PNAS

Supplementary Information for

The highest-elevation frog provides insights into mechanisms and evolution of defenses against high ultraviolet radiation

Ting-Ting Fu, Yan-Bo Sun, Wei Gao, Cheng-Bo Long, Chun-Hua Yang, Xin-Wang Yang, Yi Zhang, Xin-Qiang Lan, Song Huang, Jie-Qiong Jin, Robert W. Murphy, Yun Zhang, Ren Lai, David M. Hillis, Ya-Ping Zhang, Jing Che

Yun Zhang, Ren Lai, David M. Hillis, Ya-Ping Zhang, Jing Che jointly directed this work. Correspondence should be addressed to J.C. (chej@mail.kiz.ac.cn), Y.-P.Z. (zhangyp@mail.kiz.ac.cn), D.M.H. (dhillis@austin.utexas.edu), R.L. (rlai@mail.kiz.ac.cn), Y.Z. (zhangy@mail.kiz.ac.cn).

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Supplementary Text

Genome sequencing and assembly of *N. parkeri*

Briefly, genomic DNA was extracted from liver with QIAGEN® Genomic DNA extraction kit. Total RNA was extracted from each of tissues above using TRIzol reagent (Invitrogen Corp., Carlsbad, CA) and RNeasy Mini Kit (Qiagen, Chatsworth, CA). For long-read sequencing, SMRTbell libraries with fragment size of 20 kb were constructed with SMRTBell template preparation kit 1.0 (Pacific Biosciences, Menlo Park, CA, USA). The libraries were sequenced by the PacBio Sequel system (Pacific Biosciences) with 25 SMRT cells. short-read sequencing data were obtained based on a paired-end library with short insert sizes of about 350 bp. Three Hi-C libraries were prepared following the standard Hi-C library protocol. Then we used BGISEQ-500 platform to sequence the libraries with paired-end 100 bp reads. For transcriptome sequencing, a single pooled RNA sample was prepared by mixing equal volumes of the RNA extracted from each tissue. Total RNA was synthesized to the cDNA using Clontech SMARTer PCR cDNA Synthesis Kit. The DNA was used for SMRTbell library construction. Qualified sequencing data were obtained by sequencing this library using PacBio Sequel platform. To obtain consensus full-length isoforms, we performed SMRT analysis through Reads of insert, Classify, Cluster, and Quvier.

Based on 328.62 Gb clean short-read sequencing data, the genome size of *N. parkeri* was estimated to be 2.47 Gb through the k-mer (k = 17) depth frequency distribution analysis. In total, we generated 237.27 Gb of PacBio long-read data (21.25 million reads) with N50 > 17 Kb for most cells. The preliminary assembled genome was approximately 2.46 Gb in size and contig N50 was 2.32 Mb (Table S2). After polishing the primary assembly using both PacBio long-read data and BGISEQ short-read data, we assembled a 2.47 Gb genome of *N. parkeri*, with a contig N50 of 2.34 Mb. Assessment of the genome quality using BUSCO (1) showed that the assembly included 93.5% of the orthologue genes, including 87.9% of the complete and 5.6% fragmented genes (Table S3). Approximately 394.87 Gb clean data were generated from the Hi-C libraries. We then constructed a chromosome-level scaffold genome with these data using Lachesis. Finally, we obtained 13 pseudo-chromosomes, which is the first chromosome-level high-elevation amphibian genome with scaffold N50 of 268.57 Mb (Table S4). Our chromosome-level assembly of *N. parkeri* genome was more than 250 times better both in contig and scaffold N50 than the previous assembly version, which used only Illumina data (contig N50: 2.34 Mb versus 8.1 Kb, scaffold N50: 268.57 Mb versus 1.05 Mb), and resulted in a larger assembled genome size (2.47 Gb versus 2.1 Gb). In addition, compared to other published genomes of anurans frogs, our *N. parkeri* genome assembly exhibited the highest contig N50 and the third highest scaffold N50 (just smaller than *Leptobrachium leishanense* and *L. ailaonicum*; Table S5).

Genome annotation of *N. parkeri*.

In total, 1.46 Gb of repetitive sequences were identified in the *N. parkeri* genome using combined structure-based and homology-based methods, accounting for 58.90% of the genome. Most of the repeated sequences were transposable elements (TEs), occupying 55.20% (1.36 Gb) of the assembly (Table S8). We identified four major types of TEs, which were 574.37 Mb of DNA transposons (23.23% of the genome), 622.33 Mb of long interspersed nuclear elements (25.17%; LIENs), 28.04 Mb of short interspersed nuclear elements (1.13%; SINEs), and 366.12 Mb of long terminal repeats (14.81%; LTRs), respectively (Table S8). Other types and unknown TEs accounted for 0.00097% and 0.17%, respectively (Table S8).

