Article type: Original Article

Demography, genetic diversity and expansion load in the colonizing species *Leontodon longirostris* (Asteraceae) throughout its native range.

Running Title: Expansion load in *Leontodon longirostris*

Manuel de Pedro[1], Miquel Riba[1,2], Santiago C. González-Martínez[1,3], Pedro Seoane[4,5,6], Rocío Bautista[6,7], M. Gonzalo Claros[4,5,6,7,8], Maria Mayol[1,*]

[1] CREAF, Cerdanyola del Vallès 08193, Spain

[2] Univ. Autònoma Barcelona, Cerdanyola del Vallès 08193, Spain

[3] INRAE, Univ. Bordeaux, UMR BIOGECO, 33610 Cestas, France

[4] Department of Molecular Biology and Biochemistry, Universidad de Málaga, and Institute for Mediterranean and Subtropical Horticulture (IHSM-CSIC-UMA), Málaga, E-29071, Spain

[5] CIBER de Enfermedades Raras (CIBERER), Málaga, E-29071, Spain

[6] Institute of Biomedical Research in Malaga (IBIMA), IBIMA-RARE, Málaga, E-29010, Spain

[7] Andalusian Platform for Bioinformatics-SCBI, Universidad de Málaga, Málaga, E-29590, Spain

[8] Institute for Mediterranean and Subtropical Horticulture "La Mayora" (IHSM-UMA-CSIC), Málaga, E-29010, Spain

*Corresponding author: Maria.Mayol@uab.es

Abstract

Unraveling the evolutionary processes underlying range expansions is fundamental to understand the distribution of organisms, as well as to predict their future responses to environmental change. Predictions for range expansions include a loss of genetic diversity and an accumulation of deleterious alleles along the expansion axis, which can decrease fitness at the range-front (expansion load). In plants, empirical studies supporting expansion load are scarce, and its effects remain to be tested outside a few model species. *Leontodon longirostris* is a colonizing Asteraceae with a widespread distribution in the Western Mediterranean, providing a particularly interesting system to gain insight into the factors that can enhance or mitigate expansion load. In this study, we produced a first genome draft for the species, covering 418 Mbp (~53% of the genome). Although incomplete, this draft was suitable to design a targeted sequencing of ~1.5 Mbp in 238 *L. longirostris* plants from 21 populations distributed along putative colonization routes in the Iberian Peninsula. Inferred demographic history supports a range expansion from southern Iberia around 40,000 yrs BP, reaching northern Iberia around 25,000 yrs BP. The expansion was accompanied by a loss of genetic diversity and a significant increase in the proportion of putatively deleterious mutations. However, levels of expansion load in *L. longirostris* were smaller than those found in other plant species, which can be explained, at least partially, by its high dispersal ability, the self-incompatible mating system, and the fact that the expansion occurred along a strong environmental cline.

Keywords

Colonizing species; demographic history; range expansions; expansion load; *Leontodon longirostris*; single nucleotide polymorphism.

INTRODUCTION

Population expansions and range shifts are common processes in response to environmental change, often involving complex interactions between demographic and genetic processes. Thus, they have been extensively studied in ecology and evolution (see Chuang & Peterson, 2016 for a review). Theoretical works have demonstrated that the demographic processes taking place during range expansions translate into particular footprints at the genomic level (Excoffier, Foll, & Petit, 2009). On the one hand, the strong and consecutive bottlenecks create gradients of decreasing diversity along the expanding range (Austerlitz, Jung-Muller, Godelle, & Gouyon, 1997). On the other hand, the low population densities and strong genetic drift at the expansion front can promote "surfing" of new and standing variants that quickly fix and spread over large regions (Edmonds, Lillie, & Cavalli-Sforza, 2004; Klopfstein, Currat, & Excoffier, 2006). This surfing effect can result in an accumulation of deleterious mutations due to inefficient selection, leading to a steady decrease in mean fitness at the expansion front, the so-called "expansion load" (Peischl, Dupanloup, Kirkpatrick, & Excoffier, 2013).

Over the past decades, many phylogeographic studies in plants and animals from the Northern Hemisphere have supported some of these predictions. For example, an overall decline in genetic diversity is often found along postglacial colonization routes (Hewitt, 2000, 2004). Direct experimental evidence for the accumulation of deleterious mutations following range expansions is, however, very limited (e.g. Bosshard et al., 2017 in bacteria), and indirect empirical evidence mainly comes from a number of studies of the out-of-Africa expansion of human populations (e.g. Fu, Gittelman, Bamshad, & Akey, 2014; Henn et al., 2016; Peischl, Dupanloup, Bosshard & Excoffier, 2016). In plants, studies supporting expansion load are restricted to species with substantial genomic

resources such as the model tree *Populus trichocarpa* (Zhang, Zhou, Bawa, Suren, & Holliday, 2016), the annual herb *Mercurialis annua* (González-Martínez, Ridout, & Pannell, 2017), and three Arabidopsis satellite systems (all Brassicaceae), i.e. *Arabis alpina* (Laenen et al., 2018), *Eutrema salsugineum* (Wang et al., 2018) and *Arabidopsis lyrata* (Willi, Fracassetti, Zoller, & Van Buskirk, 2018). Because both low genetic diversity and a high frequency of deleterious alleles can constrain the ability of populations to adapt to novel environmental conditions, it is thus relevant in the context of current climate change to investigate the prevalence of genetic surfing and expansion load in other plants with distinct life-history traits.

Theoretical models have shown that deleterious alleles can persist in range-front populations for thousands of generations after the end of the expansion (Peischl et al., 2013). Nevertheless, the impact and magnitude of the expansion load depends on a number of factors. In general, expansion load increases under conditions that favor genetic surfing, such as small population size, high growth and low migration rate from neighbor populations (Klopfstein et al., 2006; Peischl et al., 2013). Conversely, mechanisms favoring increased migration to range-front populations throughout the course of range expansion, such as the evolution of higher dispersal rates at the leading edge ("spatial sorting"; Shine, Brown, & Phillips, 2011), could reduce genetic drift and "rescue" populations from incurring expansion load (Peischl & Gilbert, 2020; Tomasini & Peischl, 2020). In the same way, when range expansions occur along environmental gradients, the expansion is slowed by the need for colonizing populations to adapt to the novel local environment. This allows more time for migrants to reach the range-front, restoring genetic diversity, and for natural selection to reduce the frequency of deleterious alleles (Gilbert et al., 2017). Then, species experiencing expansions along heterogeneous habitats and with

high dispersal capacity may prevent and/or mitigate the accumulation of deleterious mutations in range-front populations.

Species colonizing disturbed and newly human-made habitats are ideal systems to gain insight into the factors that can enhance or mitigate expansion load. Colonizing plants can spread quickly into ecosystems where they have not been before, so they are expected to experience frequent demographic and spatial expansions. Some attributes of colonizing plants, such as founder events involving small numbers of individuals and/or high growth rates, are expected to favor gene surfing and the accumulation of deleterious alleles. By contrast, colonizing plants display high dispersal capabilities that can mitigate the accumulation of deleterious mutations in range fronts by reshuffling genetic diversity from neighboring populations. Moreover, most colonizing species form temporary seed banks allowing them to survive during unfavorable seasons. By increasing the effective size of populations, seed banks might also play a key role to mitigate genetic drift and load following range expansions. Finally, colonizing species usually expand their range along environmental gradients, which may require quick adaptive responses (e.g., Montague, Barrett, & Eckert, 2008; Colautti & Barrett, 2013). The need to adapt to the novel environmental conditions found in the expanding front can slow the expansion rate and, consequently, the accumulation of deleterious mutations (Gilbert et al., 2017).

Leontodon longirostris (Finch & P.D. Sell) Talavera (≡ *Leontodon saxatilis* subsp. *rothii* Maire ≡ *Thrincia hispida* Roth) (Asteraceae, Cichorieae) is a common colonizing species in the Western Mediterranean Basin that grows in abandoned agricultural fields, therophytic grasslands, roadsides, and other disturbed spaces on a variety of soils (Talavera, Talavera, & Sánchez, 2015). It is also widely naturalized in other regions with Mediterranean climate such as Chile, parts of the United States and southern Australia, where is considered to be invasive (CGP; Groves et al., 2003). *Leontodon longirostris*

usually behaves as an annual plant, and it is a self-incompatible outcrossing species pollinated by generalist insects (Ruiz de Clavijo, 2001). As some other Asteraceae, it produces dispersible and non-dispersible diaspores, a mixed strategy traditionally interpreted as useful for increasing survival in highly unpredictable habitats (Venable & Lawton, 1980). Central achenes bear well-developed dispersal structures (i.e., a pappus) and exhibit limited dormancy, while peripheral achenes lack a pappus and remain in a dormant state on the ground, forming a temporary seed bank (Ruiz de Clavijo, 2001). This enables the spreading of offspring in space and time, since the central achenes are easily dispersed into new habitats, while the peripheral ones can persist in the established populations until the arrival of favorable conditions for germination.

The species is widely distributed throughout the Iberian Peninsula, a territory characterized by a marked latitudinal gradient of temperature and precipitation. It has one of the smallest genomes within the Asteraceae (Vallès et al., 2013), and a high-quality reference transcriptome assembly is already available for the sister species *L. saxatilis* Lam. (≡ *Thrincia saxatilis* (Lam.) Holub & Moravec, ≡ *Leontodon taraxacoides* (Vill.) Mérat, nom. illeg.) (Hodgins et al., 2014), which greatly facilitates population genomic studies. Then, it provides a particularly good system to gain insight into the factors that can enhance or mitigate expansion load, and thus contribute to the current debate on the potential for species' ranges to shift in response to climate change. With this goal in mind, we generated a genome draft assembly and annotation for *L. longirostris* and used it to design target capture probes to address the following questions: 1) What is the specific demographic history of the species in the Iberian Peninsula? 2) If range expansions took place, can we still detect the predicted loss in genetic diversity and increase in genetic load along the expansion axis in this widespread colonizing plant? And 3) How the

accumulation of deleterious mutations in front-range populations compares to what is observed in plant species with other life-history traits?

MATERIAL AND METHODS

Sampling for DNA extraction

Leaves from 7 to 20 plants (mean = 12) of *L. longirostris* were collected at 21 localities following three south-to-north transects across the Iberian Peninsula (western, central and eastern, Fig. 1, Table S1), resulting in a total sample size of 248 individuals. An additional sample from northeastern Iberia was used to generate the genome draft. We also sampled 20 individuals of the close relative *L. saxatilis* from two localities in northern Iberia to be used as outgroup (Fig. 1). High quality genomic DNA was isolated from 50 to 100 mg of dry leaf material using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following standard protocols.

