Conserv Genet (2016) 17:1333–1344
DOI 10.1007/s10592-016-0865-y

RESEARCH ARTICLE

# Genomic diversity and geographical structure of the Pyrenean desman

**Marina Querejeta[1] · Jorge González-Esteban[2] · Asunción Gómez[3] · Angel Fernández-González[4] · Pere Aymerich[5] · Joaquim Gosálbez[5] · Lídia Escoda[1] · Javier Igea[1,6] · Jose Castresana[1]**

Received: 19 February 2016 / Accepted: 7 July 2016 / Published online: 11 July 2016


**Abstract** The Pyrenean desman (*Galemys pyrenaicus*) is a small semi-aquatic mammal endemic to the Iberian Peninsula. The species has recently experienced a strong decline and some of its populations are severely threatened with extinction. To help in the preservation of this species, it is critical to understand its genetic structure and main evolutionary units, as these may have specific local adaptations and could be of great conservation value. Sequencing reduced representation libraries (ddRAD) from 26 specimens selected from across its entire range, we obtained around 45,000 loci per specimen and 1185 single nucleotide polymorphisms. Heterozygosity varied substantially among individuals from different areas. Interestingly, specimens from the southeastern Pyrenees had some of the lowest proportions of heterozygous positions inferred from genome-wide data in mammals so far. In addition, we estimated a tree reflecting genomic divergence, performed a principal component analysis, and carried out a Bayesian analysis of the population structure. Combined evidence supported the existence of five distinct genomic clusters largely coincident with the main mountain ranges where the species occurs, with few specimens presenting relevant admixture levels. There was good correspondence between these populations and the mitochondrial lineages detected in a previous study, yet substantial differences in some areas demonstrate the importance of performing genomic analysis to reveal the whole population history. Although the analysis of further specimens is necessary to better characterize the distribution of the different evolutionary units, the distinctive geographical structure of this species revealed by the genomic data should be considered in future conservation plans.

**Keywords** *Galemys pyrenaicus* · ddRAD · SNPs · Conservation genomics · Genetic structure

**Electronic supplementary material** The online version of this article (doi:10.1007/s10592-016-0865-y) contains supplementary material, which is available to authorized users.

✉ Jose Castresana
jose.castresana@csic.es

[1] Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37, 08003 Barcelona, Spain

[2] Desma Estudios Ambientales S.L, Ukulu 11, 31791 Sunbilla, Navarra, Spain

[3] Tragsatec, Área de Biodiversidad, Calle Julián Camarillo 6, 28037 Madrid, Spain

[4] Biosfera Consultoría Medioambiental S.L, Calle Candamo 5, 33012 Oviedo, Spain

[5] Departament de Biologia Animal, Universitat de Barcelona, Avinguda Diagonal 643, 08028 Barcelona, Spain

[6] Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK

## Introduction

Genomic analyses have enormous potential in the conservation of threatened species (Ouborg et al. 2010; Steiner et al. 2013; Shafer et al. 2015). In particular, the study of large numbers of genetic markers allows robust conclusions to be drawn regarding population structure and gene flow, which is essential knowledge for delimiting conservation units (Crandall et al. 2000; Funk et al. 2012). The identification of different genetic clusters within species is especially important because these clusters may possess singular characteristics and adaptations to their particular

habitats (Orsini et al. 2013; Lanier et al. 2015). In addition, the detection of low genetic diversity in specific populations is crucial because it may be related to loss of adaptive variation and reduced fitness (Tallmon et al. 2004; Allendorf et al. 2010). Although mitochondrial sequences and microsatellites have long been used for these purposes, SNPs and sequences obtained through next-generation sequencing (NGS) techniques have the advantage of providing massive quantities of genetic data, thus revealing more precise details of the population structure and variability within a species and admixture levels of specimens (Morin et al. 2009; Vonholdt et al. 2011; Xue et al. 2015; Kjeldsen et al. 2015). However, genome-wide studies used to address conservation problems are still very scarce and the actual potential of these new sequencing techniques for improving the management of endangered species is still to be determined.

The Pyrenean desman (*Galemys pyrenaicus*) is a small semi-aquatic mammal endemic to the northern part of the Iberian Peninsula. It lives in clean and oxygenated rivers, a habitat generally found in mountain areas (Palmeirim and Hoffmann 1983). Some populations have experienced a strong decline over recent decades, with the most southerly populations, in the Central System, having suffered the most significant decreases (Nores et al. 2007; Gisbert and Garcia-Perea 2014). There are several factors that have been suggested to explain these population decreases such as water pollution, water extraction, habitat fragmentation caused by hydroelectric plants and reservoirs, and predation by the invasive American mink (*Neovison vison*) escaped from fur farms (Fernandes et al. 2011). The Pyrenean desman is legally protected in the four countries where it is present (Spain, Portugal, France and Andorra) and currently appears as "Vulnerable" on the IUCN Red List (Fernandes et al. 2011). Furthermore, the population of the Central System was recently catalogued by the Spanish Government as "In danger of extinction", which is the highest protection category. If reinforcement of some populations becomes necessary in the future, it is important to have sufficient information on the population structure and conservation units of the species in order to do so without altering the genetic identity of the different populations. This knowledge is essential because, even if the Pyrenean desman occupies a relatively small geographic range, its populations are not homogeneous. First, two subspecies, *pyrenaicus* and *rufulus*, were described according to differences in coloration and size (González-Esteban et al. 1999; Lopez-Fuster et al. 2006). Furthermore, a strong phylogeographic structure was revealed with mitochondrial data, as explained below (Igea et al. 2013). Finally, the distribution of the Pyrenean desman encompasses a large range of climatic and hydrological regimes, and the existence of local adaptations cannot be discarded. Therefore, outbreeding depression after admixture of divergent populations might occur (Lynch 1991; Tallmon et al. 2004; Allendorf et al. 2010).

In a previous work, Igea et al. (2013) studied the phylogeography of the Pyrenean desman using mitochondrial markers. One of the principal conclusions was the existence of a marked phylogeographic structure in which two main groups (A and B) were further divided into two subgroups. The resulting four mitochondrial lineages (A1, A2, B1 and B2) had an allopatric distribution and a likely glacial origin. The populations with largest genetic diversity were detected in the most occidental part of the distribution, suggesting that one of the most important glacial refugia was in this area. Eight nuclear introns permitted corroboration of these conclusions about genetic diversity although their variability was too low to add complementary information about population structure (Igea et al. 2013). The contact zones between the main mitochondrial groups, one located in the Cantabrian Mountains and the other one in the Iberian Range, indicated an almost complete absence of spatial mixing between these maternal lineages after postglacial recolonization (Igea et al. 2013). This is a very peculiar situation in both contact zones, where no apparent barriers to the dispersal of desmans seem to exist. Since this phylogeographic analysis was based on mitochondrial genes, of exclusive maternal inheritance, it is possible that analyzing large numbers of nuclear loci could reveal additional details of the overall genetic structure, as demonstrated with other species (Godinho et al. 2008; Nater et al. 2011).

