iScience, Volume *27*

Supplemental information

Genomic insights into the Montseny brook newt (*Calotriton arnoldi*), a Critically Endangered glacial relict

Adrián Talavera, Marc Palmada-Flores, Bernat Burriel-Carranza, Emilio Valbuena-Ureña, Gabriel Mochales-Riaño, Dean C. Adams, Héctor Tejero-Cicuéndez, Anna Soler-Membrives, Fèlix Amat, Daniel Guinart, Francesc Carbonell, Elena Obon, Tomàs Marquès-Bonet, and Salvador Carranza

Table S1. Summary of genome assembly metrics of *Calotriton arnoldi* reference genome (aCalArn1) and the other available Caudata assemblies. aCalArn1 shows higher contiguity at contig level (N50). Gb = Gigabase; Mb = Megabase; bp = basepairs.

Metrics	Assembly		
	aCalArn1	AmbMex60DD	Pw_v4.2
Assembly size (no Ns) (Gb)	22.89	28.20	19.26
Number of Contigs	47,990	211,437	65,627,701
Number of Scaffolds	47,257	27,157	106,895
Contig N50 (Mb)	1.60	0.22	1.14
Scaffold N50 (Gb)	0.00165	1.20571	-
Contig L50	4,166	35,415	-
Scaffold L50	4,025	11	-
Shortest / longest contig (bp)	33 / 11,568,535	1085 / NA	71 / 7,104,246
% of bp in contigs of range $<10^2/10^2$-$10^3/10^3$-$10^4/10^4$-$10^5/>10^5$	0 / 0 / 2 / 29 / 68	-	12 / 36 / 43 / 5 / 4

Table S2. Assembly assessment of *Calotriton arnoldi* reference genome under VGP standards. Structural accuracy, base accuracy (regarding k-mer completeness) and functional completeness are within VGP standards. BUSCO(S) = Single copy orthologs using BUSCO v5.2.2 with vertebrata ofb10 database.

Quality Category	Quality Metric	VGP standard	aCalArn1
Structural accuracy	False duplications	0.2-5.0%	1.84%
Base accuracy	Base pair QV	39-43	35.2 (80.9% ct QV > 30 ; 11.4% ct QV >39)
Base accuracy	K-mer completeness	87-98%	97.54%
Functional completeness	Genes (BUSCOs (S))	82-98%	82.3%

Table S3. Summaries of Generalized Least Squares (GLS) models describing the dependence of *Calotriton* spp. demographic trends on temperature or ice sheet volume along Late Pleistocene and Holocene while accounting for temporal autocorrelation. Degrees of freedom (Df), Sum of Squares Type I (SS), the coefficient of determination (R^2), F-values, the effect size (Z-score, based on the distribution of F-values obtained from permutations) and p-values are showcased. Related to Figure 2.

Species	Ind. variable	Df	SS	R^2	F	Z	*p*-value
C. arnoldi	temperature	1	1930869998	0.851	793.67	10.26	<0.001
C. asper	temperature	1	3816711796	0.248	124.17	6.26	<0.001
C. arnoldi	ice volume	1	1739244848	0.767	456.32	8.96	<0.001
C. asper	ice volume	1	1583891610	0.103	43.21	4.53	<0.001

Table S4. Fastsimcoal2 best-fitting run point estimations for each of the four models. Effective population sizes (N_X) mean "haploid" individuals. N_A, N_B and N_O stands for current population sizes of Eastern Montseny, Western Montseny and Pyrenean brook newts, respectively. N_{AB} is population size before Montseny lineages split, and N_{ABO}, *Calotriton* population size before species divergence. Lineage divergence (T_{DIV}) and *Calotriton arnoldi*'s bottleneck (T_{BOT}) are measured in 4-year generations ago and migration rates (m_x) in "haploid" migrants per generation. $m_{arnoldi}$ stands for within species migration between Montseny lineages, whereas m_{asper} for ancestral migration from the Pyrenees into Montseny. The best-fitting model (in bold) showcases both migration patterns. Related to Table S5 and Figure S4.

	N_A	N_B	N_0	N_{AB}	N_{0AB}	T_{DIV}	T_{BOT}	$m_{arnoldi}$	m_{asper}
1	2,194	8,481	1,351,926	185,546	3,219,370	1,966	21,463	-	-
2	671	9,046	1,557,947	659875	3,340,096	330,190	15,171	3.32E-04	-
3	1,205	8,137	341,330	7,474	2,598,606	1,644	15,078	-	20.2
4	**973**	**6,599**	**443,717**	**283,022**	**2,900,358**	**12,226**	**16,096**	**2.84E-04**	**5.98E-04**

Table S5. Fastsimcoal2 estimated composite likelihoods from best runs and fitting comparisons between models. The best-fitting model (in bold) implies both ancestral migration from the Pyrenees into the Montseny and within-Montseny secondary contacts. Related to Figure S4.

Model	Max $\log_{10}(Lhood_i)$	No. of parameters (k)	AIC_i	ΔAIC_i	Model normalized relative likelihood (w_i)
1	-6,133	7	28,255	2,024.444	0
2	-6,073	8	27,985	1,754.642	0
3	-5,733	8	26,419	188.092	1.433E-41
4	**-5,692**	**9**	**26,231**	**0**	**1**

Table S6. Heterozygosity inferred from low coverage WGS from 17 *Calotriton arnoldi* (7 from the Eastern and 10 from the Western Montseny lineage) and 5 *C. asper* representing five mitochondrial lineages. Diversity is higher in the Pyrenean species than in the Montseny populations. Within Montseny, the Western lineage has in general slightly higher values than the Eastern one.

