REGULATORY CHANGES IN PTERIN AND CAROTENOID GENES UNDERLIE BALANCED COLOR POLYMORPHISMS IN THE WALL LIZARD

Pedro Andrade[1,2,#], Catarina Pinho[1,#], Guillem Pérez i de Lanuza[1], Sandra Afonso[1], Jindřich Brejcha[3,4,5], Carl-Johan Rubin[6], Ola Wallerman[6], Paulo Pereira[1,2], Stephen J. Sabatino[1], Adriana Bellati[7], Daniele Pellitteri-Rosa[7], Zuzana Bosakova[8], Ignas Bunikis[9], Miguel A. Carretero[1], Nathalie Feiner[10], Petr Marsik[11], Francisco Paupério[2], Daniele Salvi[1,12], Lucile Soler[13], Geoffrey M. While[14,15], Tobias Uller[10], Enrique Font[5], Leif Andersson[6,16,17,*], Miguel Carneiro[1,2,*]

SUPPORTING INFORMATION

CONTENT

SUPPLEMENTARY METHODS

Figures S1 – S7

Tables S1 – S9

www.pnas.org/cgi/doi/10.1073/pnas.1820320116

SUPPLEMENTARY METHODS

Animal material

The majority of animals used in experiments were captured in the wild in the Pyrenees region of Southern France (Figure 1B). The individual used for the genome assembly was sampled from the locality of Llívia (42°27'N, 1°58'E). Wall lizard samples for whole-genome resequencing came from Angostrina (42°28'N, 1°57'E) and Tor de Querol (42°27'N, 1°53'E). An additional set of samples from Tor de Querol were used for tissue harvesting for phenotypic characterization and gene expression analysis. Permits for sampling and sacrificing these lizards were provided by the Prefecture des Pyrénées Orientales (Arrêté nº 2016-2-09) and the Servei de Biodiversitat i Protecció dels Animals (SF/474). Sampling localities of animals used for amplicon sequencing are listed in SI Appendix, Table S8.

Phenotypic characterization

Microscopy. The ultrastructure, distribution, and relative abundance of chromatophores was studied with light and transmission electron microscopy (TEM) following standard procedures (1). Pieces of skin (ca. 2 mm^2) from focal regions (throat, belly, and ventral tail) of all morphs of *P. muralis* (orange, n=7; white, n=7; yellow, n=4; orange-yellow, n=1; orange-white, n=1), sampled from Llívia (42°27'N, 1°58'E), were excised immediately after sacrifice, placed in Karnovsky's fixative (2% paraformaldehyde, 2.5% glutaraldehyde in 0.1 M PB buffer, pH = 7.3) and stored at 4°C. Samples were washed with 0.1M PB, postfixed with 2% osmium tetra oxide (in 0.1M PB solution), washed with 2% uranyl acetate in 70% ethanol solution, dehydrated in an increasing ethanol series, washed in propylenoxid solution, and transferred to resin (Durcupan, Sigma). Polymerized resin blocks were cut on a Leica

UCT Ultracut ultramicrotome. Ultra-thin sections were put on mesh grids, stained with lead citrate, and both observed and photographed on a FEI Tecnai Spirit G2 TEM equipped with a digital camera (Soft Image System, Morada) and image capture software (ITEM). Magnification at TEM ranged from 1,250 to 43,000x depending on the structures observed. The intensity of the electron beam was adjusted to be in the optimal range for different magnifications.

Biochemical analyses. The carotenoid and pteridine content in the skin of all morphs was determined using chromatographic methods. Samples of integument (ca. 12 mm^2) from the throat and belly of individuals of each morph (orange, n=4; white, n=4; yellow, n=5; orange-yellow, n=4; orange-white, n= 5) were excised, immediately cleaned mechanically to remove muscle and connective tissue, washed briefly with PB to get rid of potential contamination from blood or body fluids, divided in two halves to be analyzed separately for carotenoids and pteridine derivatives, and frozen at -20°C until analyses.

Carotenoids were extracted with 0.5 mL ethyl acetate for 4 days at room temperature in complete darkness. The extracts were then evaporated to dryness by a stream of nitrogen at 27°C and stored at -18°C. Prior to the analyses, samples were diluted in 200µl ethyl acetate. Standards of lutein, zeaxanthin, asthaxanthin, and canthaxanthin were purchased from Sigma Aldrich (Munich, Germany). Stock solutions of the external standards were prepared at a concentration of 0.1 mg/mL by dissolving in ethyl acetate. Carotenoids were determined using UPLC coupled with PDA and MS detector. The UPLC system Dionex Ultimate 3000 (Thermo Fischer, USA) consisted of autosampler, binary pump, and diode-array (PDA) detector. The column (Kinetex C18 RP, 2.6 mm, 150 x 2.1 mm; Phenomenex, USA) maintained at 35°C was used for separation. Acetonitrile (A), methanol/water 1:1 v/v (B) and

a mixture of tert-Butyl methyl ether/acetonitrile/methanol 86:86:8 v/v/v (C) were used as mobile phases for gradient elution with a constant flow rate of 0.2 mL/min. Chromatograms were monitored at 445 and 472 nm. The identity of the carotenoids was confirmed by High Resolution Accurate Mass (HRAM) Q-TOF mass spectrometer (IMPACT II, Bruker Daltonik, Germany) coupled to the UPLC/PDA system. Samples were measured with electrospray (ESI), as well as atmospheric pressure chemical ionization (APCI) in positive mode.

Pteridine derivatives were extracted with dimethyl sulfoxide (DMSO) following a previously published procedure (2). Standards of 6-biopterin, D-neopterin, leucopterin, pterin, pterin-6-carboxylic acid, and riboflavin were purchased from Sigma Aldrich (Munich, Germany). Isoxanthopterin and xanthopterin were obtained from Fluka (Buchs, Switzerland). Drosopterin was prepared following (3). Stock solutions of the external standards were prepared at a concentration of 0.1 mg/mL by dissolving them in DMSO. The working solution of the mixture of all the studied pteridine derivatives in DMSO was prepared at a concentration of 0.01 mg/ml from the stock solutions. All chromatographic measurements were carried out on a HPLC system Agilent series 1290 coupled with a Triple Quad 6460 tandem mass spectrometer (Agilent Technologies, Waldbronn, Germany). For data acquisition, the Mass Hunter Workstation software was used. A ZIC®-HILIC (4.6 × 150 mm, 3.5 μm) column, based on zwitterionic sulfobetaine groups, was chosen (Merck, Darmstadt, Germany). The chromatographic and detection conditions were adapted from (4). The isocratic elution at a flow rate of 0.5 ml/min with the mobile phase consisted of acetonitrile/5mM ammonium acetate buffer, pH 6.80 at a volume ratio of 85:15 (*v/v*), was used for the separation of all the studied compounds with exception of drosopterin. Since drosopterin exhibits higher polarity compared with other pterin compounds, we performed a

run at a volume ratio 55:45 (*v/v*) and flow rate of 0.8 ml/min. The tandem mass spectrometry (MS/MS) measurements were performed in the selected reaction monitoring mode (SRM) with positive ionization. SRM conditions used for MS/MS detection are listed in SI Appendix, Table S6. For compounds marked by an asterisk, the MS/MS conditions were adopted from (4). Mann-Whitney U tests were used to compare the concentration of carotenoids and pterin derivatives between morphs.

Reference genome sequencing and assembly

Long read sequencing and assembly. To generate a reference genome sequence for the common wall lizard, we sequenced a yellow male individual from the Pyrenees region using Pacific Biosciences (PacBio) single molecule real time sequencing (SMRT). The remaining material of this specimen was deposited in the collection of the Natural History Museum from the University of Porto (Museu de História Natural e da Ciência da Universidade do Porto). After dissection, the different tissues were stored at -80ºC until DNA preparation. Pure genomic DNA (10 ug) was obtained from muscle tissue and fragmented to 20 kb using Hydroshear DNA shearing device (Digilab). The sheared fragments were size-selected for 7-50 kb size window on Blue Pippin (Sage Science). The sequencing libraries were prepared following the standard SMRT bell construction protocol and sequenced on 100 PacBio RSII SMRT cells using the P6-C4 chemistry. Raw data was imported into SMRT Analysis software 2.3.0 (PacBio) and filtered for subreads longer than 500 bp or with polymerase read quality above 75. A *de novo* assembly of filtered subreads was generated using *FALCON* assembler version 0.4.0 (5). We used the following configuration file:

input_fofn = input.fofn
input_type = raw
length_cutoff = 8000

length_cutoff_pr = 8000
pa_concurrent_jobs = 24
cns_concurrent_jobs = 36
ovlp_concurrent_jobs = 36
pa_HPCdaligner_option = -v -dal64 -t16 -e.70 -l1000 -s1000
ovlp_HPCdaligner_option = -v -dal64 -t32 -h60 -e.96 -l500 -s1000
pa_DBsplit_option = -x500 -s200
ovlp_DBsplit_option = -x500 -s200
falcon_sense_option = --output_multi --min_idt 0.70 --min_cov 4 --local_match_count_threshold 2 --max_n_read 200 --n_core 8 --output_dformat
overlap_filtering_setting = --max_diff 100 --max_cov 100 --min_cov 1 --bestn 10 --n_core 8

Short read sequencing and assembly correction: To improve the accuracy of the genome sequence generated by *FALCON*, we sequenced the same individual at 30X coverage using Illumina reads. The assembly was corrected using *Pilon* (6) after mapping the short Illumina reads against the contigs obtained from the PacBio sequencing by means of *bwa-mem* (v0.7.5a-r405) (7).

Chicago library preparation and sequencing. Two Chicago libraries from an additional individual sampled from the same locality were prepared by Dovetail Genomics (https://dovetailgenomics.com/) as described by (8). Unlike HiC libraries (see below), Chicago libraries are generated from proximity ligation of chromatin assembled in vitro rather than chromatin obtained from in vivo sources. This reduces confounding biological signal, such as telomeric clustering or chromatin looping. Briefly, for each library, ~500 ng of high molecular weight genomic DNA (mean fragment length = 150 kb) was reconstituted into chromatin *in vitro* and fixed with formaldehyde. Fixed chromatin was digested with *DpnII*, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using

NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina platform. The number and length of read pairs produced for each library was: 122 million, 2x151 bp for library 1; 145 million, 2x151 bp for library 2. When combined, these Chicago library reads provided 437.8X physical coverage of the genome (1-50 kb pairs).

HiC library preparation and sequencing. Two HiC libraries from the same individual used for the Chicago libraries were prepared by Dovetail Genomics as described by (9). Briefly, chromatin was fixed in place with formaldehyde in the nucleus and then extracted, for each library, independently. Fixed chromatin was digested with *DpnII*, the 5' overhangs filled in with biotinylated nucleotides, and then free blunt ends were ligated. After ligation, crosslinks were reversed and the DNA purified from protein. Purified DNA was treated to remove biotin that was not internal to ligated fragments. The DNA was then sheared to ~350 bp mean fragment size and sequencing libraries were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries were sequenced on an Illumina HiSeq platform. The number and length of read pairs produced for each library was: 104 million, 2x151 bp for library 1; 134 million, 2x151 bp for library 2. These HiC library reads provided 11,977.9X physical coverage of the genome (1-50kb pairs).

Scaffolding the assembly with HiRise. The input *de novo* assembly, shotgun reads, Chicago library reads, and HiC library reads, were used as input data for *HiRise* software, a software pipeline designed specifically for using proximity ligation data to scaffold genome

assemblies (8). This analysis was carried out by Dovetail Genomics. An iterative analysis was conducted. First, Shotgun and Chicago library sequences were aligned to the draft input assembly using a modified *SNAP* read mapper (http://snap.cs.berkeley.edu). The separations of Chicago read pairs mapped within draft scaffolds were analyzed by *HiRise* to produce a likelihood model for genomic distance between read pairs, and the model was used to identify and break putative misjoins, to score prospective joins, and make joins above a threshold. After aligning and scaffolding Chicago data, Dovetail HiC library sequences were aligned and scaffolded following the same method. After scaffolding, PacBio shotgun sequences were used to close gaps between contigs. Finally, in order to improve the sequence accuracy of the gapped filled assembly, two rounds of sequence polishing were performed using the *Arrow* consensus calling algorithm (https://github.com/PacificBiosciences/GenomicConsensus). Scaffolds were named as chromosomes from larger to smaller, except for the Z-chromosome (see below).

