11 Myr old; Carrasco-Ñuñez et al. 1989) and isolated mountain (fig. 2), and 3) data on vascular plants distributions showing that the Cerro Zamorano contains a high number endemic species or species restricted to it and to neighbor mountains in the SMOr (Rzedowski et al. 2012).

Regarding *B. alpina* ingroup populations, samples from topographically isolated mountains are expected to be genetically more differentiated than populations separated by less shallow elevations. During the Pleistocene climate fluctuations, the spatial distribution of climate variation did not undergo substantial latitudinal changes in Central Mexico, but it did undergo altitudinal shifts (Metcalfe 2006). During glacial periods cold temperatures existed at lower altitudes than today, allowing alpine grasslands to occur down to 2,500 masl, approximately 1,000 m below their current interglacial range (Lozano-García et al. 2005; Metcalfe 2006; Vázquez-Selem and Heine 2011). By performing altitudinal migrations involving only short horizontal distances, species from alpine grasslands of the TMVB are expected to have persisted relatively in situ, with altitude being the main variable influencing

possible habitat connectivity, and thus gene flow, among mountains in the past (Toledo 1982; Graham 1999). The subset of putative orthologous RAD-loci within *B. alpina* ingroup supports this expectation because the populations that are topographically more isolated (Cofre de Perote and Malinche, fig. 2) present the highest F_{ST} values (0.085–0.122 and 0.068–0.100, respectively, supplementary table S4, Supplementary Material online) and have the highest number of private alleles (1,067 and 485, respectively, supplementary table S1, Supplementary Material online). Genomic differentiation was significant among all populations, with F_{ST} values typically greater than 0.05, (supplementary table S4, Supplementary Material online), with all populations exhibiting low frequency alleles (fig. 3*b*), as expected for old and stable populations. These genetic patterns support the hypothesis that *B. alpina* populations were able to survive in situ through several Pleistocene climate fluctuations. Similar conclusions have been reached for animal taxa of the TMVB using more traditional population genetic and phylogeographic approaches (e.g., McCormack et al. 2008; Bryson et al. 2011, 2012).

Berberis moranensis grows at lower elevations than *B. alpina* from the TMVB (supplementary table S1, Supplementary Material online). Interestingly, the Cofre de Perote population of *B. alpina* and *B. moranensis* exhibits similar F_{ST} values against *B. alpina* ingroup populations (0.085–0.122 and 0.076–0.138, respectively; table 1). However, the differentiation of Cofre de Perote is driven by a high number of private alleles (1,067, table 1), whereas *B. moranensis* has fewer private alleles (194, table 1) but presents 267 RAD-loci that are presumed to be private paralogs, approximately twice the number than in Cofre de Perote (174, fig. 4*c*). Morphologically there are similar leaf characters (e.g., rachis and number of leaflets; supplementary table S1, Supplementary Material online) between *B. alpina* ingroup populations and *B. moranensis*. This phenomenon could be explained by different scenarios of hybridization, ancestry or selection favoring duplicated loci. However, it is not possible to assess these kinds of hypotheses with our current geographical sampling of *B. moranensis*.

Paralogous Loci as a Source of Genomic Differentiation

A central finding of this study is that there are quantitative differences in the distribution of potential paralogous loci among populations and species: *B. moranensis* and the population likely representing a different species (Cerro Zamorano) have a high number of private paralogs (fig. 4), and the populations in the *B. alpina* ingroup that are more differentiated for presumed orthologous loci also present a larger number of presumed private paralogs (fig. 6).

Examining the distribution of paralogous loci among populations and species is relevant because 1) gene duplication might lead to functionally relevant, ecologically significant polymorphisms (Moore and Purugganan 2005); and 2) the divergent evolution of recently duplicated genes can lead to postzygotic isolating barriers within existing species (Bikard et al. 2009). Testing whether the former phenomena were consequential for genome divergence among our *Berberis* species would require analyzing the identified paralogous loci with a more detailed understanding of their genomic context and potential function. However, the paralogous loci found here are already an extra source of evidence for the genomic differentiation among our *Berberis* taxa. First, the fact that the population of *B. moranensis* had more paralogous loci than the most differentiated population of *B. alpina* (Cofre de Perote) shows that *B. moranensis* is more differentiated from *B. alpina* than what would be inferred from the PCoA or the F_{ST} values. This highlights that paralogous loci can be an important source of genomic differentiation among closely related, ecologically divergent, and partially sympatric plant lineages. Second, the distribution of potential paralogous loci among our *Berberis* species is congruent with the expectation that the independent occurrence of gene duplication within lineages should lead to different species presenting a unique set of paralogs that originated after the speciation event (Lynch and Conery 2000). This has also been shown for species of *Arabidopsis* (Moore and Purugganan 2003) and *Drosophila* (Zhou et al. 2008) so in the case of our *Berberis* species it highlights that Cerro Zamorano population is indeed likely to be a different species. The rate of gene duplication could not be estimated due the uncertainty about divergence in the absence of gene flow between our populations and species, as well as lack of calibration points and reliable nuclear mutation rates for our *Berberis* data. Nevertheless, the independent accumulation of paralogs seems to be linearly correlated with the differentiation estimated from orthologous loci (fig. 6) although the number of private potential paralogous of Cerro Zamorano seems an underestimate based on the trajectory of the previous points. This could be an effect of Cerro Zamorano species being too divergent, leading to the existence of paralogs of older origin that our method would have filtered.

