Levels and patterns of genetic diversity differ between two closely related endemic *Arabidopsis* species

Julie Jacquemin[1]*, Nora Hohmann[2], Matteo Buti[3], Alberto Selvaggi[4], Thomas Müller[1,5], Marcus A. Koch[2] and Karl J. Schmid[1]*

[1] Crop Biodiversity and Breeding Informatics; Institute of Plant Breeding, Seed Science and Population Genetics; University of Hohenheim; Stuttgart; Germany

[2] Biodiversity and Plant Systematics, Centre for Organismal Studies (COS), Ruprecht-Karls-University Heidelberg; Heidelberg; Germany

[3] Istituto Agrario di San Michele all'Adige; San Michele All'Adige; Italy

[4] Istituto Piante da Legno e l'Ambiente; Torino; Italy

[5] Present address: Department of Plant and Microbial Biology, University of Zurich, Switzerland

*Corresponding authors

Corresponding authors
Julie Jacquemin
Karl Schmid

Crop Biodiversity and Breeding Informatics (350b)
Institute of Plant breeding, Seed Science and Population Genetics
Universität Hohenheim
Fruwirthstrasse 21
70599 Stuttgart
Germany
fax: +49 711 459-24458
email: karl.schmid@uni-hohenheim.de; julie.jacquemin@uni-hohenheim.de

Keywords. *Arabidopsis cebennensis*, *Arabidopsis pedemontana*, genetic diversity, population structure, endemics conservation, genotyping-by-sequencing.

Running title. Genetic diversity in endemic *Arabidopsis*

Abstract

Theory predicts that a small effective population size leads to slower accumulation of mutations, increased levels of genetic drift and reduction in the efficiency of natural selection. Therefore endemic species should harbor low levels of genetic diversity and exhibit a reduced ability of adaptation to environmental changes. *Arabidopsis pedemontana* and *Arabidopsis cebennensis*, two endemic species from Italy and France respectively, provide an excellent model to study the adaptive potential of species with small distribution ranges. To evaluate the genome-wide levels and patterns of genetic variation, effective population size and demographic history of both species, we genotyped 53 *A. pedemontana* and 28 *A. cebennensis* individuals across the entire species ranges with Genotyping-by-Sequencing. SNPs data confirmed a low genetic diversity for *A. pedemontana* although its effective population size is relatively high. Only a weak population structure was observed over the small distribution range of *A. pedemontana*, resulting from an isolation-by-distance pattern of gene flow. In contrary, *A. cebennensis* individuals clustered in three populations according to their geographic distribution. Despite this and a larger distribution, the overall genetic diversity was even lower for *A. cebennensis* than for *A. pedemontana*. A demographic analysis demonstrated that both endemics have undergone a strong population size decline in the past, without recovery. The more drastic decline observed in *A. cebennensis* partially explains the very small effective population size observed in the present population. In light of these results, we discuss the adaptive potential of these endemic species in the context of rapid climate change.

Introduction

A main concern in the current debate on ecological effects of climate change is whether populations and species can adapt fast enough to keep up with the rapid rate of environmental changes (Salamin *et al.* 2010). Mitigation of biodiversity loss requires knowledge on ecological and evolutionary responses of populations to habitat changes, and on the conditions that allow a recovery of declining populations (Gonzalez *et al.* 2012). Plant endemic species are an important component of biodiversity, particularly in biodiversity hotspots, where they make up a large proportion of the local flora (Myers *et al.* 2000). With a limited geographic distribution that is often tied to specific habitats, endemic species are more vulnerable to environmental changes as they frequently depend on the existence of particular biotic and abiotic interactions (Thomas *et al.* 2004). For this reason endemic biodiversity should be a central focus of conservation efforts.

Because of their limited distribution range, endemic species tend to have small census population sizes and smaller effective population sizes, N_e (Freville *et al.* 2001; Strasburg *et al.* 2011). Theory predicts that a small effective population size leads to slower accumulation of mutations, increased levels of genetic drift and reduction in the efficiency of natural selection (reviewed by Ellstrand & Elam 1993). According to the neutral theory of molecular evolution, genetic diversity levels at neutral sites reflect a balance between the mutational input per generation and the loss of genetic variation due to genetic drift (Kimura 1983). All else being equal, species with smaller population sizes should thus harbor lower levels of neutral genetic diversity (Ramos-Onsins 2004; Leimu *et al.* 2006; Leffler *et al.* 2012). The chance of new, potentially advantageous

mutations appearing is also reduced compared to larger populations. Additionally, hard selective sweeps from new mutations and soft selective sweeps from standing variation are less likely, reducing the rate of adaptation (Lanfear *et al.* 2014). Furthermore, genetic drift may lead to the fixation of mildly deleterious mutations which are not efficiently removed because of weak purifying selection (Kimura 1983; Cao *et al.* 2011; Xue *et al.* 2015), leading potentially to a mutational meltdown (Lynch *et al.* 1993). Consequently, the direct and indirect effects of a small effective population size have a strong influence on the evolutionary dynamics and the adaptive potential of a species (Charlesworth 2009). Continuous adaptation in changing environments requires sufficient and appropriate genetic variation that provides extreme genotypes capable of surviving intense stress conditions and allow the population persistence (Reed *et al.* 2011; Bell 2013). Low levels of genetic diversity, reduced average fitness and limited adaptive potential have been indeed observed for plant populations with small population size (Pluess & Stöcklin 2004; Hensen & Oberprieler 2005; Leimu *et al.* 2006; Michalski & Durka 2007; Leimu & Fischer 2008). A meta-analysis demonstrated that genetic erosion significantly contributes to the extinction risk of plant species, beside short term demographic and ecological processes (Spielman *et al.* 2004). To evaluate the adaptive potential and demographic responses of endemic populations in a rapidly changing environment, it is therefore necessary to estimate levels of genetic diversity and to infer the evolutionary processes (genetic drift, gene flow, mutation, mating system) that have shaped the pattern of genetic variation (Reed *et al.* 2011).

In the genus *Arabidopsis*, the two endemic species *Arabidopsis cebennensis* (DC.) and *Arabidopsis pedemontana* (Boiss.) provide an excellent model of species with restricted ranges.

Arabidopsis thaliana and its close relatives have become model plant species in population genetics and the data accumulated for the whole genus enable an interspecific comparative approach of genetic diversity levels (Mitchell-Olds & Schmitt 2006; Clauss & Koch 2006). The whole genus reflects the diversity of plant distribution ranges: while *A. thaliana* is distributed world-wide, *A. cebennensis* and *A. pedemontana* have the smallest distribution range in the genus (O'Kane & Al-Shehbaz 1997; Koch *et al.* 2008). *Arabidopsis cebennensis* is restricted to the mountainous Massif Central region in Southern France at elevations ranging from 900 to 1500 m a.s.l. Highly disjunct populations occur over an area of about 11,000 km^2 in the Cevennes, Cantal, Aveyron and Ardèche regions. *Arabidopsis pedemontana* occupies a much smaller distribution range of 50 km^2 in the Piedmont region of the northwestern Italian Alps, at altitude ranging from 1,300 to 2,200 m a.s.l. All known populations of this species occur on the two sides of a single mountain ridge located between the valleys Po and Pellice. *Arabidopsis pedemontana* is included in the red list of Italian and Piedmont Floras under the category "critically endangered' according to the IUCN definition (The IUCN Species Survival Commission 2004), whereas *A. cebennensis* is not protected but occurs mostly in national and regional protected reserve areas. The two species are closely related to each other and their common ancestor is genetically distinct from other lineages in the genus (Koch & Matschinger 2007; Hohmann *et al.* 2014). They are perennial diploids and presumably self-incompatible, with a strong tendency for vegetative reproduction by clonal growth (Hohmann *et al.* 2014). Both species occupy a very specific niche in riverine habitats (streams and waterfalls), with a semi-continental mountainous climate (cold winters). Adaptation to this specific habitat led to a very specialized ecology, and phenotypically

differentiate them from their relatives which are found in different habitats like meadows or rocky outcrops. Previous studies revealed a highly reduced diversity for chloroplast DNA, ribosomal DNA and microsatellite loci in *A. cebennensis* and *A. pedemontana* compared to other *Arabidopsis* species (Koch & Matschinger 2007; Hohmann *et al.* 2014). Due to this reduced level of genetic variation, these surveys have not provided much information on genetic structure. Whole genome approaches are likely more promising.

The present work aims to reveal the extent and structure of genetic variation in the two endemic *A. cebennensis* and *A. pedemontana* on a genome-wide scale. We used Genotyping-By-Sequencing (GBS, Elshire *et al.* 2011) to obtain genome-wide SNPs for estimating population genetic parameters. GBS and other reduced-representation sequencing approaches (Davey *et al.* 2011) have rapidly become important tools for the study of genetic diversity, adaptation and conservation (Narum *et al.* 2013; Huang *et al.* 2014; Gonçalves da Silva *et al.* 2015; Xue *et al.* 2015). As the genomes of both endemic species have not yet been sequenced, we used *A. lyrata*, which is the most closely related species with a sequenced genome, as a reference for mapping and SNP calling. Additionally we analyzed the structure of genetic variation at the population level based on a set of microsatellite loci and plastid DNA sequences from the *trnLF* region, which have been used previously to characterize *A. cebennensis* and *A. pedemontana* genetic diversity within the whole genus *Arabidopsis* (Hohmann *et al.* 2014).

Based on an extensive sampling representing 90% and 40% of all known locations for *A. pedemontana* and *A. cebennensis*, respectively, we characterized the genetic diversity and the contemporary effective population size of both endemic species over their entire distribution

range. We also investigated how their demographic history may have led to the observed levels of genetic variation, and tested whether it is consistent with models of declining population sizes. In light of these new results, we discuss the adaptive potential of the two endemic species in the context of the probable environmental changes to come.

Materials and Methods

More detailed information on the materials and methods is available in Text S1.