Sex-chromosome identification: Lacertid lizards such as *P. muralis* are known to possess genetic sex determination (ZZ/ZW) (10). To identify the Z-chromosome, we compared sequence coverage between males and females using DNA pools. We sampled 24 individual females (with a balanced representation of the different morphs) from one of the localities used for the mapping of the color morphs (locality 2, SI Appendix, Table S4 and see section "Whole-genome resequencing of color morphs" below for details on DNA and library preparation). The resulting library was sequenced to an average coverage of 65.5X (SI Appendix, Table S4). To create a pool of male samples we merged all sequence alignment files generated from locality 2 with all five morphs combined. The merged pool of males had

an average coverage of 78.6X. The initial trimming and mapping of the reads to the genome was done as described below for the whole genome resequencing dataset.

Coverage per position for the pools of females and males was calculated using *SAMtools* (v0.1.19-44428cd) (11) *depth*. We removed positions with: i) mapping quality below 40, ii) with coverage lower than half of the genome-wide average, and iii) with coverage higher than the double of the genome-wide average. We then obtained a ratio, both for males and females, of the average coverage for each chromosome divided by the average coverage of the genome.

Quality assessment. We assessed quantitively the completeness of the assembled genome using *BUSCO* (version v3.0.2b) (12,13). Genes contained in *BUSCO* sets for each major lineage are selected from orthologous groups present as single-copy genes in at least 90% of the species. We ran *BUSCO* searches against the tetrapod_odb9 gene dataset.

Genome annotation

Transcriptome sequencing. To obtain empirical information to assist with the genome annotation, we conducted RNA-sequencing (RNA-seq) of five tissues (brain, duodenum, muscle, skin, and testis). The tissues were collected from a third male individual sampled in the same locality as the individual used for the reference genome assembly. After dissection, the tissues were snap frozen in liquid nitrogen and stored at -80ºC until RNA extraction. RNA was extracted using the RNeasy Mini Kit with RNase-Free DNase Set (Qiagen). Prior to library construction, RNA integrity, concentration, and quality were assessed using the Agilent Tapestation 2200, Nanodrop, and Qubit RNA (Thermo Fisher Scientific). RNA libraries were prepared using the Illumina TruSeq RNA Sample Preparation kit following the

manufacturer's instructions. The libraries were sequenced on an Illumina Hiseq1500 using 2x125 bp reads. Statistics summarizing the data are given in SI Appendix, Table S3. RNA-seq data from two embryos at 31-somite stage, incubated at 15 or 24°C, were obtained from a previous publication (14) and combined with the newly generated data.

We employed two complementary approaches to assemble *de novo* the transcriptome of the common wall lizard, which we used alongside other sources of information to annotate the reference genome. First, we generated an assembly of normalized reads using the *Trinity* package (v2.2.0) (15) and default settings. This approach was conducted both for each tissue separately and for a combined read file merging all the data. Second, we mapped the RNA-seq reads to the reference genome sequence by means of the *HISAT2* aligner (16) and then used *Cufflinks* (v2.2.1) (17) to obtain genome-guided transcriptome assemblies for each tissue independently. The individual assemblies were merged into a master transcriptome using the tool *Cuffmerge*.

Protein database. To provide guidance to the putative coding sequence and structure of annotated features in our reference genome, we obtained high-confidence protein sequence evidence from several sources. First, we extracted 551,705 protein sequences from the Uniprot-Swissprot (18) reference data set (downloaded on 2016-08). This non-redundant collection contains only manually annotated and reviewed proteins. Second, we queried the Uniprot database for protein sequences belonging to the family Lacertidae and extracted a total of 4,794 sequences. Finally, we queried the same Uniprot database and obtained 19,334 protein sequences from the species *Anolis carolinensis*, which is the most comprehensively annotated lizard genome.

Repeat masking. We started by creating a species-specific library of repeats by means of the software *RepeatModeler* (http://www.repeatmasker.org/RepeatModeler/). As repeats can be part of actual protein-coding genes, the candidate repeats modelled by *RepeatModeler* were vetted against our proteins set (minus transposons) to exclude any nucleotide motif stemming from low-complexity coding sequences. Based on the created library, the identification of repeat sequences across the genome was performed using *RepeatMasker* (v4.0.3) (19) and *RepeatRunner* (http://www.yandell-lab.org/software/repeatrunner.html). *RepeatRunner* is a program that integrates *RepeatMasker* results with protein-based information from *BLASTX*. It improves the efficacy of repeat identification by identifying highly divergent repeats, or portions of repeats. It also helps identifying divergent protein coding portions of retro-elements and retro-viruses.

Ab-initio training. We opted for an *ab-initio* based annotation strategy combining gene predictions with the available evidence data. The use of multiple gene finders in general improves the genome annotations, therefore gene predictions were computed in several complementary ways. We used *GeneMark* as *ab-initio* predictor due to its effective training on fungal and eukaryote genomes (20,21). We used the algorithm *GeneMark-ES_ET* (v4.3), which integrates information from mapped RNA-seq into the training process. Additionally, we used *AUGUSTUS* (v2.7) (22) and *SNAP* (23). Both algorithms were trained by means of a profile created by running the pipeline *MAKER* (v3.00.0) (24) one time based on the protein evidence described above and the RNA-seq transcript information generated for several tissues. From this gene build, we created a training gene set by selecting the best gene models based on the following criteria: 1) the genes had to be complete (i.e. start/stop codons mandatory); 2) no similarity over 85% was allowed among genes; 3) AED scores (Annotation

Edit Distance) had to be lower than 0.1; and 4) genes had to be at a minimum distance of 1,000 bp from each other. In total, 2,828 genes were selected for the *AUGUSTUS* training process.

Gene build. High-confidence gene models were computed using the *MAKER* software by combining evidence-based data (protein homology, transcripts, and repeats) with *ab-initio* profiles. An evidence-guided build was computed by allowing the *MAKER* software to construct gene models directly from both aligned transcript sequences and reference proteins. Evidence builds generally closely reflect the information provided by the available sequence data and try to synthesize consensus transcript structures. However, this approach is vulnerable to missing or incomplete sequence material, as well as the noise level into the transcriptome data.

The *ab-initio* evidence driven build was based on the initial evidence annotation by using *MAKER* alignments, together with specifically trained *AUGUSTUS*, *SNAP* and *GeneMark ab-initio* profiles. The aim of this approach was to perform a second run with all gene models and to replace any gene locus where a longer putative CDS could be predicted by the gene finder or fill in gene predictions where sufficient evidence was lacking from the construction of evidence models.

Statistical evaluation of the final annotation was performed with an in-house (SciLife) Perl script. To improve the annotation, we also ran an in-house script to improve the open reading frame (ORF) start and end, since we noticed that a non-negligible proportion of genes were fragmented and missing from the BUSCO post-annotation validation

ncRNA annotation. In addition to protein-coding genes, we also performed an annotation of ncRNA. tRNA were predicted through tRNAscan (v1.3.1) (25). For annotation of other types of broadly conserved ncRNA, we used as the main source of information the RNA family database *Rfam* (v11) (26). *Rfam* provides curated co-variance models, which can be used together with the *Infernal* package (27) to predict ncRNAs in genomic sequences. The set of co-variance profiles was limited to only include broadly conserved, eukaryotic ncRNA families.

Functional annotation and gene name inference. Functional inference for genes and transcripts was performed using the translated CDS features of each coding transcript. Each predicted protein sequence was blasted against the Uniprot-Swissprot reference data set in order to retrieve the gene name and the protein function. Each predicted sequence was also blasted against *InterProscan* (v5.7-48) (28) in order to retrieve *Interpro* (29), *Pfam* (30), *GO* (31), *MetaCyc* (32), *KEGG* (33) and *Reactome* (34) information. Finally, using the output from those analyses and the *ANNIE* annotation tool (http://genomeannotation.github.io/annie), we extracted and reconciled relevant meta data into predictions for canonical protein names and functional predictions.

Whole-genome resequencing of color morphs

Sampling strategy. Genome-wide polymorphism data was obtained using whole-genome resequencing of DNA pools. We sampled male individuals of the five color morphs at two localities in the eastern Pyrenees (Angostrina and Tor de Querol; SI Appendix, Table S4). We also sampled an additional set of wall lizards from a location in Northern Italy (45°13'N, 9°13'E) to serve as an outgroup. For each lizard we removed the tip of the tail for genomic

DNA extraction, which was performed using the EasySpin kit (Citomed) followed by a RNAse A treatment. DNA quality and concentration were assessed using a NanoDrop spectrophotometer, agarose gels, and the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific). Individuals were pooled by morph and locality in equimolar concentrations for sequencing. The number of individuals included in each pool varied from 9 to 21.

Whole-genome resequencing and quality control. Each pool was sequenced to an effective coverage of 15-18-fold using 2x125 bp reads on an Illumina Hi-seq 1500 (SI Appendix, Table S4). The read data files were initially checked for sequence quality statistics with *FastQC* (v1.7.119) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). Next, we trimmed the reads to remove adapters, low-quality bases, and flanking regions of each read with *Trimmomatic* (v0.35) (35) using the following parameters: *TRAILING*=15, *SLIDINGWINDOW*=4:20, *MINLEN*=30. Trimmed sequence files were reanalyzed with *FastQC* before proceeding with further analyses.

Read mapping and SNP calling. Reads were mapped to our *de novo* reference genome assembly (PodMur1.0) using *bwa-mem* with default settings. Alignment and coverage statistics (SI Appendix, Table S4) were calculated with *SAMtools* and custom scripts. To reduce noise in downstream analyses, we eliminated reads with mapping quality lower than 40 and with overlapping portions between forward and reverse read pairs. This prevented cases in which an allele from an overlapped portion of the same molecule could be considered twice.

These filtered mapped reads were used for variant calling using *FreeBayes* (v1.0.2-33-gdbb6160) (36). *FreeBayes* is a variant detection method that, instead of a traditional

alignment-based single-position approach, uses a Bayesian framework to reconstruct haplotypes and call small variants such as SNPs, indels, and other complex variants. *FreeBayes* was run with the 10 pools from both localities using default parameters and setting the ploidy to twice the number of diploid individuals for each pool. Compound variants were decomposed into single variants using the program *decompose-blocksub* from the *vt* tool set (37).

Population genomics

Genetic structure. To infer patterns of population structure among sampled localities and among color morphs we constructed a neighbor-joining tree. We excluded positions with less than 15 reads in any of the pools, followed by a random down sampling of reads in all positions to a maximum coverage of 15X. These filtering steps ensure that low coverage positions are not incorporated in the analysis and that differences in sequencing depth among pools are not biasing estimates of genetic differentiation. These steps were performed using *SAMtools* and *PoPoolation2* (38). From the remaining set of variants, we randomly chose 250,000 SNPs to calculate Nei's standard genetic distance (39) between each pair of pools based on allele frequencies estimated from allele counts. We then used the distance matrix as input to generate a Neighbor-Joining tree in the *neighbor* program in *PHYLIP* (v3.696) (40).

Nucleotide variation and differentiation. To compare levels and patterns of genetic diversity between localities and morphs, we summarized the allele frequency spectrum using Tajima's *D* (41) and genetic diversity using nucleotide diversity (π) (42). Genetic differentiation was

summarized using F_{ST}. As before, reads with a mapping quality lower than 40 and bases with a sequencing quality lower than 20 were not considered. Both π and Tajima's D were calculated in non-overlapping windows of 10,000 bp using *PoPoolation* (43). The same window strategy was used for F_{ST} calculated by means of *PoPoolation2*. We restricted this analysis to positions with a minimum coverage of 2/3 and a maximum coverage of twice the average coverage of each pool. A genome-wide estimate for each pool was obtained by averaging the values for all windows.