We integrated *de novo* prediction-based method, homology-based method and RNA sequence-based method to annotate the protein-coding genes. *De novo* prediction was performed using Augustus (2). Homologous gene sets of the protein gene sets from *Homo sapiens*, *Gorilla gorilla*, *Macaca mulatta*, *Mus musculus*, *Rhinopithecus roxellana* were searched in the *N. parkeri* genome with BLAST+. For RNA sequences-based method, the transcriptome sequences were mapped to the *N. parkeri* genome using GMAP (3). Finally, genes predicted from the above methods were integrated by EVidenceModeler (EVM) (4) into a non-redundant consensus of gene sets composing of 22,884 protein-coding genes (Table S6), with an average exon number per gene of 7.78. The average transcript length, average CDS length, average exon length and average intron length was 31,131 bp, 1,382 bp, 177 bp, and 4,381 bp, respectively (Table S7). These statistical indicators of the *N. parkeri* genome showed similar distribution patterns compared to

closely related species (Fig. S5), suggesting dependability of the annotation. Furthermore, a total of 22,788 predicted genes (accounting for 99.58% of total genes) were successfully annotated in at least one of the public databases. There were 21,378 (93.42%), 22,619 (92.00), and 22,629 (99.10%) genes annotated to SwissProt, TrEMBL, and NR database, respectively; 19,005 (83.05%) and 11,768 (51.42%) genes got positive hits in the KEGG and GO databases, respectively (Table S6). Four types of non-coding RNAs were identified of *N. parkeri* genome from homology-based and structure-based annotation, including 220 miRNAs, 1,509 rRNAs, 330 snRNAs, and 19,627 tRNAs, respectively (Table S9).

References

1. Simão F. A., Waterhouse R. M., Ioannidis P., Kriventseva E. V., Zdobnov E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212 (2015).

