Supplementary materials for:

“Parthenogenesis in *Darevskia* lizards - a rare outcome of common hybridisation, not a common outcome of rare hybridisation”, by Susana Freitas, Anja Marie Westram, Tanja Schwander, Marine Arakelyan, Çetin Ilgaz, Yusuf Kumlutas, D. James Harris, Miguel A. Carretero and Roger K. Butlin.

Supplemental Methods:

**RNA extraction**

Total RNA was extracted from four samples; a tail-tip from a male of the sexual species *D. valentini*, brain and ovaries from one individual and tail-tip from another individual of the parthenogenetic *D. unisexualis*. Given the expected differential expression depending on the tissue, tissue samples from the same individual were analysed separately. The tail-tips were collected in the field and preserved in RNAlater. Ovaries and brain were used from a sacrificed animal, and were preserved in RNAlater immediately after dissection. All samples were then kept at -80°C until extraction.

RNA was extracted using a Trizol (Invitrogen) extraction method, following (Westram et al., 2014). All extractions were performed in duplicates. RNA was preserved at -80°C prior to library preparation. Total RNA quality and concentration were determined using the Agilent 2100 BioAnalyzer and the sample used for library preparation was selected according to quality parameters.

**RNA reads processing**

RNAseq libraries for the four samples were prepared in Edinburgh Genomics, Edinburgh, UK where they were then sequenced in a single lane, using an Illumina HiSeq 2000 machine (100 bp paired-end reads; insert size around 80 bp). Read quality, before and after filtering, was assessed with FastQC (Andrews, 2010). Removal of contaminant substrings and size/quality filtering were performed with Scythe (Buffalo, 2011) and Sickle (Joshi and Fass, 2011), respectively. All reads shorter than 50 bp and/or with Ns were discarded, with the remaining parameters at their default options.

**RNA de novo assembly**

A *de novo* assembly for the contigs selection was constructed with Velvet (1.2.10) (Zerbino and Birney, 2008) and Oases (version 0.2.08; http://www.ebi.ac.uk/~zerbino/oases/) (Schulz et
Two assemblies were done, one with the reads from the tail-tip of *D. valentini* (hereafter VT) and the other with all reads from the ovaries and the brain of *D. unisexualis* (hereafter UBUO), to focus on transcripts that might be absent from the tail-tip tissues. The same parameters were used in both de novo assemblies, with initial Velvet assemblies using k-mer length values from 25 to 65 and a step increase of 8 [25, 33, 41, 49, 57, 65], and a final assembly of the initial runs was performed with Oases with a k-mer length of 31 and a coverage cut-off of 4. Redundancy of the assembly was decreased with CD-Hit (version:1.3.1) (Li and Godzik, 2006), setting a minimum similarity of 0.95 and a word length of 8.

**RNA: Contig selection and probe design**

To map the reads against the reference transcriptome, Stampy version 1.0.23 (Lunter and Goodson, 2011) was selected due to its high sensitivity for indels and divergent reads. All four samples (tail-tip from a sexual male *D. valentini*, and brain, ovaries and tail-tip from a parthenogenetic *D. unisexualis*) against both de novo assemblies. Platypus version 0.7.9.1 (Rimmer et al., 2014) was used to call the variants mapped. Only reads mapped with PHRED quality over 20 were selected, so that there is 1% or less probability of a wrongly identified SNP. Heterozygosity for each sample was calculated in the vcf outputs using VCFtools (0.1.16) (Danecek et al. 2011) -het option (Danecek et al., 2011).

In order to find many, individually-informative markers that are long enough and with sufficient variable sites, we aimed to select contigs to give a total capture target of approximately 500 kbp. The longest isoform per contig was selected. To simplify downstream analyses, contigs with indels were eliminated, and only contigs with SNPs were selected. For the specific *D. unisexualis* tissues (brain and ovaries), the *D. unisexualis* de novo assembly was also used to select variable loci and the same method was applied. The most variable contigs were selected from both assemblies and added to the target set until the required total target size was reached. To construct a dataset with markers at different rates of evolution, we also randomly selected sequences that did not present any variation in the backward alignment.