Plant material

All samples were collected between 2010 and 2014 in France and Italy (Fig. 1, Supporting information, Hohman *et al.* 2014). For GBS, genomic DNA was extracted using a modified CTAB protocol (Saghai-Maroof *et al.* 1984). For samples collected more recently, DNA was extracted with the Genomic Micro AX Blood Gravity kit (A&A Biotechnology, Gdynia, Poland). DNA concentration and quality was checked with agarose gel electrophoresis and Qubit 2.0 Fluorometer (Life Technologies). In the GBS analysis, 53 *A. pedemontana* and 28 *A. cebennensis* samples were used.

Genotyping-by-Sequencing (GBS)

Double-digest GBS libraries were constructed based on the design of (Poland *et al.* 2012) with ApeKI as rare-cutting and HindIII as common restriction enzyme. Adapters and primers (Metabion) were taken from (Elshire *et al.* 2011). Adapters were modified to introduce 58

different barcodes and sticky ends corresponding to ApeKI and HindIII cut sites (Fig. S1). Five µl of each samples with different barcodes were pooled after adapter ligation. Three pools were prepared, one with 24 *A. pedemontana* individuals, one with 29 *A. pedemontana* individuals, and one pool including all 28 *A. cebennensis* samples. The pooled samples were purified with QIAquick PCR purification kit (Qiagen, Hilden/Germany) and PCR amplified. The GBS libraries were run on BluePippin (Sage science) for selection of DNA fragments ranging between 200-350 bp. Each of the three pools was sequenced on one lane of the same flowcell, on an Illumina HiScanSQ with single end sequencing and 105 cycles. With a genome size of around 250 Mbp for both species (Lysak *et al.* 2009; Hohmann *et al.* 2014) the targeted coverage per site was between 20 and 30x.

Sequence data analysis and SNP calling

Raw reads were processed with custom Python scripts, bwa (Li & Durbin 2009) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). All reads with ambiguous 'N' nucleotides and reads with low quality values (< 90% bases with Q>20) were discarded. Read length after barcode and end-trimming was 90 bp (Supporting information). The pre-processed reads were aligned to the genome of *Arabidopsis lyrata* strain MN47 (Hu *et al.* 2011) with pBWA (Peters *et al.* 2012) with 10 mismatches allowed. Biallelic SNPs were called with SAMtools (Li *et al.* 2009) and custom Python scripts from aligned reads with a minimum mapping quality of 1. The vcf file was parsed to filter out SNPs with a coverage of at least 30x and at most 3000x,

whereby at least five reads had to confirm the variant nucleotide. Five *A. pedemontana* and seven *A. cebennensis* individuals with less than 5,000 SNPs were excluded to reduce the proportion of missing values (Supporting information). The SNPs were further filtered using vcftools (Danecek *et al.* 2011) to include only intra-specific *A. pedemontana* and *A. cebennensis* polymorphic sites (Table S2). Two additional files were created including only SNPs with data for at least 50% and 70% of the sampled individuals (Table S2). SNP calling was conducted additionally for ten different subsets of 21 randomly chosen *A. pedemontana* individuals, in order to compare the number of SNPs observed for *A. cebennensis* and *A. pedemontana* when working with the same sample size. The total number of sites (polymorphic and non-polymorphic) and variants (SNPs and indels) were also calculated, applying the same filters described above, allowing us to determine the percentage of polymorphic loci for both species.

Structure and genetic diversity analysis

Population structure was inferred using SNPs with data for at least 50% of the sampled individuals. The optimal number of clusters was obtained using ADMIXTURE (Alexander *et al.* 2009) with the cross-validation procedure (--cv) with K ranging from 1 to 9. Ten iterations with different seed values were completed and compared for homogeneity. Discriminant Analysis of Principal Components (DAPC; Jombart *et al.* 2010) was conducted using the adegenet R package, first using the function find.clusters to determine the optimal number of clusters (K), with $K \leq 10$. 25 and 60 principal components (PCs) were kept to explain the variance in *A. cebennensis* and *A. pedemontana* respectively. Phylogenetic networks were generated in

Splitstree (Huson & Bryant 2006) after vcf files were converted into IUPAC coded (Cornish-Bowden 1985) FASTA files. Distances were calculated using the Uncorrected_P method, ignoring ambiguous states, and the network was generated using the Neighbor-net distances transformation (Bryant & Moulton 2004) with 100 bootstrap replicates. A network was also constructed for the two species and including the reference *A. lyrata* as outgroup, using SNPs called between the three species. To investigate the extent to which the genetic distance between populations was affected by spatial structure (dispersal limitation), and to test the isolation-by-distance (IBD) hypothesis (Slatkin 1993), a Redundancy Analysis (RDA) was performed (Borcard *et al.* 1992) using the R package ade4 (Text S1). The presence of migration events between *A. cebennensis* large populations was tested by running TreeMix v1.12 (Pickrell & Pritchard 2012) with *A. pedemontana* as outgroup, and four migration events allowed. One thousand bootstrap replicates were generated by resampling blocks of 500 SNPs.

All population genetic parameters were calculated from SNPs with data for at least 70% of the sampled individuals. The percentage of missing data was calculated using the R package adegenet 1.4-2 (Jombart & Ahmed 2011). Nucleotide diversity (π) was calculated for each SNP and then averaged over the total number of sites to obtain an average nucleotide diversity per bp, using the formula in Begun *et al.* (2007). Watterson's estimator (θ_w) was calculated to compare the number of segregating sites between the populations. Nei's gene diversity (or proportion of heterozygosity expected, H_{exp}) as well as the proportion of heterozygosity observed (H_{obs}) were calculated using adegenet. The sum of expected heterozygosity per polymorphic sites was divided by the total number of sites to calculate the mean expected heterozygosity. F_{st} was calculated per site with the

R package pegas 0.6 (Paradis 2010), using the formula of Weir & Cockerham (1984). Mean values of F_{st} were calculated over all sites. The partitioning of genetic variability among and within populations in *A. cebennensis* was analyzed with an AMOVA in Arlequin v.3.5, using 10,000 permutations and pairwise difference as distance calculation method.

Demographic analysis

The long-term demographic history of the two species was investigated with δaδi v1.7.0 (Gutenkunst *et al.* 2009). We examined five one-dimension models for each species independently, one neutral and four population size change models (Fig. S2). The standard neutral model (SNM) assumes a constant population size (ancestral population size N_a). For the simple exponential size change model 1a, the population size has changed instantaneously at time T in the past to yield a present population size of N. Model 1b assumes that the population size has changed exponentially since time T in the past, leading to a contemporary population size N. Model 1c assumes that the species have first experienced an instantaneous size change, yielding a population size of N_0, followed by an exponential population size change, going from N_0 to N in time T. In Model 1d, we expand Model 1a by incorporating a second instantaneous size change event. At time $T + T_0$ in the past, the population goes through a size change of depth N_0, and then recovers to relative size N. The last two models allow us test the probability of a past bottleneck event followed by either recovery or further decrease of the population size. Additionally we tested three two-populations models in which *A. cebennensis* and *A. pedemontana* diverged from an ancestral population (Fig. S2). In model 2a, at T_{pc} x $2N_a$ generations ago, the two species split,

yielding a stable contemporary population size N_c for *A. cebennensis* and N_p for *A. pedemontana*. In model 2b, the split is followed by exponential size change of both new species, going from N_{c0}/N_{p0} to N_c/N_p during the time T_{pc} since the divergence. In model 2c, we incorporated a bottleneck event in the ancestral equilibrium population before the divergence of the species. The depth and duration of the bottleneck are denoted by N_0 and T_0 respectively. The Python script implementing the models is provided as Supporting information.

The models were fitted to the SNPs dataset with data for at least 70% of the sampled individuals. The data joint site frequency spectrum (SFS) was estimated by δaδi, and projected down to 15 individuals for *A. cebennensis*, nine for *A. cebennensis* Cantal population alone, and 40 individuals for *A. pedemontana*, which provided 6,583, 3,981, and 10,422 SNPs for the single-population models. For the two-population models, the data was projected down to a sample of 14 and 30 individuals for *A. cebennensis* and *A. pedemontana* respectively, which provided a large number of segregating sites (10,442). The SFS were folded. Parameters for each model were optimized with the upper and lower bounds, arbitrary start values and grid sizes indicated in Supporting information. Multiple optimizations runs were evaluated for each model to ensure the convergence of the optimized parameters. To estimate parameter's confidence intervals (CIs), 100 replicate pseudo-data sets were generated by bootstraping the SNP data by 1 Mb regions on the *A. lyrata* scaffolds to account for putative linkage between the SNPs. CIs were estimated using the SFS of the bootstraped data sets. The log composite likelihood of each model was calculated using δaδi's *ll_multinom* function and the likelihood ratio test was used to test if the differences in the likelihood values between the models were significant.

Microsatellite and plastid DNA sequence analysis

In a previous work, 148 *A. cebennensis* individuals from ten populations and 40 *A. pedemontana* individuals from nine populations were genotyped with seven microsatellite loci and sequences from the *trnLF* region (Hohmann *et al.* 2014). Detailed experimental protocols are available in this study. We carried out a population structure analysis of the two endemic species based on this dataset. Geographic locations of the populations analyzed, total numbers of alleles and mean number of allele per locus for the seven microsatellite loci are indicated in Fig. S3 and Supporting information. Microsatellite genotypes were analyzed using STRUCTURE v.2.3.4 (Pritchard *et al.* 2000; Hubisz *et al.* 2009). Ten replicates were run for each K-value and a burn-in-period of 1 x 105 and 2 x 105 iterations was used. The option 'admixture model' was used in combination with 'correlated allele frequencies'. The estimation of the optimal K number of populations (ranging from 1 to 5) was calculated using the R-script Structure-sum (Ehrich 2006). Input files for CLUMPP were generated with STRUCTURE HARVESTER (Earl & VonHoldt 2012), alignments of replicate runs were conducted in CLUMPP (Jakobsson & Rosenberg 2007), and the mean of 10 runs were visualized.