In this work, we aimed to study the population structure of the Pyrenean desman using genomic sequences and SNPs obtained through NGS techniques. Genomic studies of threatened species such as the Pyrenean desman present several difficulties arising from the shortage of samples, the low quality of some non-invasive samples and the small quantity from each sample usually available for DNA extraction. We tried to overcome these difficulties by improving certain laboratory protocols and bioinformatic analyses. Genomic libraries were generated using a double digest restriction associated DNA (ddRAD) protocol (Peterson et al. 2012), a reduced representation library approach that allows a large number of specific genomic fragments to be sequenced, but with several modifications to be able to start with small quantities of DNA. We also used full sequence data to perform essential quality tests of the library sequences, determine the sex of individuals, and estimate the proportion of heterozygous positions. Finally, we studied the population structure of the species using SNP data. The number of samples available for this work was small and some important geographical locations could not be covered with enough sample size. However,

we were able to test these NGS techniques and uncover the global population structure of the species, which could be useful for developing conservation strategies.

## Materials and methods

### Samples

We used 26 samples of the Pyrenean desman from different geographical locations and mitochondrial clades (Table S1; Fig. S1), 22 of which came from DNA extractions performed in a previous phylogeographic study of the species (Igea et al. 2013). Most of these 22 samples came from very small amounts of fresh tissue obtained during previous survey works of the species but three of them belonged to dry tissue from different biological collections (Table S1). From the areas where a higher number of samples was available (Occident and Pyrenees), a maximum of 8 samples were used to avoid overrepresentation of these populations. From other areas, all samples from which enough DNA was available were used. Unfortunately, for a relevant area like the Central System only one DNA sample could be used. Four additional samples were obtained from specimens captured in a survey commissioned by the Regional Government of La Rioja (Spain), in 2011. These were used to complete the sampling in the Iberian Range, which included few samples from the previous study. A small piece of tail tip was taken from the captured specimens and stored in ethanol, and the animals were released back in the wild. The work to obtain samples for genetic analyses was performed with permit number A/2011/52 issued by the Regional Government of La Rioja; it was carried out following national and international regulations, and all necessary steps were taken to prevent any damage to the specimens. DNA was extracted from these samples as in Igea et al. (2013).

### Library construction and sequencing

The DNA library was constructed using the ddRAD protocol (Peterson et al. 2012), with several modifications to allow for smaller quantities of initial DNA. The experiments were performed in series of 12 samples each time, repeating samples with low sequence yield in subsequent experiments. To avoid bias that could be related to specific sequencing runs, samples were randomly selected from different geographical areas for each experiment. In order to identify each specimen in the analyses, we used 12 different P1 adapters, each with a different 5 nucleotide barcode, and one common P2 adapter. Using the single-stranded oligonucleotides and the corresponding complementary molecules provided in Peterson et al. (2012), an annealing reaction was carried out in order to produce double-stranded adapters.

The DNA concentration was measured either by absorbance at 260 nm (with Nanodrop) or using qPCR (see supplementary information). From each sample, a quantity of genomic DNA varying between 200 and 250 ng was double digested using restriction enzymes EcoRI and MspI, in a 20 µl volume. After incubating overnight, the enzymatic reaction was heat inactivated over a 30-min period at 80 °C. We then added (with no previous cleaning step in order to avoid DNA loss) a 80 µl ligation mix, including T4 DNA ligase, P1 adapter (which binds to the EcoRI overhangs), and P2 adapter (which binds to the MspI overhangs), and the solution was incubated overnight at 23 °C. The ligation reaction was heat inactivated at 65 °C over a 10-min period.

We then mixed all the ligation reactions in a single tube. This mix was concentrated in 30 µl using the MinElute PCR Purification Kit (QIAGEN). The entire 30 µl pool was run on a 1 % low-melting agarose gel for 1 h and was then selected by size (300–400 bp) using a fresh razor blade. The gel purification was eluted in 30 µl using the MinElute Gel Extraction Kit (QIAGEN).

Subsequently, a PCR amplification of the size-select sample was performed to add Illumina adapter sequences, using the oligonucleotides described in Peterson et al. (2012). Phusion high-fidelity DNA polymerase (New England Biolabs) amplifications with a number of cycles ranging from 14 to 24, depending on the intensity of initial PCR products, were carried out. A total of 4 µl of the size-select samples pool was used as template. To increase the concentration of the libraries and to minimize bias, a total of 4–7 PCRs were performed. Then, we combined and concentrated the PCR products into a 30 µl volume using the MinElute PCR Purification Kit. Finally, to get rid of small fragments, the library was a run in a precast EX 2 % agarose gel using an E-Gel system (Invitrogen). The band corresponding to the library was then extracted with the Qiaquick Gel Extraction Kit (QIAGEN) in 20 µl.

The library size was checked with a Bioanalyzer 2100 (Agilent) and sequenced using a MiSeq (Illumina) in the Genomics Core Facility at the Pompeu Fabra University. Sequencing was carried out using the MiSeq Reagent Kit v3 in a single read length of 150 cycles.

### Sequence processing

We used the Stacks 1.10 package (Catchen et al. 2011, 2013) to process the sequences obtained. First, the *process_radtags* program was used to filter out reads with low-quality sequences or malformed restriction sites or barcodes, and to separate reads belonging to different

samples according to the barcodes. After this step, sequences of the same sample from different runs were combined. Then *ustacks* was used to assemble loci in each sample, with a minimum number of sequences for each sample at a locus (minimum stack depth or m) of 3 and a maximum number of differences between reads (M) of 2. The mean sequence coverage of each specimen was then calculated as the number of assembled reads divided by the number of assembled loci. The loci of all the specimens were subsequently merged with *cstacks*, allowing for the maximum number of differences between reads (n) of 2, and a catalog of loci and sequences was then created with *sstacks*. Then the *populations* program was used to save two different output files for downstream analyses. First, the SNPs were saved from complete loci, i.e., those represented in all samples (r = 1). In addition, assembled loci represented in at least 14 samples (r = 0.51) were saved in FASTA format.

Using a series of custom scripts, different quality checks and processing steps were performed on the sequences obtained in FASTA format. First, incompatible loci (i.e., those with more than two alleles) were eliminated from the FASTA file. In addition, loci with less than 140 bp were also removed.

It has been suggested that the Stacks algorithm may allow for the presence of duplicate loci, which could cause problems for correct assembling (Sovic et al. 2015). To detect these, we performed a Blast search (Altschul et al. 1997) of all the loci against themselves, using one sequence per locus and an e-value of $1e^{-30}$. Different e-values may lead to different stringencies in the search. In this case, a stringent value was used so that only duplicate loci with a high similarity, which would be the most problematic, were detected.

Since there is no *G. pyrenaicus* reference genome available, we needed to detect possible contaminant sequences such as those coming from bacteria present in the samples. To do so, we performed an additional Blast search of one sequence per locus against the GenBank nucleotides database. In this case, a less stringent e-value of $1e^{-10}$ was used to ensure that any contamination was detected.

Sequences detected by Blast searches as duplicates or contaminants were eliminated and this set of filtered loci was used for further downstream analyses.

In addition, from this set of filtered loci, a new set of complete SNPs was constructed for comparison with the initial set saved by Stacks, using only loci previously genotyped by Stacks. For the new set, we selected from each locus the SNP with the largest minimum allele frequency.