Genetic mapping. To test for the association between each color morph and specific genomic regions, we calculated allele frequency differentiation at each previously identified SNP between each relevant comparison. We filtered out SNPs that had coverage higher than two times the average coverage of each pool in at least one of the pools and restricted the minimum coverage to 5x (SNPs with less than this coverage in at least one of the pools were removed). The *PoPoolation2* software was then used to calculate pairwise allele frequency differences between all the pools in each population, for each of these main contrasts based on the phenotypic data:

1. Presence or absence of orange pigmentation: pure orange and mosaic morphs vs. pure white and pure yellow morphs
2. Presence or absence of yellow pigmentation: pure yellow and mosaic yellow morphs vs. white and mosaic-white morphs. For this contrast we removed the pure orange pools – this was done to account for the possibility that the stronger orange coloration could be masking the yellow coloration.

3. Mosaic patterning or uniform coloration: mosaic morphs vs. pure orange morphs.

For each SNP we calculated the median value of allele frequency difference (ΔAF) for each relevant comparison. To reduce the possibility of detecting false positive SNPs in association with each phenotype, we excluded positions that were within 5 bp at both directions from an identified indel using *PoPoolation2* and removed positions that had a minor allele frequency lower than 0.1 in the whole dataset. We then used a sliding window approach (20 SNPs and steps of 5 SNPs) to identify regions with consistent differentiation across many SNPs and avoid spurious associations from individual SNPs derived from erroneous read mapping or stochasticity in allele frequency estimates from pool sequencing.

To estimate whether windows displayed higher allele frequency differentiation than expected by chance, we conducted for each contrast 1000 permutation tests in which we randomized the median allele frequency difference but maintained the genomic positions as in original data. We then applied the same sliding approach and recorded the top value for each permutation. Using this approach, we considered a significant allele frequency difference a value in the original data that was higher than the top 1% of the obtained distribution by permutation. A SNP of high delta within the top windows in each contrast was genotyped by Sanger sequencing, including for the mosaic phenotype where no windows had ΔAF values above the significance threshold.

We further conducted a Cochran-Mantel-Haenszel (CMH) test (44,45). CMH is a repeated test of independence, which assesses consistent allele frequency changes across biological or technical replicates. We used the same filtering steps as for the ΔAF analysis (minimum 5x coverage, maximum double coverage, removal of SNPs within 5 bp around indels, removal of SNPs with minor allele frequency below 0.1), and conducted the test in

PoPoolation2. In order not to violate the assumption of independence of the method, several replicates for each association were performed, and in each replicate a given pool was used only once:

Presence or absence of orange pigmentation:

- orange vs. white + orange-white vs. yellow
- orange vs. white + orange-yellow vs. yellow
- orange-white vs. white + orange-yellow vs. yellow
- orange-white vs. white + orange vs. yellow
- orange-white vs. orange-yellow + orange vs. yellow
- orange-yellow vs. white + orange-white vs. yellow

Presence or absence of yellow pigmentation:

- yellow vs. orange-white + orange-yellow vs. white
- orange-yellow vs. orange-white + yellow vs. white

Mosaic patterning:

- orange vs. orange-white
- orange vs. orange-yellow

To combine the results for each association test, we summarized *p*-values using the Fisher's combined probability test (46), and performed the same sliding window analysis as for the ΔAF analysis (20 SNPs and steps of 5 SNPs).

Confirmation of association by individual genotyping

We confirmed the genotype-phenotype associations in individuals from the Pyrenees through individual-based genotyping using Sanger sequencing. We designed primers to amplify a fragment overlapping the regions associated with phenotypes upstream of *SPR* and *BCO2* (SI Appendix, Table S7). Within each fragment, we tested the association using a single variant showing high allele frequency differentiation between morphs (orange vs. non-orange; yellow vs. white). For the orange locus that variant was a 38-bp insertion relative to the reference (chr9:77,999,995-77,999,996 bp), and for the yellow locus it was a SNP (chr15:26,161,682 bp). A significant association between these variants and their respective phenotypes was tested using the R package *SNPassoc (*v1.9-2) (47). These same variants were genotyped in additional individuals belonging to other sub-lineages of the common wall lizard. We also designed primers to amplify an amplicon around the structural variants upstream of *SPR* in 15 orange individuals (SI Appendix, Table S7).

Nanopore sequencing

DNA-preparation. To investigate structural variation around the candidate loci, we sampled skeletal muscle from an orange-white individual (from locality 2) and extracted high-molecular weight DNA using a modified phenol-chloroform extraction (https://www.protocols.io/view/ultra-long-read-sequencing-protocol-for-rad004-mrxc57n). We ground the tissue with a pre-frozen mortar and pestle over dry ice and incubated the tissue for 5 h in TE with proteinase K and RNase A. After digestion, an equal volume of phenol-chloroform-isoamyl alcohol (25:24:1) was added, mixed gently for 10 min and centrifuged at 4,500 rpm for additional 10 min. This step was repeated a second time using an equal volume of chloroform-isoamyl alcohol (24:1). The aqueous phase was mixed with 1:10

volume of ammonium acetate (5M) and 2:1 of 100% ice-cold ethanol and mixed gently. DNA was then physically removed from the solution and placed in a tube with 70% ethanol, which was centrifuged at 10,000 g. The washed DNA pellet was dried and then dissolved in 1x TE buffer. Dissolved DNA was subjected to mechanical shearing to 15 kb and 25 kb fractions using a Megaruptor2 instrument (Diagenode).

Oxford Nanopore library generation and sequencing. The sheared DNA-fractions were cleaned up using Ampure XP beads (Beckman Coulter) and the DNA was eluted in nuclease-free water. We produced two Oxford Nanopore libraries for sequencing in a MinION device, starting from 7.6 µg of DNA each (15 kb and 25 kb fractions, separately) using the SQK-LSK108 kit, with DNA-repair and DNA end-prep combined according to the one-pot protocol (https://www.protocols.io/view/one-pot-ligation-protocol-for-oxford-nanopore-libr-k9acz2e). After the combined FFPE DNA repair/end repair step we cleaned the repaired/end-prepped DNA using Ampure XP beads and eluted in nuclease-free water. Adapter ligation, and all following library preparation steps, was carried out using the official protocol (1D Genomic DNA by ligation) from Oxford Nanopore (www.nanoporetech.com). We loaded 0.8 and 2 µg, onto two different FLO-MIN106 (r9.4.1) flow cells for the 15 kb and 25 kb libraries, respectively. Sequencing was run for 48 h and we obtained 2,320,498 sequence reads with an average length of 8,730 bp, resulting in an average genome-wide coverage of 9.7X (84.9% of the total reads mapped against the reference genome). The resulting raw data fast5 files were base-called using *guppy* (v0.3.0) (Oxford Nanopore).

Analysis of long read sequencing data. The resulting sequence files were subjected to adapter trimming using *downpore* (v0.2) (J. Teutenberg, downloaded from:

https://github.com/jteutenberg/downpore). Trimmed files were mapped to our reference assembly using *NGMLR* (v0.2.6) (48) with default parameters. Since our objective was to investigate structural variation in the regions associated with pigmentation, we retrieved reads that mapped within and around *SPR* and *BCO2*. We manually realigned the reads in BioEdit (v7.2.5) (49) to construct a consensus sequence based on the plurality rule (i.e. for each position we obtained the most frequent state, even if not found in the majority of sequences), using a script by Joseph Hughes (University of Glasgow, downloaded from https://github.com/josephhughes) and aligned the consensus to the reference genome sequence. The results were visualized using *Jalview* (v2.10.4b1) (50).

Annotation of SNPs and indels

Variant annotation was performed using the genetic variant annotation and effect prediction toolbox *SnpEff* (v4.3) (51). We specifically searched for SNP and indel variants with potential functional significance around the candidate regions and that have the potential to alter protein structure, such as frameshift, nonsynonymous, stop, and splice site mutations.

Gene expression and allelic imbalance analysis

Quantitative real-time PCR analyses of gene expression. Gene expression was quantified using quantitative PCR. We sampled 26 individuals belonging to all five morphs (six white, six orange, five yellow, five orange-white and four orange-yellow) from Llívia (42°27'N, 1°58'E). After dissection, several organs (skin, brain, liver and muscle) were harvested, snap-frozen in liquid nitrogen, and stored at -80ºC until RNA extraction. RNA was extracted using the RNeasy Mini Kit (Qiagen). cDNA was generated by reverse transcribing ~1 μg of RNA using the GRS cDNA Synthesis Kit (GRiSP) following the manufacturer's protocols. For the

analysis on the skin, we focused only on the pure morph animals, to avoid possible variation in expression across skin patches of mosaic animals.

We designed primers located on different exons to minimize amplification from contaminant genomic DNA for the following genes, *SPR*, *BCO2*, *PTS*, and *18S* (housekeeping gene used for expression normalization across samples) (SI Appendix, Table S7). Three independent assays were performed for each biological replicate on a CFX96 Touch Real-Time PCR Detection System using iTaq Universal SYBR Green Supermix (Bio-Rad Laboratories). Cq (quantification cycle) values of the three technical replicates were then averaged and the expression of each focal gene for each sample was normalized to the expression of *18S* using a -ΔCq approach (52). Finally, we compared the normalized expression of each gene between color morphs in a pairwise manner using the Mann-Whitney U test. For the test of *PTS* expression on the skin, one individual of the white morph was excluded, as it showed only residual PCR amplification possibly explained by a mutation in the primer binding site.

Quantification of allelic imbalance: We also assessed levels of gene expression by analyzing allele imbalance. Since the association does not extend to the coding region in *SPR* and weakly so in *BCO2*, we could not examine a diagnostic variant between pigmentation alleles to quantify allelic expression. Thus, we cannot deduce which specific allele is preferentially expressed at these loci due to the lack of linkage disequilibrium between the non-coding sequences showing an association and SNPs located in the transcript. However, if the color polymorphisms are controlled by cis-regulatory mutations affecting the expression of *SPR* and/or *BCO2*, we expect to see allelic imbalance in expression in individuals that are heterozygous *O/o* and *Y/y*. For all individuals that among our cohort of samples used for the

qPCR analysis were heterozygous *O/o* and *Y/y*, we screened part of the coding sequence to detect polymorphisms that could be used to quantify the relative expression of each allele. We then designed primers to amplify small fragments from cDNA overlapping these polymorphisms. Primers were located on exon-exon junctions to minimize amplification from genomic DNA and were carefully placed to avoid overlapping existent polymorphisms (SI Appendix, Table S7).

The allelic expression was quantified by sequencing on a MiSeq (MiSeq v3 600-cycle kit, 2x300 bp reads). Amplification of the cDNA template with 5' labeled primers and pre-processing of the reads prior to mapping to the genome was done with the same protocol as described in more detail below ("*Amplicon sequencing overlapping regions of association*"). To calculate the relative proportion of alleles expressed in the skin of each individual for each transcript, we counted the number of reads corresponding to the reference and alternative alleles. Finally, allelic imbalance was tested using chi-square tests (a proportion of 1:1 was used as the null hypothesis).

Amplicon sequencing overlapping regions of association

Sampling and PCR amplification: We sequenced amplicons (~550 bp) overlapping the regions of association for the two loci recovered in the genome-wide association mapping. We sequenced a set of samples (n=48) that included common wall lizard samples from our primary study location in the eastern Pyrenees (n=16) and other *Podarcis* species (n=32) that are known to present ventral color polymorphism (further details can be found on SI Appendix, Table S8). Genomic DNA was extracted from tail-tip samples using the EasySpin kit (Citomed) followed by a RNAse A treatment.