After contig selection from the VT and UBUO assemblies, the sets were blasted against each other using the megablast option in blastn (Camacho et al., 2013) with a minimal value of 1e^-3 in order to eliminate possible repeated sequences selected simultaneously from both assemblies. The sum of the selected contigs length amounted to 575,561 bp in total.

Probe design for the reference loci and capture sequence was outsourced to RAPiD Genomics (FL, USA). Probes were designed to span across the two transcript contigs data sets (575,561 bp), to be 120 bp long, tiled across the whole length of the contig, and with a 40 bp shift to give 3x coverage. 10,000 probes were needed to cover the ~500kb capture target intended.
Sequence Capture: Sample selection and read processing

A total of 69 individuals belonging to 14 sexual *Darevskia* species and three outgroup individuals (two *Iranolacerta brandtii* and one *Podarcis liolepis*) were sequenced after genomic DNA fragmentation and ligation with barcoded Illumina-compatible adapters to the resulting fragments (Table S1). See below for information on parthenogenetic individuals. Enrichment PCR and capture reactions with the custom-designed probes preceded the sequencing of the libraries.

Read filtering was performed with Cutadapt (Martin, 2011) eliminating all reads shorter than 40 bp and with PHRED quality lower than 20. Stampy (Lunter and Goodson, 2011) was used to map the reads against the reference *de novo* assembly used to design the probes. A substitution rate was set regarding the distance of the different species to the reference transcriptome, higher for phylogenetically distant species and lower for phylogenetically close species. Accordingly, individuals were divided into three groups: Rudis group (with SR=0.01), non-Rudis group (SR = 0.03) and the outgroup (SR = 0.1). In cases of multiple mappings that are almost (but not exactly) the same, the SR parameter helps to choose the best mapping by influencing the mapping quality (MAPQ) score. Best SR values were chosen when comparing number of variants (after SNP calling) and flagstat values for the alignments. Samtools (Li et al., 2009) was used to convert to bam, sort, index and calculate basic statistics of the alignment (flagstat option). The MarkDuplicates option in Picard tools was used to remove PCR duplicates, which can be high given the capture sequence protocol. SNP calling was used with FreeBayes (Garrison and Marth, 2012) using minimum coverage of two per position, calling the four best alleles (for high substitution rate loci, given we are using different species and expect more variation than within population) and without the population priors. Read coverage was calculated with vcftools and plotted against divergence with reference species *D. valentini* used to design the probes (we did not use the second reference species because of its hybrid ancestry). Post-vcf filtering was performed with VCFTools (Danecek et al., 2011) and vcflib (https://github.com/vcflib/vcflib) selecting variants with genotype quality > 30, coverage between 5 and 400 for ingroup species. Ingroup individuals *D. parvula* [12738], *D. valentini* [12641] and *D. valentini* [16280] were filtered separately since they had higher coverage; the selected coverage interval was 5-800 read depth. Mean depth by individual was estimated using VCFTools (0.1.16) (Danecek et al. 2011). Outgroup individuals (*Iranolacerta brandtii*) were filtered with the same parameters as the ingroup except for coverage (DP = 2-400). The outgroup *P. liolepis atrata* was removed from the phylogenetic analyses because of low coverage. Only SNPs were used in the downstream analyses, hence indels and other types of variants were removed. Multiple nucleotide positions (MNP) identified by FreeBayes were deconstructed with vcfallelicprimitives option from vcflib.

Sequence Capture: Phylogenetic inference

Fasta files for each marker (i.e. each captured transcript) were made from the vcf file using a custom python script (https://github.com/SusanaFreitas/capture-data). Fasta files were then
trimmed to remove missing data, allowing for maximum of 10% missing genotypes for each position in each alignment. Species tree was constructed with SVDQuartets 1.0 using the 300 markers with the highest number of SNPs, hence more phylogenetically informative (Chifman and Kubatko, 2015). After all individuals were assigned (and confirmed) to different species with the phylogenetic inference, diversity parameters (nucleotide diversity and Tajima’s D) on the sexual species were estimated using VCFtools (0.1.16) (Danecek et al. 2011).