Species	Locality	Sample ID	Population	1st-scaffold heterozygosity	Whole-genome heterozygosity	Non-repetitive whole-genome heterozygosity
C. arnoldi	Torreferrussa	OR193	F1 A	0.0000680	0.0001005	0.0000645
C. arnoldi	Eastern Montseny	D3247	A1	0.0000520	0.0000815	0.0000613
C. arnoldi	Eastern Montseny	D3158	A1	0.0001214	0.0001115	0.0000701
C. arnoldi	Eastern Montseny	D3116	A2.1	0.0000420	0.0000717	0.0000869
C. arnoldi	Eastern Montseny	SPM12051608	A2.2	0.0000441	0.0000784	0.0000672
C. arnoldi	Eastern Montseny	SPM12051602	A3	0.0000560	0.0000887	0.0000736
C. arnoldi	Eastern Montseny	SPM12051601	A3	0.0000436	0.0000749	0.0000769
C. arnoldi	Western Montseny	D3139	B1	0.0000610	0.0000948	0.0000790
C. arnoldi	Western Montseny	D3146	B1	0.0000505	0.0000912	0.0000783
C. arnoldi	Western Montseny	SPM13032411	B2	0.0000504	0.0000906	0.0000729
C. arnoldi	Western Montseny	SPM11111011	B2	0.0000573	0.0001008	0.0000758
C. arnoldi	Western Montseny	D3065	B3	0.0000426	0.0000850	0.0000824
C. arnoldi	Western Montseny	SPM10042119	B3	0.0000598	0.0000946	0.0000871
C. arnoldi	Western Montseny	D3085	B4	0.0000453	0.0000885	0.0000867
C. arnoldi	Western Montseny	D3086	B4	0.0000567	0.0000939	0.0000829
C. arnoldi	Western Montseny	D3152	B5	0.0000630	0.0000991	0.0000811
C. arnoldi	Western Montseny	D3215	B5	0.0000412	0.0000843	0.0001018
C. asper	Estanys de Juclar	SPM004169	O1	0.0011390	0.0012977	0.0009025
C. asper	Vidrà	19030701	O2	0.0009626	0.0011434	0.0009148
C. asper	Valle de Pineta	SPM004784	O3	0.0009621	0.0013441	0.0011240
C. asper	Puerto de Portalet	SPM004128	O4	0.0012150	0.0015156	0.0012200
C. asper	Montsec de Rúbies	B7	O5	0.0008774	0.0011348	0.0010517

Table S7. Observed heterozygosity (H_O), expected heterozygosity (H_E) estimates and inbreeding coefficient (G_{IS}) per population for all *Calotriton arnoldi* (A: Eastern Montseny, B: Western Montseny) and three *C. asper* populations inferred from ddRADseq data with GenoDive. H_O was higher for *C. asper* than for *C. arnoldi*, and within Montseny, slightly higher for Western than Eastern Montseny. G_{IS} was, in all cases but one, negative, i.e. heterozygosity was higher than expected. Related to Figure 4A.

Population	H_O	H_E	G_{IS}
A1	0.008	0.007	-0.145
A2.1	0.007	0.006	-0.105
A2.2	0.008	0.007	-0.127
A3	0.009	0.008	-0.120
B1	0.011	0.009	-0.123
B2	0.010	0.009	-0.109
B3	0.009	0.007	-0.232
B4	0.011	0.009	-0.132
B5	0.010	0.009	-0.107
O2.1	0.055	0.042	-0.315
O2.2	0.049	0.037	-0.343
O3	0.121	0.124	0.022

Table S8. Mean±SD number of Runs of Homozygosity (ROHs) in the largest 50 scaffolds (11.6-7.1 Mbp) of 16 wild *Calotriton arnoldi* (10 from Western and 6 from Eastern Montseny lineages), 5 *C. asper* individuals, and 1 *C. arnoldi* from the breeding programme (F1, Eastern Montseny lineage) from lcWGS. Related to Figure 4B.

	Mean number of ROH ± SD			
ROH length (Mb)	**Eastern Montseny (n=6)**	**Western Montseny (n=10)**	**Pyrenees (n=5)**	**Breeding center (n=1)**
0.5-1	125.0±21.80	145.0±9.99	35.0±6.63	128
1-2	91.3±8.21	88.3±6.90	1.0±1.00	102
2-3	24.8±8.59	17.0±4.24	0.0±0.00	25
>3	9.0±4.82	3.0±1.49	0.0±0.00	10

Table S9. Population-pairwise fixation indices (F'_{ST}) for all *Calotriton arnoldi* (A: Eastern Montseny, B: Western Montseny) and three *C. asper* populations (O2-3) below diagonal and respective p-values above thereof. Population comparisons between Western and Eastern Montseny yielded F'_{ST} values similar to between-species comparisons. Related to Figure 4C.

	A1	**A2.1**	**A2.2**	**A3**	**B1**	**B2**	**B3**	**B4**	**B5**	**O2.1**	**O2.2**	**O3**
A1	-	0.184	0.148	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.001	<0.001
A2.1	<0.001	-	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.007	0.001
A2.2	<0.001	0.066	-	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.003	<0.001
A3	0.133	0.22	0.223	-	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.001	<0.001
B1	0.892	0.89	0.892	0.89	-	0.078	<0.001	0.029	0.283	<0.001	0.005	<0.001
B2	0.895	0.894	0.896	0.894	<0.001	-	<0.001	<0.001	<0.001	<0.001	0.002	<0.001
B3	0.914	0.919	0.918	0.913	0.179	0.183	-	<0.001	<0.001	<0.001	0.001	<0.001
B4	0.886	0.882	0.885	0.885	0.015	0.086	0.212	-	<0.001	<0.001	0.001	<0.001
B5	0.891	0.889	0.891	0.89	0.011	0.064	0.162	0.021	-	<0.001	0.003	<0.001
O2.1	0.998	0.997	0.997	0.998	0.997	0.997	0.998	0.997	0.997	-	0.006	<0.001
O2.2	0.999	0.998	0.998	0.999	0.998	0.998	0.999	0.998	0.998	0.181	-	0.012
O3	0.993	0.99	0.992	0.993	0.99	0.993	0.994	0.992	0.991	0.603	0.486	-

Table S10. Mean±SD of 20 cross validation (CV) errors from Admixture analyses with two different datasets. For each dataset, Admixture was run for K values ranging from 1 to the number of different sampled populations, i.e. twelve for the *general* dataset with outgroup and nine for the *C. arnoldi* dataset (Table S12). Results support clear differences in between Eastern Montseny, Western Montseny and Pyrenean brook newts (K3, *general* dataset). Related to Figure 5A.