### Sex determination

Sixteen samples were initially sexed by qPCR amplification of a Y-linked fragment (see supplementary information). Using the nine males and seven females with sex determined in this way, we searched, within the assembled loci, for putative Y chromosome loci in order to sex the rest of specimens. For this analysis, the Stacks *populations* program was run with r = 0.25 to ensure the assembly of Y-chromosome loci from males. We then selected all loci present in the seven known males and in none of the nine females, and in which the two sequences assembled by Stacks were identical. The application of these conditions expected for Y-chromosome genes resulted in the initial identification of 50 loci. One of the loci was inconsistent with the others (it was found in two specimens that did not have any of the other loci) and was consequently eliminated. The discriminating value between the sexes of the remaining 49 loci was tested using the sequenced reads (just filtered with *process_radtags*). For this purpose, we performed a Blast search of the reads of each sample against the likely Y-linked loci using very stringent criteria to avoid finding duplicates in the X chromosome ($1e^{-74}$).

### Proportion of heterozygous positions in each specimen

To ensure adequate coverage for individual heterozygosity rate estimation (Buerkle and Gompert 2013), a new dataset of assembled loci was generated with Stacks using a minimum stack depth for each sample of 9. This new set of sequences was processed as described above, and loci with duplicate and contaminant sequences were eliminated. The proportion of heterozygous positions for each sample was estimated from the sequences in FASTA format as the number of variable positions for the sample in all loci divided by the total length of the loci.

### Genomic tree

We estimated a genomic tree from the variable loci using two different phylogenetic methods, as in Igea et al. (2015). First, we reconstructed a distance tree from the matrix of the average genomic divergence between specimens (Gronau et al. 2011; Freedman et al. 2014). To summarize the divergence of the two separate alleles, we calculated pairwise distances between all specimens using formula 8.2 in Freedman et al. (2014). Distances were corrected for multiple substitutions using the Jukes-Cantor formula. The resulting matrix of pairwise distances was used to reconstruct a distance tree using the Fitch program in the Phylip package (Felsenstein 1989). Another tree was obtained from the distances calculated using formula 8.1 in Freedman et al. (2014), where a more conservative estimate of the differences at each position is computed, but the results were very similar (not shown).

In addition, we reconstructed a nuclear genomic tree by concatenating all loci. We performed several concatenations where the order of each allele pair was randomly changed. From each concatenation set, a maximum-likelihood tree was estimated using RaxML version 7.4.2 (Stamatakis 2006) with a general time reversible plus invariable sites model.

Tree figures were prepared with SplitsTree (Huson and Bryant 2006). In these figures, the tree branches were oriented to maximize the match with the geographical location of the populations without altering the tree topology.

### Principal component analysis and population structure

Principal component analysis (PCA) was performed with SNPrelate (Zheng et al. 2012) from the genetic covariance matrix of the SNP genotype data. The axes of the plot were oriented to maximize the match of the positions of the samples in the plot and their geographical locations.

Population structure and admixture were estimated from the SNP data sets with the program Structure 2.3.4, which implements a Bayesian model-based clustering method (Pritchard et al. 2000) using the admixture and correlated allele frequency models, and with no prior information on population origin. A total of 500,000 generations were run after a burn-in of 50,000 generations with a number of clusters (K) ranging from 2 to 7 ($K = 1$ was not used because it gave unrealistically low posterior probability values). We performed 10 different runs for each K value in order to plot the trend of the estimated posterior probability of the data, Ln P(D) (Pritchard et al. 2000). In addition, the optimal K value was estimated using the $\Delta K$ method (Evanno et al. 2005) as the inflection point of the Ln P(D) curve before reaching an horizontal or slightly increasing plateau. All 10 runs gave very similar results for meaningful K values and only the first run was shown in the table of admixture proportions and figures.

## Results

### Genome fragments obtained

A total of 62,551,837 sequence reads passed the initial quality filters, with an average of 2,405,840 reads for each of the 26 *G. pyrenaicus* specimens analyzed (Table S2). Different strategies were used to assemble these reads, as indicated in Fig. S2. Under a default minimum stack depth in the Stacks program ($m = 3$), 2,130,583 reads and 69,111 loci were assembled on average per sample, resulting in a mean depth of coverage of 28. When loci of different specimens were merged and only loci with at least 14 samples were selected, 57,212 loci remained, with a total of 8,326,212 bp potentially sequenced for each sample. However, the samples differed in the resulting number of assembled loci, ranging from 56,378 (99 % of the total loci) to 20,256 (35 %), with an average of 45,446 (79 %) (Table S2; Fig. S3). The average sequenced length per sample was 6,201,426 bp. Samples represented in few loci may have slightly degraded DNA but this did not prevent their use in further analyses.

The number of variable loci in this set was 16,742. The plot of the frequency of SNPs at each position along the sequence read provides a measure of the sequencing quality (Sovic et al. 2015). There is a slight trend in increasing polymorphisms towards the end of the read (Fig. S4), which could be related to the decrease in quality characteristics of Illumina reads. However, given the small proportion of positions affected, these errors are not likely to influence the results.

In order to evaluate the quality of the assembled loci, we applied further tests and filters. First, one sequence of each locus was selected to perform a Blast search against all other sequences. This way, 3563 duplicate loci were detected.

Next, a Blast search against the GenBank database was performed to estimate the amount of contaminants and other problematic sequences in our libraries. Of the 57,212 assembled loci, 6655 gave a significant hit to some sequence in GenBank. Of these, 21 belonged to non-Craniates and were considered contaminants: 16 sequences were found to belong to various bacterial species, 4 to nematodes and there was 1 unclassified sequence. There was also a hit to one mitochondrial fragment from *G. pyrenaicus*, not appropriate to be analyzed with nuclear data. All other hits corresponded to nuclear genes of Craniates (mostly mammals) and were considered endogenous because Craniate species are unlikely to be degrading organisms or parasites in these tissues (and it is expected that a few divergent sequences give a Blast hit outside mammals). The rest of sequences, giving no hit, probably belonged to non-coding DNA and other fast-evolving regions of the genome. When duplicate and other problematic loci were removed (some loci were detected in both analyses), we obtained a new data set with 53,637 loci, 15,242 of which were variable.