Genome draft and annotation

Short-read data are provided on BioProject PRJNA648858 and correspond to 1/2 lane of Illumina HiSeq 2000 (2 x 100 bp) reads from one library with 300- to 500-bp insert size, performed at GATC Biotech, Konstanz, Germany. Raw reads were preprocessed and filtered using SEQTRIMNEXT V2.0.67, a next-generation sequencing-evolved version of SEQTRIM (Falgueras et al., 2010), with default parameters to remove adapters, low-quality base calling, PCR duplicates, short or empty insert sequences, and contaminants (including microorganisms, organelles and plasmids). Useful, pre-processed reads were then connected, when overlapping, using *k*-mer frequencies to conform a long read using COPE V1.1.3 (Connecting Overlapped Pair-End reads; Liu et al., 2012) since average insert size

was smaller than the sum of the two read length. Reads were then assembled using three different assemblers, RAY v2.3.1 (Boisvert, Raymond, Godzaridis, Laviolette, & Corbeil, 2012), SOAPDENOVO2 v2.40 (Luo et al., 2012) and VELVET v1.2.10 (Zerbino & Birney, 2008), with *k*-mers fixed at 31, 43, 57, 71, and a complete range combination from 24 to 71. Scaffolds were obtained from each assembling procedure and submitted to a gap-filling step with GAPCLOSER v1.12 (Li et al., 2010), reusing the useful reads, to provide the final set of contigs and scaffolds with as few indeterminations and gaps as possible. Useful reads were then mapped to the resulting contigs and scaffolds using BOWTIE V2.2.9 as a measure of quality and representativity. The assembly produced by SOAPDENOVO2 using the *k*-mer range 24-71 produced a slightly smaller number of contigs and scaffolds than other assemblies (410,019 contigs + scaffolds), but also longest contigs and scaffolds and a highest mapping rate, and therefore it was chosen as the final genome draft for subsequent analyses (Claros et al., 2020; see below).

Gene prediction in the final set of contigs and scaffolds was performed depending on their length. Scaffolds ≥ 10 kbp were annotated using MAKER v2.31.6 (Cantarel et al., 2008) trained with the available transcripts of *L. saxatilis* produced by Hodgins et al. (2014) and the full-length plant proteins found in the UniProtKB database. Gene models were then converted in coding sequences and their tentatively coded protein annotated with FULL-LENGTHERNEXT v0.9.8 (P. Seoane, N. Fernández-Pozo & M.G. Claros, in preparation; available at http://www.scbi.uma.es/ingebiol/commands/full_lengther_next). The remaining contigs and scaffolds, whose length cannot presumably contain a complete gene model, were loaded into our GENEASSEMBLER v0.1.0 pipeline (available at https://rubygems.org/gems/gene_assembler; Seoane-Zonjic et al., 2016) with the aim of joining contigs and/or scaffolds containing sequences from the same gene to generate and annotate fragmented gene models based on the full-length proteins of *Helianthus annuus*

found at the UniProtKB database. These "chimaeric" gene models were also annotated with the UniProtKB proteins used as model. Sequences of the 410,019 scaffolds and contigs conforming the *L. longirostris* genome draft, their MAKER-based gene models of the longest (>10 kb) fragments, their functional annotation, and the sequence of the "chimaeric", GENEASSEMBLER-based gene models reconstructed from short (<10 kb) genome draft fragments are publicly available with DOI: 10.6084/m9.figshare.12706247.v3.

Probe design, library construction, and targeted sequencing

We used the annotated draft genome to select ~1.59 Mbp for targeted sequencing in 248 individuals of *L. longirostris* and 20 of *L. saxatilis*. The detailed selection procedure was the following. We first focused on the 536 gene models with a functional ortholog in the protein databases that were located on the longest scaffolds (lengths ≥ 10 kbp; see Results), where more accurate gene predictions and functional annotations were expected. We refined this dataset discarding those scaffolds containing duplicated annotations and putative gene models with uncharacterized functions. For the remaining gene models, we assessed their biological functions by using the information available at UniProt, and prioritized those involved in processes that can result in adaptive variation (e.g., resistance to cold and drought, resistance to pathogens, phenological traits such as germination and flowering, etc.) over those related to biological functions common to distinct organisms (e.g., DNA replication). This resulted in a dataset of 93 scaffolds containing 249 distinct gene models. We repeated the same procedure with the 472 annotated gene models located on scaffolds with shorter lengths (< 10 kbp; see Results), and retained 50 additional gene models, so the final dataset to be investigated in the whole population sample consisted of 299 gene models located on 143 contigs or scaffolds with a total length of ~1.59 Mbp.

Approximately one third of the 1.59 Mbp selected corresponded to genic (coding and non-coding) regions (0.58 Mbp), while the remaining two thirds were intergenic (upstream and downstream) regions (1.01 Mbp).

The selected contigs and scaffolds were sequenced in the 268 individuals sampled (including outgroups) by a targeted sequence capture approach using a custom SeqCap EZ design (Roche NimbleGen Inc, Basel, Switzerland) followed by next-generation sequencing of captured regions on an Illumina platform. Probes were designed by Roche Tech-Support in Madison (WI, USA) starting from the selected target sequences. Library preparation and targeted sequencing were outsourced to IGA Technology Services (Udine, Italy). Libraries were constructed by using KAPA DNA Library Preparation Kit (Roche, NimbleGen Inc.) following the manufacturer's protocol and enrichment performed using NimbleGen solution-based SeqCap EZ probe libraries kit. Cluster generation, template hybridization, isothermal amplification, linearization, blocking and denaturization, and hybridization of the sequencing primers were then performed on Illumina cBot and flow cell HiSeq SBS V4 50 cycle kit, loaded on HiSeq 2500 Illumina sequencer producing 50 bp single reads. The CASAVA version of the Illumina Pipeline 1.8.2 was used for base calling and demultiplexing. Adapters were masked using CUTADAPT v3.0 (Martin, 2011). Masked and low-quality bases were filtered using ERNE-FILTER v2.1.1 (Del Fabbro, Scalabrin, Morgante, & Giorgi, 2013).

The Genome Analysis Toolkit GATK was used for SNP calling, following GATK best practices for SNP filtering and the following filtering expression: "MQ0> = 4 && ((MQ0 / (1.0 * DP)) > 0.1) || DP < 10.0 || Q < 50.0 || QD <1.5 || FS > 60". SNP data were further filtered using VCFTools v0.1.15 (Danecek et al., 2011) and vcffilter (vcflib C++ library) for mean coverage across all samples between 20 and 250, a maximum level of missing data per SNP of 10% and a maximum level of missing information per individual

of 25%. Ten individuals of *L. longirostris* were discarded with this filter, leaving a final sample size of 238 individuals from 21 populations (Table 1). The dataset was additionally filtered removing those SNPs with a significant heterozygote excess in ≥ 3 populations to minimize the impact of putative paralogous loci on data analysis. Finally, only biallelic SNPs were retained, leaving a final dataset of 168,733 polymorphic SNPs distributed along intergenic regions (63%), exons (13%) and non-coding sections of genes (24%; Table S2).

Genetic diversity and population structure

We applied two different approaches to infer population genetic structure in *L. longirostris* at the regional level using the full SNP dataset (168,733 SNPs). First, the existence of discrete clusters was explored using the Bayesian clustering method implemented in FASTSTRUCTURE v1.0.4 (Raj, Stephens, & Pritchard, 2014). Three independent runs for each K were performed, from $K = 1$ (no structure) to $K = 10$, and averaged Q values (i.e., the individual assignment probability to each of the K groups) were used to draw pie charts using QGis v2.14.0-Essen (Quantum GIS Development Team, 2016). Second, given that model-based clustering methods tend to overestimate the number of discrete clusters when genetic variation is continuously distributed across the landscape (e.g., under isolation by distance; Meirmans, 2012), a principal component analysis (PCA) was performed using PLINK v2.00a (Chang, Chow, Tellier, Vattikuti, Purcell, & Lee, 2015) with default parameters.

For subsequent analyses (i.e., population genetic diversity and differentiation, neutrality tests and demographic inference), we refined the SNP dataset and used only those SNPs for which the state of the allele (ancestral vs. derived) could be inferred by comparison with *L. saxatilis* (116,946 SNPs). Nucleotide diversity θ_π (Tajima, 1983) and Tajima's D neutrality test (Tajima, 1989) were computed using MSTATSPOP

(https://bioinformatics.cragenomica.es/numgenomics/people/sebas/software/software.html) on concatenated sequences, both for each population and considering all populations as a whole. In addition, the efficacy of selection evaluated as the ratio of non-synonymous to synonymous nucleotide diversity ($\theta_{\pi N}/\theta_{\pi S}$) was computed based on the 14,381 SNPs from coding regions. The percentage of heterozygous sites was calculated at the population level with VCFTools v0.1.15 (Danecek et al., 2011). Since departures from random mating could be indicative of restricted migration and/or changes in the mating system, two processes that are expected to decrease genetic diversity and increase expansion load, average inbreeding coefficients within (F_{IS}) and between (F_{ST}) populations were obtained using ARLEQUIN v3.5.2.2 (Excoffier & Lischer, 2010) and their significance was evaluated with 10,100 permutations. Isolation-by-distance (IBD) was evaluated according to Rousset (1997) by testing the correlation between the matrix of pairwise [$F_{ST}/(1- F_{ST})$] and the matrix of geographical distances (logarithmic scale) using the Mantel test implemented in ARLEQUIN v3.5.2.2, with 10,100 permutations. Finally, the relationship between genetic diversity with latitude and longitude was tested using Pearson correlations in R v3.4.4 (R Core Team, 2018).

Demographic history

We used two complementary approaches based on the unfolded site frequency spectrum (SFS) to infer the most likely demography of *L. longirostris*. On the one hand, the model-flexible approach implemented in STAIRWAY PLOT v2 (Liu & Fu, 2015) was applied to infer the demographic history of each population separately. This method is not restricted to a specific demographic model, so it can infer significantly more detailed demographic history than model-constrained methods, being more suitable for demographic analysis where no previous knowledge is available. We assumed a per-generation mutation rate of

6.5×10^{-8} per base pair, that was the mutation rate inferred considering a demographically stable population ("standard model") in FASTSIMCOAL2 (see below), and a generation time of one year (as the species usually behaves as an annual herb). Median estimates of the effective population size (N_e) and confidence intervals were estimated with the built-in bootstrap function using 200 subsets of the input data.

However, the method implemented in STAIRWAY PLOT does not allow testing the goodness-of-fit of the expected SFS with the observed SFS. Then, we conducted a second complementary approach to fit specific demographic models by maximum likelihood to the unfolded SFS using FASTSIMCOAL2 v2.6.0.3 (Excoffier, Dupanloup, Huerta-Sánchez, Sousa, & Foll, 2013). Since similar and simple demographic histories were inferred with STAIRWAY PLOT for all populations, four basic demographic models of population change were compared (Fig. 2). In the first three, the parameters were restricted to fit particular demographic scenarios: (1) a "standard model" for a demographically stable population, defined by a single parameter, the current population size; (2) a "bottleneck + expansion model", characterized by a reduction in population size followed by a population growth; and (3) an "expansion + bottleneck model", with a period of population growth followed by a decrease in population effective size. A fourth 3-epoch model (4) with no assumptions about the past demographic events (i.e., no parameter restriction) was also included. This model consisted of three effective population sizes allowed to change at two different times in the past (Fig. 2). As in the STAIRWAY PLOT approximation, the mutation rate was inferred from the "standard model" (6.5×10^{-8}) and the generation time was set to one year.