2. Stanke M., Schöffmann O., Morgenstern B., Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7, 62–73 (2006).

3. Wu T. D., Watanabe C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875 (2005).

4. Haas B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9, R7.1–R7.22 (2008).

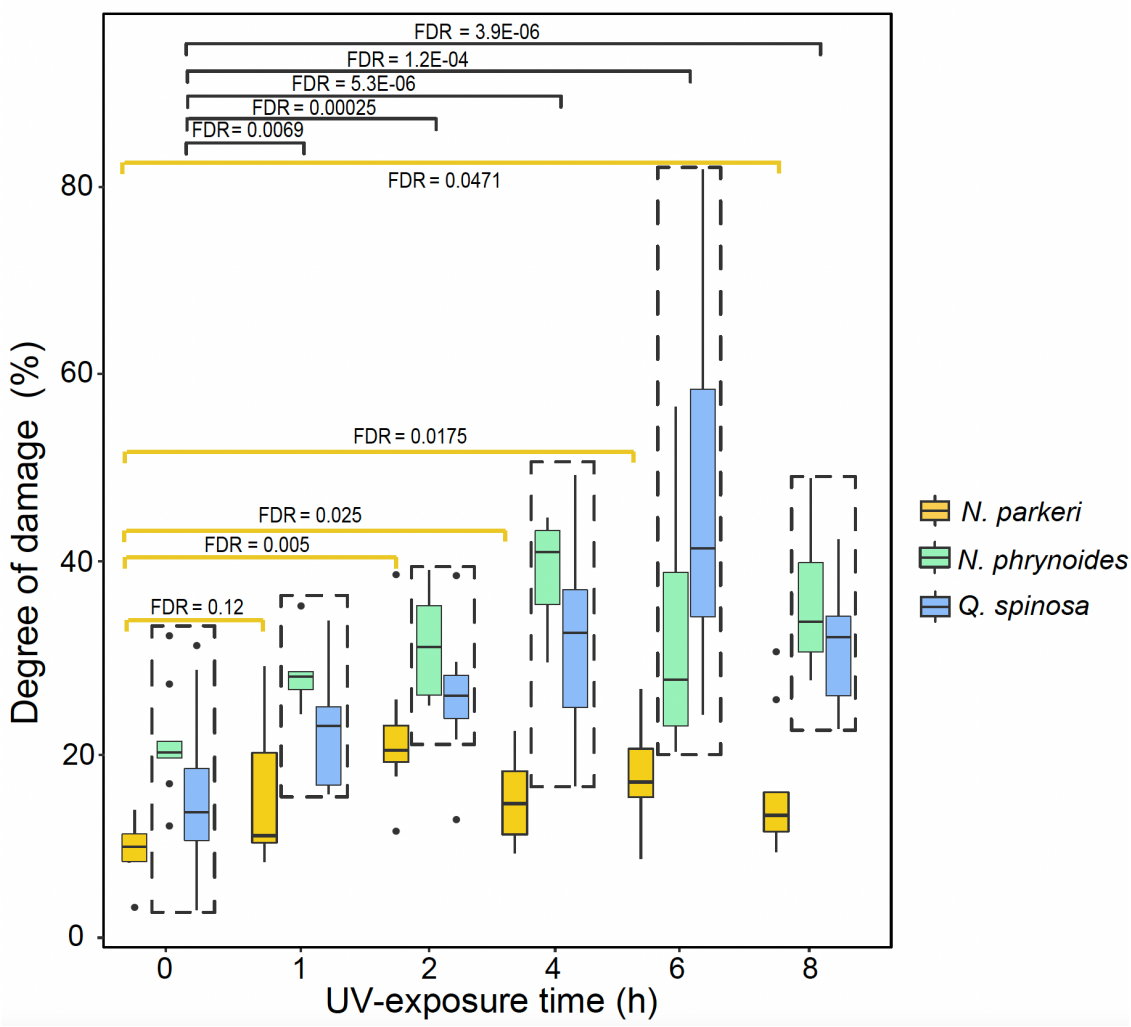

Fig. S1. Quantification of the degree of damage separately in skin of *N. parkeri* (yellow), *N. phrynoides* (green) and *Q. spinosa* (blue) throughout UV exposure based on histological sections. At each time point, at least 9 samples of each species (3 individuals x 3 fields per individual) were used for quantification. Degree of damage of UV-exposed groups was separately compared with the unexposed group (0 h) in each species for a two-tailed *t*-test (*N. phrynoides* and *Q. spinosa* were combined for testing). Box boundaries represent 1st and 3rd quartiles, middle line represents median, and whiskers extend to the nearer of the data extremes.

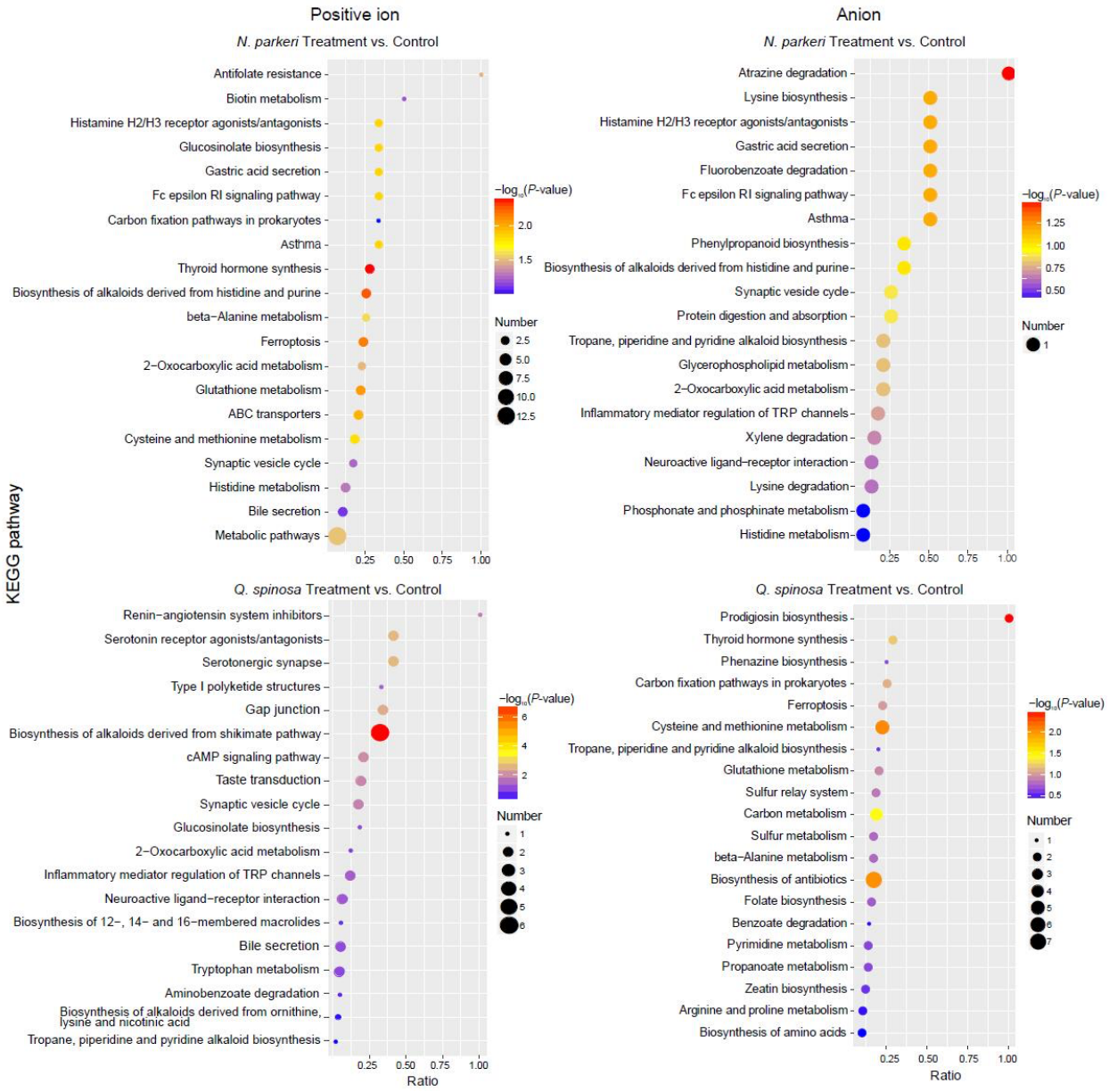

Fig. S2. Top 20 enrichment terms of the KEGG pathway of differently expressed metabolites (DEMs) between Treatment group (UV exposed for 8 h) and Control group (unexposed) based on a LC-MS/MS system. The identified metabolites were divided into positive ion (left tunnel) and anion (right tunnel). Ratio = (number of DEMs)/ (number of total identified metabolites). Higher ratios indicate higher enrichment. Color indicates the *P*-value of hypergeometric test, with blue indicating higher significance. Dot size represents number of identified DEMs.