**Parthenogenetic hybrids: STRUCTURE analysis**

52 individuals putatively belonging to 6 parthenogenetic taxa were included in the capture sequencing (Table S2) and read data were processed as above. The ancestry of the parthenogens was investigated using the Bayesian multilocus clustering analysis implemented in STRUCTURE v2.3.4 (Pritchard et al. 2000; Hubisz et al. 2009). The rate of change in the log probability of the data between successive $K$ values in STRUCTURE (Pritchard, Stephens, and Donnelly 2000) and Delta $K$ (Evanno, Regnaut, and Goudet 2005), as implemented in STRUCTURE HARVESTER (Earl and vonHoldt 2011), were calculated from STRUCTURE runs performed for $K$ values between three and 15. Each $K$ value was run for 1 million generations with a burn-in of 30000, allowing for admixture between the clusters, and ten replicates were performed using only the sexual species dataset. Values from different replicate runs were combined with CLUMPP (Jakobsson and Rosenberg 2007). The sexual species were used as ‘learning samples’ (PopFlag = 1) (Murgia et al. 2006) to define the cluster membership when, subsequently, diploid parthenogenetic individuals were included in the dataset. Cluster membership was assigned according to the optimal number of clusters calculated previously ($K$), and excluding the clusters of parthenogens (since only sexual species were used as ‘learning samples’). The remaining STRUCTURE parameters were set as in Freitas et al. (2019).
Supplemental Results:

RNA: Reads and SNP detection

Between 36 and 42 million paired-end reads were generated per library and 94% were kept after trimming, adaptor removal and size and quality filtering. Oases assembles transcripts and groups them into contigs, which correspond to groups of transcripts put together according to similarity. Transcripts within contig groups may represent different isoforms of the same gene, but alternatively could simply represent different alleles, or even versions with sequencing error. The VT assembly comprised 245,539 transcripts distributed over 26,866 contigs with an average length of 1588 bp and N50 length of 2315. After CD-Hit redundancy analysis, 103,605 transcripts distributed over 26,899 contigs were left with a mean contig length of 1447 bp and N50 length of 2095. The UBUO assembly comprised 690,263 transcripts distributed over 58,106 contigs with an average length of 1588 bp and N50 length of 2469. After CD-Hit redundancy analysis, 289,870 transcripts distributed by 54,563 contigs were left with an average contig length of 1556 bp and a maximum length of 24,289 bp and N50 length of 2544.

RNA: Contig selection

To select the most variable contigs, combined reads from all samples were mapped against both the VT the UBUO assemblies. Around 80% of the *D. valentini* reads and between 65 - 72% of the *D. unisexualis* reads mapped to the VT reference. The yield of mapping regarding the UBUO assembly was significantly lower than for the VT assembly, with only around between 51 - 53% of the reads successfully mapping against that reference. The SNP calling of all reads against the VT assembly resulted in a total of 721,238 variants. To avoid probe alignment problems, contigs with indels longer than 2 bp were discarded. In the end, 773 contigs were selected, presenting in total 2372 SNPs. The same approach was followed in the UBUO assembly. For this *de novo* assembly a total of 3765 contigs were selected, presenting a total of 7981 variants.

From each of the two lists of contigs, the most variable contigs with a minimum length of 500 bp were selected. This resulted in 193 contigs from the VT assembly (163,534 bp in total) and 400 contigs (265,335 bp) from the UBUO assembly. After adding the sequences that did not present any SNP, (23 from the VT assembly and 10 from the UBUO assembly) we finally had 215 contigs from the VT assembly and 410 from the UBUO assembly.

Sequence Capture: SNP calling and Phylogenetic inference

SNP calling of the *Darevskia* species against the reference contigs resulted in a total of 64150 SNPs. Probes successfully captured all contigs, however one recovered no SNP variation. SNP count varied from 1 – 379 per contig (0.12 – 24%). A total of 10318 SNPs from 453 contigs
were recovered in all individuals (with 0% missing data), and when a maximum of 10% missing data was allowed, 37092 SNPs from 582 contigs were recovered.