Each model was run 100 replicated times, with 100,000 coalescent simulations for the calculation of the composite likelihood and 20 expectation-conditional maximization (ECM) cycles (see Table S3 for search ranges). The run with the maximum likelihood in

each scenario was retained and used to obtain point estimates of the different demographic parameters. The best runs in each model were then compared using AIC and the relative likelihood (Akaike's weight of evidence) to select the most probable demographic scenario, as in Excoffier et al. (2013). Confidence intervals (95%) were obtained for each parameter of the best scenario using parametric bootstrap replicates. One hundred SFS were simulated from the run with the maximum composite likelihood, and then parameters were re-estimated performing 20 runs per simulated SFS. The runs with the highest maximum likelihood per simulated SFS were then used to define the 95% CI values (Excoffier et al., 2013).

Finally, to infer the possible origin and direction of a range expansion, we applied the method of Peter & Slatkin (2013) to detect asymmetries in the 2D site frequency spectrum between pairs of populations. The directionality index (ψ) was calculated using the R package rangeExpansion v.0.0.0.9000 (freely available at https://github.com/BenjaminPeter/rangeexpansion).

Expansion load

To evaluate the potential effect of range expansions on genetic load, the 116,946 SNPs with known allelic state (ancestral vs. derived) were classified into four categories according to the predicted effect of the variant change, using SNPEFF v4.3t (Cingolani et al., 2012) and the genome draft as a reference: HIGH (e.g. loss of start and stop codons, 299 SNPs), MODERATE (e.g. non-synonymous mutations, 7,562 SNPs), LOW (e.g. synonymous mutations, 7,642 SNPs) and MODIFIER (e.g. non-coding variants or variants affecting non-coding parts of genes, 101,443 SNPs). We then retained the SNPs included in the first three categories (15,503 SNPs) for the estimation of genetic load, discarding the SNPs located in non-coding regions where the impact of the variant change is difficult to

predict. About 96% of these SNPs were located on the largest scaffolds (≥ 10 kbp), where predictions of gene models were expected to be more accurate.

We used several proxies to evaluate the impact of the expansion on the spread and fixation of deleterious alleles. First, we counted the total number of fixed derived mutations at the population level for each SNPEFF category (HIGH, MODERATE and LOW). Second, we computed the additive and recessive genetic load at the individual level and tested for differences between three different groups of populations. The additive genetic load, defined as the number of potentially deleterious derived alleles per individual, was obtained by counting the derived mutations classified as HIGH and MODERATE in homozygous (x2) and heterozygous (x1) state. The recessive genetic load was obtained as the number of potentially deleterious derived alleles (i.e. HIGH and MODERATE) in homozygous state. Both measures were scaled by derived mutations with LOW effect (which are likely to be neutral or just slightly deleterious) in order to evaluate differences in the efficacy of purifying selection at individual level.

The groups of populations were defined using the information obtained in the demographic analyses, which suggested that *L. longirostris* progressively expanded its range from south to north in the Iberian Peninsula (see Results for further details). To determine whether this range expansion has resulted in a higher genetic load due to relaxed selection in range-front populations (i.e. expansion load), we compared the additive and recessive genetic load among three groups of populations according to their geographical position and the starting time of the demographic expansion inferred with STAIRWAY PLOT (see Results). The first group (*core*) included those populations from southern latitudes that have been estimated to be present for more than 35,000 yrs BP (population's ID: 10, 20, 23, 25, 31, 35, 45, 55). The second (*intermediate*) was composed by populations at intermediate latitudes that were estimated to colonize its current range during the past

29,000-33,000 yrs (population's ID: 04, 05, 60, 63, 68, 83). Finally, the third group (*front*) comprised those populations that reached northern latitudes not before 28,000 yrs ago (population's ID: 01, 27, 67, 69, 72, 74, 78). Averaged values of recessive and additive genetic load were computed for each group and significant differences among them were evaluated using Mann-Whitney non-parametric tests. In addition, to distinguish the effect of range expansion on more ancient (pre-existing) variation vs. new (more recent) mutations, we compared the averaged additive and recessive genetic load obtained for sites with derived mutations shared between *core*, *intermediate* and *front* groups of populations (3,910 SNPs) with those obtained for private sites within each group (5,015 SNPs in the *core*, 1,489 SNPs in the *intermediate* and 1,565 SNPs in the *front*).

RESULTS

Leontodon longirostris genome draft

From the original 186,552,004 paired-end reads, 13,701,592 were discarded, mainly due to their chloroplastic origin. Most remaining reads were still paired-end (91.13% or 170,006,980 reads), indicating that they were suitable for genome assembling. After 20 assembling approaches using different assemblers, *k*-mers and consolidation strategies, SOAPDENOVO2 using the *k*-mer range 24-71 presented the best compromise since it had a high mapping rate (82.14%) and produced a reasonable amount of scaffolds and contigs (Table 2). It provided an assembly length of 418 Mb covering ~53% of the genome (total size: 0.78 Gb, Vallès et al., 2013). Low N50 (1,532 bp) indicated a fragmented genome draft, which was an expected result due to the absence of long paired-end reads among the input read for assembling. Gene predictions were obtained with different bioinformatic approaches for the 1,007 longest scaffolds and the remaining "short" fragments of the

399 genome draft (see Material and Methods). From long scaffolds, 853 gene models were
400 inferred suing 536 different proteins; 202 of the models contained the complete ORF;
401 additionally, other 169 gene predictions coded for an unknown protein (Table 2).
402 Regarding the small fragments, further 472 "chimaeric" gene models were reconstructed
403 (Table 2), only five using the same protein than those from the longest scaffolds (P45739,
404 Q8LSM7, E3SU13, F8R6K3 and P85200). Therefore, at least 1,003 different genes were
405 identified in the *L. longirostris* genome draft (DOI: 10.6084/m9.figshare.12706247.v3)
406 based on known proteins, a number that can be extended to 1,172 when the 169 putative
407 genes without ortholog are also considered. A brief consultation in PANTHER (DOI:
408 10.1038/s41596-019-0128-8), where only 471 orthologs were recognized, indicated that
409 the most important groups of functions corresponded to "catalytic activity" (GO:0003824)
410 and "metabolic process" (GO:0008152), including at least 113 and 103 of the predicted
411 genes, respectively.
412
413 ***Genetic diversity and population structure***
414 Based on the targeted sequencing data, all the partitions in FASTSTRUCTURE (K = 2-10)
415 supported a clear separation between southern and northern populations, with populations
416 of admixed composition at the intersection of both groups (Fig. S1). However, no discrete
417 groups were found in the PCA, where the distribution of individual genotypes along the
418 first axis (PCA1; 25.3% of the variation) mostly reflected a south-to-north latitudinal cline
419 (Fig. 3). A longitudinal separation of northernmost populations was also evident in the
420 PCA2 axis (10.6% of the variation). The third and forth axes (9.1% and 8.8% of the
421 variation, respectively) revealed the separation of some particular populations (e.g., 35, 78)
422 or groups of geographically close populations (e.g., 23, 45, 83; Fig. S2).

At the population level, genetic diversity estimates based on the refined SNP dataset with known allelic state (116,946 SNPs) revealed a significant negative correlation with latitude, both for the number of heterozygous sites (r = -0.79, $p < 0.001$; Fig. 4A) and the overall nucleotide diversity (r = -0.92, $p < 0.001$; Fig. 4B). The number of polymorphic loci, as well as the nucleotide diversity present at synonymous and non-synonymous sites, were also reduced in northern (*front*) populations as compared to the southern (*core*) ones (Table 1). The ratio of non-synonymous to synonymous nucleotide diversity ($\theta_{\pi N}/\theta_{\pi S}$), however, did not show any geographic trend, with mean values being very similar for all groups of populations (*core*: 0.222 ± 0.012; *intermediate*: 0.220 ± 0.019; *front*: 0.228 ± 0.012; Table 1). Only two populations departed from Hardy-Weinberg equilibrium, supporting random mating within populations (F_{IS} range: -0.072 − 0.115; Table 1). Among populations, we found low to moderate pairwise F_{ST}, but most pairs were significantly greater than zero ($p < 0.001$; Table S4). Pairwise [$F_{ST}/(1- F_{ST})$] increased with the logarithm of the geographical distance, supporting an isolation-by-distance pattern (r = 0.56, $p < 0.001$; Fig. S3).

Demographic history

STAIRWAY PLOT modeling revealed that all populations were characterized by demographic expansions, as expected in a colonizing plant. Demographic expansions were also suggested by the negative Tajima's *D* obtained in all populations (Table 1). Nevertheless, the onset of the expansion varied among populations following a latitudinal gradient (r = -0.90, $p < 0.001$; Fig. 1), suggesting a northward spatial expansion of the species across the Iberian Peninsula. The oldest expansions were inferred for southernmost populations and took place around 40,000 yrs BP, reaching mid and high latitudes of Iberia around 30,000 and 25,000 yrs BP, respectively (Fig. 1). The directionality index ψ also

448 pointed to a range expansion towards the north, with the most likely origin located in
449 southern Iberian Peninsula (population 35, $p < 0.001$, Fig. S4). After a phase of growth
450 following colonization, most populations maintained large and constant effective sizes
451 until recently (N_e : 50,000-100,000), although some of them experienced a demographic
452 bottleneck around 20,000 yrs ago (05, 25, 35, 83). During the last 2,000 yrs, however, a
453 strong demographic decline was evident in most populations (15 out of 21), reaching in
454 some cases current effective sizes as small as 5,000 (Fig. 1). In contrast, three populations
455 (23, 27, 67) seem to have experienced a recent exponential growth resulting in much
456 higher current N_e (>150,000).

457 Statistical comparisons of the four demographic models tested in FASTSIMCOAL2
458 strongly supported the 3-epoch model with no parameter restriction (Table S5). In each
459 population the inferred parameters were generally in accordance with the events inferred
460 with STAIRWAY PLOT (Tables 3 and S5). The best model for 13 populations (01, 05, 10, 20,
461 23, 25, 27, 31, 35, 45, 63, 67, 83) consisted of two consecutive demographic expansions,
462 while the remaining eight (04, 55, 60, 68, 69, 72, 74, 78) showed a demographic expansion
463 followed by a decrease in population size occurring recently (~50-2,700 yrs BP), as
464 already detected with STAIRWAY PLOT. In some cases (e.g., populations 05, 20, 31, 35),
465 however, FASTSIMCOAL2 failed to detect the recent demographic decline that was evident
466 with STAIRWAY PLOT, probably because of the simplicity of the models (restricted to 3-
467 epochs). Regarding time estimates, more recent demographic events were inferred with
468 FASTSIMCOAL2 compared to STAIRWAY PLOT, but the oldest expansions were still detected
469 in the *core* (~25,000-32,000 yrs BP), being progressively more recent for *intermediate*
470 (~22,000-26,000 yrs BP) and *front* (~17,000-22,000 yrs BP) populations.