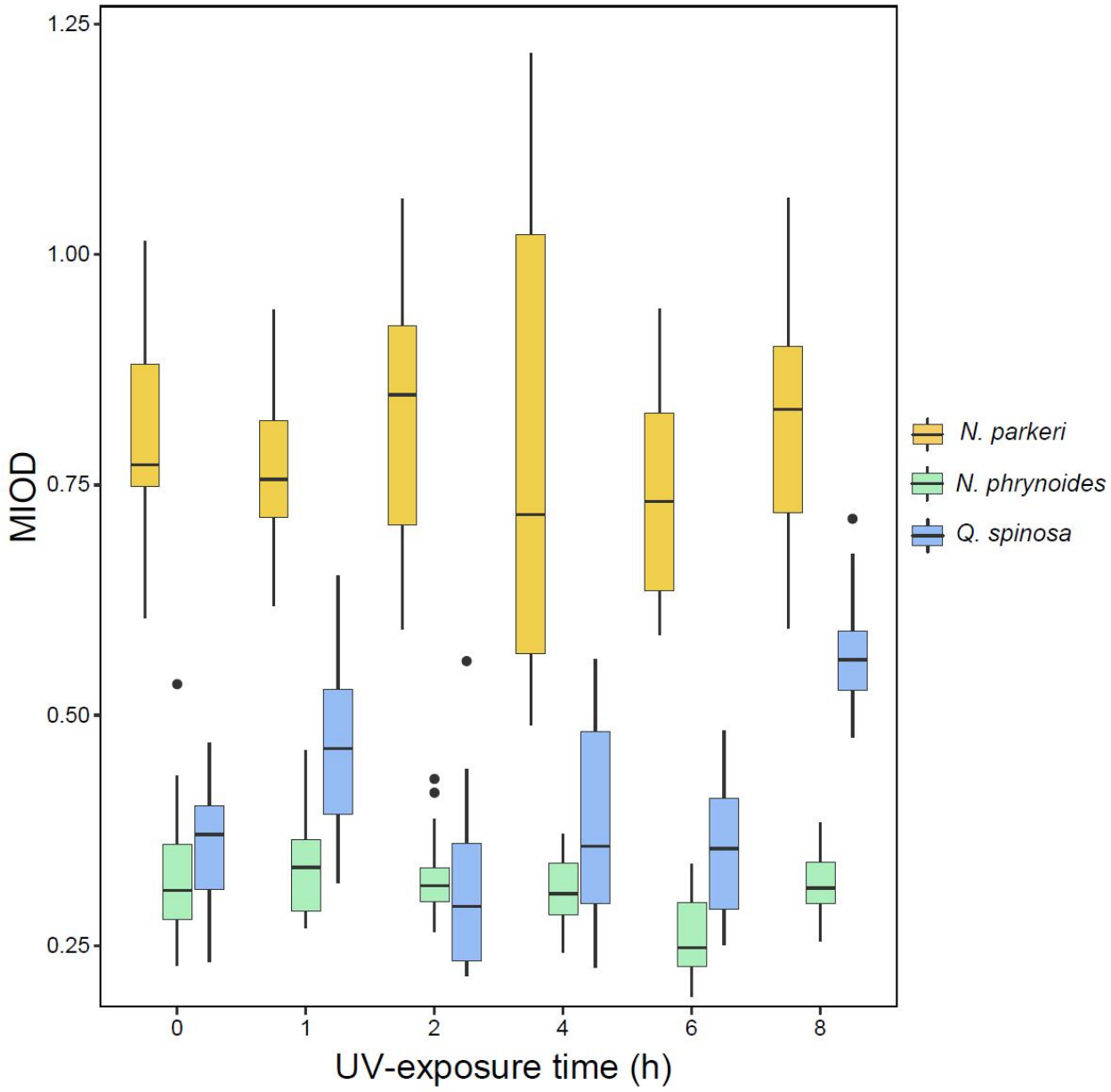

Fig. S3. Changes of the skin pigment content (MIOD) throughout UV exposure in *N. parkeri* (yellow), *N. phrynoides* (green) and *Q. spinosa* (blue) based on histological sections. At each time point, at least 9 samples of each species (3 individuals x 3 fields per individual) were used for quantification. Box boundaries represent 1st and 3rd quartiles, middle line represents median, and whiskers extend to the nearer of the data extremes.