Amplification of the target regions was done via a two-step PCR protocol based on (53). Briefly, a first PCR with 5'-tailed primers served to amplify the target region, followed by a second PCR to attach barcoding sequences to the amplified DNA (SI Appendix, Table S7). The first PCR reaction was prepared with approximately 25 ng DNA, 5 μL 2x Qiagen MasterMix, 0.4 μL of 10 pM of each primer and 3.2 μL PCR-grade water, and was run under the following conditions: 1) an initial denaturing step of 95ºC for 15 min; 2) 5 touch-down cycles with 95ºC denaturing for 30 s, a 68-64ºC annealing temperature touchdown for 30 s and 72ºC extension temperature for 45 s; 3) 35 cycles with 95ºC denaturing for 30 s, a 64ºC annealing step for 30 s and 72ºC extension for 45 s; 4) a final extension at 60ºC during 20 min. Before conducting the second PCR we purified the resulting DNA with a standard bead cleaning protocol (a 0.7:1 bead-to-sample volume ratio was used). We set up the second (barcoding) PCR reaction using 2 μL of purified PCR product, 5 μL 2x Qiagen MasterMix, 1 μL of a mix of individually labeled primers with P5/P7 binding sites and 1 μL of PCR-grade water. The following program was used for the barcoding PCR: 1) an initial denaturing step of 95ºC for 15 min; 2) 10 cycles with 95ºC denaturing for 5 s, a 55ºC annealing temperature step for 20 s and a 72ºC extension for 45 s; 3) a final extension at 60ºC during 20 min. After a second bead cleaning step, all samples were pooled at equimolar concentrations. The pooled library was loaded onto a MiSeq (MiSeq v3 600-cycle kit, 2x300 bp reads) at a concentration of 13.5 pM.

Sequence processing and haplotype inference: After demultiplexing the raw reads, we removed low quality reads, low quality bases, and primer sequences with *Trimmomatic* using the following parameters: *TRAILING*:15, *HEADCROP*:22, *SLIDINGWINDOW*:4:20, *MINLEN*:30. Since the sequence length of paired-end reads combined was larger than the

fragment size, paired-end reads were merged into a single sequence with the *fastq-join* utility (54). To filter out non-target amplified DNA, we mapped the new merged reads using *bwa-mem* with default parameters to a small reference sequence containing the two target regions and approximately 100 bp around each locus. The reads that mapped to each of the two regions were then reconverted into separate fastq files and the fastqs were randomly down-sampled to 10% of the total number of reads to decrease computational times in subsequent analyses. Both approaches were carried out with *picard* (http://broadinstitute.github.io/picard).

Given that the divergence between haplotypes was extremely high (see main manuscript), traditional mappers and SNP callers produced erroneous datasets. Therefore, to obtain variant and haplotype information from each individual/locus we used a custom python script (available upon request). Briefly, the script begins by defining an average sequence similarity threshold between a subset of all unique reads present in the dataset (reads are mapped pairwise using dynamic programming; 10% of reads were used by default). If the average sequence similarity within the subset exceeded 99.5% (equivalent to only one or two mutations in a 500 bp amplicon), this was used as the reference value. After defining the threshold, all unique reads from the dataset appearing at least five times were sorted by frequency of occurrence. Starting from the most frequently occurring unique read (which is considered one of the haplotypes), the program sequentially maps the following unique reads to the first read using dynamic programming with two possible outcomes: 1) if the sequence similarity between the two reads exceeds the average threshold that was previously estimated (indicating a very similar sequence), the second read is discarded and assigned to the haplotype; 2) if the sequence similarity between the two reads is lower than the average threshold that was previously estimated (indicating a divergent sequence), the

second read is kept as a new haplotype for the following pairwise comparisons. This approach led to most samples either having one or two representative sequences. For ambiguous samples with three or more estimated haplotypes, we manually curated the result by analyzing the mapping files using *IGV2.4.10* (http://www.broadinstitute.org/igv). Finally, we manually realigned the sequences in *BioEdit* (v7.2.5) to build our final dataset.

Intra- and inter-specific haplotype networks: We used *popart* (v1.7) (55) to construct and plot median-joining haplotype networks and calculated values of d_{xy} and Tajima's *D* for each locus using *DnaSP* (v6.11.01) (56). To explore patterns of sequence evolution in a broader phylogenetic context, we constructed neighbor-net networks using *SplitsTree* (v4.14.2) (57) using an extended dataset with additional sequences from other *Podarcis* species. A neighbor-net approach was used because (unlike traditional phylogenetic trees) it considers processes like recombination and hybridization. To simplify the visual representation of the phylogenetic networks we used the following criteria to subsample the alignment dataset: 1) since the initial within-population median-joining network analysis detected 2-3 highly divergent haplogroups, we randomly selected one sequence from each haplogroup to represent *P. muralis*; 2) from each of the other species (for each tree, orange locus and yellow locus) we randomly selected two sequences from individuals with the relevant pigmentation (orange or yellow) and two sequences from individuals without the pigment. We checked for potential biases introduced by this subsampling by repeating the analysis with other sequences multiple times, and confirmed that the general pattern remained qualitatively unchanged. Indels were not considered in both analyses.

Haplotype divergence: We estimated pairwise sequence divergence between haplotypes of the *SPR* and *BCO2* amplicons within *P. muralis* and between *P. muralis* and six additional *Podarcis* species. In addition, we performed the same calculations for 31 amplicons randomly distributed in the genome (SI Appendix, Table S9). These amplicons ranged from 255 bp to 667 bp (average = 410 bp) (58-63). Part of the data was obtained from a previous publication (64), but we PCR and Sanger sequenced some amplicons to complete the data for the six additional species. PCR protocols were carried out according to (64).

REFERENCES FOR SI

1. Haisten DC, Paranjpe D, Loveridge S, Sinervo B (2015) The cellular basis of polymorphic coloration in common side-blotched lizards, *Uta stansburiana*. *Herpetologica* 71:125-135.
2. Krajíček J, et al. (2014) Capillary electrophoresis of pterin derivatives responsible for the warning coloration of Heteroptera. *J Chromatogr A* 1336:94-100.
3. Ferré J, Silva FJ, Real MD, Ménsua JL (1986) Pigment patterns in mutants affecting the biosynthesis of pteridines and xanthommatin in *Drosophila melanogaster*. *Biochem Genet* 24:545-569.
4. Kozlík P, et al. (2013) Hydrophilic interaction liquid chromatography with tandem mass spectrometric detection applied for analysis of pteridines in two Graphosoma species (Insecta: Heteroptera). *J Chromatogr B* 930:82-89.
5. Chin CS, et al. (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 13:1050-1054.
6. Walker BJ, et al. (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963.
7. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.
8. Putnam NH, et al. (2016) Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* 26:342-350.

9. Lieberman-Aiden E, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326:289-293.
10. Rovatsos M, et al. (2016) Conservation of sex chromosomes in lacertid lizards. Mol Ecol *25:*3120-3126.
11. Li, H *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.
12. Simão FM, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210-3212.
13. Waterhouse RM, et al. (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 35:543-548.
14. Feiner N, Rago A, While GM, Uller T (2018) Signatures of selection in embryonic transcriptomes of lizards adapting in parallel to cool climate. *Evolution* 72:67-81.
15. Grabherr MG, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644-652.
16. Kim D, Langmead B, Salzberg SL (2015) HISAT: A fast spliced aligner with low memory requirements. *Nat Methods* 12:357-360.
17. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7:562-578.
18. Magrane M, UP Consortium (2011) UniProt Knowledgebase: A hub of integrated protein data. *Database* 2011:bar009.
19. Tarailo-Graovac M, Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinforma* 2009:UNIT 4.10.
20. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* 18:1979-1990.
21. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 33:6494-6506.
22. Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19:ii215-ii225.

23. Korf I (2004) Gene finding in novel genomes. *BMC Bioinformatics* 5:59.
24. Holt C, Yandel M (2011) MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491.
25. Lowe TM, Eddy SR (1996) TRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955-964.
26. Nawrocki EP, et al. (2015) Rfam 12.0: Updates to the RNA families database. *Nucleic Acids Res* 43:D130-D127.
27. Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29:2933-2935.
28. Jones P, et al. (2014) InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30:1236-1240.
29. Hunter S, et al. (2012) InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Res* 40:D306-D312.
30. Finn RD, et al. (2013) Pfam: The protein families database. *Nucleic Acids Res* 42:D222-D230.
31. Ashburner M, et al. (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25:25-29.
32. Caspi R, et al. (2018) The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 46:D633-D639.
33. Kanehisa M, et al. (2014) Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res* 42:D199-D205.
34. Fabregat A, et al. (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res* 46:D649-D655.
35. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114-2120.
36. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv* 2012:1207.3907v2.
37. Tan A, Abecasis GR, Kang HM (2015) Unified representation of genetic variants. *Bioinformatics* 31:2202-2204.
38. Kofler R, Pandey RV, Schlötterer C (2011) PoPoolation2: Identifying differentiation

between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* 27:3435-3436.

39. Nei M (1972) Genetic Distance between Populations. *Am Nat* 106:283-292.
40. Felsenstein J (1989) PHYLIP - Phylogeny inference package. *Cladistics* 5:164-166.
41. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.
42. M Nei, *Molecular evolutionary genetics* (Columbia University Press, New York, 1987).
43. Kofler R, et al. (2011) Popoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* 6:e15925.
44. Cochran WG (1954) Some methods for strengthening the common χ 2 tests. *Biometrics* 10:417-451.
45. Mantel N, Haenszel W (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 22:719-748.
46. Fisher R (1925) *Statistical methods for research workers*. (Oliver & Boyd, Edinburgh).
47. González JR, et al. (2007) SNPassoc: An R package to perform whole genome association studies. *Bioinformatics* 23:654-655.
48. Sedlazeck FJ, et al. (2018) Accurate detection of complex structural variations using single molecule sequencing. *Nat Met* 15:461-468.
49. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* 41:95-98.
50. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25:1189-1191.
51. Cingolani P, et al. (2014) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6:80-92.
52. Schmittgen TD, Livak KJ (2008) Analyzing real-time PCR data by the comparative CT method. *Nat Protoc* 3:1101-1108.
53. McInnes JC, et al. (2017) Optimised scat collection protocols for dietary DNA metabarcoding in vertebrates. *Methods Ecol Evol* 8:192-202.

54. Aronesty E (2013) Comparison of Sequencing Utility Programs. *Open Bioinforma J* 7:1-8.
55. Leigh JW, Bryan D (2015) POPART: Full-feature software for haplotype network construction. *Methods Ecol Evol* 6:1110-1116.
56. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol Biol Evol* 34:3299-3302.
57. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23:254-267.
58. Salvi D, Harris DJ, Kaliontzopoulou A, Carretero MA, Pinho C (2013) Persistence across Pleistocene ice ages in Mediterranean and extra-Mediterranean refugia: Phylogeographic insights from the common wall lizard. BMC *Evol Biol* 13:147.
59. Chiari Y, et al. (2004) New evidence for parallel evolution of colour patterns in Malagasy poison frogs (*Mantella*) *Mol Ecol* 13:3763-3774.
60. Gamble T, Bauer AM, Greenbaum E, Jackman TR (2008) Evidence for Gondwanan vicariance in an ancient clade of gecko lizards. *J Biogeogr* 35:88-104.
61. Godinho R, Domingues V, Crespo EG, Ferrand N (2006) Extensive intraspecific polymorphism detected by SSCP at the nuclear C-mos gene in the endemic Iberian lizard *Lacerta schreiberi*. *Mol Ecol* 15:731-738.
62. Pinho C, Harris DJ, Ferrand N. (2008) Non-equilibrium estimates of gene flow inferred from nuclear genealogies suggest that Iberian and North African wall lizards (*Podarcis* spp.) are an assemblage of incipient species. *BMC Evol Biol* 8:63.
63. Pinho C, et al. (2010) New primers for the amplification and sequencing of nuclear loci in a taxonomically wide set of reptiles and amphibians. *Conserv Genet Resour* 2:181-185.
64. Pereira C, et al. (2013) Twenty-one new sequence markers for population genetics, species delimitation and phylogenetics in wall lizards (*Podarcis* spp.) *BMC Res Notes* 6:299.