K	***general***	***no_outgroup***
1	0.4926±0.0003	0.3645±0.0004
2	0.0999±0.0002	0.1662±0.0003
3	**0.0663±0.0128**	0.1695±0.0011
4	0.0573±0.0034	0.1728±0.0006
5	0.0564±0.0034	**0.1773±0.0009**
6	0.0557±0.0049	0.1813±0.001
7	0.0546±0.0043	0.1857±0.0011
8	0.0571±0.007	0.1909±0.0012
9	0.0536±0.0045	0.1956±0.0012
10	0.0542±0.0045	-
11	0.0561±0.0067	-
12	0.0561±0.008	-

Table S11. Loci count using different clustering thresholds and assembly methods of ddRADseq data with ipyrad. Reference-assisted assembly yielded the highest number of retrieved loci.

Assembly	Clustering threshold	RAD loci
de novo	0.85	118,993
de novo	0.89	133,846
de novo	0.92	147,214
de novo	0.95	162,745
Reference-assisted	-	233,171

Table S12. ddRADseq datasets produced, with the number of samples and species comprised, retrieved SNPs, the filters applied and the analyses in which they have been used. For more information about filters see STAR Methods: Method Details: Processing of ddRADseq data.

Dataset	Samples	Taxa	Data amount	Used for	Postprocessing filtering
general	132	*Calotriton* spp.	9,470 uSNPs	PCA, Admixture, Genodive, prabclus	Common markers; 75% individual & 15% genotype missingness allowed; MAC=1; removing monomorphic sites and extreme heterozygote excess; keep only biallelics; LD-pruning 0.5 correlation within 50-SNP windows.
no_outgroup	115	*Calotriton arnoldi*	1,339 uSNPs	PCA, Admixture, divMigrate	Common markers; 30% genotype missingness allowed; MAC = 1; removing monomorphic sites and extreme heterozygote excess; keep only biallelics; LD-pruning 0.5 correlation within 50-SNP windows.
no_missingness	7	*Calotriton* spp.	8,139 uSNPs	SNAPP	0% genotype missingness allowed; removing monomorphic sites and extreme heterozygote excess; keep only biallelics; LD-pruning 0.5 correlation within 50-SNP windows
loci	132	*Calotriton* spp.	1,402,750 sites	RAxML-NG	15% genotype missingness allowed.
no_relatives	88	*Calotriton* spp.	8,841 uSNPs	Stairwayplot2, Fastsimcoal2	1st degree or closer relatives excluded; common markers; 15% genotype missingness allowed; removing monomorphic sites and extreme heterozygote excess; keep only biallelics; LD-pruning 0.5 correlation within 50-SNP windows.

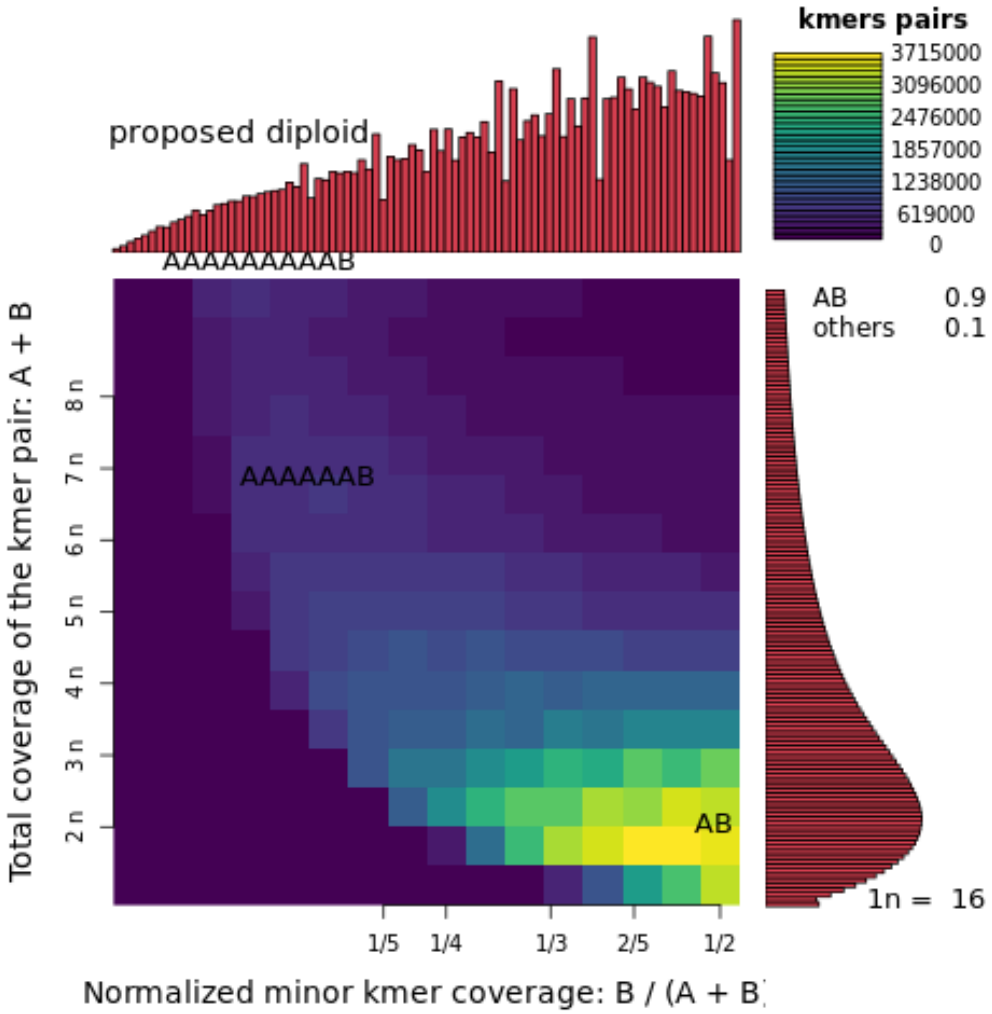

Figure S1. Smudgeplot based on 22-mer database to check the diploidy of *Calotriton arnoldi*. Most of the proportion of k-mers follow a proportion AB (0.9) indicating a diploidy for the species. It does not exclude the possibility of a specific chromosome having a different ploidy than the rest.