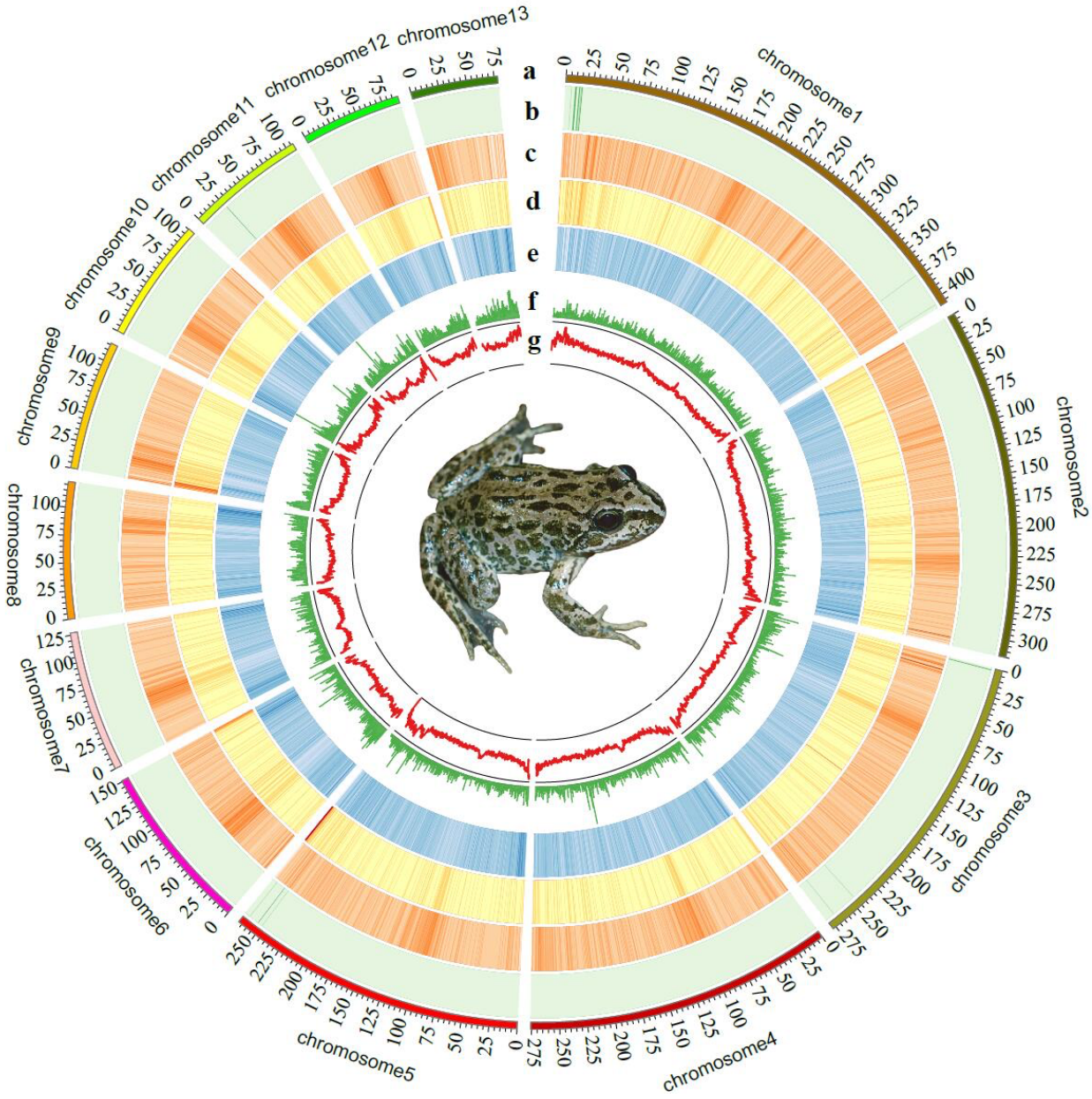

Fig. S4. Genome characteristics of *N. parkeri*. From outer circle to inner circle: **a,** Lines represent 13 chromosomes; **b,** Distribution of short interspersed nuclear elements (SINE); **c,** Distribution of long interspersed nuclear elements (LINE); **d,** Distribution of long terminal repeats (LTR); **e,** Distribution of DNA transposons; **f**, Distribution of GC content; **g,** Distribution of coding density.