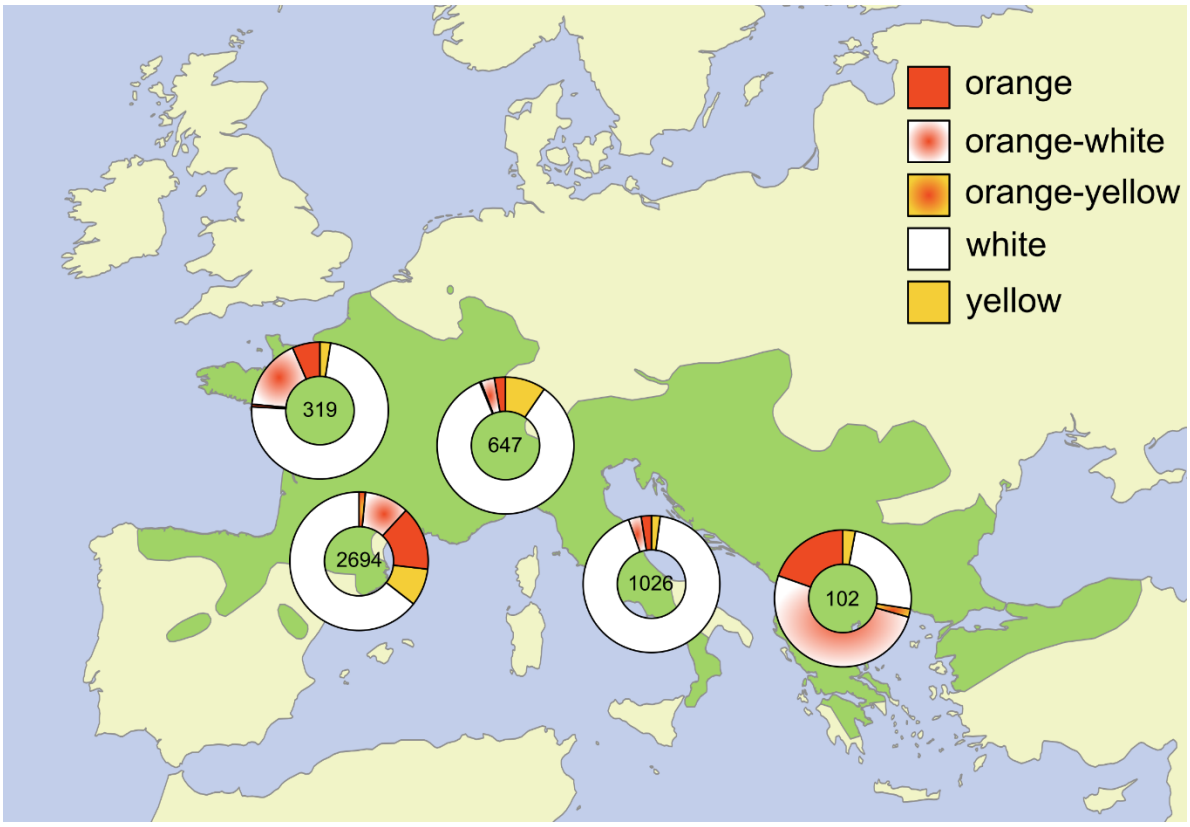

Figure S1. Distribution of the common wall lizard (*Podarcis muralis*) in Europe (green shading), with information on morph frequencies in several locations, showing the wide distribution of this polymorphism across the species' distribution. Morph frequencies vary at a regional level as well as on a micro-geographic scale. The number of individuals sampled in each locality is indicated.

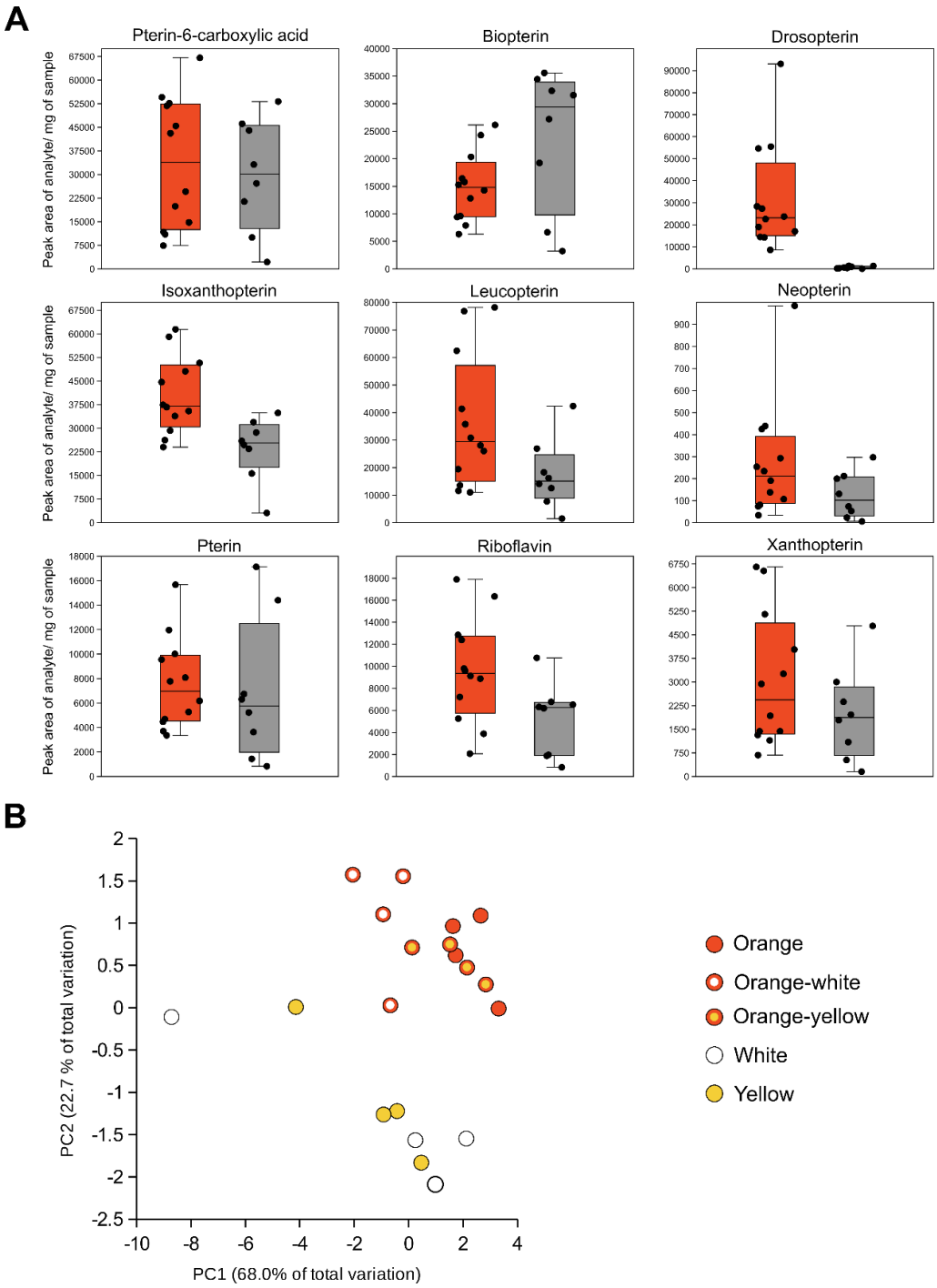

Figure S2. Quantification of pterin compounds (HPLC-MS/MS) in animals expressing orange, yellow and white coloration. **(A)** Results for individual pterin compounds. Orange, orange-white, and orange-yellow individuals are collapsed in a single group (orange), and the same for individuals of the white and yellow morphs (grey). **(B)** Principal component analysis using all pterins as variables, indicating that individuals expressing orange coloration (orange and both mosaic morphs) differ from white and yellow individuals in their overall pterin profile.

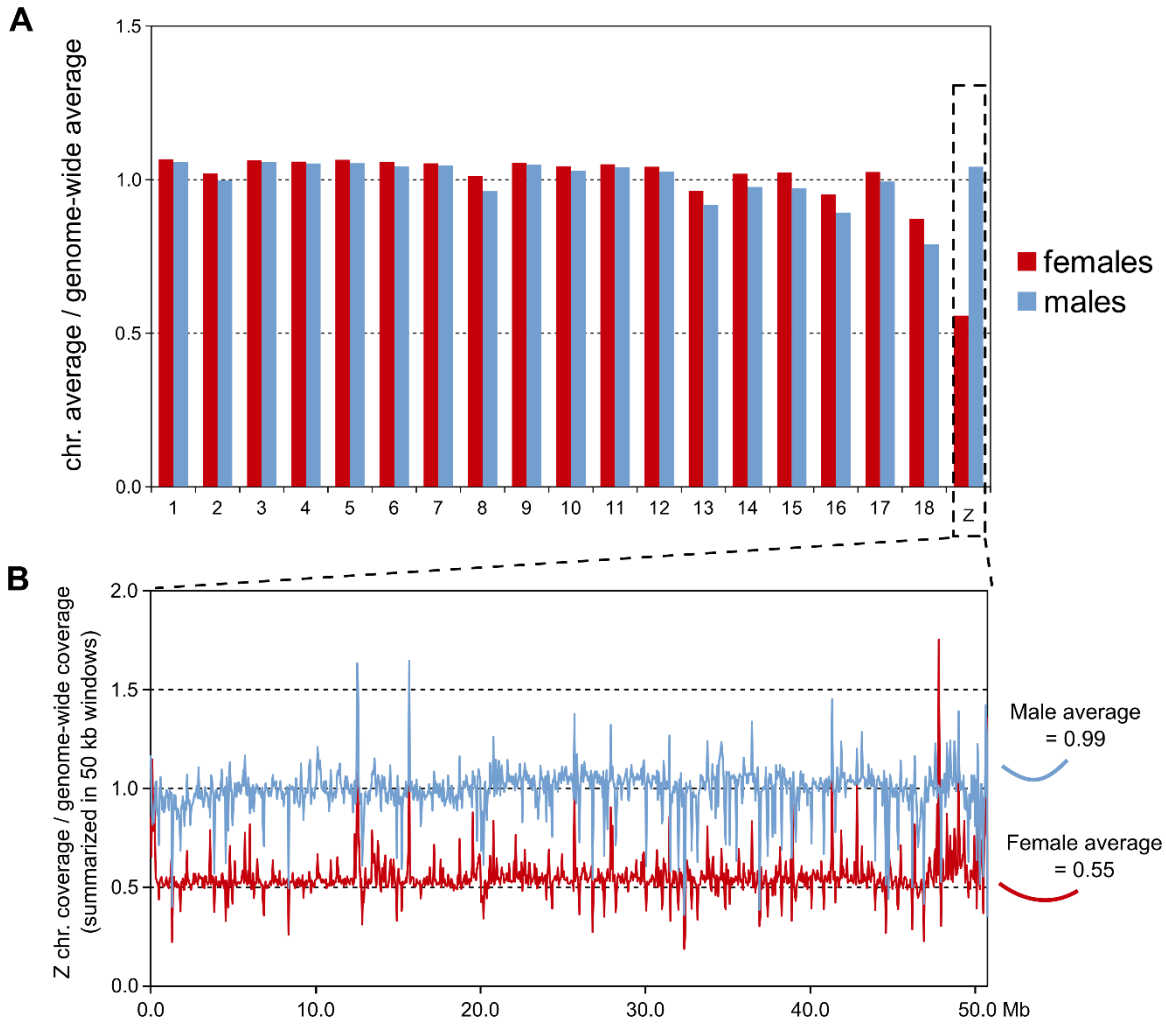

Figure S3. Identification of the Z chromosome in the common wall lizard *de novo* reference assembly using pool-sequencing data of females (ZW) and males (ZZ). **(A)** Ratio between the average coverage of each of the 19 chromosomes and the average genome-wide coverage. The average was calculated per position. Consistent with a ZW sex determination, coverage for one scaffold, that we named Z, was reduced by half in females when compared to males. **(B)** The ratio between the average number of reads along the candidate Z chromosome and the average number of reads genome-wide. The average coverage was summarized in 50 kb windows.