471

472 ***Expansion load***

The number of derived mutations that were fixed at the population level was extremely low in the three categories considered (i.e., HIGH, MODERATE and LOW). Specifically, we found three derived mutations with MODERATE effect that were fixed in population 35 (*core*), and another three with LOW effect, each one fixed in populations 01 (*front*), 04 (*intermediate*) and 23 (*core*), respectively.

The comparisons among groups of populations revealed that the average number of potentially deleterious mutations (i.e., with HIGH and/or MODERATE effect) carried by each individual decreased along the expansion route (Table 4), both for the sites shared between the *core*, *intermediate* and *front* populations (presumably pre-existing variation), and for those that were private (presumably more recent mutations). Nevertheless, once scaled by neutral variation (i.e., variants with LOW effect), additive and recessive genetic load were invariably higher in populations from the *front* as compared to *core* and *intermediate*, suggesting that purifying selection is relaxed in the expansion front (Table 4). For shared mutations, the values of additive genetic load increased from 0.52 ± 0.06 and 0.53 ± 0.07 in *core* and *intermediate* populations, respectively, to 0.60 ± 0.09 in those of the *front* (Table 4). A similar pattern was observed for shared mutations found in the homozygous state (recessive genetic load), with significantly higher values in the group of *front* populations (0.55 ± 0.37) compared to the rest (0.39 ± 0.15 and 0.37 ± 0.16 for *core* and *intermediate*, respectively). Additive genetic load was also higher for mutations only found in the *front* (1.63 ± 0.79) compared to those only found in the *intermediate* (1.35 ± 0.63) and the *core* (1.16 ± 0.31) populations. Recessive genetic load for private sites was not calculated due to the small number of mutations. It must be noted, however, that standard deviations of genetic load were particularly high in the *front* (Table 4), indicating a non-equal contribution of all individuals to the observed values.

DISCUSSION

Despite the vast amount of theory dealing with the genetic consequences of range expansions, empirical studies in plants are scarce and most are restricted to model species, limiting generalization. Here, we extended these studies to *L. longirostris*, a non-model Asteraceae that is widely distributed through the Iberian Peninsula. We produced the first genome draft for the species, which provided enough genomic information for subsequently studying nucleotide diversity, demographic history and expansion load via targeted sequencing. We found a northward expansion of the species, which was accompanied by a loss of genetic diversity and an increase in the proportion of putatively deleterious mutations (expansion load), as predicted by theory. Remarkably, signatures of range expansions were still noticeable after thousands of generations. Nevertheless, expansion load did not extend to all the individuals at the range-front, and deleterious mutations were not fixed at the population or the group level, suggesting that several factors could have mitigated the genomic signatures of range expansions in this species, as discussed below.

A first genome draft for Leontodon longirostris

The genome draft for *L. longirostris* obtained in the present study (DOI: 10.6084/m9.figshare.12706247.v3), even if incomplete (a maximum of 1,172 genes were identified), low-covering (only ~53% of the genome was assembled) and fragmented (low N50, Table 2), was reliable (read mapping rate of 82.14%) and contained enough and suitable genomic information for further applications. The targeted sequencing approach stemming from this genome draft enabled us to infer the demographic history and expansion load of the species in the Iberian Peninsula, but it could also support population

genetic studies oriented to identify candidate genes underlying ecologically important traits in *L. longirostris*, as well as for comparative genomic studies with other *Asteraceae*. Moreover, this draft genome is a base reference that can be further improved, for instance, by the addition of long paired-end reads.

Demographic history of Leontodon longirostris in the Iberian Peninsula

The history *L. longirostris* in the Iberian Peninsula was dominated by both demographic and spatial expansions. Demographical models suggested a progressive expansion of the species' range from south to north, pointing to a possible colonization from northern Africa. The proximity of the Iberian and African land plates, and the sea level drop occurred during the last glacial period, have facilitated intermittent plant migrations across the Strait of Gibraltar, being more frequent in the case of short-lived herbs and pioneer shrubs (Rodríguez-Sánchez, Pérez-Barrales, Ojeda, Vargas, & Arroyo, 2008; Lavergne, Hampe, & Arroyo, 2013). Within the Asteraceae, for instance, *Hypochaeris* sect. *Hypochaeris* and *Helminthotheca*, closely related to *Leontodon*, originated in western North Africa and then expanded through the Strait of Gibraltar to the Iberian Peninsula during the Pleistocene (Ortiz, Tremetsberger, Stuessy, Terrab, García-Castaño, & Talavera, 2009; Tremetsberger et al., 2016). Similarly, a North African origin with subsequent migration into the Iberian Peninsula has been postulated for a relict lineage of the model plant *Arabidopsis thaliana* (e.g., Toledo, Marcer, Méndez-Vigo, Alonso-Blanco, & Picó, 2020). Future studies including North African samples of *L. longirostris* could clarify whether the origin of the Iberian samples is located or not in this area.

Contrary to the prevailing pattern of multiple refugia that is commonly found for many plants in Iberia ("refugia-within-refugia" model, Gómez & Lunt, 2007), we did not find divergent lineages within the species. The separation of southern and northern

548 populations inferred by FASTSTRUCTURE is not supported by any other evidence and is
549 probably due to the poor performance of model-based clustering methods in the presence
550 of continuous patterns of genetic differentiation (e.g., "isolation-by-distance"; Frantz,
551 Cellina, Krier, Schley, & Burke, 2009). Instead, after the initial colonization of southern
552 Iberia, the species seems to have expanded progressively northward for several millennia.
553 During the Late Pleistocene (125,000-11,500 yrs BP), steppes and grasslands were
554 frequent in the Iberian landscapes (González-Sampériz et al., 2010), as well as large
555 herbivore fauna (Álvarez-Lao & Méndez, 2016). These conditions could have facilitated
556 the spread of *L. longirostris*, allowing at the same time substantial gene flow between
557 different sectors along the expansion axis, precluding the formation of genetically distinct
558 groups. High rates of genetic exchange seem to have occurred over long periods of time, as
559 supported by the low differentiation among neighboring populations and the large effective
560 population size (N_e > 50,000 individuals) inferred for most populations, with only few of
561 them showing signals of demographic bottlenecks matching the Last Glacial Maximum
562 (around 20,000 yrs BP). Moreover, the gradual increase of forest biomes occurred with
563 deglaciation (14,000 yrs BP) in many parts of Europe (Binney et al., 2017) did not have a
564 significant impact on the size of the populations, which have remained stable until very
565 recent times. During the last two millennia, however, a sharp demographic decline was
566 inferred in some populations (Fig. 1, Table 3), suggesting a rapid decrease of suitable
567 habitats in much of the species distribution, presumably associated to the agricultural
568 expansion and increasing urbanization.
569
570 ***Genetic consequences of range expansion***
571 A first consequence of the northward range expansion in *L. longirostris* was a remarkable
572 loss of genetic diversity. Despite the relatively small area covered by the expansion

(~1,000 km from south to north) and the intrinsic characteristics of the species (e.g., high dispersal ability, outcrossing mating system), the reduction in the number of polymorphic sites (Table 1) was comparable to that found in other short-lived plants at much wider geographical scales (e.g., *Mercurialis annua*, González-Martínez et al., 2017). Moreover, while populations of *M. annua* from the expanded ranges exhibited slightly lower values of nucleotide diversity for synonymous and non-synonymous sites than those of the core, the decrease was substantial in *L. longirostris* (Table 1). These results could be unexpected given the long time elapsed since the expansion (more than 20,000 generations) and support the idea that the genetic effects of range expansions can be maintained over thousands of generations (Peischl et al., 2013), even in the presence of frequent gene flow. Interestingly, though dispersal seems to have a significant role to attenuate the negative effects of gene surfing on the fixation of deleterious alleles (see below), it has not been enough to fully restore the loss of genetic diversity resulting from founder events occurred during the colonization process.

A second consequence of the northward range expansion in *L. longirostris* was an accumulation of genetic load (i.e. expansion load), as reflected in a higher proportion of putatively deleterious to non-deleterious mutations in *front* than *core* populations. This was true for both private and shared mutations, pointing to processes that took place during the range expansion itself (and excluding other demographic processes such as recent bottlenecks). The observed accumulation of deleterious relative to non-deleterious mutations in the range front is consistent with the prediction of relaxed purifying selection in the expanding edge (e.g., Peischl et al., 2013). This is not supported, however, by the ratio of non-synonymous to synonymous mutations ($\theta_{\pi N}/\theta_{\pi S}$), i.e. the efficacy of selection, which was found to be similar at different parts of the species' range (Table 1). Interestingly, the discordance between these two quantities has been reported both in plants

598 (*Mercurialis annua*; González-Martínez et al., 2017) and humans (reviewed in Lohmueller,
599 2014), having been attributed to the sensitivity of $\theta_{\pi N}/\theta_{\pi S}$ to non-equilibrium conditions
600 (e.g., Gravel, 2016). In our case, such a discrepancy could also reflect the inclusion of
601 splice sites located at the exon-intron boundaries in the calculation of genetic load.
602 Although splice sites are crucial for proper splicing and some of them are under strong
603 selective constraints, others are less conserved between species and evolve under weak
604 selection, resulting in a substantial fraction of suboptimal nucleotides (genetic load) at
605 these specific positions (Denisov et al., 2014).

606 Although significant, the levels of expansion load in *L. longirostris* were smaller
607 than those found in other plant species. For instance, an increase of derived deleterious
608 variants that are fixed in range-front populations have been reported for *Mercurialis annua*
609 (González-Martínez et al., 2017) and *Arabis alpina* (Laenen et al., 2018). In contrast, the
610 number of potentially deleterious mutations fixed in populations of *L. longirostris* was
611 virtually non-existent, both in the *core* and the *front*. These differences are probably
612 associated to the intensity of the bottlenecks experienced during the expansion, which were
613 severe in the case of *M. annua* and *A. alpina*, and almost absent in *L. longirostris*.