### Sex determination with sequenced reads

Using 49 probable Y-chromosome loci identified with a subset of specimens of known sex, we determined the sex of all samples by Blast, searching the sequenced reads against these loci. We found that the number of hits per million reads was either less than 3 or more than 326 in the

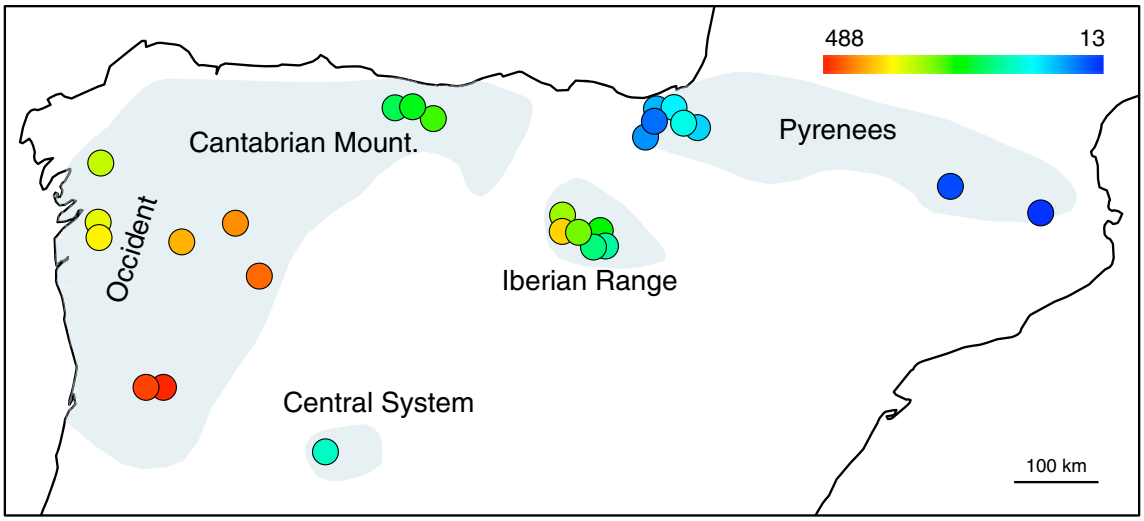


**Fig. 1** Map plotting *color*-coded heterozygosity rates in different specimens of the Pyrenean desman. The *color scale* is in number of heterozygous positions per million bases. The *shadowed* area approximately represents the current species distribution according to different sources. (Color figure online)

different samples. Therefore, these 49 loci were perfectly discriminated between the sexes and clearly Y-linked, resulting in 13 males and 13 females (Table S3).

### Proportion of heterozygous positions

To estimate the individual heterozygosity rate with loci of adequate coverage for accurate genotyping, we considered the assembled loci with a minimum stack depth for each sample of 9. This rendered 32,955 loci with at least 14 samples, totaling 4,796,424 bp of sequence length (Fig. S2). The heterozygosity rate estimated from these sequences revealed large differences between specimens, with values ranging from 0.000013 to 0.000488, that is, from 13 to 488 heterozygous positions per million bases, which is more than an order of magnitude of difference (Table S4). Furthermore, when the heterozygosity was plotted on the map (Fig. 1), it was observed that the specimens with the highest values were those from the occidental zone whereas much lower values came from the Pyrenees. In particular, the southeasternmost extent of the Pyrenees showed the lowest heterozygosity rates.

### Genomic tree

Genomic tree reconstruction was performed using two sets of variable loci: without filtering duplicates and contaminants, and with filtered loci (Fig. S2). A genomic tree was first obtained from the 16,742 unfiltered variable loci (2,436,857 bp) previously determined with a default minimum stack depth (Fig. 2). This tree clearly revealed the existence of five clades that grouped specimens from the same geographical region: Pyrenees, Occident, Cantabrian Mountains, Iberian Range and Central System. A similar phylogeny was obtained by maximum likelihood of the concatenated loci (Figure S5). The trees obtained after filtering duplicates and contaminants were also similar (not shown).

### PCA and population structure

Two sets of SNPs were also used for PCA and population structure analysis (Fig. S2). First, the output from the Stacks program rendered 1185 SNPs present in all samples. Second, after filtering duplicate and contaminant loci, a set of 1053 SNPs remained. The results were similar for the two sets and only those for the largest set are shown.

PCA based on these SNPs distinguished the same five groups as the genome tree, with the former more clearly showing the separation of the Central System population due to a large discrimination in the second PCA axis (Fig. 3).

We also used the Structure program with an admixture model. A plot of the estimated posterior probability of the data versus the number of clusters, K, shows that, after $K = 5$, this likelihood does not notably increase and does not even converge in some runs (Fig. S6). In addition, according to the Evanno method (Evanno et al. 2005), the optimal K value is 5. Figure 4 shows the map of the different specimens and the admixture proportions of each of them as determined with $K = 5$ (Table S5). The five clusters basically coincide with those obtained in the genome tree and the PCA analysis. A strong population structure is observed, with different populations occupying

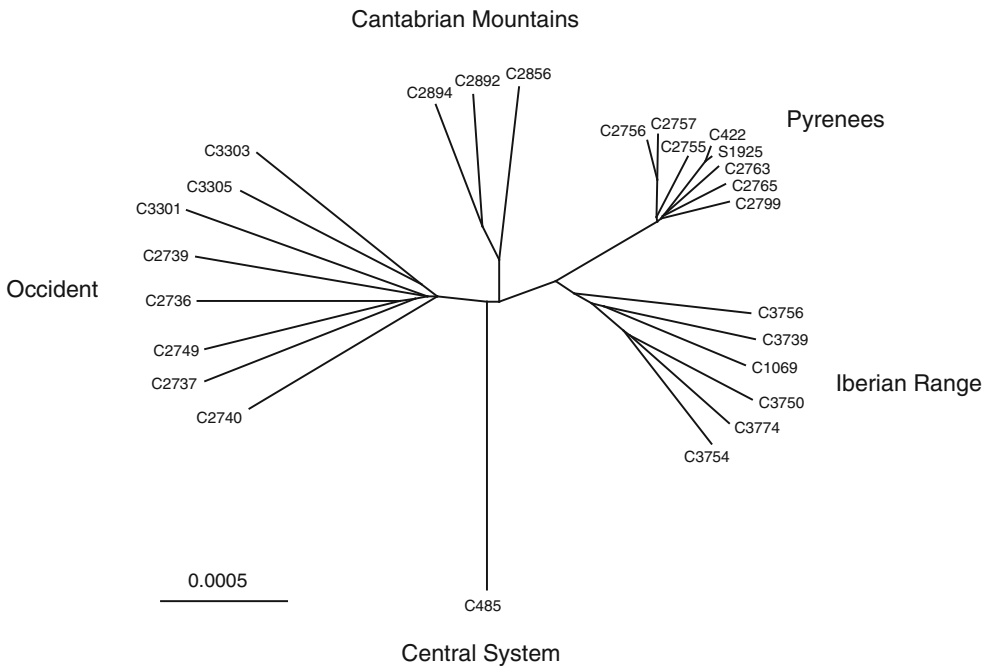


**Fig. 2** Genomic tree obtained from the genetic distances of the 16,742 variable loci (2,436,857 bp). The *scale* is in substitutions per position

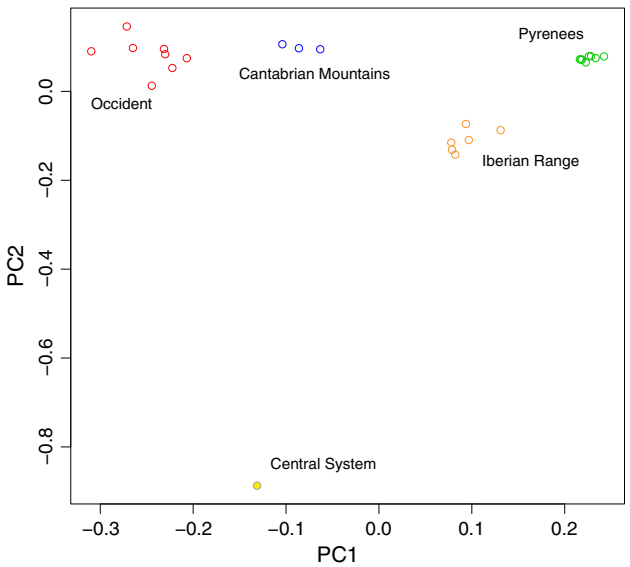


**Fig. 3** Principal component analysis (PCA) based on the SNP data

separate areas of the distribution range of the species. In addition, only four specimens showed relevant admixture levels (>5 %). Three specimens from the northwestern part of the Iberian Range had components of the Pyrenean genome and, two of them, also had elements of the Cantabrian Mountains genome. In addition, one specimen from the Cantabrian Mountains showed components from both Pyrenean and Occidental clusters (Table S5).