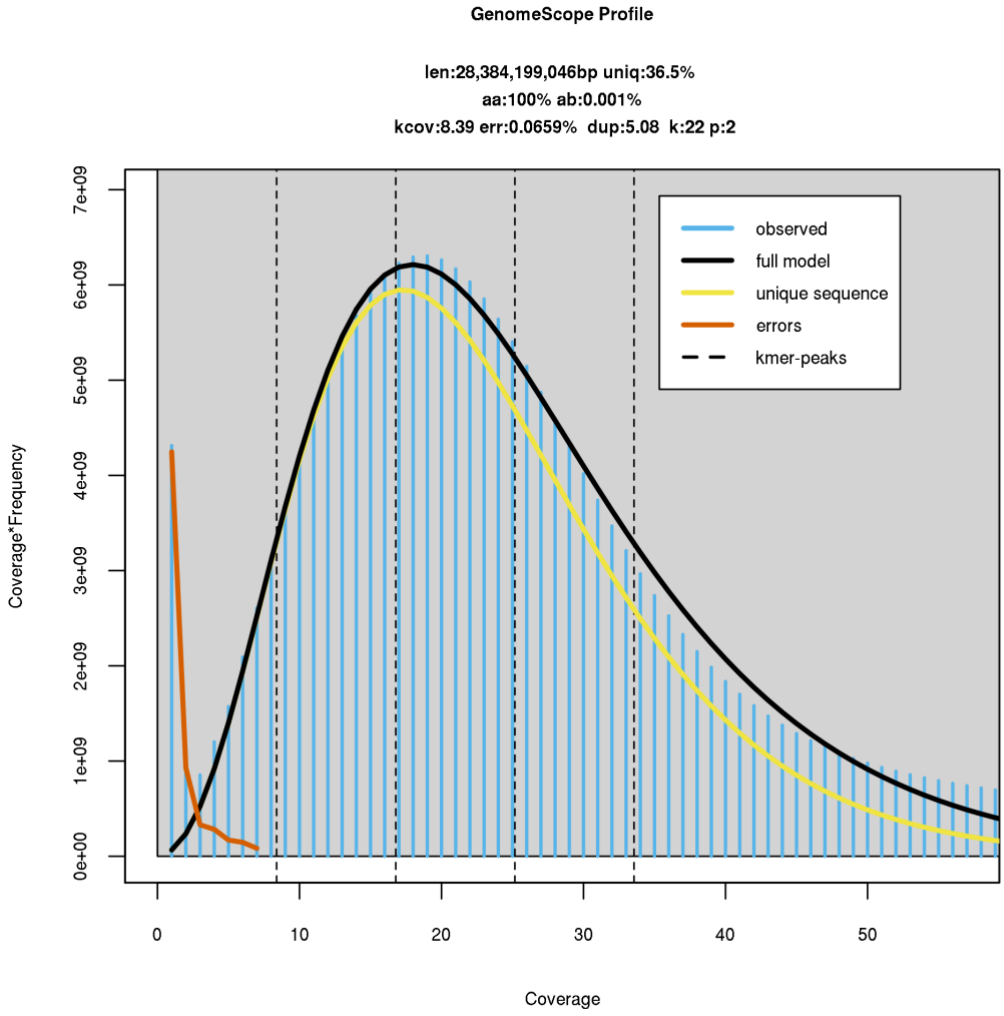

Figure S2. GenomeScope v2.0 analysis graph profile employed to estimate genome haploid length, proportion of repeats and heterozygosity of *Calotriton arnoldi*. The k-mer genome profiling shows one homozygous peak (2n) at around 19X coverage, providing a highly homozygous genome estimate (ab: 0.001%). Observed k-mer profile fits within the diploid model with a typical error crest at low k-mer counts. Estimated genome size is >28 Gb (len: 28,384,199,046 bp) with more than 63% of repetitive content (uniq: 36.5%).