614 There are several not exclusive factors that could have mitigated the severity of
615 bottlenecks during range expansions in *L. longirostris*. First, the northward expansion has
616 been a slow process, taking around 15,000 years, and occurred along a marked
617 environmental gradient. The current climate difference between the temperate northern
618 portion of the Iberian Peninsula and the drier central and southern parts has existed since
619 the Middle Pleistocene (González-Sampériz et al., 2010), or even earlier (Jiménez-Moreno,
620 Fauquette, & Suc, 2010), so it is likely that individuals at the colonization front have been
621 forced to adapt to novel conditions during the colonization process, slowing the rate of
622 expansion. Recently, Gilbert et al. (2017) showed that when range expansions are slowed

623 by the need to locally adapt, the severity of genetic drift is reduced, since slow expansions
624 provide more opportunities for high-fitness alleles to reach the colonization front through
625 migration from the core, restoring genetic diversity and increasing the efficacy of selection.
626 A slow colonization also facilitates gene flow among different sectors of the expansion
627 axis; although we have shown that gene flow was not enough to restore genetic diversity in
628 *front* populations, it may still have mitigated the impact of expansion load, preventing
629 demographic collapse (Peischl & Excoffier, 2015; Peischl et al., 2016).

630 Second, experimental studies in both southern (Ruíz de Clavijo, 2001) and northern
631 populations (García, 2004) of the species have shown that, despite its ability to self-
632 fertilize, seed set is dramatically reduced after selfing, suggesting that strong self-
633 incompatibility mechanisms have been maintained along the expansion axis. Thus, it is
634 unlikely that populations in the colonization front were founded by a small number of
635 migrants, as compatible crosses require distinct S-alleles (Allee effect). Allee effects have
636 been found to increase effective population size at the front of expanding populations,
637 lowering the role of genetic drift and gene surfing (Hallatschek & Nelson, 2008). In fact,
638 theoretical models support that the maintenance of self-incompatibility in colonizing
639 species is strongly linked to the evolution of high dispersal rates that compensate for their
640 incapacity of founding new populations from single or few individuals (Pannell & Barrett,
641 1998; Dornier, Munoz, & Cheptou, 2008). Consequently, much of the genetic diversity
642 initially lost by founder effects in colonization fronts could be restored, helping selection to
643 lower the severity of expansion load. In this sense, a recent study in *Arabidopsis lyrata* ssp.
644 *petraea* showed that, despite a sharp decline in effective population size during a
645 postglacial colonization, the allelic diversity at the self-incompatibility locus has remained
646 similar in *core* and *front* populations, suggesting that high migration rates have been
647 promoted to avoid mate limitation (Takou et al., 2019). Interestingly, the authors also

found that the number of deleterious mutations in core and front populations has remained unchanged, suggesting that gene flow could also have buffered the effect of founder events on genetic load.

Third, *L. longirostris* is able to establish temporary seed banks, which can buffer genetic diversity losses. While the central achenes possess a well-developed pappus facilitating dispersal, peripheral ones exhibit low dispersal ability but have a thick pericarp that delay germination, so they can remain temporarily in the dormant state on the ground (Ruíz de Clavijo, 2001). Beyond optimizing reproductive success in highly dynamic or disturbed habitats (Cohen, 1966), seeds stored in the soil could have contributed to minimize the loss of genetic diversity and the fixation of putatively deleterious mutations in *front* populations by increasing their effective population size and buffering some genotypes from local extinctions.

Expansion load and fitness

The impact of potentially deleterious mutations on individual performance and long-term persistence of range-front populations is difficult to predict. However, the fact that they have persisted for thousands of generations (i.e. > 20,000 yrs) suggests that they could have relatively small fitness costs. Alternatively, the accumulation of deleterious mutations observed in *front* populations could be an indirect effect of adaptation. Geographic clines in life-history traits commonly evolve when expansions occur across environmental gradients (e.g., Colautti & Barrett, 2013), and the steep environmental cline from South to North Iberia may have provided opportunities for adaptation during the species expansion. Paradoxically, strong positive selection may counteract purifying selection in neighboring genomic regions and thus maintain deleterious variants at higher frequency than expected from their detrimental fitness effect (Hartfield & Otto, 2011). The hitchhiking of

673 deleterious alleles along with positively selected variants has been reported in a variety of
674 organisms, including humans (Schrider & Kern, 2017), dogs (Marsden et al., 2016) and
675 plants (Zhang et al., 2016), among others. The potential interaction between the
676 accumulation of potentially deleterious mutations and local adaptation in *L. longirostris*
677 deserves further exploration through common garden or reciprocal transplants, as well as
678 the analysis of selective footprints at the molecular level.
679
680 **ACKNOWLEDGEMENTS**
681
682 We kindly thank C. García-Barriga and I. Regalado for their support in sample collection
683 and laboratory assistance. This work was supported by POREXPAN project (CGL2014-
684 53120-P) from the Spanish Ministry of Economy and Competitiveness (MCIU) and the
685 Catalan Government "Consolidated Research Group 2017SGR1006". Funding from grants
686 RTA2017-00054-C03-03, TIN2017-88728-C2-1-R and AGL2017-83370-C3-3-R are also
687 acknowledged. M. De Pedro was supported by a Ph.D. grant (BES-2015-074251) from the
688 Spanish Ministry of Economy and Competitiveness. This work would not have been
689 possible without the computer resources and the technical support provided by the
690 Plataforma Andaluza de Bioinformática of the University of Málaga.
691

REFERENCES

Álvarez-Lao, D. J., & Méndez, M. (2016). Latitudinal gradients and indicator species in ungulate paleoassemblages during the MIS 3 in W Europe. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 449, 455-462.

Austerlitz, F., Jung-Muller, B., Godelle, B., & Gouyon, P. H. (1997). Evolution of coalescence times, genetic diversity and structure during colonization. *Theoretical Population Biology*, 51, 148-164.

Binney, H., Edwards, M., Macias-Fauria, M., Lozhkin, A., Anderson, P., Kaplan, J. O., … Zernitskaya, V. (2017). Vegetation of Eurasia from the last glacial maximum to present: Key biogeographic patterns. *Quaternary Science Reviews*, 157, 80-97.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology*, 13, R122.

Bosshard, L., Dupanloup, I., Tenaillon, O., Bruggmann, R., Ackermann, M., Peischl, S., & Excoffier, L. (2017). Accumulation of deleterious mutations during bacterial range expansions. *Genetics*, 207, 669-684.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., & Yandell. M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18, 188-196.

CGP: The Composite Genome Project (2020, July 8). *Priorities for research, education and extension in genomics, genetics, and breeding of the Compositae* [White paper]. Retrieved from The Composite Genome Project, University of California, Davis: https://compgenomics.ucdavis.edu/cwp/white_paper_draft_09_12_07.pdf

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4, 7.

Chuang, A., & Peterson, C. R. (2016). Expanding population edges: theories, traits, and trade-offs. *Global Change Biology*, 22, 494-512.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Lu, X. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3. *Fly*, 6, 80-92.

Claros, M. G., Seoane, P., Bautista, R., González-Martínez, S., Riba, M., Mayol, M., & De Pedro, M. (2020). *Leontodon longirostris genome draft 418Mb*. figshare. Dataset. https://doi.org/10.6084/m9.figshare.12706247.v3

Cohen, D. (1966). Optimizing reproduction in a randomly varying environment. *Journal of Theoretical Biology*, 12, 119-129.

741 Colautti, R. I., & Barrett, S. C. H. (2013). Rapid adaptation to climate facilitates range
742 expansion of an invasive plant. *Science*, 342, 364-366.
743
744 Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … 1000
745 Genomes Project Analysis Group (2011). The variant call format and VCFtools.
746 *Bioinformatics*, 27, 2156-2158.
747
748 Del Fabbro, C., Scalabrin, S., Morgante, M., & Giorgi, F. M. (2013). An extensive
749 evaluation of read trimming effects on Illumina NGS data analysis. *PLoS ONE*, 8,
750 e85024.
751
752 Denisov, S. V., Bazykin, G. A., Sutormin, R., Favorov, A. V., Mironov, A. A., Gelfand, M.
753 S., & Kondrashov, A. S. (2014). Weak negative and positive selection and the drift load
754 at splice sites. *Genome Biology and Evolution*, 6, 1437-1447.
755
756 Dornier, A., Munoz, F., & Cheptou, P.-O. (2008). Allee effect and self-fertilization in
757 hermaphrodites: reproductive assurance in a structured metapopulation. *Evolution*, 62,
758 2558-2569.
759
760 Edmonds, C. A., Lillie, A. S., & Cavalli-Sforza, L. L. (2004). Mutations arising in the
761 wave front of an expanding population. *Proceedings of the National Academy of*
762 *Sciences of the United States of America*, 101, 975-979.
763
764 Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust
765 demographic inference from genomic and SNP data. *PLoS Genetics*, 9, e1003905.
766
767 Excoffier, L., Foll, M., & Petit, R. J. (2009). Genetic consequences of range expansions.
768 *Annual Review of Ecology, Evolution, and Systematics*, 40, 481-501.
769
770 Excoffier, L., & Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to
771 perform population genetics analyses under Linux and Windows. *Molecular Ecology*
772 *Resources*. 10, 564-567.
773
774 Falgueras, J., Lara, A. J., Fernández-Pozo, N., Cantón, F. R., Pérez-Trabado, G. & Claros,
775 M. G. (2010). SeqTrim: a high-throughput pipeline forpre-processing any type of
776 sequence read. *BMC Bioinformatics*, 11, 38.
777
778 Frantz, A. C., Cellina, S., Krier, A., Schley, L., & Burke, T. (2009). Using spatial Bayesian
779 methods to determine the genetic structure of a continuously distributed population:
780 clusters or isolation by distance? *Journal of Applied Ecology*, 46, 493-505.
781
782 Fu, W., Gittelman, R. M., Bamshad, M. J., & Akey, J.M. (2014). Characteristics of neutral
783 and deleterious protein-coding variation among individuals and populations. *The*
784 *American Journal of Human Genetics*, 95, 421-436.
785
786 García, H. (2004). *Consecuencias ecológicas de la sucesión sobre algunas características*
787 *poblacionales e individuales de Leontodon taraxacoides (Vill.) Mérat* (Unpublished
788 Master's Thesis). Autonomous University of Barcelona, Barcelona, Spain.
789

Gilbert, K. J., Sharp, N. P., Angert, A. L., Conte, G. L., Draghi, J. A., Guillaume, F., … Whitlock, M. C. (2017). Local adaptation interacts with expansion load during range expansion: maladaptation reduces expansion load. *The American Naturalist*, 189, 368-380.

Gómez, A., & Lunt, D. H. (2007). Refugia within refugia: patterns of phylogeographic concordance in the Iberian Peninsula. In: Weiss, S. & Ferrand, N. (eds) *Phylogeography of Southern European Refugia*, pp. 155-188. Springer, Dordrecht, NL.

González-Martínez, S. C., Ridout, K., & Pannell, J. R. (2017). Range expansion compromises adaptive evolution in an outcrossing plant. *Current Biology*, 27, 2544-2551.

González-Sampériz, P., Leroy, S. A. G., Carrión, J. S., Fernández, S., García-Antón, M., Gil-García, M. J., … Figueiral, I. (2010). Steppes, savannahs, forests and phytodiversity reservoirs during the Pleistocene in the Iberian Peninsula. *Review of Palaeobotany and Palynology*, 162, 427-457.