Since determining the optimal K value is usually difficult (Pritchard et al. 2000), we also examined different aspects of the genetic structure with K values ranging from 2 to 5 (Fig. S7). When only 2 clusters are considered, there is a basic subdivision between Pyrenean specimens (with components present in adjacent populations) and the other populations. With 3 clusters, a new population is distinguished in the Iberian Range. With 4 clusters, samples from the Cantabrian Mountains give rise to a new population. Finally, with 5 clusters, the only sample of the Central System becomes distinct.

## Discussion

### NGS libraries and quality tests performed

The quantity and quality of initial DNA material is usually a limitation when applying genetic techniques, including NGS, to endangered species. This makes developing protocols that require low amounts of starting material an important first step in these studies. By removing some cleaning steps, we were able to prepare libraries starting with as little as 200 ng of DNA, in the low range of initial material recommended in a previously developed protocol (Peterson et al. 2012). However, it should be taken into account that working with low amounts of starting DNA, and particularly if it is of low quality, may affect the representation of all loci in the final library. For this reason, we evaluated the quality of the library sequences by calculating the proportion of total loci sequenced and properly assembled for each specimen (Table S2; Fig. S3). For this estimation, we used loci present in at least 14 of the 26

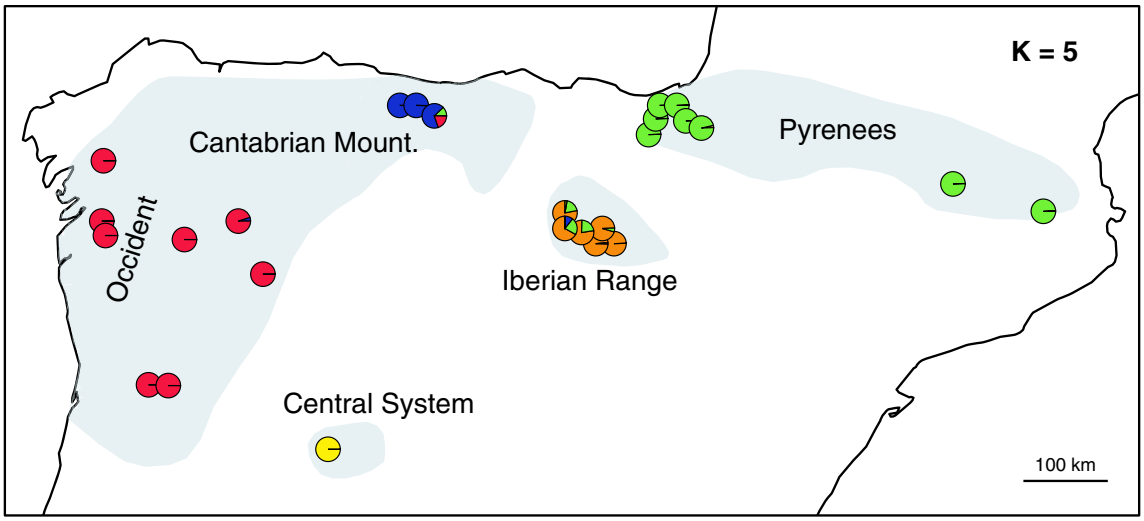


**Fig. 4** *Map* plotting *color*-coded admixture proportions of each specimen as determined with structure and K = 5. (Color figure online)

samples processed (57,212 loci). A few specimens did not even have 50 % of these loci. The two samples with lowest percentages of loci corresponded to dry skins from old collections (IBE-C2892 and IBE-C2894, prepared in 1989 and 1988, respectively) and therefore some DNA degradation may have caused this problem. Specimen IBE-C1069 belonged to an individual found dead in the field. Individual sequencing runs of this sample always gave a low number of reads, but the combination of three runs allowed the assembly of most loci. In the end, all the specimens in our final set, even those with the smallest number of assembled loci, provided reliable and coherent information in the different analyses carried out.

As with many other non-model species, no *G. pyrenaicus* genome is available for reference, making it necessary to assess additional aspects of the quality of the anonymous loci obtained. There are two main sources of potential problems with the sequences obtained: duplicate sequences and contaminants. Duplicate loci, if very similar, may present problems when assigning different alleles to the correct locus (Sovic et al. 2015). We therefore conducted a Blast search of all sequences against themselves and detected that 6 % of all loci were duplicate, although not all of these might necessarily present assembling problems. The genetic structure and variability results were similar when filtering and not filtering out these loci, indicating that duplications were not an important issue in this case. However, this test using full sequence data should still be performed when a new species is studied, or a novel protocol that selects for different loci is used.

We also used full sequence information of loci to explore if contamination with sequences from bacteria, parasites or other non-endogenous sources could be problematic. A Blast search of our final assembled loci against GenBank produced 6655 significant hits and, of these, we only detected 16 sequences belonging to bacteria and 4 to nematodes. This was a small number of contaminant sequences, as expected due to the low number of suboptimal samples we used.

The results obtained after performing a Blast search of the sequenced reads of each sample against a set of 49 putative Y-chromosome loci indicated that these loci were perfectly discriminating between the sexes. Detected Y-chromosome loci may also be used to evaluate cross-contamination between samples, since female samples are expected to have no hits for any of these loci (Green et al. 2009). This contamination may arise during the handling of samples or library preparation, especially after pooling samples. All females in our work showed very low levels of hits for Y-linked genes (less than 3 per million reads; Table S3), indicating the suitability of our approach to library preparation. The sexing results, with equal numbers of males and females, also showed a lack of sex bias in the selected specimens.