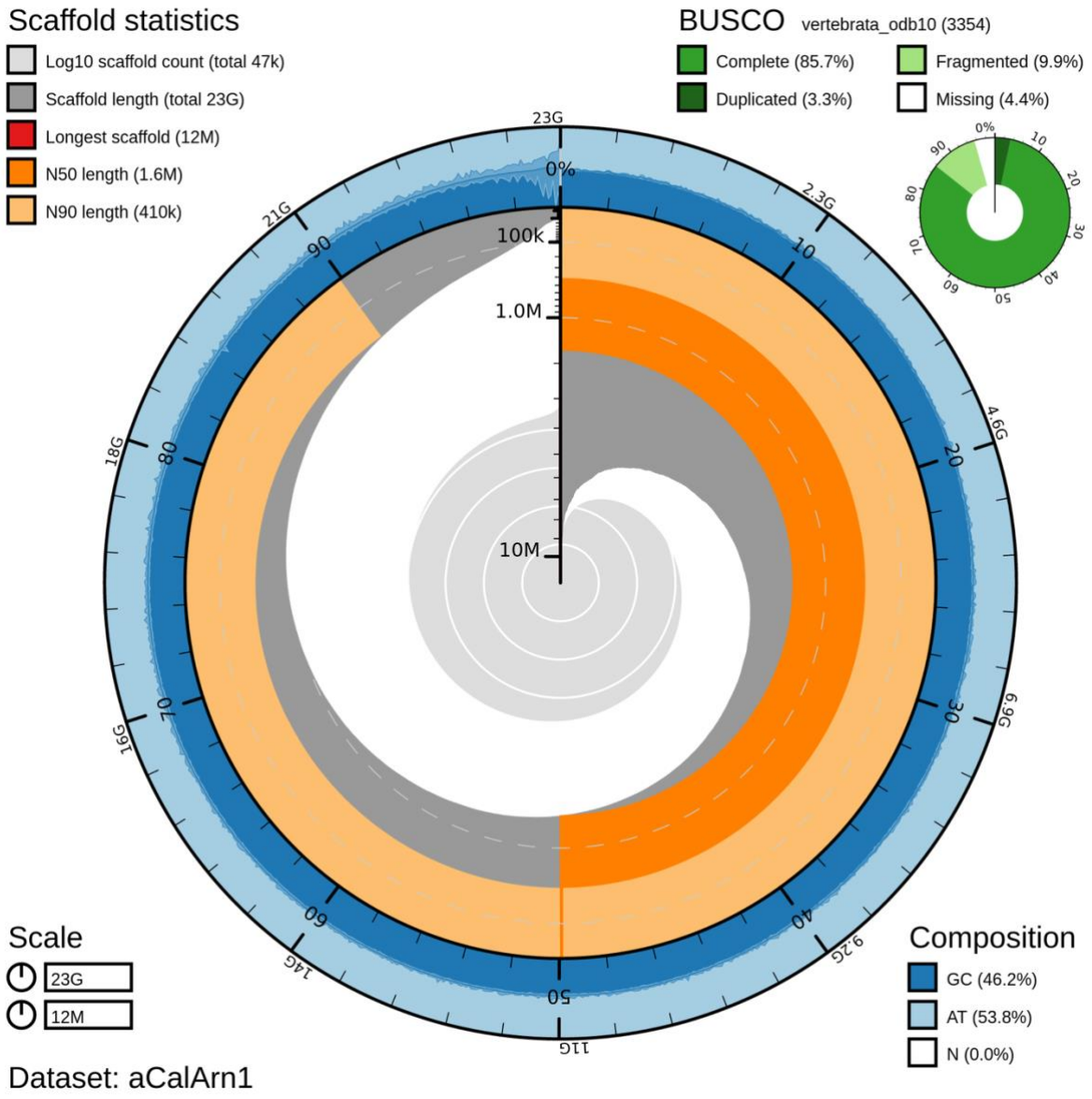

Figure S3. Snail plot of *Calotriton arnoldi* reference genome assembly presented in this study (aCalArn1) summarizing some contiguity and completeness metrics. M=Mega bases; G=Giga bases; k=kilo bases. The scaffolds contained in the assembly are shown ordered by length in the inner circle of the snail plot. The sequences within the N50 (1.6M) and N90 (410k) length marks are painted in dark or light orange, respectively. In the outer circle, the GC content of the sequences in represented in blue, whereas AT content in pale blue. Top left corner shows contiguity statistics like the longest scaffold (12M) or Scaffold length (total 23G) which indicates the assembly size. Top right corner represents the assembly genetic completeness using BUSCO scores with odb10 vertebrata database with 3354 genes, with 85.7% of the database genes found as Single-Copy Complete genes. Bottom left has a legend with the length in bases represented in the snail plot, the perimeter (23G) and the radius (12M, i.e., the longest scaffold).