Gravel, S. (2016). When is selection effective? *Genetics*, 203, 451-462.

Groves, R.H. (Convenor), Hosking, J. R.; Batianoff, G. N.; Cooke, D. A.; Cowie, I. D., Johnson, R. W., … Waterhouse, B.M. (2003). *Weed Categories for Natural and Agricultural Ecosystem Management*. Department of Agriculture, Fisheries and Forestry, Australian Government.

Hallatschek, O., & Nelson, D. R. (2008). Gene surfing in expanding populations. *Theoretical Population Biology*, 73, 158-170.

Hartfield, M., & Otto, S. P. (2011). Recombination and hitchhiking of deleterious alleles. *Evolution*, 65, 2421-2434.

Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., … Bustamante, C. D. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 440-449.

Hewitt, G. M. (2000). The genetic legacy of the Quaternary ice ages. *Nature*, 405, 907-913.

Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society of London B*, 359, 183-195.

Hodgins, K. A., Lai, Z., Oliveira, L. O., Still, D. W., Scascitelli, M., Barker, M. S., … Rieseberg, L. H. (2014). Genomics of Compositae crops: reference transcriptome assemblies and evidence of hybridization with wild relatives. *Molecular Ecology Resources*, 14, 166-77.

Jiménez-Moreno, G., Fauquette, S., & Suc, J. P. (2010). Miocene to Pliocene vegetation reconstruction and climate estimates in the Iberian Peninsula from pollen data. *Review of Palaeobotany and Palynology*, 162, 403-415.

Klopfstein, S., Currat, M., & Excoffier, L. (2006). The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, 23, 482-490.

Laenen, B., Tedder, A., Nowak, M. D., Toräng, P., Wunder, J., Wötzel, S., … Slotte, T. (2018). Demography and mating system shape the genome-wide impact of purifying selection in *Arabis alpina*. *Proceedings of the National Academy of Sciences of the United States of America*, 115, 816-821.

Lavergne, S., Hampe, A., Arroyo, J. (2013). In and out of Africa: how did the Strait of Gibraltar affect plant species migration and local diversification? *Journal of Biogeography*, 40, 24-36.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., …Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, 20, 265-272.

Liu, X., & Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra. *Nature Genetics*, 47, 555-559.

Liu, B., Yuan, J., Yiu, S.-M., Li, Z., Xie, Y., Chen, Y., … Luo, R. (2012). COPE: an accuratek-mer-based pair-end reads connection toolto facilitate genome assembly. *Bioinformatics*, 28, 2870-2874.

Lohmueller, K. E. (2014). The distribution of deleterious genetic variation in human populations. *Current Opinion in Genetics & Development*, 29, 139-146.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. 1, 18.

Marsden, C. D., Ortega-Del Vecchyo, D., O'Brien, D. P., Taylor, J. F., Ramirez, O., Vilà, C., … Lohmueller, K. E. (2016). Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. *Proceedings of the National Academy of Sciences of the USA*, 113, 152-157.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17, 10-12.

Meirmans, P. G. (2012). The trouble with isolation by distance. *Molecular Ecology*, 21, 2839-2846.

Montague, J. L., Barrett, S. C. H., & Eckert, C. G. (2008). Re-establishment of clinal variation in flowering time among introduced populations of purple loosestrife (*Lythrum salicaria*, Lythraceae). *Journal of Evolutionary Biology*, 21, 234-245.

Ortiz, M. A., Tremetsberger, K., Stuessy, T. F., Terrab, A., García-Castaño, J. L., & Talavera, S. (2009). Phylogeographic patterns in *Hypochaeris* section *Hypochaeris* (Asteraceae, Lactuceae) of the western Mediterranean. *Journal of Biogeography*, 36, 1384-1397.

Pannell, J. R., & Barrett, S. C. H. (1998). Baker's Law revisited: reproductive assurance in a metapopulation. *Evolution*, 52, 657-668.

Peischl, S., Dupanloup, I., Bosshard, L., & Excoffier, L. (2016). Genetic surfing in human populations: from genes to genomes. *Current Opinion in Genetics & Development*, 41, 53-61.

Peischl, S., Dupanloup, I., Kirkpatrick, M., & Excoffier, L. (2013). On the accumulation of deleterious mutations during range expansions. *Molecular Ecology*, 22, 5972-5982.

Peischl, S. & Excoffier, L. (2015). Expansion load: recessive mutations and the role of standing genetic variation. *Molecular Ecology*, 24, 2084-2094.

Peischl, S. & Gilbert, K. J. (2020). Evolution of dispersal can rescue populations from expansion load. *The American Naturalist*, https://doi.org/10.1086/705993.

Peter, B. M., & Slatkin, M. (2013). Detecting range expansions from genetic data. *Evolution*, 67, 3274-3289.

QGIS Development Team (2016). *QGIS Geographic Information System*. Open Source Geospatial Foundation Project. http://qgis.osgeo.org

R Core Team (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197, 573-589.

Rodríguez-Sánchez, F., Pérez-Barrales, R., Ojeda, F., Vargas, P., & Arroyo, J. (2008). The Strait of Gibraltar as a melting pot for plant biodiversity. *Quaternary Science Reviews*, 27, 2100-2117.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from *F*-Statistics under isolation by distance. *Genetics*, 145, 1219-1228.

Ruiz de Clavijo, E. (2001). The role of dimorphic achenes in the biology of the annual weed *Leontodon longirrostris*. *Weed Research*, 41, 275-286.

Schrider, D. R., & Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, 34, 1863-1877.

Seoane-Zonjic, P., Cañas, R. A., Bautista, R., Gómez-Maldonado, J., Arrillaga, I., Fernández-Pozo, N., … Ávila, C. (2016). Establishing gene models from the *Pinus pinaster* genome using gene capture and BAC sequencing. *BMC Genomics*, 17, 148.

Shine, R., Brown, G. P., & Phillips, B. L. (2011). An evolutionary process that assembles phenotypes through space rather than through time. *Proceedings of the National Academy of Sciences of the USA*, 108, 5708-5711.

940 Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations.
941 *Genetics*, 105, 437-460.
942
943 Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA
944 polymorphism. *Genetics*, 123, 585-595.
945
946 Takou, M., Hämälä, T., Steige, K. A., Koch, E., Dittberner, H., Yant, L., … de Meaux, J.
947 (2019). Maintenance of adaptive dynamics in a bottlenecked range-edge population that
948 retained out-crossing. http://dx.doi.org/10.1101/709873.
949
950 Talavera, S., Talavera, M., & Sánchez, C. (2015). Los géneros *Thrincia* Roth y *Leontodon*
951 L. (Compositae, Cichorieae) en Flora Iberica. *Acta Botanica Malacitana*, 40, 344-364.
952
953 Toledo, B., Marcer, A., Méndez-Vigo, B., Alonso-Blanco, C., & Picó, F. X. (2020). An
954 ecological history of the relict genetic lineage of *Arabidopsis thaliana*. *Environmental*
955 *and Experimental Botany*, 170, 103800.
956
957 Tomasini, M., & Peischl, S. (2020). When does gene flow facilitate evolutionary rescue?.
958 *Evolution*, 74, 1640-1653.
959
960 Tremetsberger, K., Ortiz, M. A., Terrab, A., Balao, F., Casimiro-Soriguer, R., Talavera,
961 M., & Talavera, S. (2016). Phylogeography above the species level for perennial species
962 in a composite genus. *AoB PLANTS*, 8, plv142.
963
964 Vallès, J., Canela, M. A., Garcia, S., Hidalgo O., Pellicer, J., Sánchez-Jiménez, I., Siljak-
965 Yakovlev, S., Vitales, D., & Garnatje, T. (2013). Genome size variation and evolution
966 in the family Asteraceae. *Caryologia: International Journal of Cytology,*
967 *Cytosystematics and Cytogenetics*, 66, 221-235.
968
969 Venable, D. L., & Lawlor, L. (1980). Delayed germination and dispersal in desert annuals:
970 escape in space and time. Oecologia, 46, 272-282.
971
972 Wang, X.-J., Hu, Q.-J., Guo, X.-Y., Wang, K., Ru, D.-F., German, D. A., … Liu, J.-Q.
973 (2018). Demographic expansion and genetic load of the halophyte model plant *Eutrema*
974 *salsugineum*. *Molecular Ecology*, 27, 2943-2955.
975
976 Willi, Y., Fracassetti, M., Zoller, S., & Van Buskirk, J. (2018). Accumulation of
977 mutational load at the edges of a species range. *Molecular Biology and Evolution*, 35,
978 781-791.
979
980 Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly
981 using de Bruijn graphs. Genome research, 18, 821-829.
982
983 Zhang, M., Zhou, L., Bawa, R., Suren, H., & Holliday, J. A. (2016). Recombination rate
984 variation, hitchhiking, and demographic history shape deleterious load in poplar.
985 *Molecular Biology and Evolution*, 33, 2899-2910.
986

DATA ACCESSIBILITY

All raw sequencing data are available at the National Centre for Biotechnology (NCBI) under the BioProject Accession no. BioProject PRJNA648858. Contigs, scaffolds and functional annotation of the *L. longirostris* genome draft are publicly available at FigShare with DOI: 10.6084/m9.figshare.12706247.v3. The vcf file with SNP data is archived in FigShare with DOI: 10.6084/m9.figshare.12903848.v1.

AUTHOR CONTRIBUTIONS

S.C.G.M., M.M. and M.R. conceived and designed the study. M.G.C, R.B. and P.S. produced the genome draft and performed bioinformatics analyses. M.D.P., S.C.G.M. and M.M. designed the gene capture and analyzed the data. M.M. drafted the manuscript. All the authors contributed to editing and revising the manuscript.

Table 1. Sample size (*N*) and genetic diversity statistics for the populations of *Leontodon longirostris* included in this study. The number of polymorphic sites, F_{IS} and Tajima's *D* (Tajima, 1989) were based on the 116,946 SNPs with known allelic state. Nucleotide diversity θ_π (Tajima, 1983) per site × 10^{-3} was obtained for synonymous ($\theta_{\pi S}$) and non-synonymous ($\theta_{\pi N}$) sites using a subset of 14,381 SNPs found in exons. Names and geographical location of the populations are provided in Table S1. Details for *core*, *intermediate* and *front* groups of populations are provided in the main text.