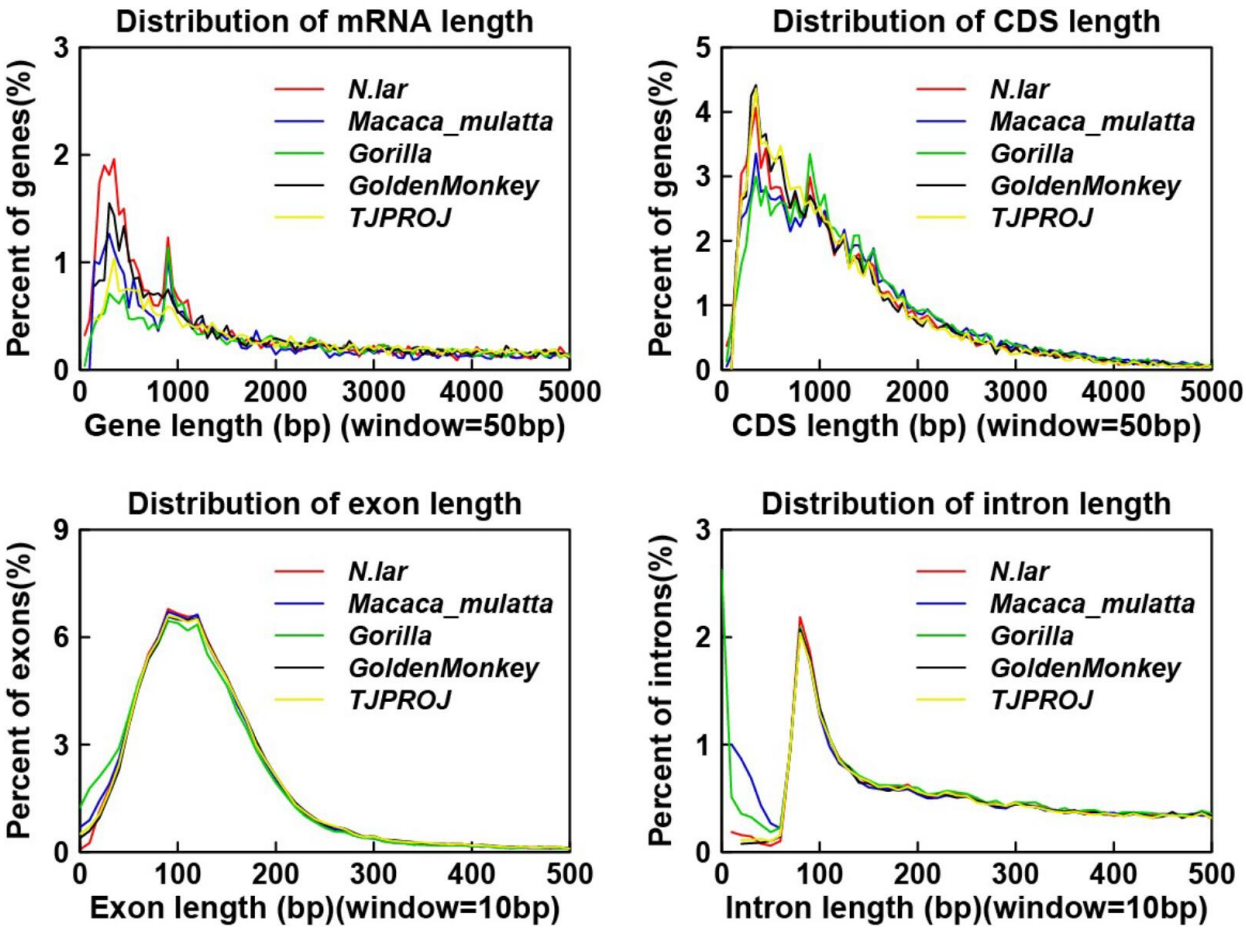

Fig. S5. Comparing the distribution of mRNA length, CDS length, exon length and intron length in *N. parkeri* genome with those in other species.