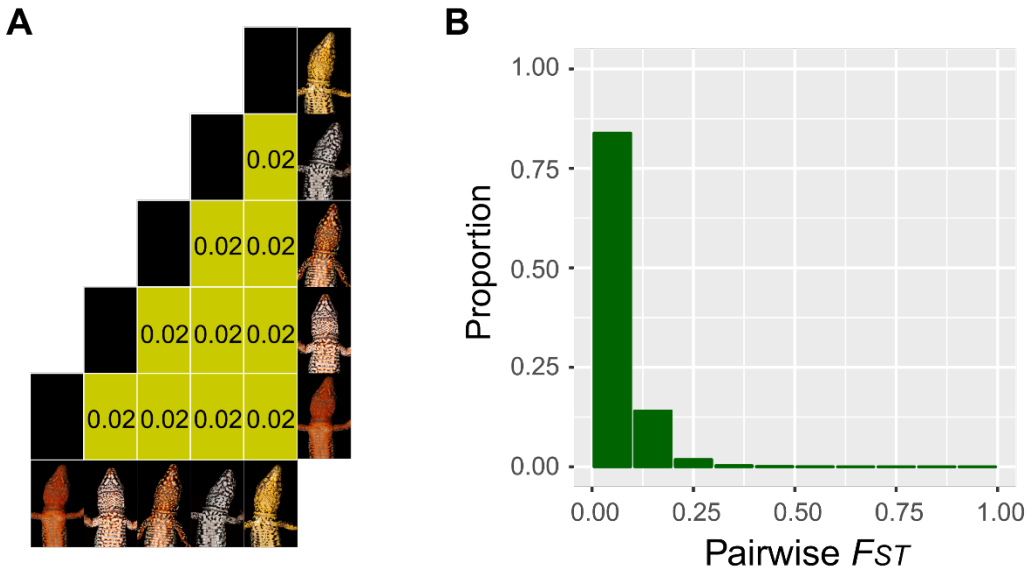

Figure S4. Genetic differentiation among the five color morphs of the common wall lizard calculated across the genome in non-overlapping 10 kb windows. **(A)** Pairwise F_{ST} values between the five color morphs. **(B)** Distribution of pairwise F_{ST} values between morphs in the population.

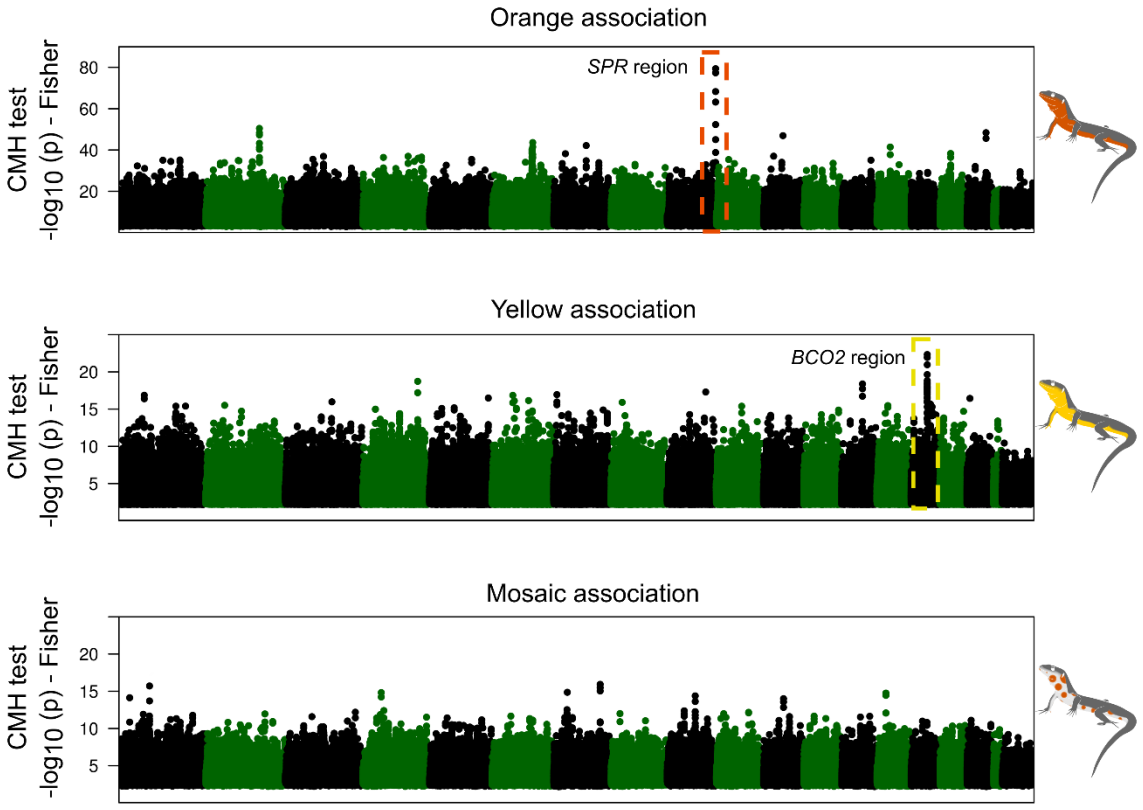

Figure S5. Genome-wide results from a Cochran-Mantel-Haenszel (CMH) test of association. Results were summarized using a sliding-window approach. Each dot represents a 20-SNP window with a step of five SNPs.

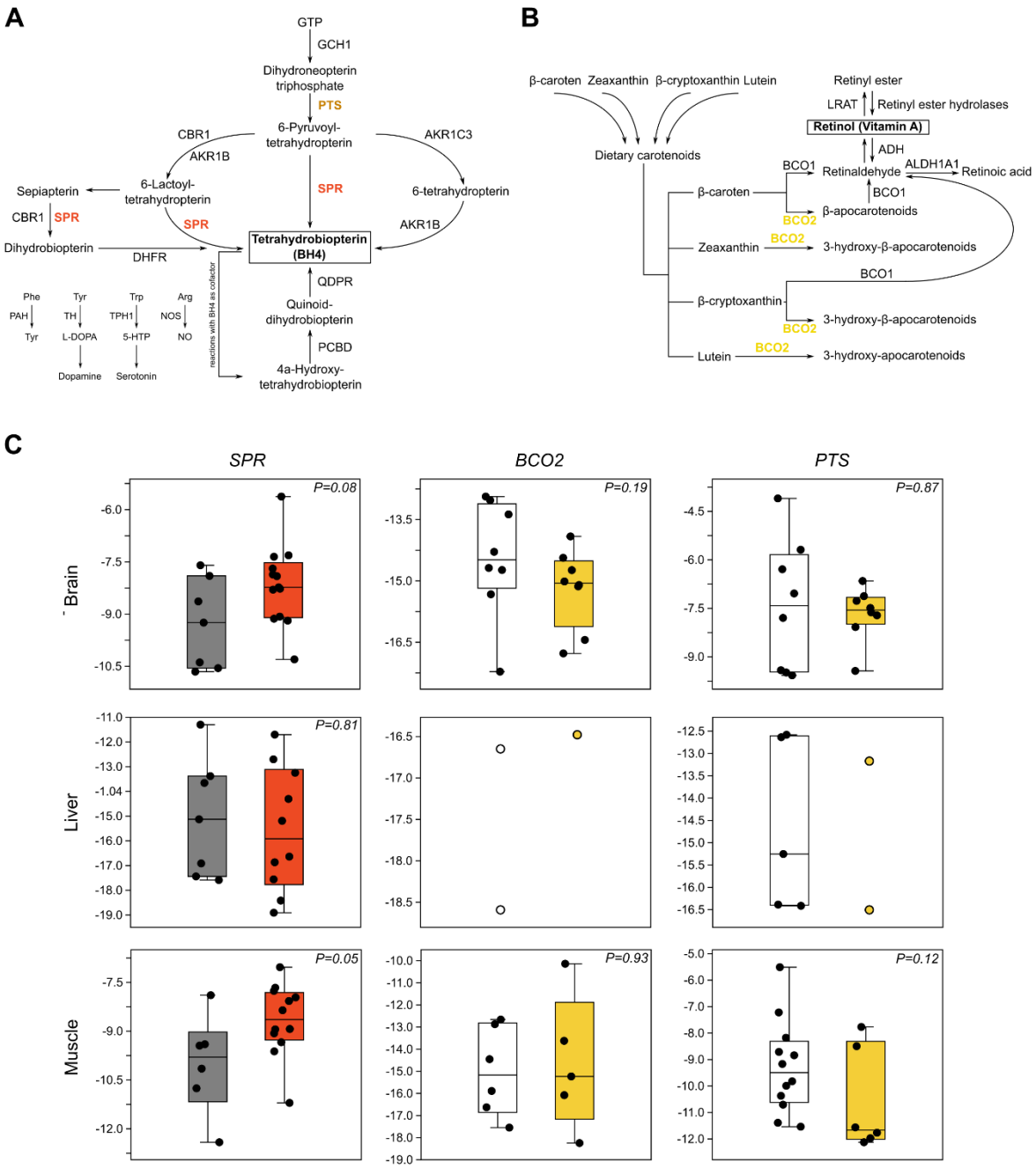

Figure S6. Pterin and carotenoid genes near the regions of maximum association are involved in vital metabolic processes. **(A)** Simplified tetrahydrobiopterin (BH4) biosynthesis pathway, evidencing the crucial roles of SPR and PTS in many of the major conversion steps, in particular of the main *de novo* synthesis pathway. BH4 is an essential co-factor in several enzymatic reactions, notably in those producing neurotransmitters (dopamine, serotonin) and nitric oxide, as well as the metabolism of phenylalanine. **(B)** Simplified view of the main reactions involved in the metabolism of precursor dietary carotenoids, with emphasis on the role of *BCO2*. **(C)** Gene expression patterns of *SPR*, *BCO2* and *PTS* in brain, liver and muscle, based on qPCR, between animals with orange/non-orange pigmentation and

white/yellow pigmentation (y axis shows -ΔCq values, normalized to 18S gene expression). Many samples failed amplification for *BCO2* and *PTS* in the liver, which is likely due to low expression levels of these genes or alternative splice variants specific to this organ. It can also be explained by degradation of the mRNA since RNA integrity numbers for liver were on average lower than that for the other tissues.