Population code	***N***	**Number of polymorphic sites**	F_{IS}	**Tajima's *D***	$\theta_{\pi N}$	$\theta_{\pi S}$	$\theta_{\pi N}/\theta_{\pi S}$
10	10	28,440	-0.027	-1.511	1.64	6.85	0.239
20	20	41,957	0.015	-1.765	1.53	6.67	0.229
23	10	26,124	0.019	-1.233	1.47	6.49	0.227
25	7	22,235	-0.003	-1.444	1.42	6.17	0.230
31	14	30,422	0.069*	-1.550	1.35	6.49	0.208
35	10	27,782	0.028	-1.320	1.43	6.96	0.206
45	14	27,323	0.030	-1.297	1.30	5.77	0.225
55	10	22,635	-0.002	-1.189	1.25	6.00	0.209
Core	95	88,098	-	-2.110	1.47	6.55	0.222 ± 0.012
04	9	20,790	0.070	-1.494	1.23	5.64	0.219
05	9	23,307	0.041	-1.542	1.22	5.66	0.216
60	9	19,773	0.060	-1.288	1.28	5.28	0.242
63	9	21,371	0.044	-1.464	1.19	5.01	0.237
68	10	20,295	-0.072	-1.263	1.12	5.95	0.189
83	13	25,995	0.042	-1.414	1.24	5.67	0.219
Intermediate	59	60,400	-	-2.084	1.24	5.56	0.220 ± 0.019
01	6	14,383	0.115*	-0.921	1.30	6.29	0.207
27	14	25,194	0.041	-1.733	1.14	5.00	0.229
67	17	23,002	-0.006	-1.393	1.11	5.00	0.223
69	17	17,373	0.039	-1.293	1.02	4.10	0.247
72	10	16,402	0.054	-1.373	1.01	4.29	0.235
74	10	18,229	0.000	-1.320	1.10	4.92	0.224
78	10	17,795	0.053	-1.386	1.05	4.60	0.228
Front	84	52,330	-	-2.028	1.10	4.81	0.228 ± 0.012
Overall	238	116,946	-	-2.249	1.32	5.83	0.226

* $p < 0.01$

Table 2. Summary of the assembly features and the gene prediction for the *Leontodon longirostris* draft genome (DOI: 10.6084/m9.figshare.12706247.v3).

Draft genome features	***Values or counts***
Number of scaffolds + contigs	410,019
Scaffolds > 10 kb	1,007
Longest scaffold (bp)	53,452
N50 (bp)	1,532
Total length (Mbp)	418
Mapping rate (%)	82.14
Gene predictions in the genome draft	
In long scaffolds	853
Unique orthologs	536
Complete unique orthologs	202
Putative coding without ortholog	169
In short scaffolds + contigs	472
Common model IDs in short and long scaffolds	5
Total different gene models	1,172

Table 3. Parameters inferred with FASTSIMCOAL2 under the best supported demographic scenario in each population. Confidence intervals (95%) are within brackets. White rows indicate populations showing two consecutive demographic expansions. Grey rows indicate populations showing a demographic expansion followed by a recent decrease in population size. Times are given in number of generations, which can be translated directly to years in annual *Leontodon longirostris*, ignoring the effects of a possible seedbank. *Ne* ANC: ancestral effective population size; *Ne* INT: effective population size between ancestral and current periods; *Ne* CUR: current effective population size; T ANC: ancestral time of population size change; T REC: recent time of population size change. Each specific parameter is illustrated in Fig. 2. Details on the comparison with the remaining demographic models analyzed with FASTSIMCOAL2 are provided in Table S5.

Population code	N_e ANC	N_e INT	N_e CUR	T ANC	T REC
Core					
10	**139** (54-7,819)	**4,378** (3,670-55,916)	**50,074** (48,930-52,170)	**29,540** (28,834-33,909)	**27,431** (11,989-28,135)
20	**587** (219-861)	**24,374** (20,172-27,724)	**74,663** (71,780-77,735)	**29,893** (29,003-31,101)	**10,869** (9,280-12,480)
23	**32** (4-255)	**30,659** (27,687-32,110)	**66,684** (47,371-87,335)	**31,787** (31,215-32,362)	**3,209** (2,322-7,692)
25	**127** (9-397)	**24,433** (18,663-30,846)	**64,308** (58,205-79,076)	**28,825** (27,962-29,435)	**10,110** (6,382-13,949)
31	**63** (9-406)	**16,503** (13,433-21,930)	**54,209** (51,982-56,862)	**28,020** (26,933-28,515)	**12,764** (10,281-14,704)
35	**2,371** (1,643-2,769)	**24,699** (15,183-32,925)	**56,414** (53,308-63,466)	**27,144** (25,746-30,105)	**11,010** (6,599-15,197)
45	**40** (8-596)	**18,536** (15,898-25,140)	**37,935** (36,702-39,914)	**29,504** (27,745-29,796)	**12,507** (7,514-14,908)
55	**1,168** (986-1,365)	**42,571** (40,153-54,632)	**9,990** (2,278-24,639)	**24,862** (24,136-25,436)	**626** (91-4,588)
Intermediate					
04	**358** (258-508)	**42,506** (40,534-62,941)	**32,822** (27,539-38,983)	**21,628** (21,306-22,065)	**2,107** (808-13,782)
05	**9** (3-321)	**12,461** (8,977-21,992)	**52,506** (50,459-56,382)	**25,531** (24,369-25,639)	**15,568** (10,570-17,397)
60	**98** (24-185)	**37,711** (36,505-55,403)	**19,717** (13,746-33,705)	**22,269** (21,920-22,649)	**630** (313-11,958)
63	**21** (3-240)	**16,294** (8,782-28,457)	**41,617** (40,208-45,328)	**24,286** (23,596-24,653)	**14,803** (7,664-18,383)
68	**88** (16-186)	**38,790** (35,992-51,671)	**19,654** (10,538-24,016)	**24,383** (23,931-24,740)	**2,705** (816-7,011)
83	**6** (5-186)	**18,908** (14,040-26,503)	**40,772** (39,541-43,491)	**25,426** (24,900-25,811)	**13,070** (8,262-16,201)
Front					
01	**29** (2-270)	**12,112** (4,109-35,358)	**31,251** (29,676-39,736)	**21,183** (20,327-21,491)	**15,677** (1,693-18,783)
27	**68** (4-199)	**22,271** (17,077-25,785)	**47,718** (44,665-51,429)	**21,665** (21,162-22,066)	**7,558** (5,702-10,475)
67	**5** (2-73)	**26,198** (21,342-26,688)	**47,053** (27,690-62,429)	**21,530** (21,014-21,673)	**413** (274-11,827)

69	**411** (304-697)	**22,089** (21,618-46,247)	**8,334** (5,661-16,835)	**17,091** (16,349-17,491)	**386** (190-6,834)
72	**72** (7-178)	**29,059** (27,667-37,830)	**5,738** (3,723-22,244)	**18,974** (18,458-19,237)	**186** (109-6,351)
74	**350** (232-490)	**34,394** (33,320-42,841)	**1,384** (1,133-21,568)	**20,225** (19,738-20,490)	**51** (41- 3,628)
78	**107** (29-220)	**29,809** (28,784-57,489)	**18,529** (12,449-27,911)	**19,734** (19,337-20,068)	**471** (304-13,664)

1025

Table 4. Scaled additive and recessive genetic load (D/ND) per individual in *core*, *intermediate* and *front* populations. Mean and standard deviation (SD) values were obtained for the full number of sites with derived mutations where it was possible to predict the impact of a variant change (All; 15,503 SNPs), for sites with derived mutations shared among the three groups (Shared; 3,910 SNPs), and for sites with private derived mutations within each group (Private; 5,015 SNPs in the *core*, 1,489 SNPs in the *intermediate* and 1,565 SNPs in the *front*). Significant differences among groups ($p < 0.05$) were evaluated using Mann-Whitney non-parametric tests and are indicated with different letters. The scaled recessive genetic load for sites with derived private mutations was not calculated due to small number of mutations. Deleterious mutations (D) include variants classified within the HIGH and MODERATE categories using SnpEff, while non-deleterious mutations (ND) include those within the LOW one (see text for further details).

	Additive genetic load			**Recessive genetic load**		
Group	**Deleterious mutations (D)**	**Non-deleterious mutations (ND)**	**D/ND**	**Deleterious mutations (D)**	**Non-deleterious mutations (ND)**	**D/ND**
All						
Core	303.87 ± 43.80	508.42 ± 68.45	0.60 ± 0.06 a	47.43 ± 23.87	117.77 ± 57.77	0.42 ± 0.16 a
Intermediate	244.08 ± 49.70	407.97 ± 82.95	0.60 ± 0.07 a	31.29 ± 16.47	78.27 ± 37.29	0.40 ± 0.17 a
Front	220.80 ± 50.86	339.25 ± 81.19	0.66 ± 0.09 b	33.98 ± 18.94	66.02 ± 43.67	0.60 ± 0.34 b
Shared						
Core	203.95 ± 29.72	396.48 ± 49.91	0.52 ± 0.06 a	40.11 ± 19.42	108.74 ± 52.10	0.39 ± 0.15 a
Intermediate	184.17 ± 37.20	349.37 ± 73.12	0.53 ± 0.07 a	27.36 ± 15.07	73.86 ± 35.37	0.37 ± 0.16 a
Front	176.87 ± 41.09	299.90 ± 69.96	0.60 ± 0.09 b	29.29 ± 17.47	62.45 ± 42.20	0.55 ± 0.37 b
Private						
Core	53.20 ± 16.13	48.73 ± 18.46	1.16 ± 0.31 a	3.14 ± 4.24	2.76 ± 4.43	-
Intermediate	22.98 ± 9.87	18.51 ± 8.94	1.35 ± 0.63 a	1.15 ± 1.90	1.19 ± 2.10	-
Front	16.51 ± 7.05	11.81 ± 7.13	1.63 ± 0.79 b	1.14 ± 1.84	0.64 ± 1.53	-

FIGURE CAPTIONS

FIGURE 1. Geographic distribution of the populations sampled and inferred demographic history of *Leontodon longirostris* across the Iberian Peninsula. Estimated changes in population size were obtained separately for each population with the STAIRWAY PLOT method using 116,946 SNPs with known allelic state (see text for further details). Black lines in plots show the changes in median effective population size (*N*e) for the last 40,000 years, and shaded areas indicate the 95% confidence intervals. The bottom right figure shows the correlation between the inferred start of expansion (in thousands years before the present) and the geographical position (latitude) of each population. Details on the populations are given in Tables 1 and S1. The two populations of *Leontodon saxatilis* used as outgroup are indicated by black stars.

FIGURE 2. Schematic representation of the four alternative demographic models tested using FASTSIMCOAL2 (see Table S3 for details on the search ranges of the parameters). Note that for the model (4) there were no restrictions regarding the direction of the population size change, as it allowed for either increases or decreases in population size. *Ne* ANC: ancestral effective population size; *Ne* INT: effective population size between ancestral and current periods; *Ne* CUR: current effective population size; T ANC: ancestral time of population size change; T REC: recent time of population size change.

FIGURE 3. Principal component analysis (PCA) of the genetic variation found in 238 individuals of *Leontodon longirostris* in the Iberian Peninsula (based on 168,733 SNPs). Each individual is represented by a colored point. Note that the distribution of individuals along the PCA 1 axis mostly reflects their latitudinal position from south to north. Population details can be found in Tables 1 and S1.