## Extremely low genomic heterozygosity rate in the Pyrenean desman

The individual heterozygosity rate estimated from the genomic data has been shown to be a relevant parameter for assessing the genetic diversity of species (Fig. S8). For example, the heterozygosity rate in primates varies from 800 heterozygous positions per million bases in non-African humans, eastern lowland gorillas, bonobos and western chimpanzees to around 1000 in African humans and 2000 in central chimpanzees, western lowland gorillas and both orangutan species (Prado-Martinez et al. 2013). Some highly endangered taxa in which this parameter has been

estimated from their complete genomes include the giant panda, with 1350 heterozygous positions per million bases (Li et al. 2010), the Amur tiger, with 490, the snow leopard, with 230 (Cho et al. 2013), and the cheetah, with 200 (Dobrynin et al. 2015). More recently, heterozygosity rates estimated in seven specimens of the Channel Island fox, a species that inhabits isolated islands with extremely small population sizes, varied between 14 and 408 (Robinson et al. 2016). In comparison, the Pyrenean desman has one of the lowest heterozygosity rates reported in any mammal so far, with a mean of 246 heterozygous positions per million bases (Table S4). Furthermore, there are substantial differences among the individual samples. We find values as low as 13 and 20 heterozygous positions per million bases in two samples from the southeastern area of the Pyrenees. All other specimens from the Pyrenean mountains have heterozygosity rates lower than 140. The only sample from the Central System also has a very low rate, with 147 heterozygous positions per million bases. The highest values are found in the occidental population, with one sample having 488 heterozygous positions per million bases. Interestingly, the range of heterozygosity rates found in the Pyrenean desman (13–488) is very similar to that of the Channel Island fox despite the larger geographical distribution of the desman. Extreme bottlenecks in small refugia during the last glacial maximum and before the postglacial recolonization may have created in the Pyrenean desman a pattern of genomic diversity similar to that found in oceanic islands.

A note of caution is necessary, however, when comparing the heterozygosity rates of *G. pyrenaicus* with other mammals, since this value was estimated in the present work from sequence fragments obtained from a double digestion of the genome and a size selection of 300–400 bp for the library construction. Although these loci should be a random representation of the genome, we cannot rule out that another factor may cause that the heterozygosity rates from ddRAD fragments and whole genomes are not totally comparable. Therefore, a firm conclusion on the heterozygosity rates of *G. pyrenaicus* compared with other mammals should await its complete genome sequencing. However, this parameter is very important for comparing specimens and populations within this species, showing how full sequence information obtained from the ddRAD data can also be exploited to achieve crucial knowledge on genetic variation.

The low heterozygosity rates found in the Pyrenees agree with the nucleotide diversity values estimated from mitochondrial and intronic data in a previous study (Igea et al. 2013). Another study of the northern part of the Pyrenean populations based on microsatellites also detected low levels of genetic diversity in this area (Gillet 2015). These populations have, until recently, been relatively well conserved (Aymerich et al. 2001; Aymerich and Gosálbez 2014), suggesting that long-term demographic decline does not explain the low genetic diversity in the Pyrenees. Instead, postglacial recolonization after a tight bottleneck seems to be the most likely explanation (Igea et al. 2013). However, recent data show some regression of the species in the Pyrenees (Charbonnel 2015). The isolation of subpopulations in some basins and the consequent increase of inbreeding or future unforeseen environmental changes may be more detrimental for populations starting with very low genetic diversity values (Allendorf et al. 2013).

## Population structure of the Pyrenean desman: novel insights from genomic data

In a previous work, we unveiled a strong phylogeographic structure in the Pyrenean desman, with an almost total absence of spatial mixing of the determined clades (Igea et al. 2013). These conclusions were obtained with mitochondrial DNA but, given the importance of analyzing this structure with nuclear markers, we also developed several introns and sequenced them in the main populations. Unfortunately, the nuclear introns did not show enough variability for these aspects to be analyzed in depth (Igea et al. 2013). It was therefore essential to develop a higher number of variable nuclear markers with genomic techniques to obtain more robust conclusions about the geographical structure of the Pyrenean desman.

A method based on nuclear genetic distances from concatenated loci (genomic tree) and another based on allele frequencies of SNPs (PCA) both revealed five clusters largely coincident with the main geographical regions and mountain ranges: Pyrenees, Occident, Cantabrian Mountains, Iberian Range and Central System (Figs. 2, 3). As previously shown (Igea et al. 2013), and contrary to expectations for a species that is highly adapted to aquatic ecosystems, this structure is not related to river basins. This indicates that, at least during the postglacial recolonization of these areas, boundaries between river basins did not act as important barriers to the dispersal of the Pyrenean desman populations. However, neither the genome tree nor the PCA analysis determines whether any of these groups comprise admixed specimens. The Structure analysis based on an admixture model showed the same five clusters and few specimens with genomic components belonging to different populations. In fact, only four specimens showed relevant admixture levels: three from the northwestern part of the Iberian Range and one from the Cantabrian Mountains (Fig. 4). The estimated admixture is relatively small and may be an indication of the low dispersal of this species. However, the lack of sampling in contact zones prevents any conclusions being drawn at this time about the true degree of interbreeding between the different

populations. In addition, the Central System population, with only one individual analyzed here, should be studied using a larger sample size.

The clusters found with genomic data corresponded fairly well with the mitochondrial lineages detected in a previous study (Igea et al. 2013) at the initial subdivisions. Thus, the major mitochondrial groups, A and B, are largely coincident with the two populations found here with Structure at $K = 2$. However, at higher subdivisions the mitochondrial clades and the nuclear clusters showed some remarkable differences. One of the major differences was found in the Iberian Range. The previous study showed that, in the Iberian Range, there are specimens belonging to the two main mitochondrial clades, A2 and B1, but they seemed to be strictly separated into two different areas (southeast and northwest, respectively), suggesting a highly restricted dispersal of female desmans despite the existence of any obvious geographic barrier to gene flow. However, the genomic data analyzed in the present work reveals a different scenario in this area. In principle, the six specimens from the Iberian Range (three from each area and mitochondrial clade) can be assigned to the same genomic cluster, despite three of them having some components of other populations. This suggests a homogenization of the two Iberian Range phylogroups through certain degree of gene flow between the areas of the two main mitochondrial clades. This discrepancy between the mitochondrial and nuclear geographical structures may be due to dispersal differences between the sexes, although current data on movements of the Pyrenean desman prevent confirmation of any dispersal sex bias (Melero et al. 2012; Gillet 2015). Alternatively, reduced dispersal together with the smaller mitochondrial population size has been shown to enable the maintenance of a mitochondrial phylogeographic break in the absence of nuclear structure, but further work is necessary to discern different possibilities (Irwin 2002; Toews and Brelsford 2012). The Iberian Range provides a clear example of how the mitochondrial phylogeography of the Pyrenean desman showed only one part of the population history, that of females, while genomic data was crucial to reveal that the whole population in this mountain range is more homogeneous than previously suspected.

The other major difference between the mitochondrial and nuclear geographical structures was found in the Central System. While the mitochondrial phylogeography showed that specimens from this area belong to the A2 clade, also present in the Iberian Range (Igea et al. 2013), the single specimen from the Central System studied here appears clearly separate in the genomic tree and, particularly, the PCA analysis (Figs. 2, 3). The results of the Structure analysis with $K = 5$ also indicate that this individual may be part of a differentiated population (Fig. 4). However, it is obvious that assessing the genetic status of the Central System population using only a single specimen is difficult. Furthermore, other specimens may have a different genomic background and, therefore, genomic data from further samples from the Central System is critical to understand more fully the evolutionary history of *G. pyrenaicus* in this area. The robust characterization of the Central System population will be particularly important because it is the most endangered population according to current data, and an in-depth understanding of its genetic distinctiveness will be necessary for delineating informed conservation plans.