Results

GBS and polymorphisms

Forty-eight *A. pedemontana* individuals from 29 locations and 21 *A. cebennensis* individuals

287 from five locations were included in the final analysis. Only results based on SNPs with data for
288 at least 70% of the sampled individuals are presented, except if noted otherwise. The average read
289 coverage per site over all individuals was 465 for *A. cebennensis* and 1,124 for *A. pedemontana*.
290 The number of SNPs and percentage of polymorphic loci over all sites were more than two times
291 higher in *A. pedemontana* than in *A. cebennensis* (Table 1). When the SNP calling was conducted
292 on ten different subsets of 21 randomly sampled *A. pedemontana* individuals, the resulting
293 average number of SNPs was 11,858 SNPs (Table S3), which is only ~1,000 less than the number
294 of SNPs observed for the total sample of *A. pedemontana*. This shows that the different levels of
295 polymorphism in *A. cebennensis* and *A. pedemontana* are not caused by contrasting sampling
296 sizes but truly indicate a higher level of polymorphism in *A. pedemontana*.
297 The Neighbor-net network constructed for the two species and the outgroup *A. lyrata* shows a
298 clear distinction between them, with high bootstrap support for the branches leading to each
299 species (Fig. S4). *Arabidopsis pedemontana* and *A. cebennensis* are closer to each other in the
300 network. 667 common SNPs called independently in *A. cebennensis* (10%) and *A. pedemontana*
301 (5%) have shared alleles, indicating a high level of shared variation between the two sibling
302 species.
303
304 **Population structure**
305 Based on the geographic distribution of the samples (Fig. 1), our initial hypotheses were: 1)
306 Genetic variation in *A. cebennensis* should primarily cluster according to the three large regions
307 where this taxon is observed in the Massif central; 2) Distinct patterns of variation should be

observed for *A. pedemontana* between the two valleys where the species occurs, as the mountain range constitutes a natural barrier to dispersion. In the ADMIXTURE analysis the lowest cross-validation value was always found when assuming K=2 for *A. cebennensis*, followed closely by K=3 (Fig. S5). With K=2, individuals from Cantal are separated from Ardèche and Cévennes individuals. The clustering for K=3 (Fig. 2) corresponds exactly to the origin of the individuals, from the Ardèche, Cévennes and Cantal regions. For K=4, individuals from Cantal were further split into two groups, with location 139R separated from locations 140R and 141R. For *A. cebennensis*, six PCs and two discriminant functions were retained when describing clusters with DAPC, which supported the K=3 clustering. The Neighbor-Net network of *A. cebennensis* also separated the Ardèche, Cantal and Cévennes populations with strong bootstrap support (Fig. 2). Within the Cantal population, individuals from site 139R clustered apart with strong support while individuals from 140R and 141R sites were mixed. We used TreeMix to test for gene flow between the three allopatric populations of *A. cebennensis*. In the resulting maximum likelihood tree, Ardèche and Cévennes populations are grouped together with a strong support (bootstrap 99.8%; Fig. S6). Only a single migration event, from Ardèche to Cantal, was inferred, but it shows a low bootstrap value (52%) and low average weight over all bootstrap replicates (13%). This indicates no or very limited gene flow between the three *A. cebennensis* populations. The genetic assignment by STRUCTURE based on microsatellite data also resulted in an optimal value of K=3, reflecting the distribution of sampled populations in their respective geographic regions (Fig. S3), although with some level of admixture, specially between Ardèche and Cévennes individuals. However, one very small population from the Cantal region (145R) is most

329 closely associated with the genetic cluster found in the Cévennes region. This population is
330 located at an artificial creek close to a forest road side. Considering that also its chloroplast
331 genome type is identical to the Cévennes region (Fig. S3), it is more likely that we observed here
332 a very rare case of recent long-distance dispersal, maybe promoted by humans. Strong
333 phylogeographic signal in *A. cebennensis* is highly supported by the distribution of maternally
334 inherited plastid DNA types (*trnLF* loci). In the three major regions, distinct types prevail and we
335 did not find variation within a population. The Redundancy analysis (RDA) on genetic diversity
336 with spatial variables explained 92.5% of the total variance in genetic diversity. However the
337 spatial variables (distance) did not have a significant contribution ($p>0.05$), rejecting isolation by
338 distance as responsible for the strong structure observed in *A. cebennensis*.
339 For *A. pedemontana*, the ADMIXTURE analysis did not suggest population structure as the
340 cross-validation value slowly increased for K=1 to K=9 (Fig. S5). Inspection of population
341 assignment for K=2 did not confirm our hypothesis of a differentiation of *A. pedemontana*
342 individuals between the two valleys. Instead, individuals from Vallone di Fiunira (Fiunira)
343 clustered as one genotype while individuals from Comba della Gianna (Giana) presented an
344 admixture of both detected genotypes (Table S1). When describing clusters with DAPC, 15 PCs
345 and 1 discriminant functions were retained. The DAPC analysis confirmed the results of the
346 ADMIXTURE analysis (Fig. 3). Phylogenetic analysis of *A. pedemontana* confirmed a weak
347 population structure with a star-shaped network, with low support for internal edges and long
348 terminal branches (Fig. 3). However, clusters could still be observed that corresponded to the
349 geographic distribution of the samples. For example, all individuals from Valle Po were clustered,

with subgroups for the East and West part of the valley with a good bootstrap support. All individuals from Fiunira and Gianna were also grouped together, although this large cluster was not strongly differentiated from the remaining individuals and showed a weak bootstrap support. Within this group, the samples from the site 007 appeared very closely related to each other, showing the homogeneity of individuals in one location. The same was observed for the site 004. Although the microsatellite sampling for *A. pedemontana* was less complete, it contained one population that was not sampled with GBS (114; Supporting information). The genetic assignment tests for microsatellite data did not find a significant K exceeding 1, confirming the GBS results. If "LocPriors" were set and varying from 2 to 9, a significant K=2 was found, separating population 114 from the other investigated populations (Fig. S3). It is consistent with the fact that geographic distances are correlated to genetic differentiation in *A. pedemontana*, as population 114 was the only one sampled in Valley Po for this dataset. However, with the more uniformly distributed samples in the GBS analysis, no higher-level structure is apparent between valley Po and valley Pellice. RDA analysis on genetic diversity with spatial variables explained a significant contribution of 13.5% of the total variance in genetic diversity found in populations ($p<0.05$). Hence dispersal limitation due to geographic distances (isolation-by-distance model) appears to be responsible for the weak structure observed in *A. pedemontana*. The level of plastid DNA variation is very low as only the ancestral type A was found (Fig. S3).

F_{st}-based analysis of genetic differentiation

The mean pairwise genetic distance based on F_{st} values between *A. cebennensis* populations were

similar (~0.17-0.19 ± 0.007), with the greatest distance observed between Ardèche and Cantal (Fig. 2). F_{st} values were high compared to the distance between the two species (0.33 ± 0.005). Within the Cantal population, the subpopulation 139R and 140/141R displayed a much lower mean F_{st} over sites (0.05 ± 0.005). *Arabidopsis pedemontana* individuals from Fiunira/Gianna clustered together showed a mean F_{st} of 0.04 ± 0.001 with the rest of *A. pedemontana* samples, and such a weak differentiation did not support this clustering. The distributions of F_{st} values between the two species and between the three subpopulations of *A. cebennensis* were homogeneous, with high and low F_{st} observed equally all along *A. lyrata* scaffolds used as reference genome (Fig. S7).

Intraspecific genetic diversity

The nucleotide diversity per base pair averaged on total number of sites (π) was calculated for both species (Table 2). π was 0.0026 for *A. cebennensis* and 0.0040 for *A. pedemontana*. The distribution of π per SNP was homogeneous along *A. lyrata* scaffolds for both species (see Manhattan plots in Fig. S8), indicating that the genetic variability within each species is distributed uniformly along chromosomes. The three *A. cebennensis* populations displayed similar values of π (Table 2). θ_w values were almost identical to π (Table 2). *Arabidopsis cebennensis* also displayed a lower mean expected heterozygosity (H_{exp}) than *A. pedemontana* (Table 2). The mean observed heterozygosity (H_{obs}) calculated were only slightly lower than the expected values. These results showed a higher diversity overall in *A. pedemontana* sampled populations compared to *A. cebennensis*. In *A. cebennensis*, 39% of the genetic variability in the

392 species is explained by the population structure (among populations), while the rest is explained
393 by individuals within populations.
394
395 **Species demographic history**
396 To infer the history of the two species, we tested five single-population demographic models
397 using the joint site frequency spectrum (SFS) and the maximum likelihood methods implemented
398 in δaδi. Considering the western alpine and sub-alpine distribution of the two species, and the fact
399 that they occur at the highest possible elevation of the surrounding regions, we assumed they
400 diverged during a glaciation area, and survived only in refugia in higher altitude during warming
401 period, where they were already adapted to the colder conditions. Such a scenario implies a
402 strong population decline of the species somewhere in their past. For each species, we fitted
403 models 1a to 1d to test for different scenarios of population size change and compared the results
404 to the standard neutral model (SNM) (Fig. S2). We were unsuccessful in testing models
405 incorporating a split of *A. cebennensis* in three sub-populations, due to the sampling size and the
406 quantity of segregating sites obtained for each of these populations. The observed minor allele
407 frequency spectra of *A. cebennensis* and *A. pedemontana* were quite different. In *A.*
408 *pedemontana*, the declining function (larger number of rarer alleles) is consistent with mutation-
409 drift equilibrium in a stable population (log composite likelihood value LL = -334.4 for SNM).
410 *Arabidopsis cebennensis* SFS was right-shifted (biased towards intermediate frequency alleles),
411 deviating from the neutral expectation (LL = -1229.8 for SNM). This was explained by the strong
412 population structure in *A. cebennensis*. The δaδi analyses suggest that the instantaneous