Conclusion

The genomic study of paralogous loci has typically been restricted to highly annotated genomes, or requires transcriptome sequencing (e.g., Lynch and Force 2000; Kondrashov et al. 2002; Zhou et al. 2008; Bikard et al. 2009; Warren et al. 2014). Here, we have shown that GBS can be used to quantify the differential distribution of recently generated paralogs among nonmodel plant populations and species. Thus, in addition to producing large amounts of genomic data for traditional population genetics analyses, GBS methods may also be used to investigate gene duplication as a source of population genomic differentiation. As shown

here, this is possible despite short sequence reads and lack of previous genomic knowledge of the analyzed taxa.

Incorporating gene duplication to population genetics and phylogenetic analyses of GBS data could be then taken further by 1) including quantitative measurements of paralogous loci into diversity indexes, and 2) developing analytical tools, such that paralogous loci are not excluded from marker-based data sets, but incorporated into models of allele and genome divergence. This may be relevant for a broad range of taxa, but should be particularly important for plants where gene duplication plays a fundamental role in their evolution.

Supplementary Material

Supplementary figures S1 and S2 and tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Associate Editor Yves Van De Peer and two anonymous referees for their constructive comments on an earlier version of the manuscript. Part of the analyses was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at UEA. This work was supported by Consejo Nacional de Ciencia y Tecnología (grant number CONACYT 213538 to A.M.Y. and CONACYT 178245 to D.P.), by Rosemary Grant Award for Graduate Student Research from the Society for the Study of Evolution to A.M.Y. and by a Swiss National Science Foundation (grant number PP00P3_144870 to Na.A. and PZ00P3_148224 to Ni.A.).

Literature Cited

Arriaga L, Aguilar C, Espinosa D, Jiménez R. 1997. Regionalización ecológica y biogeográfica de México. Taller de la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, CONABIO, México.

Bikard D, et al. 2009. Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. Science 323:623–626.

Bryson RW, García-Vázquez UO, Riddle BR. 2012. Relative roles of Neogene vicariance and Quaternary climate change on the historical diversification of bunchgrass lizards (*Sceloporus scalaris* group) in Mexico. Mol Phylogenet Evol. 62:447–457.

Bryson RW, Murphy RW, Lathrop A, Lazcano-Villareal D. 2011. Evolutionary drivers of phylogeographical diversity in the highlands of Mexico: a case study of the *Crotalus triseriatus* species group of montane rattlesnakes. J Biogeogr. 38:697–710.

Carrasco-Ñuñez G, Milán M, Verma SP. 1989. Geología del Volcán Zamorano, Estado de Querétaro. Rev Inst Geol. 8:194–201.

Catchen J, Amores A, Hohenlohe P, Cresko W, Postlethwait JH. 2011. Stacks: building and genotyping loci de novo from short-read sequences. G3 1:171–182.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. Mol Ecol. 22: 3124–3140.

Davey JW, et al. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 12: 499–510.

Deschamps S, Llaca V, May GD. 2012. Genotyping-by-sequencing in plants. Biology 1:460–483.

Dou J, et al. 2012. Reference-free SNP calling: improved accuracy by preventing incorrect calls from repetitive genomic regions. Biol Direct. 7:17.

Eaton DAR. 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30:1844–1849.

Feschotte C, Jiang N, Wessler SR. 2002. Plant transposable elements: where genetics meets genomics. Nat Rev Genet. 3:329–341.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool. 19:99.

Flagel LE, Wendel JF. 2009. Gene duplication and evolutionary novelty in plants. New Phytol. 183:557–564.

Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol. 60:433–453.

Gayral P, et al. 2013. Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. PLoS Genet. 9:e1003457.

Graham A. 1999. The Tertiary history of the northern temperate element in the northern Latin American biota. Am J Bot. 86:32–38.

Gugger PF, González-Rodríguez A, Rodríguez-Correa H, Sugita S, Cavender-Bares J. 2011. Southward Pleistocene migration of Douglas-fir into Mexico: phylogeography, ecological niche modeling, and conservation of 'rear edge' populations. New Phytol. 189:1185–1199.

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. Mol Ecol Resour. 11:117–122.

Hohenlohe PA, Catchen J, Cresko WA. 2012. Population genomic analysis of model and nonmodel organisms using sequenced RAD tags. Methods Mol Biol. 888:235–260.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. PLoS Biol. 2:e206.

Jensen RA. 2001. Orthologs and paralogs—we need to get it right. Genome Biol. 2. interactions1002.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc R Soc Lond B Biol Sci. 279: 5048–5057.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. Genome Biol. 3. RESEARCH0008.

Langham RJ, et al. 2004. Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 166:935–945.