Despite the small number of samples we could analyze in this work, the use of massive information from nuclear loci and SNPs generated with a genomic approach allowed us to complement the genetic structure analysis of the Pyrenean desman initiated with mitochondrial DNA and nuclear introns. Genomic data corroborated the existence of a marked genetic structure in this species and supported a scenario of glacial refugia that gave rise to genetically distinct populations associated to the main mountain ranges. According to the distribution of the mitochondrial clades, a scenario of four refugia (NW of the Iberian Peninsula, Cantabrian Mountains, Central System and Basque Mountains) was proposed in Igea et al. (2013), but the current results would be compatible with the existence of another refugium situated in the Iberian Range to help explain the genomic cluster found in this area. The Iberian Peninsula was classically considered a single homogeneous refugium in Europe (Hewitt 2000) but several phylogeographic studies in the last few years evidenced that the Iberian Peninsula (and other South European peninsulas) contained multiple refugia that gave rise to distinct evolutionary lineages, supporting the "refugia within refugia" hypothesis (Gomez and Lunt 2007 and references therein). Our mitochondrial and genomic studies with the Pyrenean desman, not only provide strong evidence for the importance of the Iberian Peninsula as the origin of distinct evolutionary lineages, but also show that some of them may be of critical conservation interest.

## Conclusions

Our results show the importance of performing genomic studies for properly defining evolutionary units of conservation value. This knowledge is crucial for species with endangered populations like the Pyrenean desman. The approach we used here included the sequencing by NGS techniques of specific short genomic fragments (around 45,000 fragments on average per specimen) and the generation of 1185 SNPs present in all samples. The genomic data obtained from 26 Pyrenean desmans selected from across their entire range revealed extremely low levels of

individual heterozygosity rate, with specimens from the southeastern Pyrenees having some of the lowest values inferred from genome-wide data in mammals so far. In addition, the analysis of genetic structure showed a strong geographical structure, with five main clusters that presented low levels of admixture among them and were largely coincident with the main mountain ranges. Preliminarily, and pending further analysis of more specimens from the whole distribution area, the populations found in each main mountain range should be considered as different conservation units. There is no perfect coincidence with the previously determined mitochondrial clades, particularly in the Iberian Range, demonstrating the importance of performing genomic analysis to reveal the whole population history of the species. The laboratory and bioinformatic techniques tested and improved here open the door to perform more specific analyses with additional specimens to cover poorly sampled regions. Among these, the Central System holds the most endangered population of the species according to current data therefore necessitating detailed genomic studies. In addition, the study of contact zones between different clusters would be of great help in defining the boundaries between them. These further analyses should help us better understand the dispersal patterns of the species and the true degree of isolation of the different populations. However, the overall geographical structure already uncovered here using genomic data should be taken into account in future management plans involving the Pyrenean desman.

**Acknowledgments** We thank the BTVS-ICNF collection (Banco de Tecidos de Vertebrados Selvagens—Instituto da Conservação da Natureza e das Florestas), Xunta de Galicia, Gobierno de Navarra, Diputación Foral de Gipuzkoa, Gobierno de La Rioja, Parque Nacional de Picos de Europa, Julio Gisbert and Rosa García-Perea (Proyecto Galemia) for *Galemys pyrenaicus* samples from their respective biological collections, and Henrique Carvalho, Carla Marisa Quaresma and Carlos P. Santos (Instituto da Conservação da Natureza e das Florestas) for their help with the access to the BTVS-ICNF collection. We also thank all the people of the Genomics Core Facility at the Pompeu Fabra University for help with the library preparation and sequencing. This work was financially supported by research Projects CGL2011-22640 and CGL2014-53968-P of the "Plan Nacional I + D+I del Ministerio de Economía y Competitividad" (Spain) to J.C. with contribution of project 014/2008 of the "Convocatoria de ayudas a proyectos de investigación en la Red de Parques Nacionales" (Spain) to J.G. M.Q. was supported by fellowship BES-2012-057135 of the Ministerio de Economía y Competitividad (Spain).

## References

Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. Nat Rev Genet 11:697–709. doi:10.1038/nrg2844

Allendorf FW, Luikart G, Aitken SN (2013) Conservation and the genetics of populations, 2nd edn. Wiley, Hoboken

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402

Aymerich P, Gosálbez J (2014) El desmán ibérico *Galemys pyrenaicus* (É. Geoffroy Saint-Hilaire, 1811) en los Pirineos meridionales. In: Conservation and Management of semi-aquatic mammals of South-western Europe. Munibe Monographs. Nature Series 3. Aranzadi Society of Sciences, San Sebastian, pp 37–77

Aymerich P, Casadesús F, Gosálbez J (2001) Distribució de *Galemys pyrenaicus* (Insectivora, Talpidae) a Catalunya. Orsis 16:93–110

Buerkle CA, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? Mol Ecol 22:3028–3035. doi:10.1111/mec.12105

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) *Stacks*: building and genotyping loci *de novo* from short-read sequences. Genes Genom Genet 1:171–182. doi:10.1534/g3.111.000240

Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140. doi:10.1111/mec.12354

Charbonnel A (2015) Influence multi-échelle des facteurs environnementaux dans la répartition du Desman des Pyrénées (*Galemys pyrenaicus*) en France. Dissertation, Université de Toulouse

Cho YS, Hu L, Hou H et al (2013) The tiger genome and comparative analysis with lion and snow leopard genomes. Nat Commun 4:2433. doi:10.1038/ncomms3433

Crandall KA, Bininda-Emonds ORP, Mace G, Wayne RK (2000) Considering evolutionary processes in conservation biology. Trends Ecol Evol 15:290–295

Dobrynin P, Liu S, Tamazian G et al (2015) Genomic legacy of the African cheetah, *Acinonyx jubatus*. Genome Biol 16:277. doi:10.1186/s13059-015-0837-4

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–2620. doi:10.1111/j.1365-294X.2005.02553.x

Felsenstein J (1989) PHYLIP-phylogeny inference package (version 3.4). Cladistics 5:164–166

Fernandes M, Herrero J, Aulagnier S, Amori G (2011) *Galemys pyrenaicus*. IUCN Red List of Threatened Species Version 20112 1–3

Freedman AH, Gronau I, Schweizer RM et al (2014) Genome sequencing highlights the dynamic early history of dogs. PLoS Genet 10:e1004016. doi:10.1371/journal.pgen.1004016

Funk WC, Mckay JK, Hohenlohe PA, Allendorf FW (2012) Harnessing genomics for delineating conservation units. Trends Ecol Evol 27:489–496. doi:10.1016/j.tree.2012.05.012

Gillet F (2015) Génétique et biologie de la conservation du desman des Pyrénées (*Galemys pyrenaicus*) en France. Dissertation, Université de Liège

Gisbert J, Garcia-Perea R (2014) Historia de la regresión del desmán ibérico *Galemys pyrenaicus* (É. Geoffroy Saint-Hilaire, 1811) en el Sistema Central (Península Ibérica). In: Conservation and Management of semi-aquatic mammals of South-western Europe. Munibe Monographs. Nature Series 3. Aranzadi Society of Sciences, San Sebastian, pp 19–35