413 population size change model 1a and the exponential size change model 1b are better than the
414 SNM based on LL values and fitting plots (Fig. 4 and Fig. S2). The two models 1a and 1b fitted
415 the observed data significantly better than the dual size change model 1c (with an error alpha=
416 5%, Adjusted D-statistic = -0.79 for *A. cebennensis* and -3.75 for *A. pedemontana*, p-value = 1).
417 Model 1d fitted the observed data poorly (Fig. S2).
418 Equivalent fits were observed for models 1a and 1b, which could not be compared with a
419 likelihood ratio test (LRT) as they were not nested. Both suggested a strong population decline
420 for the two species (Fig. 4) and indicated a deeper reduction and smaller relative population size
421 for *A. cebennensis*: the effective population size of this species has been reduced to
422 approximately 0.1% of its ancestral population size N_a (Fig. 4). In contrast, the *A. pedemontana*
423 population shrunk only to half of its N_a. Assuming an average generation time of two years and
424 an overall mutation rate of 7×10^{-9} base substitutions per site per generation (Ossowski *et al.*
425 2010) we estimated the modern effective population size of *A. cebennensis* to be ≈ 140 [0-287]
426 from model 1a or ≈ 450 [365-548] from model 1b. The effective population size for *A.*
427 *pedemontana* was ≈ 55,000 [47,455-62,056] from model 1a and ≈ 50,000 [42,377-58,411] from
428 model 1b. Our estimates placed the drastic population size decrease in *A. cebennensis* either ≈
429 230 [0-689] years ago (model 1a) or ≈ 3,700 [3,044-4,262] years ago (model 1b). The decrease of
430 *A. pedemontana* population was estimated either ≈ 18,000 [9,126-27,378] (model 1a) or ≈ 41,000
431 years ago (model 1b).
432 As the population structure observed in *A. cebennensis* was not modeled in the analysis, it could
433 explain the non-optimal fitting between the models SFS and *A. cebennensis* observed SFS, the

434 larger correlated residuals for *A. cebennensis*, and the strikingly low population size estimated for
435 this species. To evaluate the effect of strong population structure on model fitting, we additionally
436 fitted all single-population models to the observed data of the Cantal population separately. This
437 population, for which we have the largest sample size (nine individuals), provides a
438 representative dataset for the whole species without effect of structure. However the SFS
439 observed for the Cantal population was also biased toward intermediate allele frequencies (Fig.
440 S2) and the single-population models did not fit the Cantal data better than the whole species
441 dataset. The structure observed previously within the Cantal population was very weak (F_{st} = 0.05
442 between clusters 139R and 140R/141R) and cannot completely explain the SFS observed. In all
443 models, the contemporary population size was as low as for the whole *A. cebennensis* dataset.
444 A second hypothesis was that *A. cebennensis* and *A. pedemontana* diverged after the
445 establishment of their common ancestor to the different refugia areas, the French sub-alpine
446 region Massif central and the alpine Piemont region in Italy. Their divergence would have
447 resulted from the absence of gene flow due to geographical barriers. This scenario assumes a
448 bottleneck would have happened in the common ancestor of the two species during a deglaciation
449 period, shortly before speciation. It was tested by fitting the two-population bottleneck model 2c
450 and comparing the results to the fitting of models 2a and 2b, which assume the divergence of the
451 two species prior to instantaneous and exponential size change respectively. All two-populations
452 models fitted the data very poorly based on LL values and fitting plots (Fig. S2). The lack of
453 outgroup information and previous assumptions on any parameters of the model, the number of
454 parameters to optimize and the potential linkage existing between our SNPs (δaδi assumes that all

455 polymorphic sites are independent) may explain why none of the tested two-populations models
456 fitted the data well. The fact that model 2c display the lowest LL is not in favor of a scenario
457 including a bottleneck in the common ancestor population. The inferred divergence time between
458 *A. cebennensis* and *A. pedemontana* was similar for models 2a and 2b (0.40 N_a generations ago).
459 From model 2b we estimated that the speciation events occurred ≈ 168.000 [159676-176484]
460 years ago.

Discussion

As endemic species have higher probabilities to display a small census and effective population size and a highly specific ecology, reduced genetic diversity could be a limiting factor for their adaptive potential in changing environments (Ellstrand & Elam 1993; Frankham 1995; Allendorf & Ryman 2002; Leimu *et al.* 2006). In this study we evaluated the extent and structure of genetic variation in the two endemic species *A. cebennensis* and *A. pedemontana*, using samples from most of their currently known locations in southern France and northwestern Italy. Their demographic history was also investigated and their present effective population size estimated in order to determine which processes shaped the modern pattern of genetic diversity.

Genetic structure

Arabidopsis cebennensis presented three clearly distinct populations corresponding to the large geographic regions Cantal, Cévennes, and Ardèche, which are represented in this study by 11, four and six samples respectively. Ardèche and Cévennes individuals were more closely related to each other. The Cantal population can be further subdivided between subpopulation 139R (4 individuals) and subpopulation 140R/141R (7 individuals). This was surprising considering that the populations 139R and 140R are geographically closer to each other (~ 4 km) than to the population 141R (~ 15 km) indicating either dispersal between the Chambeuil and the Le Siniq springs where 141R and 140R populations are respectively located, or simply shared ancestry. No reliable gene flow event could be detected between the three allopatric *A. cebennensis*

populations in their recent or past history. We assume that strong barriers of gene flow exist among these units and that they are evolving independently through genetic drift. It was not surprising that an isolation-by-distance pattern (Slatkin 1993) was not apparent as only a small proportion of the total variation is expected to be spatially correlated in case of extremely limited dispersal (Meirmans 2015). In contrast, *A. pedemontana* showed only a weak clustering of individuals, probably as a result of an IBD pattern of gene flow. Dispersal is thus maintained but limited to relatively short distance within *A. pedemontana* distribution range, certainly because of the mountainous relief of the region. While *A. pedemontana* known sampling sites are all in maximum 10 km from each other, the three *A. cebennensis* populations are ~100 km away from each other with only one known additional isolated population in Aveyron, between the Cantal and the Cévennes. These long distances, the natural barriers they include, and the ecological specialization of the species explain why the connectivity among *A. cebennensis* populations is not maintained. The mean pairwise F_{st} values pointed out the isolation of these populations as they were almost as differentiated between each other than the species is differentiated from *A. pedemontana*. However, the genetic distance between *A. cebennensis* and *A. pedemontana* (0.33) was not high compared to other intra and inter-specific comparisons in the genus using neutral markers: the average pairwise F_{st} between populations of *A. halleri* from two large units separated by the Alps, was around 0.37 (Pauwels *et al.* 2012); the average F_{st} calculated between 31 *A. halleri* and 48 *A. lyrata* individuals was 0.46 ± 0.211, although this value was conservative as it was calculated from nuclear coding sequences only (Roux *et al.* 2011); the median pairwise F_{st} values calculated by Ross-Ibarra *et al.* (2008) for six natural populations of *A. lyrata* from across

the range of the species were between 0 and 0.6 with 12 on 15 comparisons above 0.2. The median F_{st} between *A. cebennensis* and *A. pedemontana* was 0.22, which is equal or lower than many intra-specific comparisons within *A. lyrata*.

Genetic diversity

Interestingly, the number and percentage of polymorphic loci were twice as high for *A. pedemontana* as for *A. cebennensis*. The nucleotide diversity (π), Watterson's estimator (θ_w) and Nei's gene diversity (H_{exp}) were all higher for *A. pedemontana* compared to *A. cebennensis*, supporting a greater genetic diversity in the Italian species despite its smaller distribution range. Only the plastid DNA variation was lower for *A. pedemontana* with only one *trnLF* plastid region haplotype while *A. cebennensis* displayed four different types (Hohmann *et al.* 2014). *A. halleri*, *A. arenosa*, and *A. lyrata*, the three major lineages in the genus, presented 14, 32 and 30 suprahaplotypes, much more that what was observed for the two endemics. Like *A. cebennensis* and *A. pedemontana*, *A. lyrata* and *A. halleri* are diploid, outcrossing and perennial species, but they are widely distributed. Higher average π values were also found for *A. halleri* ssp. *halleri* (0.0081) and *A. lyrata* ssp. *petraea* (0.0116) (Ramos-Onsins 2004) compared to *A. pedemontana* (0.004) and *A. cebennensis* (0.0026). The average π value in *A. thaliana* was also higher (0.0054) although the species is selfing (Nordborg *et al.* 2005). With the set of microsatellite loci included here for structure analysis, the highest number of total alleles, unique alleles and rare alleles, considering only diploids, were found within widely distributed *A. lyrata* ssp. *petraea* and *A. carpatica* (*A. arenosa* lineage; Hohmann *et al.* 2014). These diversity statistics were the lowest

for *A. pedemontana* and *A. cebennensis*, as well as *A. croatica*, a third endemic species in the genus (Hohmann *et al.* 2014). Microsatellite H_{exp}, calculated overall in each major lineage, is highest in *A. lyrata* (0.562 ± 0.312) and *A. arenosa* (0.560 ± 0.311), lower in *A. halleri* (0.427 ± 0.254) and much lower in *A. pedemontana* (0.259 ± 0.167) and *A. cebennensis* (0.189 ± 0.130). When calculating mean H_{exp} over SNPs only (and not over the total number of mapped sites), we obtained similar values with 0.33 for *A. cebennensis* and 0.23 for *A. pedemontana*. Only for this comparison *A. pedemontana* does not display a higher genetic diversity compared to *A. cebennnensis*. Overall, we observe that the genetic variation level in the two endemic species is lower than for the other *Arabidopsis* species with wide distribution range and the same mating system (self-incompatibility). For comparison, the narrow endemic *Aquilegia thalictrifolia*, distributed in a few valleys of the Italian South-Eastern Alps, also displayed a higher average H_{exp} (0.68 ± 0.19), which is twice as high as for the two *Arabidopsis* endemics (Lega *et al.* 2014). In *A. cebennensis* π, θ_w and H_{exp} were low and almost similar for the three subpopulations, and similar to what was estimated for the whole species. A similar level of genetic diversity is partitioned among and within the subpopulations in this species. It shows that gene flow was too limited to ensure that variation is shared over the whole distribution of the species, and led to the strong population structure observed. Overall, both *A. pedemontana* and *A. cebennensis* display low levels of genetic variation at the taxon and population levels, as would be predicted based on their narrow geographic range.