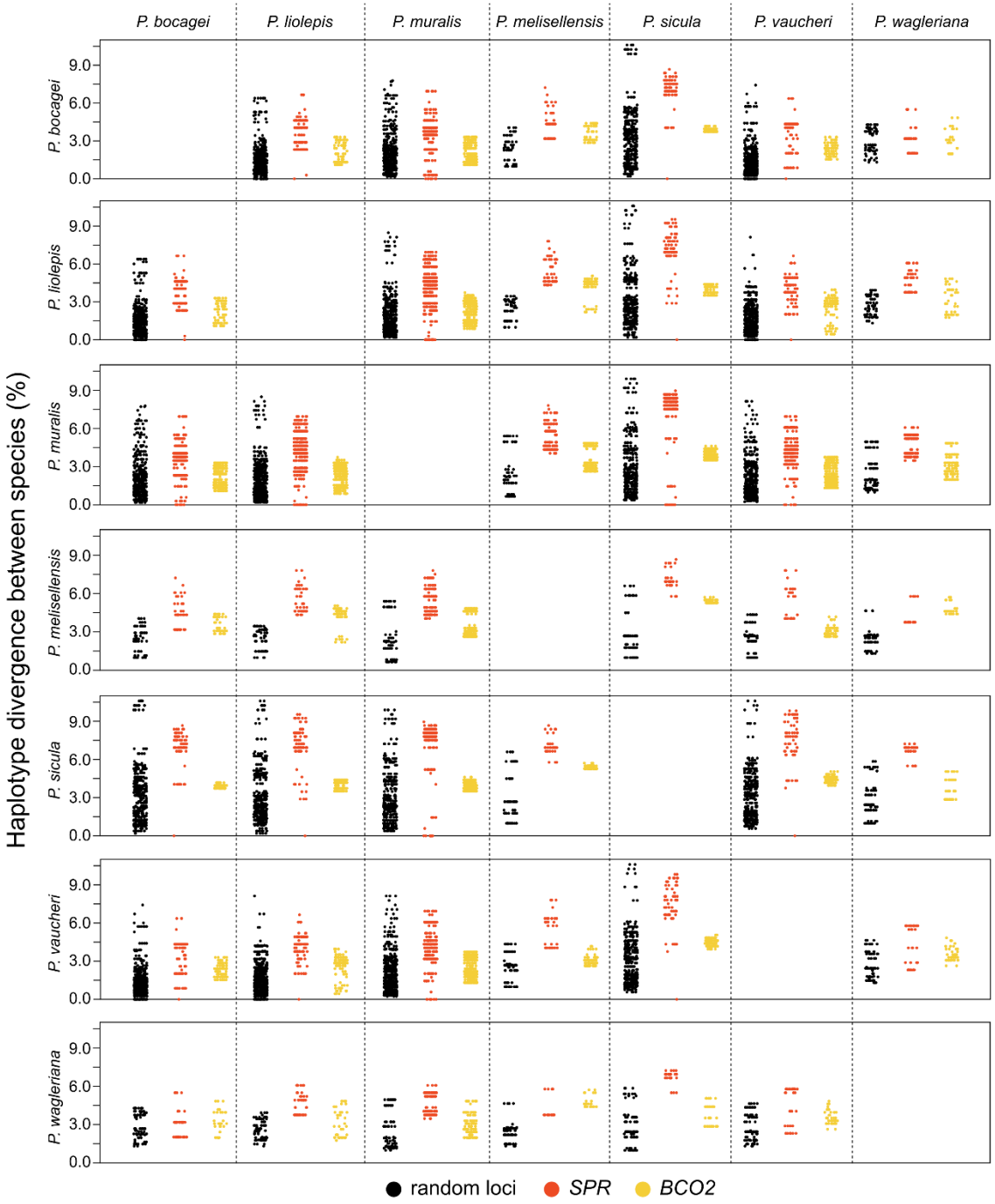

Figure S7. Pairwise nucleotide differences between haplotypes from seven *Podarcis* species for *SPR* (orange), *BCO2* (yellow), and 31 random loci (black).

Table S1. Summary statistics for our *de novo* common wall lizard (*Podarcis muralis*) reference genome assembly and annotation (PodMur 1.0)

	PodMur1.0
Genome assembly	
Contig N50 (kb)	714.6
Scaffold N50 (Mb)	92.4
Assembly size (Gb)	1.51
Number of scaffolds (>1kb)	2,122
Number of scaffolds (>10kb)	1,566
Largest scaffold (Mb)	140.2
GC content (%)	44.1
Complete BUSCOs reference (%)	93.2
Fragmented BUSCOs reference (%)	3.8
Missing BUSCOs reference (%)	3.0
Genome annotation	
Protein-coding genes	24,656
Complete BUSCOs annotation (%)	77.2
Fragmented BUSCOs annotation (%)	11.4
Missing BUSCOs annotation (%)	11.4

Table S2. Size of the 19 largest scaffolds of *de novo* reference assembly of the Common wall lizard genome (PodMur1.0).

Chromosome	Size (bp)
1	140,209,921
2	128,038,163
3	124,954,059
4	107,526,675
5	100,393,251
6	99,678,440
7	92,398,148
8	90,342,841
9	79,135,106
10	76,280,819
11	65,231,656
12	60,949,594
13	56,546,671
14	54,988,359
15	46,441,056
16	43,659,005
17	42,925,489
18	14,237,138
Z	50,871,175
Unplaced scaffolds	36,195,651
Total	**1,511,003,217**

Table S3. Summary statistics of RNA sequencing data of several tissues for genome annotation

Tissue	Number of reads	Number of mapped reads	Estimated number of transcripts[1]
Brain	230,727,821	207,227,126 (89.8%)	35,281
Duodenum	129,921,036	109,579,178 (84.3%)	32,159
Embryo (15ºC)	76,264,214	70,123,642 (92.0%)	34,820
Embryo (24 ºC)	69,718,379	63,868,404 (91.6%)	29,689
Muscle	140,540,165	129,421,702 (92.1%)	20,923
Skin	138,343,774	111,773,184 (80.8%)	21,512
Testis	220,430,106	197,029,219 (89.4%)	59,277

[1]Estimated using *Cufflinks*

Table S4. Summary statistics of the whole genome resequencing dataset

Population	Locality[1]	Morph	Number of individuals in pool	Number of reads	% of reads mapping[2]	% of reads MQ>=40	Mean depth of coverage
Pyrenees	1	Orange	14	228,390,110	98.2 (91.9)	82.8	17.1
Pyrenees	1	Orange-white	14	210,469,900	98.6 (92.3)	83.0	15.7
Pyrenees	1	Orange-yellow	15	203,437,559	98.6 (92.3)	83.1	15.2
Pyrenees	1	White	15	215,363,776	98.6 (92.4)	83.2	16.1
Pyrenees	1	Yellow	15	199,797,350	98.4 (92.1)	82.9	14.9
Pyrenees	2	Orange	20	216,047,786	98.3 (91.8)	83.1	16.3
Pyrenees	2	Orange-white	11	201,814,386	98.5 (92.0)	83.4	15.3
Pyrenees	2	Orange-yellow	9	196,309,282	98.5 (92.1)	83.3	14.8
Pyrenees	2	White	21	200,992,353	98.5 (91.8)	83.2	15.1
Pyrenees	2	Yellow	20	240,708,482	98.5 (92.0)	83.4	18.2
Italy	-	NA	20	321,742,840	98.1 (84.9)	77.1	23.5
Pyrenees (male pools merged)	2	NA	81	1,057,458,695	98.5 (92.9)	83.1	78.6
Pyrenees (female pool)	2	NA	24	755,245,078	97.9 (92.2)	83.0	65.5

[1]Locality 1 - Angostrina (42°28'N, 1°57'E); Locality 2 - Tor de Querol (42°27'N, 1°53'E).
[2]The percentage of properly paired reads are indicated in parenthesis.

Table S5. Genotyping results for the *SPR* and *BCO2* loci in other sub-lineages of the common wall lizard

Lineage[1]	**Region**	***SPR* locus**			***BCO2* locus**		
			Non-orange	**Orange**		**White**	**Yellow**
5	Northern Italy	*OO*	12	0	*YY*	0	0
		Oo	6	0	*Yy*	0	0
		oo	0	10	*yy*	3	3
9	Southern Italy	*OO*	4	1	*YY*	1	2
		Oo	0	0	*Yy*	1	0
		oo	0	0	*yy*	0	0
14	Greece	*OO*	6	0	*YY*	1	0
		Oo	0	0	*Yy*	0	0
		oo	1	4	*yy*	4	0
17	Turkey	*OO*	0	0	*YY*	3	0
		Oo	0	0	*Yy*	0	0
		oo	0	3	*yy*	0	0
Total	**-**	***OO***	**22**	**1**	***YY***	**5**	**2**
		Oo	**6**	**0**	***Yy***	**1**	**0**
		oo	**1**	**17**	***yy***	**7**	**3**

[1]Lineages according to Salvi *et al.* 2013.

Table S6. SRM conditions used for LC-MS/MS determination of the pterin compounds

Compound	Molecular weight	Precursor ion	Product ion	Fragmentor (V)	Collision energy (V)
Riboflavin	376.4	377.4	243.3	135	30
L-Sepiapterin	237.2	238.2	192.1	120	15
Pterin	163.1	164.1	119.1	100	20
Pterin-6-carboxylic acid	209.2	208.1	162.1	100	15
6-Biopterin	237.1	238.1	178.1	115	17
Isoxanthopterin	237.1	238.1	178.1	115	20
Xanthopterin	179.1	180.1	135.0	125	20
Leucopterin	195.1	196.1	140.1	120	16
Drosopterin	368.4	369.4	230.2	125	30
Erythropterin	265.1	266.1	220.1	110	8
D-Neopterin	253.2	254.2	206.2	115	14

Table S7. List of PCR primers that were used in this study for genotyping of the association, allelic imbalance, qPCR, and amplicon sequencing in *SPR* and *BCO2*

Primer name	Primer sequence[1,2]	Position	Experiment
Pmuralis_ORANGE_F	GGTAACCTGCCAGACAAGAGA	Chr9:77999736-77999756	Genotyping of association
Pmuralis_ORANGE_R	GCCTGCAAACAGACAGTCAA	Chr9:78000217-78000236	Genotyping of association
Pmuralis_YELLOW_F	GAGCTCGAAGGAATCCTGAA	Chr15:26161275-26161294	Genotyping of association
Pmuralis_YELLOW_R	TAGCCACTTTCTGCCAAACC	Chr15:26161803-26161822	Genotyping of association
Pmuralis_ORANGEindel_F	GGAAGACATCGAATGGCACT	Chr9:77997487-77997506	Structural variant analysis
Pmuralis_ORANGEindel_R	GCCTGCAAACAGACAGTCAA	Chr9:78000217-78000236	Structural variant analysis
Pmuralis_SPR_F	TGCTGCTCGTCAACAACG	Chr9:77996806-77996789	qPCR
Pmuralis_SPR_R	TGCGAGAAGAAGATCCAACA	Chr9:77994374-77994393	qPCR
Pmuralis_BCO2skin_F	GCACCAGTTTGACATTGACG	Chr15:26176709-26176728	qPCR
Pmuralis_BCO2skin_R	CTTGACATGAAGCGCTCAAA	Chr15:26176854-26176873	qPCR
Pmuralis_BCO2tissues_F	TTGACGACGGAGTTGTGAAG	Chr15:26176723-26176742	qPCR
Pmuralis_BCO2tissues_R	GTCCGTCATTTTTGGCACTT	Chr15:26176878-26179742	qPCR
Pmuralis_PTS_F	AATGCAATAACCCCAATGGA	Chr15:26202895-26202914	qPCR
Pmuralis_PTS_R	CGACTTCAGCAAAGTAGGGAAC	Chr15:26207199-26207220	qPCR
Pmuralis_18S_F	AAACGGCTACCACATCCAAG	*	qPCR
Pmuralis_18S_R	CTCGATCCCAAGATCCAACT	Chr17:41466490-41466509	qPCR
Pmuralis_SPR_AI_F	5' F tail-TGAAACCCTTCCAGACTTGG	Chr9:77994808-77994827	Allelic imbalance
Pmuralis_SPR_AI_R	5' R tail-GGAAAGGGAAGCTCTTTTCG	Chr9:77993359-77993378	Allelic imbalance
Pmuralis_BCO2_AI_F	5' F tail-CTGACCCTTGCAAGTCCATT	Chr15:26176834-26176853	Allelic imbalance
Pmuralis_BCO2_AI_R	5' R tail-GTTTTGCAGCCGGTAAACAT	Chr15:26187229-26187248	Allelic imbalance
Pmuralis_SPR_AIgeno_F	5' F tail-AGTCAGGCATCCCTCGTAAA	Chr9:77993492-77993511	Allelic imbalance genotyping
Pmuralis_SPR_AIgeno_R	GTCTGCTTGAGCCATTCCAT	Chr9:77993820-77993839	Allelic imbalance genotyping
Pmuralis_BCO2_Aigeno_F	ATCACACCCACAAACCAACC	Chr15:26179694-26179713	Allelic imbalance genotyping
Pmuralis_BCO2_AIgeno_R	GTCCTGGATTCCTGGGAGAC	Chr15:26179927-26179946	Allelic imbalance genotyping
Pmuralis_ORANGE_tailFwd	5' F tail-GGTAACCTGCCAGACAAGAGA	Chr9:77999736-77999756	Amplicon sequencing
Pmuralis_ORANGE_tailRv	5' R tail-GCCTGCAAACAGACAGTCAA	Chr9:78000217-78000236	Amplicon sequencing
Pmuralis_YELLOW_tailFwd	5' F tail-GAGCTCGAAGGAATCCTGAA	Chr15:26161275-26161294	Amplicon sequencing
Pmuralis_YELLOW_tailRv	5' R tail-TAGCCACTTTCTGCCAAACC	Chr15:26161803-26161822	Amplicon sequencing

[1] 5' F tail: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
[2] 5' R tail: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
* Designed based on transcriptome assembly.