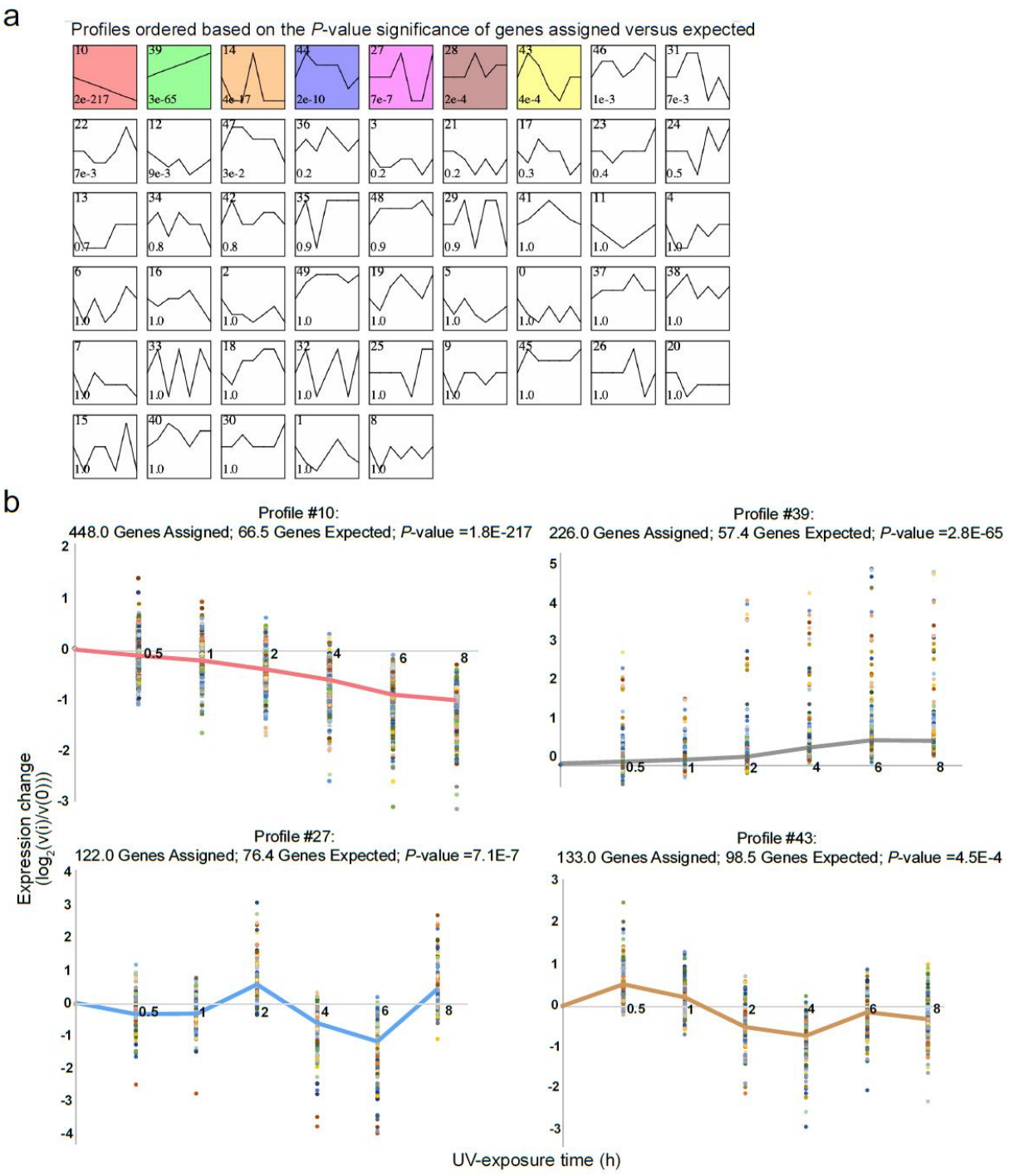

Fig. S6. Clustering of temporal expression genes by STEM in *N. parkeri* (**a**) and examples for expression patterns of gene groups (**b**). **a,** Profiles that can be significantly ($P < 0.001$) clustered are shown in color and ordered by significance. Other profiles are shown in white. The profile number is at the top-left corner and the *P*-value of significance is at the bottom left corner. **b,** Each dot represents a log_2 normalized expression change relative to 0 h (v(i)/v(0)) of an individual gene. Solid line was concatenated from the median value in gene groups of time course.