**Parthenogenetic hybrids: STRUCTURE analysis**

STRUCTURE analysis showed the best K to be 8, as expected for the number of sexual species included. Analysis of the parthenogens (Fig. S1) clearly showed a mixed ancestry for each of the parthenogenetic genomes, confirming the hybrid genome compositions described before (for review see Murphy et al. 2000). *Darevskia armeniaca* shared approximately half of its genome with *D. mixta* (maternal ancestry), and the other half with *D. valentini* (paternal ancestry). *D. unisexualis*, *D. uzzelli* and *D. bendimahiensis* shared half of their genome with *D. raddei* (maternal) and the other half with *D. valentini* (paternal). *D. dahli* shared half of its genome with *D. mixta* (maternal) and the other half with *D. portschinskii* (paternal), and finally *D. rostombekowi* shared half of its genome with *D. raddei* (maternal) and the other half with *D. portschinskii* (paternal). Even though the hybrid composition of the parthenogens was clear, STRUCTURE assigned a minor part of each parthenogen’s genome to other species besides the expected parental taxa. In some cases, this minority component was as large as 15% (for instance, in the case of some *D. unisexualis* individuals which shared with *D. rudis* almost the same amount of their genome as they shared with their paternal species, *D. valentini*). This may be explained by the inability of the STRUCTURE analysis to distinguish between closely related taxa as contributors. Gene flow between the sexual species, and consequent shared alleles between them, may have added to this difficulty. This would be consistent with our results for sexual species, showing that there has been frequent gene-flow between species in the genus, and also with other studies (for instance, Tarkhnishvili et al. (2013) have shown ongoing gene flow between *D. valentini* and *D. rudis*).

STRUCTURE results also uncovered two misidentified individuals (Dben6283 and Dros580), whose genetic assignment suggests they belong to *D. raddei* (the parthenogenetic females are morphologically very similar to the female of their maternal species, personal observation). Moreover, an individual identified as *D. dahli ‘yellow’* (Ddah9923) shared half of its genome with *D. mixta* (maternal) and the other half with *D. valentini* (paternal), instead of *D. portshinskii* as expected. *Darevskia dahli ‘yellow’* is recognized as a subspecies of *D. dahli* (Arakelyan et al. 2011), and expected to have the same parental species. We are unsure whether the individual collected here was a misidentified *D. armeniaca*, or if instead *D. dahli ‘yellow’* is yet another hybrid parthenogenetic species (instead of a local variant of *D. dahli*), but this question cannot be answered without further studies.
**Fig. S1 - STRUCTURE plot analysing the parentage of the parthenogenetic individuals**

STRUCTURE plot with the sexual and parthenogenetic individuals. Sexual species (VAL: *D. valentini*, RUD: *D. rudis*, POR: *D. portschinskii*, RAD: *D. raddei*, MIX: *D. mixta* - 1, *D. clarkorum* and *D. saxicola*, CHLO: *D. chlorogaster*, *D. steineri* and *D. defilippii*, PAR: *D. parvula* and *D. mixta* - 2, PRA: *D. praticola*) were used to inform the ancestry of the parthenogenetic hybrids (ARM: *D. armeniaca*, BEN: *D. bendimahiensis*, DAH: *D. dahli* and *D. dahli* yellow, ROS: *D. rostombekowi*, UNI: *D. unisexualis*, UZZ: *D. uzzeli*). Proposed ancestors of the parthenogens are noted as the Maternal species and the Paternal species. Samples labelled 1, 2 and 3 correspond to misidentified individuals.
Fig. S2 - Read Depth by individual

Mean read depth by individual estimated with vcf tools (of the sexual species) for the 625 markers developed in this study.