Table S8. Samples used for amplicon sequencing overlapping the regions of maximum association in *SPR* and *BCO2*

Sample code	Species	Morph	Sampling localities
PY01	*P. muralis*	Orange	Col de la Core, France
TQ53	*P. muralis*	Orange	Tor de Querol, France
PY03	*P. muralis*	Orange	Col de la Core, France
PC17	*P. muralis*	Orange	Pontchateau, France
PZ110	*P. muralis*	Orange-white	Pouzauges, France
TQ61	*P. muralis*	Orange-white	Tor de Querol, France
TQ33	*P. muralis*	Orange-yellow	Tor de Querol, France
VR01	*P. muralis*	Orange-yellow	Vitre, France
2.50	*P. muralis*	White	Montseny, Spain
PZ115	*P. muralis*	White	Pouzauges, France
TQ45	*P. muralis*	White	Tor de Querol, France
PY10	*P. muralis*	White	Col de la Core, France
2.38	*P. muralis*	Yellow	Montseny, Spain
PZ13	*P. muralis*	Yellow	Pouzauges, France
TQ51	*P. muralis*	Yellow	Tor de Querol, France
TQ57	*P. muralis*	Yellow	Tor de Querol, France
3.214	*P. bocagei*	Orange	São Mamede do Coronado, Portugal
3.216	*P. bocagei*	Orange	São Mamede do Coronado, Portugal
3.179	*P. bocagei*	White	São Mamede do Coronado, Portugal
3.19	*P. bocagei*	White	São Mamede do Coronado, Portugal
3.352	*P. bocagei*	Yellow	Gerês, Portugal
3.355	*P. bocagei*	Yellow	Gerês, Portugal
DB26804	*P. liolepis*	Orange	Llívia, Spain
LLIVIA1	*P. liolepis*	Orange	Llívia, Spain
LLIVIA2	*P. liolepis*	Orange	Llívia, Spain
LLIVIA3	*P. liolepis*	Orange	Llívia, Spain
LLIVIA19	*P. liolepis*	Orange	Llívia, Spain
LLIVIA4	*P. liolepis*	White	Llívia, Spain
LLIVIA5	*P. liolepis*	White	Llívia, Spain
RIUCERD4	*P. liolepis*	Yellow	Riu de Cerdanya, Spain
DB20827	*P. melisellensis*	Orange-white	Socerb, Slovenia
DB20835	*P. melisellensis*	Orange-white	Socerb, Slovenia
DB20822	*P. melisellensis*	Orange-yellow	Hrastovlje, Slovenia
DB20824	*P. melisellensis*	Orange-yellow	Potrec, Slovenia
DB20828	*P. melisellensis*	White	Socerb, Slovenia
DB20817	*P. melisellensis*	Yellow	Socerb, Slovenia
DB20818	*P. melisellensis*	Yellow	Hrastovlje, Slovenia
DB26186	*P. sicula*	Orange	Milazzo, Italy

DB26195	*P. sicula*	Orange-white	Salina, Italy
DB26231	*P. sicula*	Orange-white	Milazzo, Italy
DB26154	*P. sicula*	White	Milazzo, Italy
DB26171	*P. sicula*	White	Milazzo, Italy
DB26200	*P. sicula*	White	Milazzo, Italy
7.22	*P. vaucheri*	Orange	Taza, Morocco
7.41	*P. vaucheri*	Orange	Taza, Morocco
7.9	*P. vaucheri*	Orange	Taza, Morocco
7.12	*P. vaucheri*	White	Taza, Morocco
7.16	*P. vaucheri*	White	Taza, Morocco
7.19	*P. vaucheri*	White	Taza, Morocco
7.326	*P. vaucheri*	White	Ketama, Morocco
7.354	*P. vaucheri*	Yellow	Ketama, Morocco
DB26532	*P. wagleriana*	Orange-white	Favigniana, Italy
DB27409	*P. wagleriana*	Orange-yellow	Levanzo, Italy

Table S9. List of loci and primers used to amplify and sequence 31 loci randomly distributed across the genome in several color polymorphic species in the genus *Podarcis*

Locus	Primer name	Primer sequence	Reference
acm4	TgF	CAAGCCTGAGAGCAARAAGG	Gamble *et al.* 2008
	TgR	ACYTGACTCCTGGCAATGCT	Gamble *et al.* 2008
ßfibint7	BF8	CACCACCGTCTTCTTTGGAACACTG	Pinho *et al.* 2008
	BfibR	CAGGGAGAGCTACTTTTGATTAGAC	Pinho *et al.* 2008
cmos	MosF	CTCTGGKGGCTTTGGKKCTGTSTACAAGG	Godinho *et al.* 2006
	MosR	GGTGATGGCAAANGAGTAGATGTCTGC	Godinho *et al.* 2006
mc1r	MC1RF	GGCNGCCATYGTCAAGAACCGGAACC	Pinho *et al.* 2010
	MC1RR	CTCCGRAAGGCRTAGATGATGGGGTCCAC	Pinho *et al.* 2010
*nfycint16**	NFYC16F2	GCARGGACAGCAGCAGTTCAGCCAGTT	Pinho *et al.* 2010
	NFYC17R2	GCWGGCATRGTSACTTGCTGRATCTGG	Pinho *et al.* 2010
	NFYCPodF	AGCCATTGAGTTGGATTAAAAGAGG	This study
	NFYCPodR	CACTGAGGTTCCCTCTTCAATGG	This study
pdc	PDCpodF	AGTATCGCAAGCGTTGTATGCAGG	Salvi *et al.* 2013
	PDCpodR	CCCAGCAAAAAACTCCTCACTGAA	Salvi *et al.* 2013
pkm2int5	PKSQF	ACCAAAGTTGTWGATGTTGGCAGC	Pinho *et al.* 2010
	PKSQR	ATGAAGGAAGCAAACACCATGTC	Pinho *et al.* 2010
pod6b	Pod6bF	CTGGTAATGGCCCGCTATGTATGGG	Pereira *et al.* 2013
	Pod6bR	ATAAAGCTGGGAAGCTCTTGAGTCC	Pereira *et al.* 2013
pod7b	Pod7bF	GTCACTTTGGTGCTGCTCGCACAGC	Pereira *et al.* 2013
	Pod7bR	TGTAATGCTGCAACTTGGCGACACC	Pereira *et al.* 2013
pod11	Pod11F	GACTTTGGGTTCAAATCTCCACCCC	Pereira *et al.* 2013
	Pod11R	AGGTCATCTGCTTGACTGTTCTGGC	Pereira *et al.* 2013
pod12b	Pod12bF	ACCTTCTTTTGCCTACGCACGCCAG	Pereira *et al.* 2013
	Pod12bR	CTGTCCACAACACCCTTATTCTGCC	Pereira *et al.* 2013
pod13	Pod13F	GCAGTTGTTGCTGGGCTCATTTCTG	Pereira *et al.* 2013
	Pod13R	ACATGATTTTGAGGGGACGCAAACC	Pereira *et al.* 2013
pod14	Pod14F	GCTTTCCTATGAGGCTCAAGTTTGG	Pereira *et al.* 2013
	Pod14R	AGCCGACTGTCTCTAATAACTTCCC	Pereira *et al.* 2013
pod14b	Pod14bF	CTGGAGGAAGGGTAGCATGATCTCC	Pereira *et al.* 2013
	Pod14bR	CTGACAGCCGCATCAGACGTTCAGC	Pereira *et al.* 2013
pod15	Pod15F	ACTTTACATCCCATGATAGGTCTGG	Pereira *et al.* 2013
	Pod15R	TGATATAGCAGAACACCTGTGCAGC	Pereira *et al.* 2013
pod15b	Pod15bF	AATCCTGGCTAAATGCAAGCCTTGG	Pereira *et al.* 2013
	Pod15bR	GCCAGGAGAATAAGCTACTCCATCC	Pereira *et al.* 2013
pod16	Pod16F	TTCCTTTGTTACACCTTGGGAGGGGT	Pereira *et al.* 2013
	Pod16R	CTGGAGAGGGAGCAGCGGCTTCAGG	Pereira *et al.* 2013
pod17	Pod17F	TAATTGCCCATTCCCTTCGATTCCC	Pereira *et al.* 2013
	Pod17R	TGATAACCATTGCCTTCATTATGCC	Pereira *et al.* 2013
pod20	Pod20F	GAGTGCTTACAGGCTGTGAAGATGT	Pereira *et al.* 2013
	Pod20R	ATGCCGATTCAACCAAAACATGGCG	Pereira *et al.* 2013
pod21	Pod21F	TCTAGAGACCGAGTCCTTGTAAGGG	Pereira *et al.* 2013
	Pod21R	GAAACTCCTCTCCCAGAGAACGACC	Pereira *et al.* 2013
pod25	Pod25F	GTATTATCAGGCCCAGTGCTTGTGG	Pereira *et al.* 2013
	Pod25R	TGGTGGATTATCTATCATCTGCTCC	Pereira *et al.* 2013
pod31	Pod31F	AACGGCTATTTGCGGACTACAGTAG	Pereira *et al.* 2013

	Pod31R	GCAGGTCACTAGGAATATAGAAGCC	Pereira *et al.* 2013
pod33	Pod33F	ATCTGATGGGAGAGCATTCCACAGG	Pereira *et al.* 2013
	Pod33R	GTGCGCCATATTACACAGCAACTGG	Pereira *et al.* 2013
pod38	Pod38F	AGCGCTGCAACTTTCTCTGCTTCCG	Pereira *et al.* 2013
	Pod38R	GGGCATGAGTCAGGAGTAGTCACGC	Pereira *et al.* 2013
pod43	Pod43F	CCATTACGTCAAGTATTGCTAATGC	Pereira *et al.* 2013
	Pod43R	CATAGAGATTCTTATGCAGAACTGG	Pereira *et al.* 2013
pod55	Pod55F	GGATCTTTATAGGAGAGTGCAGGCC	Pereira *et al.* 2013
	Pod55R	TTCCAGATTGTGTTTATCCTGGTGG	Pereira *et al.* 2013
pod69	Pod69F	TTATAAGTGTGGGAGTAGCGAGCTG	Pereira *et al.* 2013
	Pod69R	GGAGCATTGAAAATATCCAAGATGG	Pereira *et al.* 2013
pod72	Pod72F	GAAGGGAGACGGTGTGCTATTGTCG	Pereira *et al.* 2013
	Pod72R	CCTCCTGCTCTCTCTTCCTAACACG	Pereira *et al.* 2013
rag1	Rag1PodF	AGATTCATTGCCCTCACTTGCCC	This study
	Rag1PodR	CAAATATTCCAGTTTGAGATTGGCG	This study
*rag2**	31FNVenk	TTYGGICARAARGGITGGCC	Chiari *et al.* 2004
	Lung460R	GCATYGRGCATGGACCCARTGICC	Chiari *et al.* 2004
	Lung320R	AYCACCCATATYRCTACCAAACC	Chiari *et al.* 2004
	Rag2seq	GTATGGCCCAGTCTGAGAAGTGTCC	This study
relnint61	Reln61F	GAGTMACTGAAATAAACTGGGAAAC	Pinho *et al.* 2010
	Reln62R	GCCATGTAATYCCATTATTTACACTG	Pinho *et al.* 2010

*The second set of primers was used to perform a nested PCR.