Comparison of population genetics inference methods

The SNPs from the GBS analysis and the microsatellites markers gave similar results with regard to population structure, despite the differences in sampling size. In *A. cebennensis*, the level of admixture observed between the Ardèche and Cévennes populations with the microsatellite data confirm that these two populations had more recent genetic exchanges than with the Cantal populations. Concerning genetic diversity statistics, the GBS analyses gave much lower H_{exp} values than the microsatellites markers when both polymorphic and non-polymorphic sites were accounted for. As it takes in account the number of SNPs detected over the total number of sites, this approach gives a more realistic estimation of the level of heterozygosity over the whole genome. Nonetheless the two methods of polymorphism detection gave similar results in terms of ratio of H_{exp} between the two endemic species. The genome-wide SNPs and the multiple microsatellite loci were thus equally useful in inferring population structure and comparing the level of diversity between related species in our study. One advantage of using GBS is to screen thousands of polymorphism that are subject to the full range of evolutionary processes acting across the genome (mutation, drift, selection) (Narum *et al.* 2013) and improve the precision of demographic inferences by greatly increasing the number of putatively neutral markers assayed.

Past decline of the endemics populations

Our overall neighbor-net network with *A. lyrata*, *A. cebennensis* and *A. pedemontana* confirmed previous observations that the two endemic species, although clearly separated, are sister species (Koch & Matschinger 2007; Hohmann *et al.* 2014). Additionally, 10% and 5% of the total polymorphism in *A. cebennensis* and *A. pedemontana* are shared polymorphisms, confirming the joint evolutionary history of the two species. As the two species are ecologically, morphologically

565 and in certain case phylogenetically closer to *A. halleri* (Hohmann *et al.* 2014), we assumed that
566 they diverged from a common ancestor related to *A. halleri* lineage. *A. halleri* is distributed in the
567 whole alpine chain while the endemic species occur in refuge areas west of the mountainous
568 range, at the highest altitude in the surroundings. The periphery of European Alps, in particular
569 the south east and the west of the mountain range, harbors many small refugia with hundreds of
570 endemic plant species, which result from Pleistocene glaciation cycles (Comes & Kadereit 2003;
571 Tribsch & Schonswetter 2003; Schönswetter *et al.* 2005). We therefore hypothesized that the
572 evolutionary histories of both species have been influenced by Pleistocene climate oscillations:
573 the two cold-adapted species diverged from their common ancestor in the western alps during a
574 glaciation period; with the next deglaciation phases the populations migrated or became restricted
575 to refugia in higher altitude where they could survive the warming temperatures, resulting in a
576 strong overall population size decrease; because of their adaptation to their specific refugia
577 habitat and/or competition, they could not expand back to their unknown original distribution
578 range, forming relictual populations.
579 The demographic analysis confirmed that both endemic species have undergone a strong decline
580 in the past and did not recover to the present days, although we were not able to state if this
581 decline happened through a strong bottleneck or through exponential population size decrease.
582 The population decline was particularly strong for *A. cebennensis*, which displayed a
583 contemporary effective population size N_e of 140 to 450 individuals, only 0.1% of the size of its
584 ancestral population. *A. pedemontana*'s decline was less drastic, as the species exhibits a modern
585 N_e of 50,000-55,000 individuals, half of its ancestral population size. In comparison, Roux *et al.*

(2011) estimated N_e of ~82,000 and ~79,200 for the two more widespread relatives *A. halleri* and *A. lyrata*. According to our estimates, *A. pedemontana* population started decreasing 18,000 to 41,000 years ago (ya), while *A. cebennensis*' decline was much more recent, i.e. 200 to 3,700 ya. The time of divergence of the two species was estimated at ~170,000 ya. However these estimates are doubtful for two reasons: (i) the two-populations demographic models within which the time of divergence parameter was optimized fitted the observed dataset poorly; (ii) the time values were converted in units of years by using an average generation time of two years. Because the two species are long-lived perennials, propagate vastly vegetatively, and their dispersal could be limited for some times by their specific ecology and the space competition observed in their habitat (pers. obs.), the chosen generation time as well as the divergence time and times of decline could be underestimated. The divergence of *A. cebennensis* and *A. pedemontana* falls in the medium Pleistocene (Ionien), which corroborates our hypothesis. These results fit with the time of radiation calculated for all *Arabidopsis* lineages with n = 8 chromosomes (all except *A. thaliana*) (1.63 Mya; Hohmann *et al.* 2015), and for *A. halleri* ssp. *halleri* (~335,000 ya; Roux *et al.* 2011). While *A. pedemontana*'s decline falls still in the late Pleistocene, the decline of *A. cebennensis* seems to have occured in modern time (Holocene). As the demographic models (1a and 1b) used to estimate this value did not fit *A. cebennensis* observed data SFS perfectly, and *A. cebennensis* population structure was not included in the models, we can not say with certainty that this very recent estimate is accurate. Our separate demographic analysis of *A. cebennensis* Cantal population showed that the strong population structure of *A. cebennensis* alone does not explain completely the right-shifted site frequency

607 spectrum observed for this species and the extremely low population size that was estimated.
608 While population structure could explain the excess of intermediate frequency alleles, a recent
609 strong bottleneck could explain the lack of rare alleles (Luikart *et al.* 1998) and the small
610 population size. In the future, the evolutionary history of the three *A. cebennensis* populations
611 will be further investigated with an extended and more representative sampling scheme, which
612 will also allow us to integrate the population structure into the demographic analysis. Particularly
613 we will test if the structure and the sampling size together have created a false bottleneck signal
614 for this species (Chikhi *et al.* 2010).
615
616 ***Consequence on adaptive potential***
617 In summary we described two endemics dissimilar patterns of genetic variation and
618 differentiation. Due to a potentially rapid and recent decline, and its split into three allopatric
619 populations, *A. cebennensis* currently exhibits a very low effective population size and low levels
620 of genetic diversity. Although the absence of gene flow between these three sub-populations
621 initiated their divergence by genetic drift or divergent selection (according to F_{st} values), the
622 genetic diversity among them is also low (according to π values). The probability of finding
623 strongly differentiated genotypes and phenotypes in the different geographic region where the
624 species occurs is consequently small. *Arabidopsis pedemontana* population also declined rapidly
625 in the past but not as strongly as for *A. cebennensis*, and *A. pedemontana* effective population size
626 is still relatively high. Distance-limited dispersal is ongoing within its distribution range but did
627 not contribute to create much diversity between the locations. Indeed the overall diversity level

628 remains low. The contemporary population size, levels of gene flow versus genetic drift and
629 demographic history of both species can explain why we observed higher level of diversity for *A.*
630 *pedemontana* compared to *A. cebennensis*, despite its smaller distribution range. A historically
631 larger population size in *A. pedemontana* could also have contributed to its higher level of genetic
632 diversity.
633 These first results are not optimistic regarding the adaptive potential of the two species in case of
634 environmental change. The relatively recent and on-going global climate warming could impact
635 the habitats of these two species by two means: 1) In Europe, it was already observed that climate
636 change causes a general upward migration of plants to higher altitudes where they may compete
637 with locally adapted endemic plants (Lenoir *et al.* 2008); 2) The increasing drought that could
638 result from rising temperatures and decreasing precipitation predicted to be induced by climate
639 change could particularly impact the highly specific riverine habitat of the two species. As a
640 result, the occurrence of the two species in high altitude habitats, where they can not escape the
641 increased competition, and their habitat specialization render them particularly vulnerable to
642 climate change (Gottfried *et al.* 2012). Predicting populations persistence during environmental
643 change is complex as it depends on numerous factors (Reed *et al.* 2011). However, with the
644 relatively low levels of genetic diversity observed, we can predict that the two *Arabidopsis*
645 endemic species may experience reduced relative adaptive ability and a higher risk of genetic
646 extinction. The risk could be relatively higher for *A. cebennensis*, as small population size
647 decrease the rates of adaptive evolution (Strasburg *et al.* 2011; Lanfear *et al.* 2014) and do not
648 allow the populations to sustain further decline before evolutionary rescue (Gonzalez *et al.* 2012).

649

650 # Acknowledgments

651 We thank Elizabeth Kokai-Kota for laboratory assistance, and Markus Stetter and Fabian Freund

652 for help with the population genetic analysis. Financial support was provided to K.J.Schmid by

653 the DGF Grant SCHM1354/6-1 within the DFG priority programme 'Adaptomics'.

References

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–64.

Allendorf FW, Ryman N (2002) The role of genetics in population viability analysis. In: Population Viability Analysis (eds Beissinger SR, McCullough DR), pp. 50–85. University Chicago press.

Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, e310.

Bell G (2013) Evolutionary rescue and the limits of adaptation. *Philosophical transactions of the Royal Society of London*, **368**, 20120080.

Borcard D, Legendre P, Drapeau P (1992) Partialling out the Spatial Component of Ecological Variation. *Ecology*, **73**, 1045–1055.

Bryant D, Moulton V (2004) Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, **21**, 255–65.

Cao J, Schneeberger K, Ossowski S *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, **43**, 956–963.

Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195-205.

Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA (2010) The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics*, **186**, 983–95.

Clauss MJ, Koch MA (2006) Poorly known relatives of *Arabidopsis thaliana*. *Trends in Plant Science*, **11**, 449–459.

Comes HP, Kadereit JW (2003) Spatial and temporal patterns in the evolution of the flora of the European Alpine System. *Taxon*, **52**, 451–462.

Cornish-Bowden A (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic acids research*, **13**, 3021–3030.

Danecek P, Auton A, Abecasis G *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–8.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Earl DA, VonHoldt BM (2012) STRUCTURE HARVESTER : a website and program for

visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.

Ehrich D (2006) AFLPDAT: A collection of R functions for convenient handling of AFLP data. *Molecular Ecology Notes*, **6**, 603–604.

Ellstrand NC, Elam DR (1993) Population genetic consequences of small population size: Implications for Plant Conservation. *Annual Review of Ecology and Systematics*, **24**, 217–242.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one*, **6**, e19379.