FIGURE 4. Population genetic diversity (based on 116,946 SNPs with known allelic state) in relation to latitude. A) Proportion of heterozygous sites; B) Overall nucleotide diversity θ_π (Tajima, 1983). Pearson coefficients and their significance are reported.

FIGURE 1

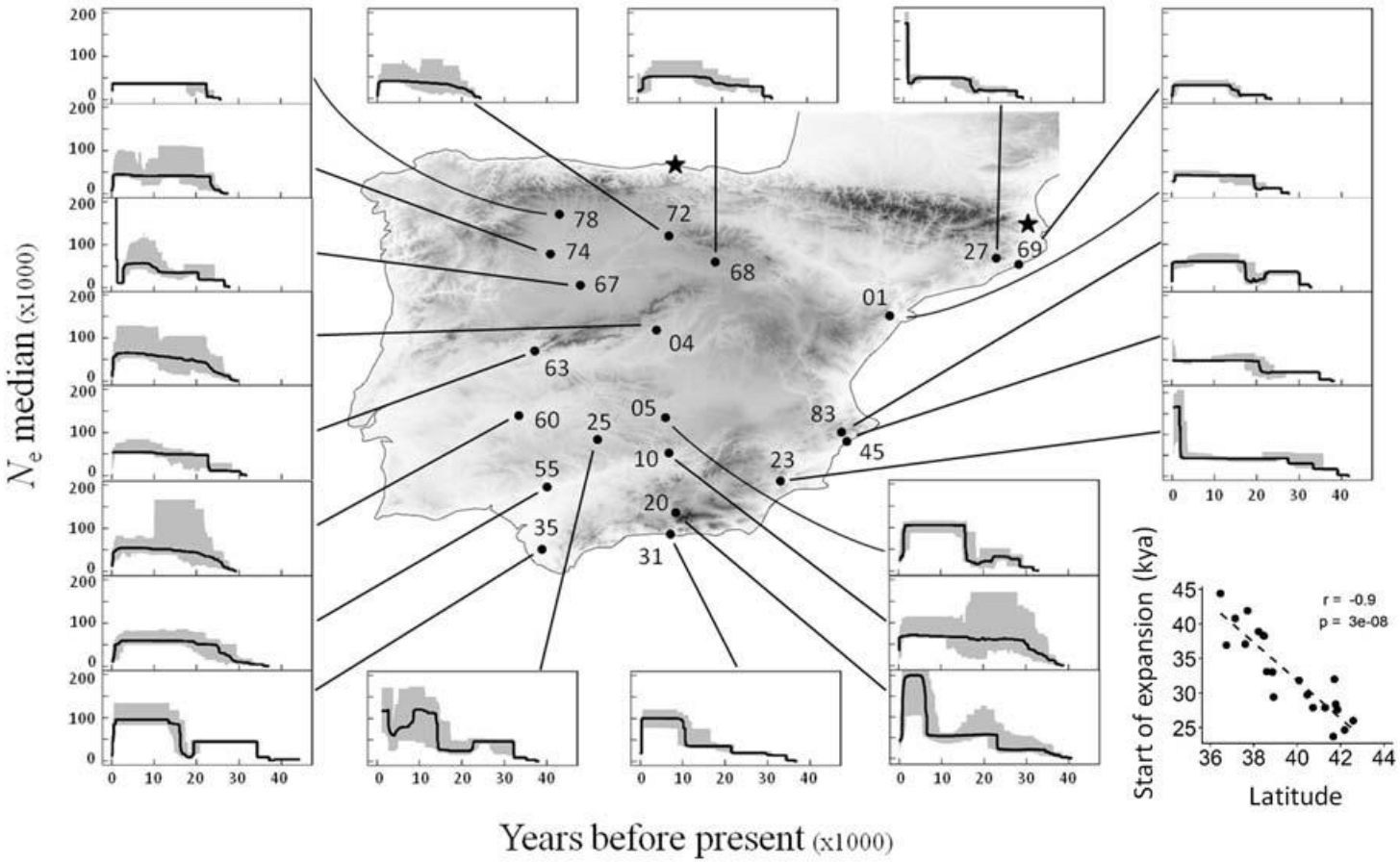
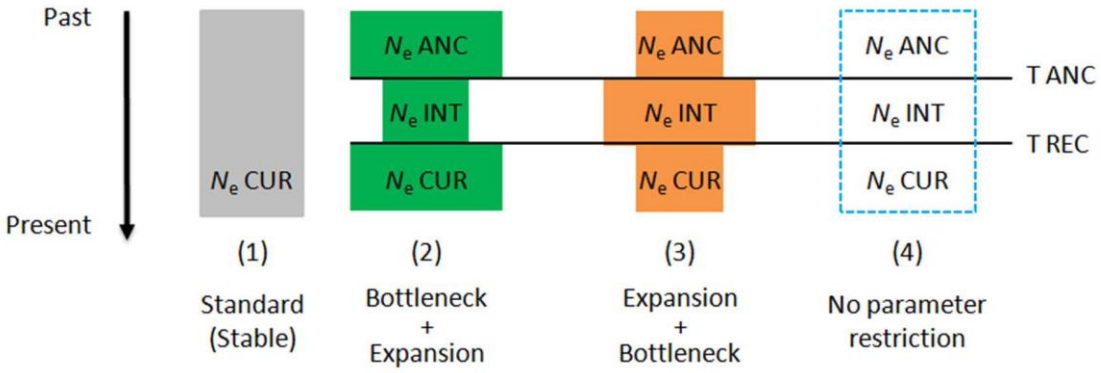
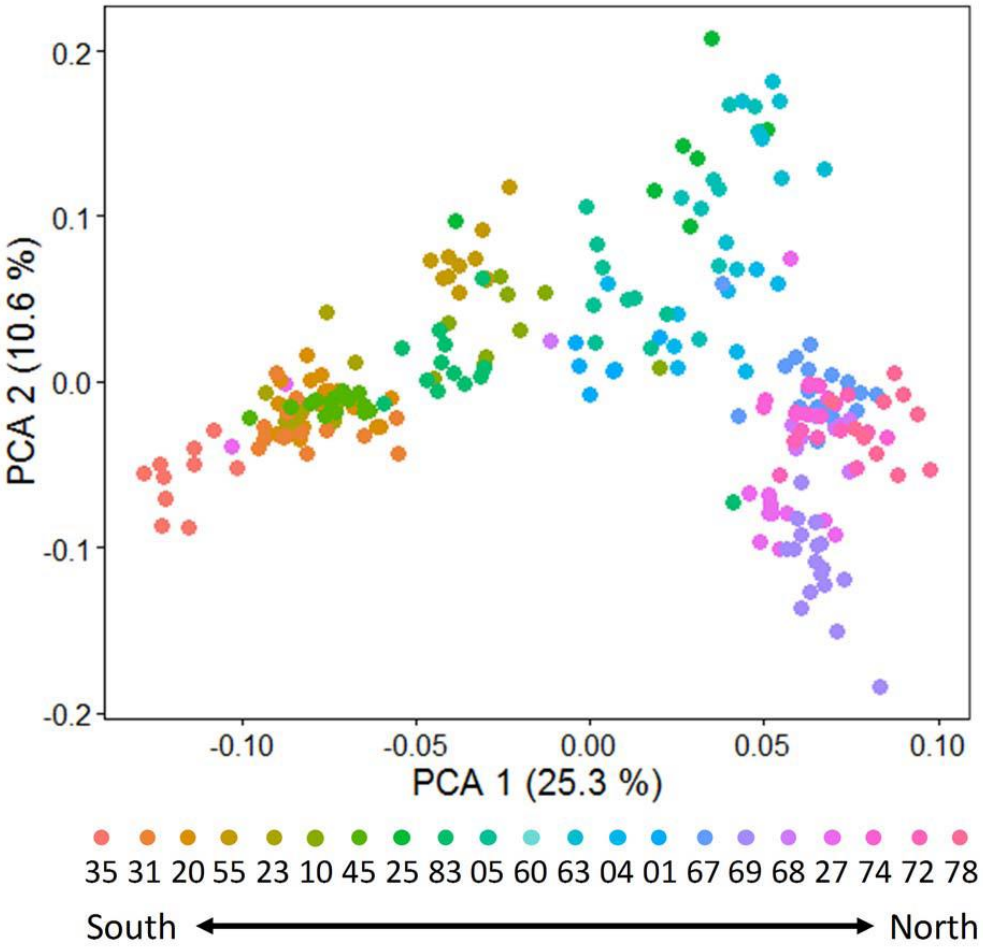

N_e median (x1000)
Years before present (x1000)
78
74
67
72
68
27
69
01
04
63
60
25
05
10
55
35
20
31
83
45
23
Start of expansion (kya)
Latitude
r = -0.9
p = 3e-08

FIGURE 2


Past
Present
N_e CUR
N_e ANC
N_e INT
N_e CUR
N_e ANC
N_e INT
N_e CUR
N_e ANC
N_e INT
N_e CUR
T ANC
T REC
(1)
Standard
(Stable)
(2)
Bottleneck
+
Expansion
(3)
Expansion
+
Bottleneck
(4)
No parameter
restriction

1077 **FIGURE 3**

1078



1079

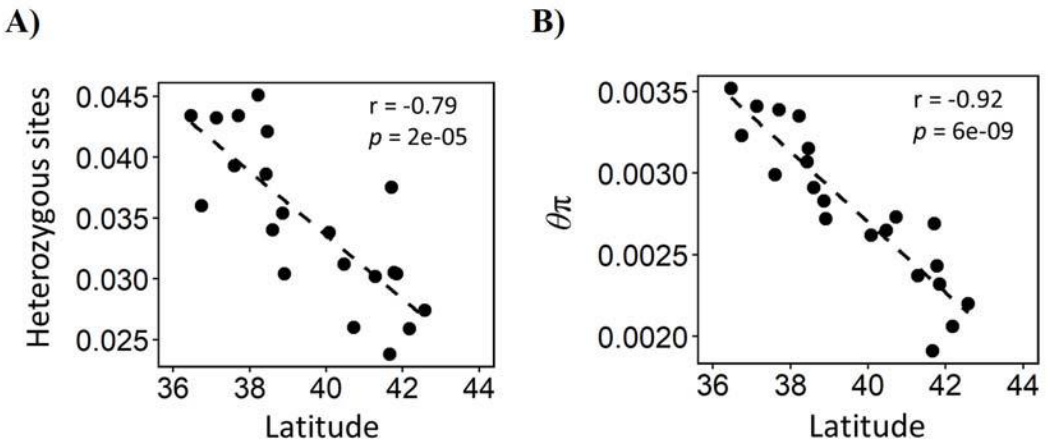
1080

1081

1082 **FIGURE 4**

1083

1084



1085

1086