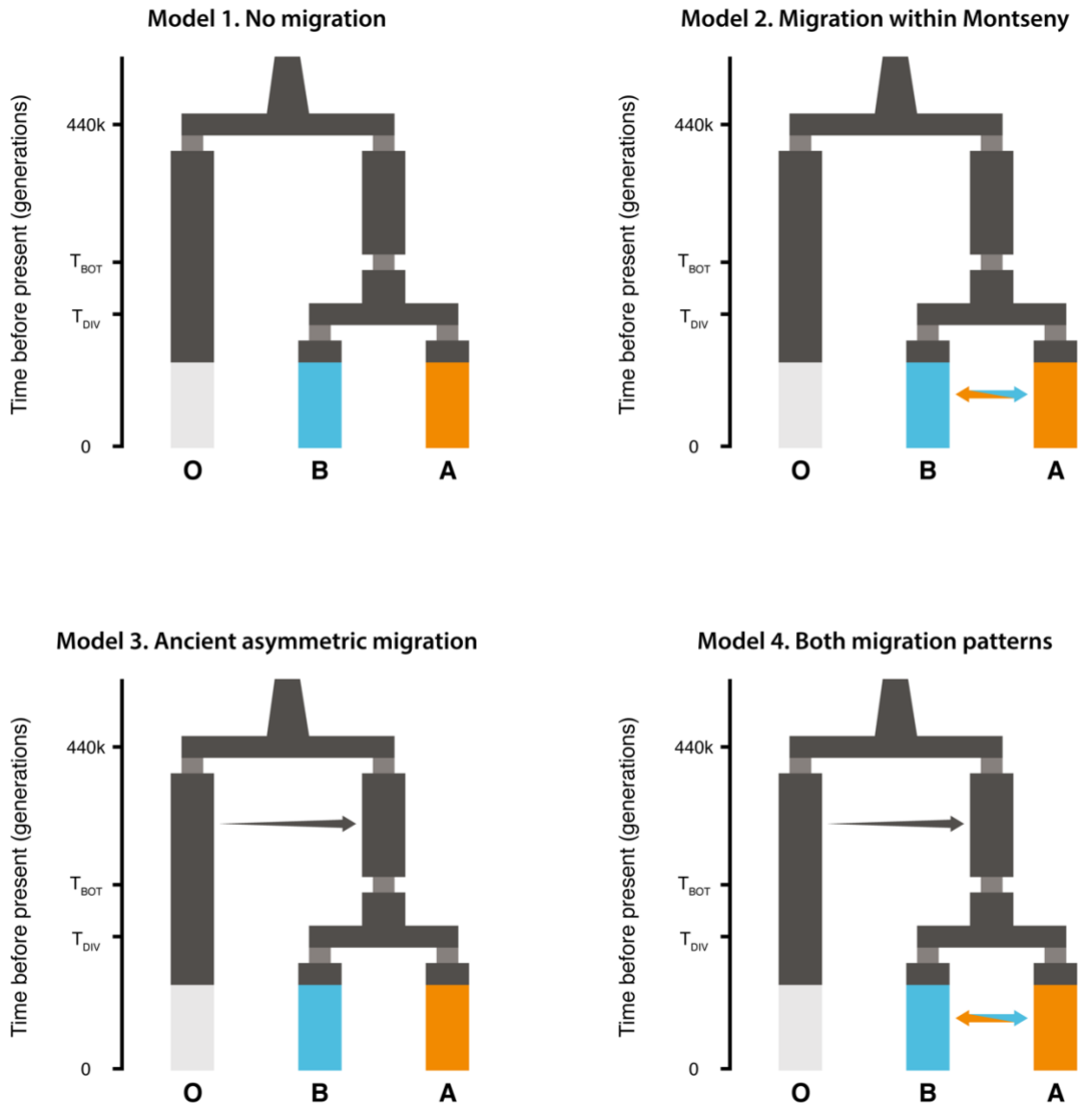

Figure S4. Fastsimcoal population tree models with same topology but allowing different migration matrices. The first model lacks migration at all, the second one is allowed for recent migration between *C. arnoldi* lineages, the third for ancient asymmetrical migration from *C. asper* to *C. arnoldi*; and the fourth model including both migrations patterns. The most likely model was number 4, allowing both types of migration. Western Montseny *C. arnoldi* in blue, Eastern Montseny *C. arnoldi* in orange, Pyrenean brook newt *C. asper* in grey. T_{BOT}, bottleneck time; T_{DIV}, *C. arnoldi* divergence time. Divergence time between *C. arnoldi* and *C. asper* was fixed in 440k generations ago.

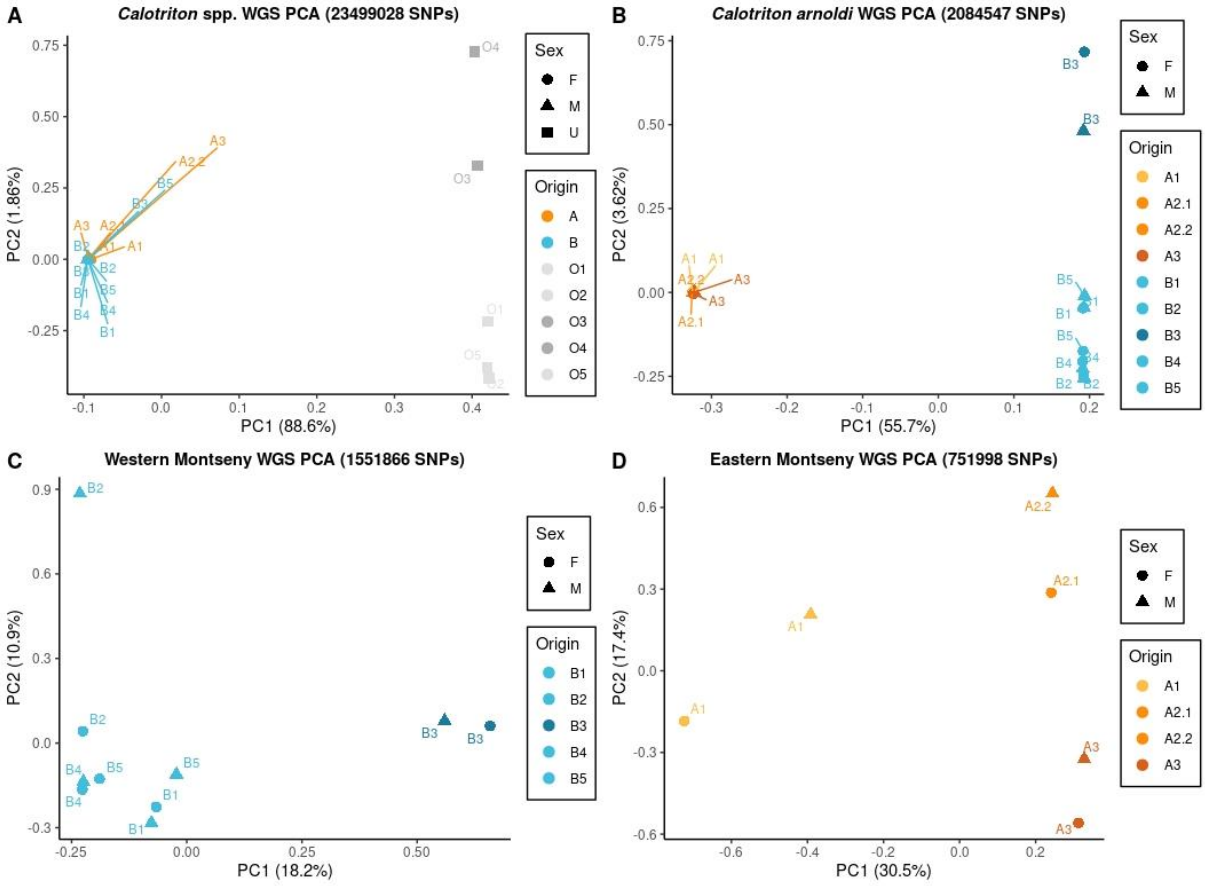

Figure S5. Hierarchical genomic PCAs from lcWGS with one male and female from each Montseny population and five representatives of *C. asper* diversity. A) shows both brook newt species with >23M SNPs. B) is restricted to Montseny, showing the split between Eastern and Western Montseny lineages along PC1. C, D) show within-lineage PCAs at the Western and Eastern submassifs, respectively.