Godinho R, Crespo EG, Ferrand N (2008) The limits of mtDNA phylogeography: complex patterns of population history in a highly structured Iberian lizard are only revealed by the use of nuclear markers. Mol Ecol 17:4670–4683. doi:10.1111/j.1365-294X.2008.03929.x

Gomez A, Lunt DH (2007) Refugia within refugia: patterns of phylogeographic concordance in the Iberian Peninsula. Phylogeography of Southern European Refugia, pp 155–188

González-Esteban J, Castién E, Gosálbez J (1999) Morphological and colour variation in the Pyrenean desman *Galemys pyrenaicus* (Geoffroy, 1811). Z Säugetierkunde 64:1–11

Green RE, Briggs AW, Krause J, Prüfer K, Burbano HA, Siebauer M, Lachmann M, Pääbo S (2009) The Neandertal genome and ancient DNA authenticity. EMBO J 28:2494–2502. doi:10.1038/emboj.2009.222

Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. Nat Genet 43:1031–1034. doi:10.1038/ng.937

Hewitt GM (2000) The genetic legacy of the Quaternary ice ages. Nature 405:907–913. doi:10.1038/35016000

Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23:254–267. doi:10.1093/molbev/msj030

Igea J, Aymerich P, Fernández-González A, González-Esteban J, Gómez A, Alonso R, Gosálbez J, Castresana J (2013) Phylogeography and postglacial expansion of the endangered semi-aquatic mammal *Galemys pyrenaicus*. BMC Evol Biol 13:115. doi:10.1186/1471-2148-13-115

Igea J, Aymerich P, Bannikova AA, Gosálbez J, Castresana J (2015) Multilocus species trees and species delimitation in a temporal context: application to the water shrews of the genus *Neomys*. BMC Evol Biol 15:209. doi:10.1186/s12862-015-0485-z

Irwin DE (2002) Phylogeographic breaks without geographic barriers to gene flow. Evolution 56:2383–2394

Kjeldsen SR, Zenger KR, Leigh K, Ellis W, Tobey J, Phalen D, Melzer A, FitzGibbon S, Raadsma HW (2015) Genome-wide SNP loci reveal novel insights into koala (*Phascolarctos cinereus*) population variability across its range. Conserv Genet 17:337–353. doi:10.1007/s10592-015-0784-3

Lanier HC, Massatti R, He Q, Olson LE, Knowles LL (2015) Colonization from divergent ancestors: glaciation signatures on contemporary patterns of genomic variation in Collared Pikas (*Ochotona collaris*). Mol Ecol 24:3688–3705. doi:10.1111/mec.13270

Li R, Fan W, Tian G et al (2010) The sequence and *de novo* assembly of the giant panda genome. Nature 463:311–317. doi:10.1038/nature08696

Lopez-Fuster MJ, Garcia-Perea R, Fernández-Salvador R, Gisbert J, Ventura J (2006) Craniometric variability of the Iberian desman, *Galemys pyrenaicus* (Mammalia: Erinaceomorpha: Talpidae). Folia Zool 55:29–42

Lynch M (1991) The genetic interpretation of inbreeding depression and outbreeding depression. Evolution 45:622–629

Melero Y, Aymerich P, Luque-Larena JJ, Gosálbez J (2012) New insights into social and space use behaviour of the endangered Pyrenean desman (*Galemys pyrenaicus*). Eur J Wildl Res 58:185–193. doi:10.1007/s10344-011-0561-7

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. Mol Ecol Resour 9:66–73. doi:10.1111/j.1755-0998.2008.02392.x

Nater A, Nietlisbach P, Arora N, van Schaik CP, van Noordwijk MA, Willems EP, Singleton I, Wich SA, Goossens B, Warren KS, Verschoor EJ, Perwitasari-Farajallah D, Pamungkas J, Krützen M (2011) Sex-biased dispersal and volcanic activities shaped phylogeographic patterns of extant Orangutans (genus: *Pongo*). Mol Biol Evol 28:2275–2288. doi:10.1093/molbev/msr042

Nores C, Queiroz AI, Gisbert J (2007) *Galemys pyrenaicus*. Atlas y libro rojo de los mamíferos terrestres de España, pp 92–98

Orsini L, Vanoverbeke J, Swillen I, Mergeay J, De Meester L (2013) Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. Mol Ecol 22:5983–5999. doi:10.1111/mec.12561

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. Trends Genet 26:177–187. doi:10.1016/j.tig.2010.01.001

Palmeirim JM, Hoffmann RS (1983) *Galemys pyrenaicus*. Mamm Species 207:1–5

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS One 7:e37135. doi:10.1371/journal.pone.0037135

Prado-Martinez J, Sudmant PH, Kidd JM et al (2013) Great ape genetic diversity and population history. Nature 499:471–475. doi:10.1038/nature12228

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, Vonholdt BM, Marsden CD, Lohmueller KE, Wayne RK (2016) Genomic flatlining in the endangered Island Fox. Curr Biol 26:1183–1189. doi:10.1016/j.cub.2016.02.062

Shafer ABA, Wolf JBW, Alves PC et al (2015) Genomics and the challenging translation into conservation practice. Trends Ecol Evol 30:78–87. doi:10.1016/j.tree.2014.11.009

Sovic MG, Fries AC, Gibbs HL (2015) AftrRAD: a pipeline for accurate and efficient *de novo* assembly of RADseq data. Mol Ecol Resour 15:1163–1171. doi:10.1111/1755-0998.12378

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690. doi:10.1093/bioinformatics/btl446

Steiner CC, Putnam AS, Hoeck PEA, Ryder OA (2013) Conservation genomics of threatened animal species. Annu Rev Anim Biosci 1:261–281. doi:10.1146/annurev-animal-031412-103636

Tallmon DA, Luikart G, Waples RS (2004) The alluring simplicity and complex reality of genetic rescue. Trends Ecol Evol 19:489–496. doi:10.1016/j.tree.2004.07.003

Toews DPL, Brelsford A (2012) The biogeography of mitochondrial and nuclear discordance in animals. Mol Ecol 21:3907–3930. doi:10.1111/j.1365-294X.2012.05664.x

Vonholdt BM, Pollinger JP, Earl DA, Knowles JC, Boyko AR, Parker H, Geffen E, Pilot M, Jedrzejewski W, Jedrzejewska B, Sidorovich V, Greco C, Randi E, Musiani M, Kays R, Bustamante CD, Ostrander EA, Novembre J, Wayne RK (2011) A genome-wide perspective on the evolutionary history of enigmatic wolf-like canids. Genome Res 21:1294–1305. doi:10.1101/gr.116301.110

Xue Y, Prado-Martinez J, Sudmant PH, Narasimhan V, Ayub Q, Szpak M, Frandsen P, Chen Y, Yngvadottir B, Cooper DN, de Manuel M, Hernandez-Rodriguez J, Lobon I, Siegismund HR, Pagani L, Quail MA, Hvilsom C, Mudakikwa A, Eichler EE, Cranfield MR, Marques-Bonet T, Tyler-Smith C, Scally A (2015) Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. Science 348:242–245. doi:10.1126/science.aaa3952

Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28:3326–3328. doi:10.1093/bioinformatics/bts606