Figure S3 - Nucleotide Diversity estimates per sexual species

Nucleotide diversity (pi) estimated over 1000 bp windows on the 625 markers developed in this study. D mixta corresponds to D. mixta - 1 individuals only.
Figure S4 - Tajima’s D estimates per sexual species
Tajima’s D estimated over 1000 bp windows on the 625 markers developed in this study on the 625 markers developed in this study. *D mixta* corresponds to *D. mixta* - 1 individuals only.
**Figure S5 - SVDquartets tree with the sexual species**

SVDquartets tree (solid line) built with the 300 most phylogenetically informative markers. Numbers on the nodes refer to bootstrap support values, and the tree is the majority-rule consensus tree over 1000 bootstrap samples. Branch lengths are not meaningful for the SVDquartets tree.
**Figure S6 - Coalescent consensus tree with the sexual species.**

Consensus tree built from the *BEAST2* analysis based on the 31 most informative markers. Branch lengths estimates are the median branch length values. Node labels correspond to the 95% highest posterior density interval of the age estimates for each node.
Figure S7 - The fraction of variance in relatedness between populations for models with 0 through 21 migration edges

The fraction of variance in relatedness between populations accounted for by phylogenetic models with 0 through 21 migration edges in the TreeMix analysis. The fraction of variance in the sample covariances accounted for by the model covariances (Pickrell and Pritchard, 2012) began to asymptote with four migration edges and there was little variation among runs at this value.
Supplemental Tables (legends):

**Table S1 - Sexual samples used in the study**

Samples that were misidentified in the field have the species corrected according to genetic results here. GPS coordinates (latitude and longitude columns) are provided whenever available.

**Table S2 - Parthenogenetic samples used for the ancestry analysis**

Detailed information for all parthenogenetic samples not included in the study, but used for the STRUCTURE analysis to confirm parentage (Fig. S1). GPS coordinates (latitude and longitude columns) are provided whenever available.

**Table S3 - Information for all ABBA/BABA tests performed here.**

First column (Z>4) corresponds to the total number of replicates per test (out of 100) that showed Z-score > 4. S1, S2, S3 and O columns correspond to the order of taxa included in the D-statistic test: S1 and S2 taxa were tested for gene flow against the S3 taxon, using the O taxon as the outgroup. ‘Z-mean’ corresponds to the mean of all Z values obtained for the 100 tests, and ‘Z_meanSD’ is the standard deviation of these Z values. ‘Dstat_mean’ corresponds to the mean of the D-statistic values obtained for the 100 replicates, and the following columns correspond to the D-statistic mean standard deviation, the minimum D- value and the maximum, respectively.
Supplemental References:

Arakelyan, Marine, Felix Danielyan, Claudia Corti, Roberto Sindaco, and Alan E. Levinton. 2011. *The Herpetofauna of Armenia and Nagorno-Karabakh*. Society for the Study of Amphibians and Reptiles. San Francisco.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, et al. 2011. ‘The Variant Call Format and VCFtools’. *Bioinformatics* 27 (15): 2156–58. https://doi.org/10.1093/bioinformatics/btr330.

Evanno, G., S. Regnaut, and J. Goudet. 2005. ‘Detecting the Number of Clusters of Individuals Using the Software Structure: A Simulation Study’. *Molecular Ecology* 14 (8): 2611–20. https://doi.org/10.1111/j.1365-294X.2005.02553.x.

Freitas, Susana N., D. James Harris, Neftalí Sillero, Marine Arakelyan, Roger K. Butlin, and Miguel A. Carretero. 2019. ‘The Role of Hybridisation in the Origin and Evolutionary Persistence of Vertebrate Parthenogens: A Case Study of Darevskia Lizards’. *Heredity* 123 (6): 795–808. https://doi.org/10.1038/s41437-019-0256-5.

Murgia C, Pritchard JK, Kim SY, Fassati A, Weiss RA (2006). Clonal Origin and Evolution of a Transmissible Cancer. *Cell* 126: 477–487.

Murphy, Robert W., Jinzhong Fu, Ross D. Macculloch, Ilya S. Darevsky, and Larissa A. Kupriyanova. 2000. ‘A Fine Line between Sex and Unisexuality: The Phylogenetic Constraints on Parthenogenesis in Lacertid Lizards’. *Zoological Journal of the Linnean Society* 130 (4): 527–49. https://doi.org/10.1111/j.1096-3642.2000.tb02200.x.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. ‘Inference of Population Structure Using Multilocus Genotype Data’, 15.

Tarkhnishvili D, Murtskhvaladze M, Gavashelishvili A (2013). Speciation in Caucasian lizards: climatic dissimilarity of the habitats is more important than isolation time. *Biological Journal of the Linnean Society* 109: 876–892.