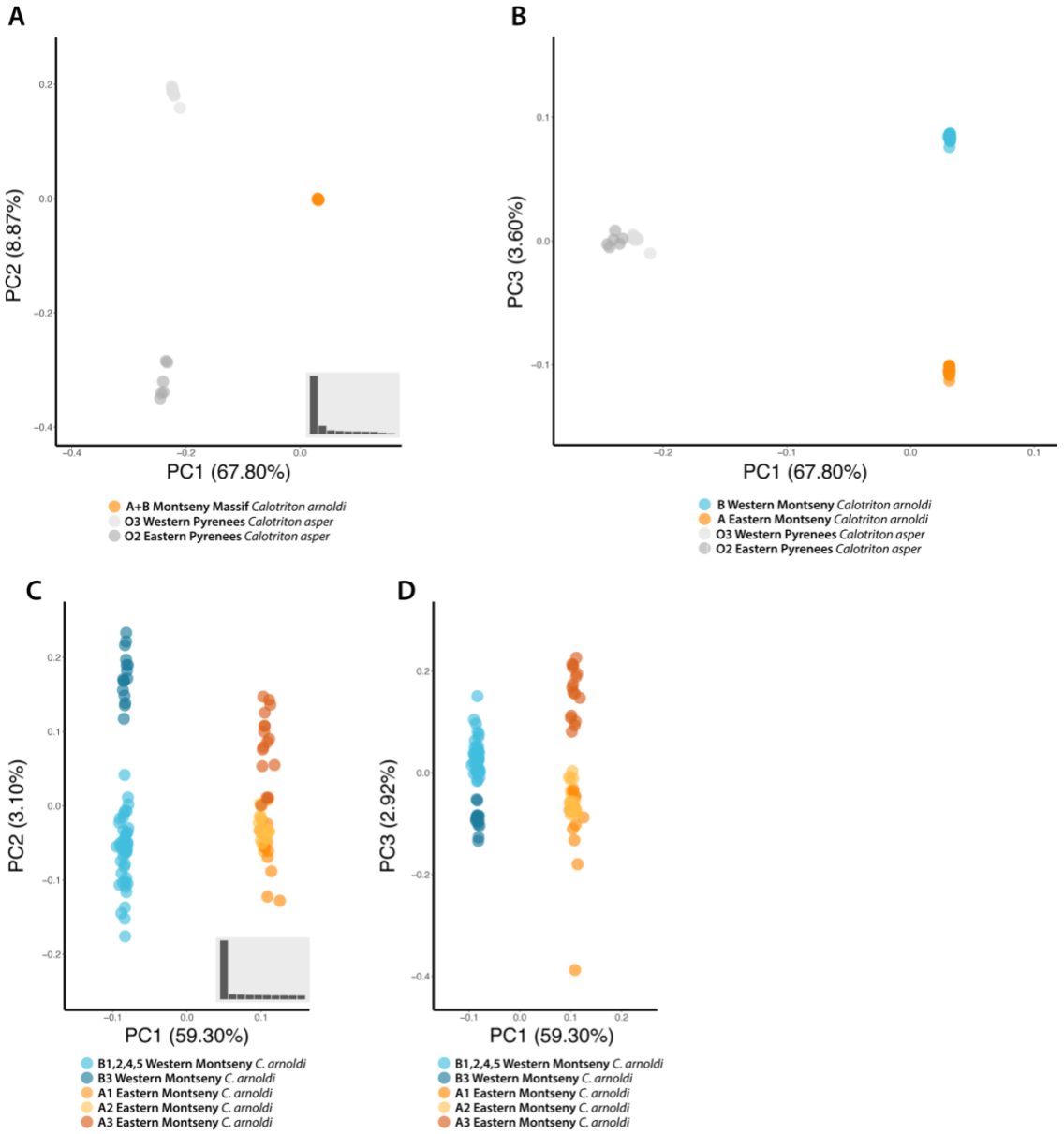

Figure S6. Genomic genus- and species-level PCAs from 132 ddRADseq samples (9,470 and 1,339 uSNPs, repectively). A and B) show, at genus level, the split between *Calotriton* species (PC1), Western-Eastern Pyrenees (PC2) and Western-Eastern Montseny (PC3). C and D) illustrate, at species level, the split between Western and Eastern Montseny (PC1), as well as within-lineage diversity (PC2-3). Histograms in A and C show the proportion of variance explained by the first ten Principal Components of the genus- and species-level PCAs, respectively.