Frankham R (1995) Conservation genetics. *Annual Review of Genetics*, **29**, 305–327.

Freville H, Justy F, Olivieri I (2001) Comparative allozyme and microsatellite population structure in a narrow endemic plant species, *Centaurea corymbosa* Pourret (Asteraceae). *Molecular Ecology*, **10**, 879–889.

Gonçalves da Silva A, Appleyard SA, Upston J (2015) Establishing the evolutionary compatibility of potential sources of colonizers for overfished stocks: a population genomics approach. *Molecular Ecology*, **24**, 564–79.

Gonzalez A, Ronce O, Ferriere R, Hochberg ME (2012) Evolutionary rescue: an emerging focus at the intersection between ecology and evolution. *Philosophical transactions of the Royal Society of London*, **368**, 20120404.

Gottfried M, Pauli H, Futschik A *et al.* (2012) Continent-wide response of mountain vegetation to climate change. *Nature Climate Change*, **2**, 111–115.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, **5**, e1000695.

Hensen I, Oberprieler C (2005) Effects of population size on genetic diversity and seed production in the rare *Dictamnus albus* (Rutaceae) in central Germany. *Conservation Genetics*, **6**, 63–73.

Hohmann N, Schmickl R, Chiang T *et al.* (2014) Taming the wild : resolving the gene pools of non-model *Arabidopsis* lineages. *BMC Evolutionary Biology*, **14**, 1–21.

Hohmann N, Wolf EM, Lysak MA, Koch MA (2015) A time-calibrated road map of Brassicaceae species radiation and evolutionary history. *The Plant Cell*, 1–16.

Hu TT, Pattyn P, Bakker EG *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, **43**, 476–481.

720 Huang P, Feldman M, Schroder S *et al.* (2014) Population genetics of *Setaria viridis*, a new
721 model system. *Molecular Ecology*, **23**, 4912–25.

722 Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with
723 the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–32.

724 Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components : a new
725 method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

726 Kimura M (1983) The Neutral Theory of Molecular Evolution. *Scientific American*, **241**, 65-97.

727 Koch MA, Matschinger M (2007) Evolution and genetic differentiation among relatives of
728 *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States*
729 *of America*, **104**, 6272–6277.

730 Koch MA, Wernisch M, Schmickl R (2008) *Arabidopsis thaliana*'s wild relatives: an updated
731 overview on systematics, taxonomy and evolution. *Taxon*, **57**, 933–943.

732 Lanfear R, Kokko H, Eyre-Walker A (2014) Population size and the rate of evolution. *Trends in*
733 *Ecology & Evolution*, **29**, 33–41.

734 Leffler EM, Bullaughey K, Matute DR *et al.* (2012) Revisiting an old riddle: what determines
735 genetic diversity levels within species? *PLoS Biology*, **10**, e1001388.

736 Lega M, Fior S, Li M, Leonardi S, Varotto C (2014) Genetic drift linked to heterogeneous
737 landscape and ecological specialization drives diversification in the alpine endemic
738 columbine. *Journal of Heredity*, **105**, 542–554.

739 Leimu R, Fischer M (2008) A meta-analysis of local adaptation in plants. *PloS one*, **3**, e4010.

740 Leimu R, Mutikainen PIA, Koricheva J, Fischer M (2006) How general are positive relationships
741 between plant population size, fitness and genetic variation? *Journal of Ecology*, **94**, 942–
742 952.

743 Lenoir J, Gégout JC, Marquet PA, De Ruffray P, Brisse H (2008) A significant upward shift in
744 plant species optimum elevation during the 20th Century. *Science*, **320**, 1768–1771.

745 Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform.
746 *Bioinformatics*, **25**, 1754–60.

747 Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and
748 SAMtools. *Bioinformatics*, **25**, 2078–2079.

749 Luikart G, Allendorf FW, Cornuet J, Sherwin WB (1998) Distortion of allele frequency
750 distributions provides a test for recent population bottlenecks. *The Journal of Heredity*, **89**,
751 238–247.

752 Lynch M, Burger R, Butcher D, Gabriel W (1993) The mutational meltdown in asexual

753 populations. *Journal of Heredity*, **84**, 339–344.

754 Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ (2009) The dynamic ups and downs of
755 genome size evolution in Brassicaceae. *Molecular Biology and Evolution*, **26**, 85–98.

756 Michalski SG, Durka W (2007) High selfing and high inbreeding depression in peripheral
757 populations of *Juncus atratus*. *Molecular Ecology*, **16**, 4715–27.

758 Mitchell-Olds T, Schmitt J (2006) Genetic mechanisms and evolutionary significance of natural
759 variation in *Arabidopsis*. *Nature*, **441**, 947–52.

760 Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GA, Kent J (2000) Biodiversity hotspots
761 for conservation priorities. *Nature*, **403**, 853–8.

762 Narum SR, Buerkle CA, Davey JW, Miller MR, Hohenlohe PA (2013) Genotyping-by-
763 sequencing in ecological and conservation genomics. *Molecular Ecology*, **22**, 2841–7.

764 Nordborg M, Hu TT, Ishino Y *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*.
765 *PLoS biology*, **3**, e196.

766 O'Kane SL, Al-Shehbaz IA (1997) A Synopsis of *Arabidopsis* (Brassicaceae). *Novon*, **7**, 323–
767 327.

768 Ossowski S, Schneeberger K, Lucas-lledó JI *et al.* (2010) The rate and molecular spectrum of
769 spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.

770 Paradis E (2010) Pegas: An R package for population genetics with an integrated-modular
771 approach. *Bioinformatics*, **26**, 419–420.

772 Pauwels M, Vekemans X, Godé C *et al.* (2012) Nuclear and chloroplast DNA phylogeography
773 reveals vicariance among European populations of the model species for the study of metal
774 tolerance, *Arabidopsis halleri* (Brassicaceae). *New Phytologist*, **193**, 916–928.

775 Peters D, Luo X, Qiu K, Liang P (2012) Speeding Up Large-Scale Next Generation Sequencing
776 Data Analysis with pBWA. *Journal of Applied Bioinformatics & Computational Biology*, **1**,
777 1–6.

778 Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide
779 allele frequency data. *PLoS genetics*, **8**, e1002967.

780 Pluess AR, Stöcklin J (2004) Genetic diversity and fitness in *Scabiosa columbaria* in the Swiss
781 Jura in relation to population size. *Conservation Genetics*, **5**, 145–156.

782 Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic
783 maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach.
784 *PloS one*, **7**, e32253.

785 Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus

786 genotype data. *Genetics*, **155**, 945–959.

787 Ramos-Onsins SE (2004) Multilocus Analysis of Variation and Speciation in the Closely Related
788 Species *Arabidopsis halleri* and *A. lyrata*. *Genetics*, **166**, 373–388.

789 Reed TE, Schindler DE, Waples RS (2011) Interacting effects of phenotypic plasticity and
790 evolution on population persistence in a changing climate. *Conservation Biology*, **25**, 56–63.

791 Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of polymorphism and demographic
792 history in natural populations of *Arabidopsis lyrata*. *PloS one*, **3**, e2411.

793 Roux C, Castric V, Pauwels M *et al.* (2011) Does speciation between *Arabidopsis halleri* and
794 *Arabidopsis lyrata* coincide with major changes in a molecular target of adaptation ? *PloS*
795 *one*, **6**, e26872.

796 Saghai-Maroof MA, Soliman KM, Jorgensen RA, Allard RW (1984) Ribosomal DNA spacer-
797 length polymorphisms in barley: mendelian inheritance, chromosomal location, and
798 population dynamics. *Proceedings of the National Academy of Sciences of the United States*
799 *of America*, **81**, 8014–8018.

800 Salamin N, Wüest RO, Lavergne S, Thuiller W, Pearman PB (2010) Assessing rapid evolution in
801 a changing environment. *Trends in Ecology & Evolution*, **25**, 692–8.

802 Schönswetter P, Stehlik I, Holderegger R, Tribsch A (2005) Molecular evidence for glacial
803 refugia of mountain plants in the European Alps. *Molecular Ecology*, **14**, 3547–3555.

804 Slatkin M (1993) Isolation by Distance in equilibrium and non-equilibrium populations.
805 *Evolution*, **47**, 264–279.

806 Spielman D, Brook BW, Frankham R (2004) Most species are not driven to extinction before
807 genetic factors impact them. *Proceedings of the National Academy of Sciences of the United*
808 *States of America*, **101**, 15261–4.

809 Strasburg JL, Kane NC, Raduski AR *et al.* (2011) Effective population size is positively
810 correlated with levels of adaptive divergence among annual sunflowers. *Molecular Biology*
811 *and Evolution*, **28**, 1569–1580.

812 The IUCN Species Survival Commission (2004) 2004 IUCN red list of threatened species: a
813 global species assessment. (eds Baillie J, Hilton-Taylor C, Stuart S). IUCN Publications
814 Services Unit, Cambridge, CB3 ODL, UK.

815 Thomas CD, Cameron A, Green RE *et al.* (2004) Extinction risk from climate change. *Nature*,
816 **427**, 145–8.

817 Tribsch A, Schonswetter P (2003) Patterns of endemism and comparative phylogeography
818 confirm palaeoenvironmental evidence for pleistocene refugia in the eastern Alps. *Taxon*, **52**,
819 477–497.

Weir B, Cockerham CC (1984) Estimating F-Statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Xue Y, Prado-Martinez J, Sudmant PH *et al.* (2015) Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, **348**, 242–5.

Data Accessibility

- GBS raw reads: Genbank NCBI submission SRP072277
- Biallelic SNPs final vcf files: DataDryad
- Demographic analysis scripts: Supporting information

Author Contributions

JJ coordinated the project, performed the GBS sequencing and population genetic analysis and wrote the manuscript. NH and MAK conducted the microsatellite and plastid DNA sequences analysis and contributed to write the manuscript. MB contributed to the sample collection and DNA extraction. AS helped with the sample collection. TM assisted with the GBS reads processing and population genetic analysis. KJS designed and coordinated the project and contributed in writing the manuscript. All authors read and approved the final manuscript.