Lewis EB. 1951. Pseudoallelism and gene evolution. Cold Spring Harb Symp Quant Biol. 16:159–174.

Lockton S, Gaut BS. 2005. Plant conserved non-coding sequences and paralogue evolution. Trends Genet. 21:60–65.

Lozano-García S, Sosa-Nájera S, Sugiura Y, Caballero M. 2005. 23,000 yr of vegetation history of the Upper Lerma, a tropical high-altitude basin in Central Mexico. Quat Res. 64:70–82.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

Mastretta-Yanes A, et al. 2014a. Data from: RAD sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Dryad Digit Repos. doi:10.5061/dryad.g52m3.

Mastretta-Yanes A, et al. 2014b. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Mol Ecol Resour. Advance Access published July 3, 2014, doi: 10.1111/1755-0998.12291.

Mastretta-Yanes A, Wegier A, Vázquez-Lobo A, Piñero D. 2011. Distinctiveness, rarity and conservation in a subtropical highland conifer. Conserv Genet. 13:211–222.

McCormack JE, Peterson AT, Bonaccorso E, Smith TB. 2008. Speciation in the highlands of Mexico: genetic and phenotypic divergence in the Mexican jay (*Aphelocoma ultramarina*). Mol Ecol. 17:2505–2521.
Metcalfe SE. 2006. Late Quaternary environments of the northern deserts and Central Transvolcanic belt of Mexico. Ann Mo Bot Gard. 93: 258–273.
Moore RC, Purugganan MD. 2003. The early stages of duplicate gene evolution. Proc Natl Acad Sci U S A. 100:15682–15687.
Moore RC, Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol. 8:122–128.
Morrell PL, Buckler ES, Ross-Ibarra J. 2012. Crop genomics: advances and applications. Nat Rev Genet. 13:85–96.
Morrone JJ, Espinosa-Organista D, Llorente-Bousquets J. 2002. Mexican biogeographic provinces: preliminary scheme, general characterizations, and synonymies. Acta Zool Mex. 85:83–108.
Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J. 2000. Biodiversity hotspots for conservation priorities. Nature 403:853–858.
Ohno S. 1970. Evolution by gene duplication. London: George Alien & Unwin Ltd.
Ornelas JF, et al. 2013. Comparative phylogeographic analyses illustrate the complex evolutionary history of threatened cloud forests of northern Mesoamerica. PLoS One 8:e56283.
Parchman TL, et al. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Ecol. 21:2991–3005.
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.
Poland JA, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253.
Poland JA, Rife TW. 2012. Genotyping-by-sequencing for plant breeding and genetics. Plant Genome J. 5:92.
Pujolar JM, et al. 2014. Genome-wide single-generation signatures of local selection in the panmictic European eel. Mol Ecol. 23:2514–2528.
R Core Team. 2012. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available from: http://www.R-project.org/.
Rounsaville TJ, Ranney TG. 2010. Ploidy levels and genome sizes of *Berberis* L. and *Mahonia* Nutt. species, hybrids, and cultivars. HortScience 45:1029–1033.
Rzedowski J, Calderón de Rzedowski G, Zamudio S. 2012. La flora vascular endémica en el estado de Querétaro. I. Análisis numéricos preliminares y definición de áreas de concentración de las especies de distribución restringida. Acta Bot Mex 99:91–104.
Schatz MC, Witkowski J, McCombie WR. 2012. Current challenges in de novo plant genome sequencing and assembly. Genome Biol. 13: 243.
Semple C, Wolfe KH. 1999. Gene duplication and gene conversion in the *Caenorhabditis elegans* genome. J Mol Evol. 48:555–564.
Stölting KN, et al. 2013. Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. Mol Ecol. 22:842–855.
Toledo V. 1982. Pleistocene changes of vegetation in tropical Mexico. In: Prance GT, editor. Biological diversification in the tropics. New York: Columbia University Press. p. editor–111.
Tovar-Sánchez E, et al. 2008. Chloroplast DNA polymorphism reveals geographic structure and introgression in the *Quercus crassifolia* × *Quercus crassipes* hybrid complex in Mexico. Botany 86:228–239.
Vázquez-Selem L, Heine K. 2011. Late Quaternary Glaciation in Mexico. In: Ehlers J, Gibbard PL, Hughes P, editors. Quaternary glaciations—extent and chronology—a closer look, Vol. 15. p. 849–861.
Volff J-N. 2004. Genome evolution and biodiversity in teleost fish. Heredity 94:280–294.
Warren IA, et al. 2014. Extensive local gene duplication and functional divergence among paralogs in Atlantic salmon. Genome Biol Evol. 6: 1790–1805.
Zamudio S. 2009a. Familia Berberidaceae. Flora Bajioy y Regiones Adyacentes Fasículo 163:1–40.
Zamudio S. 2009b. Notas sobre el Género Berberis (Berberidaceae) en México. Acta Bot Mex. 87:31–70.
Zhou Q, et al. 2008. On the origin of new genes in *Drosophila*. Genome Res. 18:1446–1455.

Associate editor: Yves Van De Peer