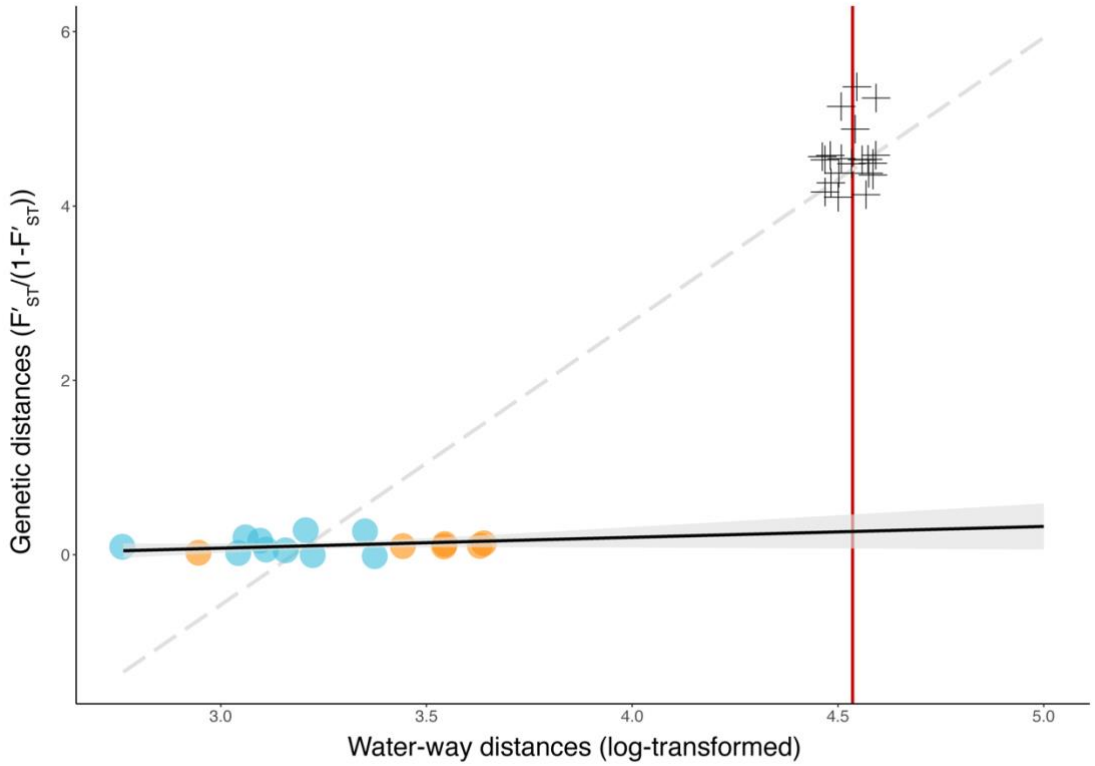

Figure S7. Relationships between genomic and log-transformed water-way geographical distances in pairs of populations of the two *C. arnoldi* lineages to explore Isolation-by-Distance patterns (IBD). Genomic distances are measured by $F'_{ST}/(1- F'_{ST})$ from ddRADseq data, whereas geographical distances are measured in log-transformed water-way distances in meters, to account for the dispersal constraints of the species. Blue and orange circles: distances between populations belonging to Western and Eastern Montseny, respectively. Black crosses: distances between populations belonging to different lineages. Black solid line: regression line fitted only for within-lineages comparisons (i.e., blue and orange circles) and its 95% CI in light grey. Dashed grey line: regression line fitted for all comparisons, either within- or between-lineages (circles and crosses), i.e., the IBD pattern needed to explain lineage differences exclusively due to distance. Red vertical line: center of between-lineages geographical distances. IBD shows the same pattern within the Western and Eastern submassifs: i.e., is explained in both lineages by the same slope (black solid line). However, genomic differences between lineages cannot be explained by the same IBD pattern (the dashed grey line is significantly different than the black one).

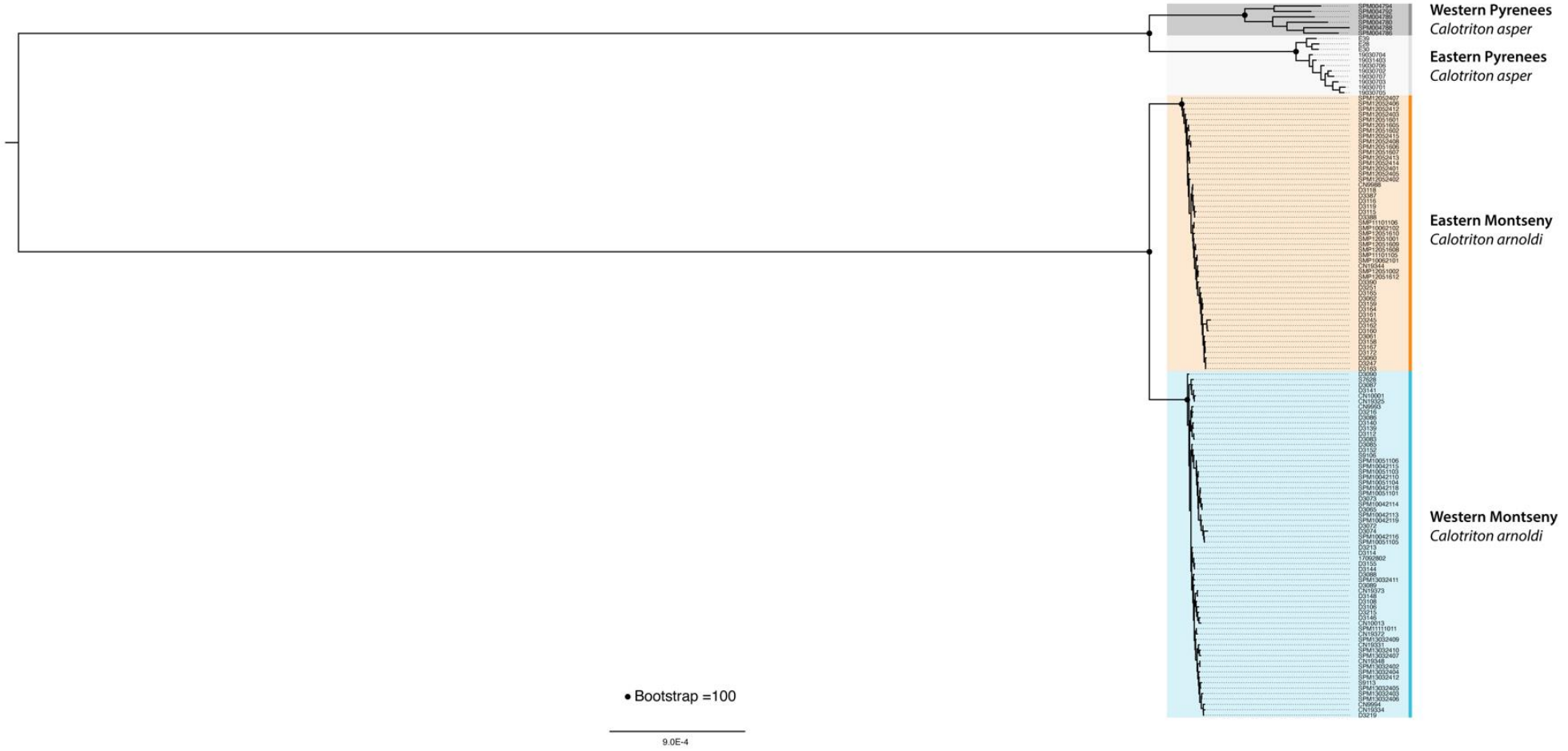

Figure S8. ML phylogeny with 132 ddRAD samples by RAxML-NG with GTRGAMMA model from 50 random and 50 parsimony-based starting trees and 1,000 bootstrap replicates. The four intra-specific lineages are monophyletic (Western and Eastern Montseny and Western and Eastern Pyrenees; black circles, bootstrap value=100).