Tables

Table 1. GBS analysis output: Number of total and polymorphic sites. Only SNPs with data for at least 70% of the sampled individuals are included.

	A. cebennensis	***A. pedemontana***
Total sites	874,955	739,868
SNPs	6,583	12,909
% polymorphic sites	0.7	1.74
Average read depth per SNPs	465	1,124
% missing data	13.7	7.9

Table 2. Intra-specific genetic diversity parameters

	A. cebennensis				***A. pedemontana***
	All	Ardèche	Cévennes	Cantal	All
π	0.0026 ± 7.00E-5	0.0018 ± 6.00E-5	0.0014 ± 6.00E-5	0.0020 ± 6.00E-5	0.0040 ± 8.00E-5
θ_w	0.0021	0.0018	0.0016	0.0018	0.0039
H_{exp}	0.0025 ± 6.00E-5	0.0016 ± 6.00E-5	0.0013 ± 5.00E-5	0.0019 ± 6.00E-5	0.0039 ± 8.00E-5
H_{obs}	0.0022 ± 7.00E-5	0.0022 ± 8.00E-5	0.0017 ± 8.00E-5	0.0024 ± 8.00E-5	0.0037 ± 9.00E-5

Figures

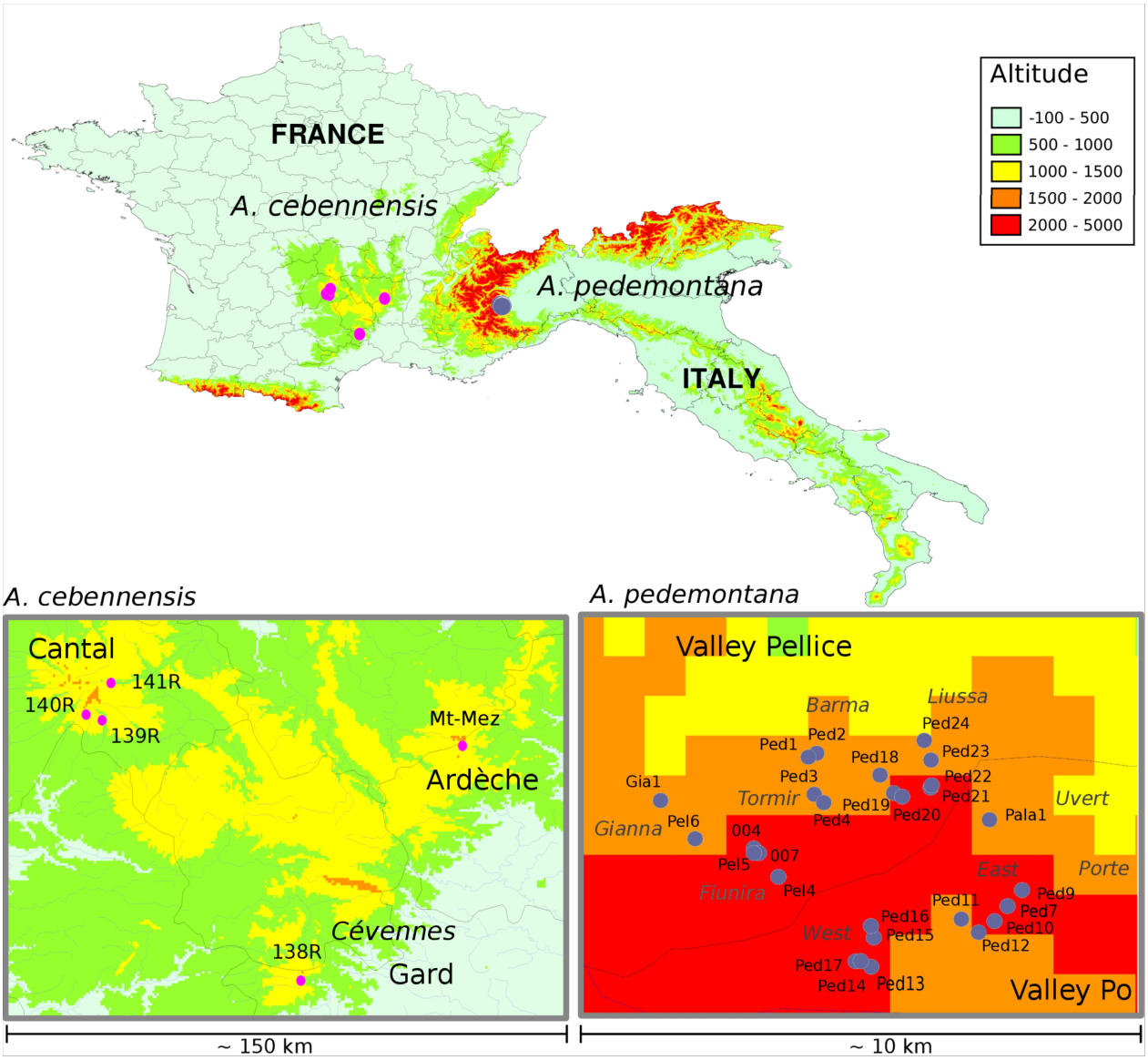

Figure 1. Distribution range and altitude of the populations and individuals sampled for *A. cebennensis* and *A. pedemontana*.

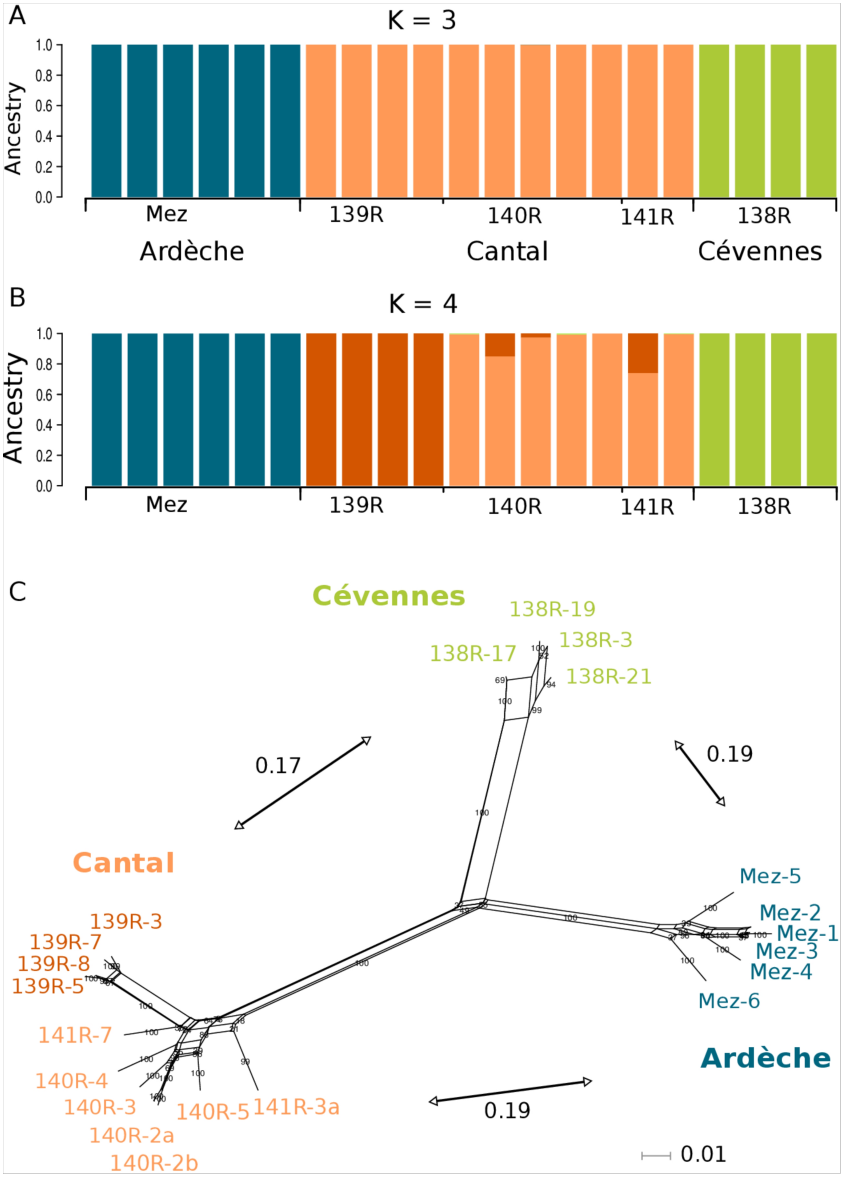

Figure 2. Population structure analysis of *A. cebennensis*: A) ADMIXTURE plots for number of clusters K=3 and K=4, B) Neighbor-net phylogenetic network. Bootstrap values are indicated on the network branches. F_{st} values between the three large populations are indicated along the arrows.

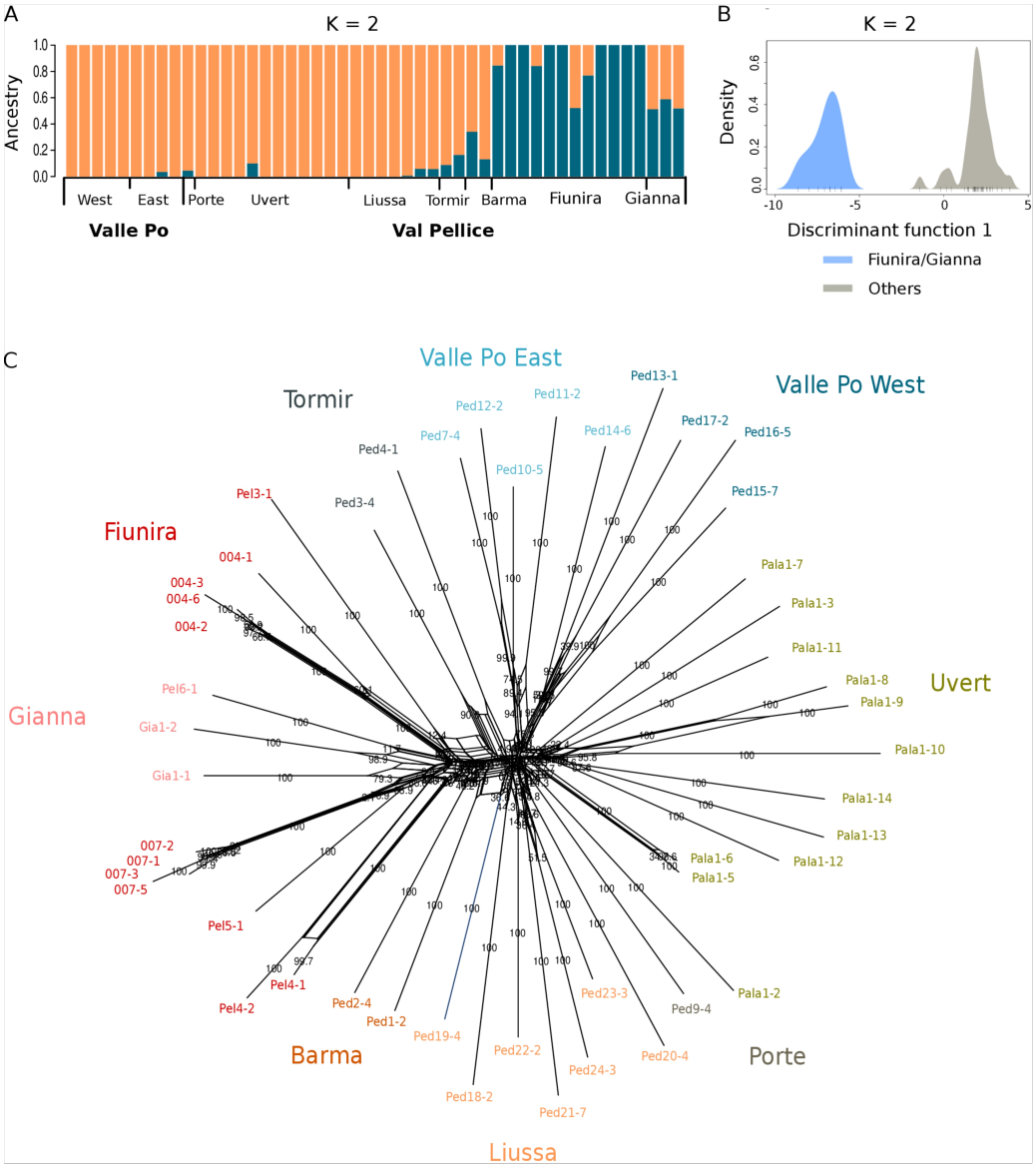

Figure 3. Population structure analysis of *A. pedemontana*: A) ADMIXTURE plots for number of clusters K=2, B) Plot of DAPC showing the first principal component of the analysis with K=2, with individuals (vertical bars) and groups (colored courbs) plotted on the plane, C) Neighbor-net phylogenetic network. Bootstrap values are indicated on the network branches.

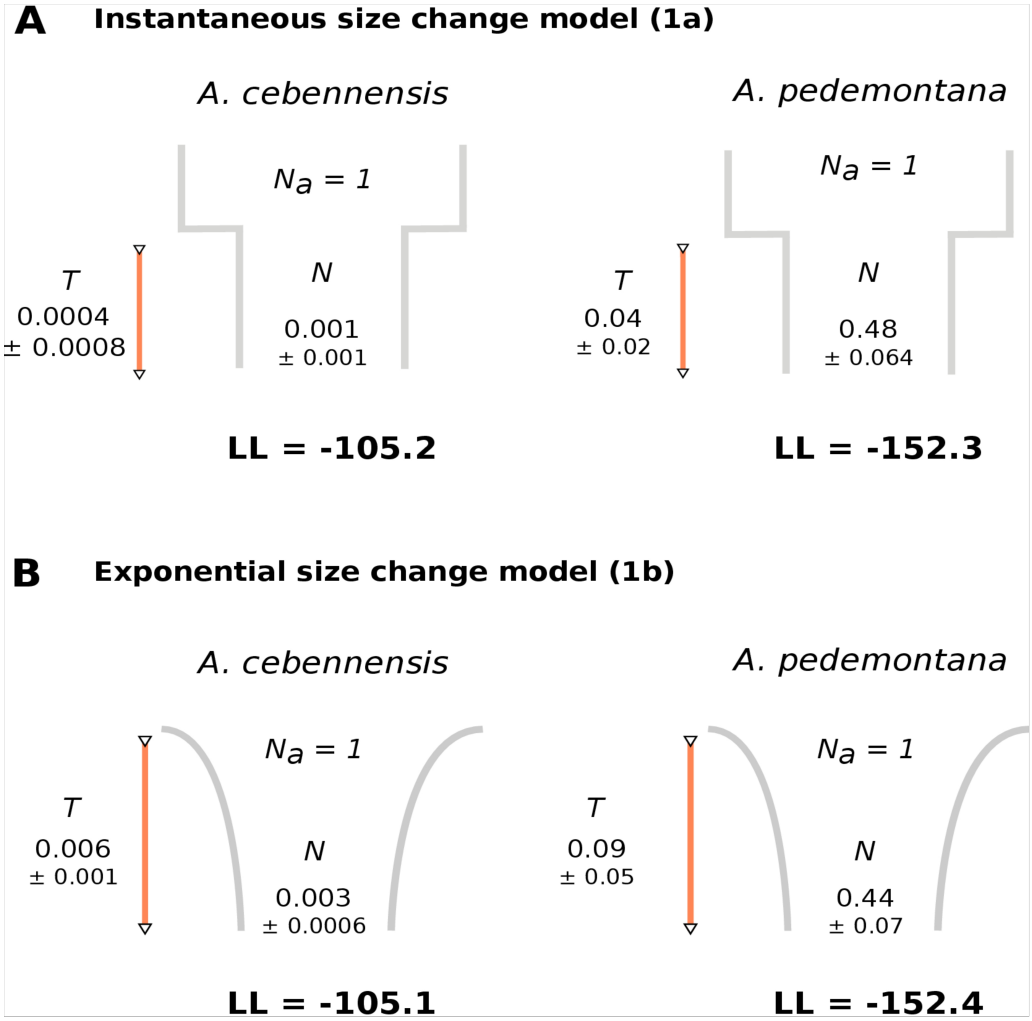

Figure 4. Representation of demographic models 1a and 1b and their optimized parameters fitted to observed data in *A. cebennensis* and *A. pedemontana*. The log likelihood values (LL) are indicated for each model.

Appendices

Figure S1. Barcode sequences and design of double-digest GBS libraries

Figure S2. Comparison of the fitted demographic models. The optimized parameters and log likelihood values (LL) are indicated for each model. The site frequency spectrum is plotted for each model in red and compared to the observed data SFS plotted in blue. Ac = *Arabidopsis cebennensis*; Ap = *A. pedemontana*.

Figure S3. Genetic assignments of *A. cebennensis* and *A. pedemontana* populations from population structure analysis with microsatellite and plastid DNA sequences: A) Summary of microsatellite analysis; B) Distribution of *A. cebennensis* populations analyzed for microsatellite variation. The color-code refers to STRUCTURE results shown with C; C) STRUCTURE plot of the microsatellite data with K=3 (highest probability) for 144 *A. cebennensis* individuals from 10 populations. The chloroplast genome type are indicated for each population below the graph (T, BG, A, BH); D) Spatial distribution of *A. pedemontana* populations analyzed for microsatellite variation. The color-code refers to STRUCTURE results shown with E; E) STRUCTURE plot of the microsatellite data with K=2 (highest probability) for 40 *A. pedemontana* individuals from 9 different localities. All individuals carried the same single chloroplast genome type (A).

Figure S4. Neighbor Net network constructed for *A. cebennensis*, *A. pedemontana* and outgroup *A. lyrata*. Bootstrap supports are indicated on the branches.

Figure S5. Cross validation values for each iteration and number of clusters in ADMIXTURE analyses: A) *A. cebennensis*; B) *A. pedemontana*.

Figure S6. TreeMix Maximum likelihood tree of *A. cebennensis* three large populations with *A. pedemontana* as outgroup.

Figure S7. Manhattan plots of F_{st} values between: A) *A. cebennensis* and *A. pedemontana* individuals; B) the three large populations of *A. cebennensis;* along *A. lyrata* scaffolds used as reference sequence.

Figure S8. Manhattan plots of nucleotide diversity values (π) within: A) all *A. cebennensis* individuals; B) all *A. pedemontana individuals;* along *A. lyrata* scaffolds used as reference sequences.

Table S1. Q-estimates (admixture proportion at the individual level, expressed as the fraction of each population that contribute to the individual genome) for *A. pedemontana* individuals, with two populations assumed in ADMIXTURE analysis. Individuals for which more than 10% of

each population contribute to the genome are highlighted in grey.

Table S2. SNPs filtering after SNP-calling analysis: 1) Number of total sites, variant sites and polymorphic sites at the successive steps of the filtering process, 2) Mean read depth per site, 3) Percentage of variant and polymorphic sites on the total number of sites; when calculating the % of polymorphic loci, only intraspecific variants and SNPs were taken in account; '>50' = only SNPs with data for at least 50% of the sampled individuals; '>70' = only SNPs with data for at least 70% of the sampled individuals.

Table S3. SNP-calling results for subsets of 21 randomly chosen *A. pedemontana* samples: numbers of intraspecific biallelic SNPs with data for at least 70% (>70) of the sampled individuals are indicated.

Text S1. Detailed Materials and Methods

Supporting information

Supporting information 1. Identification, location and collection informations for samples of *A. cebennensis* and *A. pedemontana* used in the GBS, microsatellite and plastid DNA analysis.

Supporting information 2. Summary of GBS results in number of raw and processed reads overall and per individual. The shaded rows in the second table point out the individuals with less than 5,000 SNPs in the final calling, which were removed from subsequent analysis.

Supporting information 3. Python script implementing all models used in the demographic analyses.