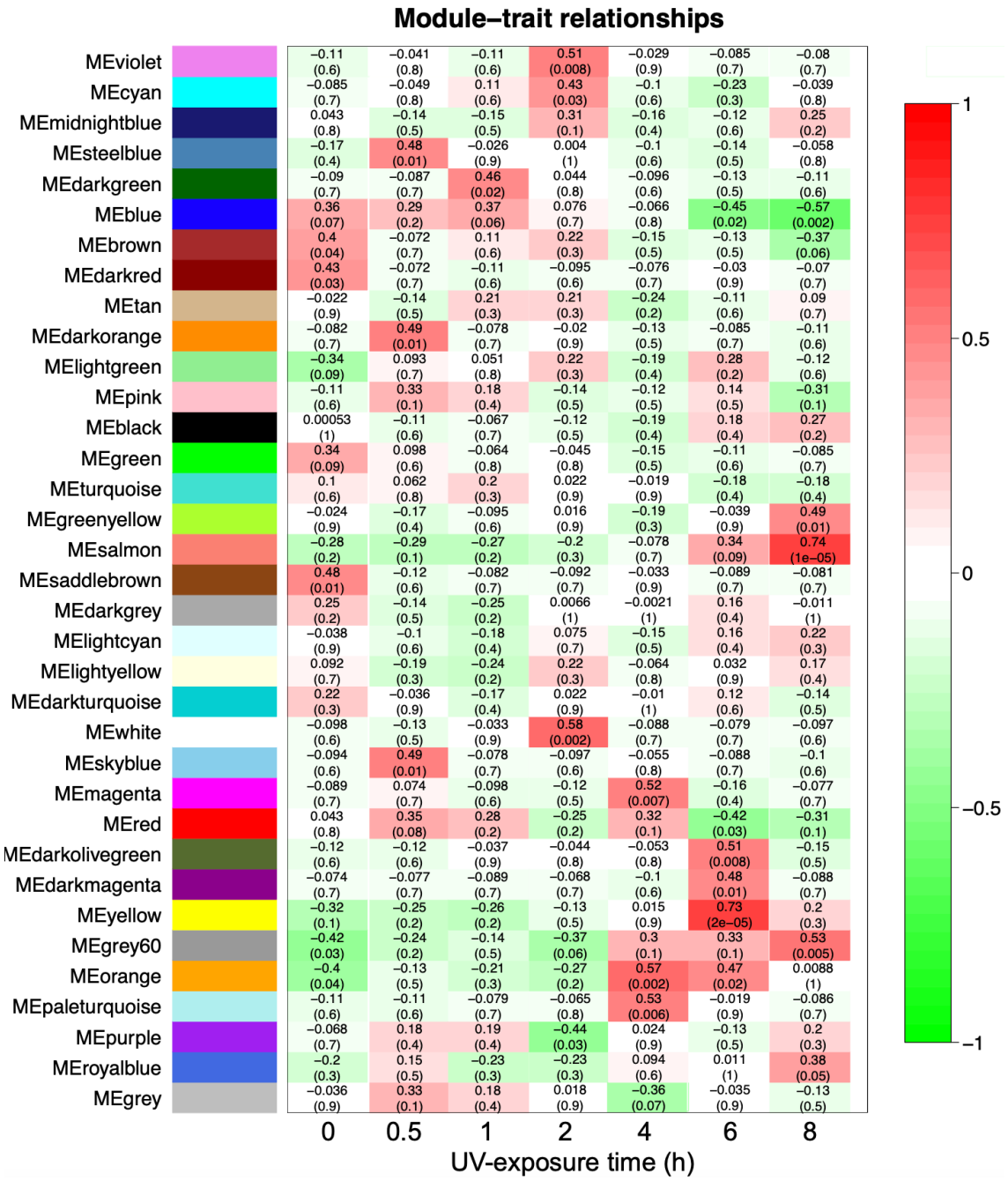

Fig. S7. Module–UV-exposure time relationships of mRNAs in *N. parkeri* according to the WGCNA analysis. The color indicates the respective correlation coefficient, as shown in the scale bar.

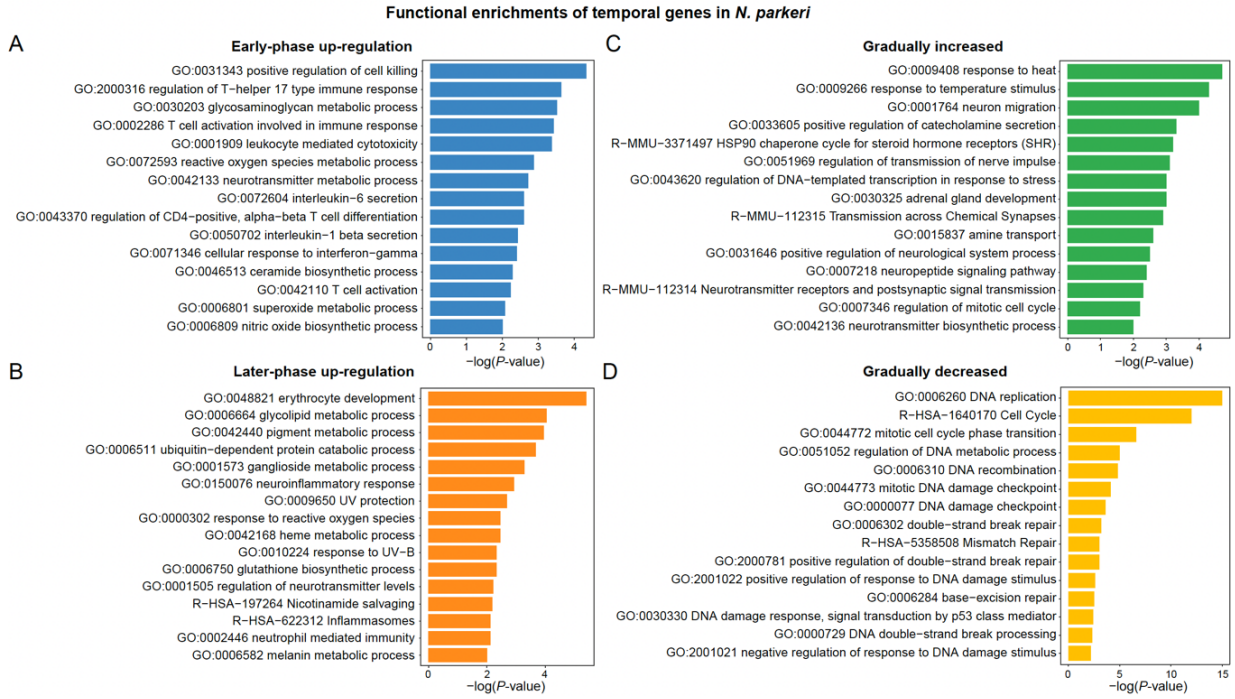

Fig. S8. Functional enrichments of genes in four temporal groups in *N. parkeri*. (*A*) Representative enrichment terms of genes in the early-phase upregulation group. (*B*) Representative enrichment terms of genes in the later-phase upregulation group. (*C*) Representative enrichment terms of genes in the gradually increased expression group. (*D*) Representative enrichment terms of genes in the gradually decreased expression group. Length of a histogram represents -*log*-normalized significance of enrichment with *P* value.

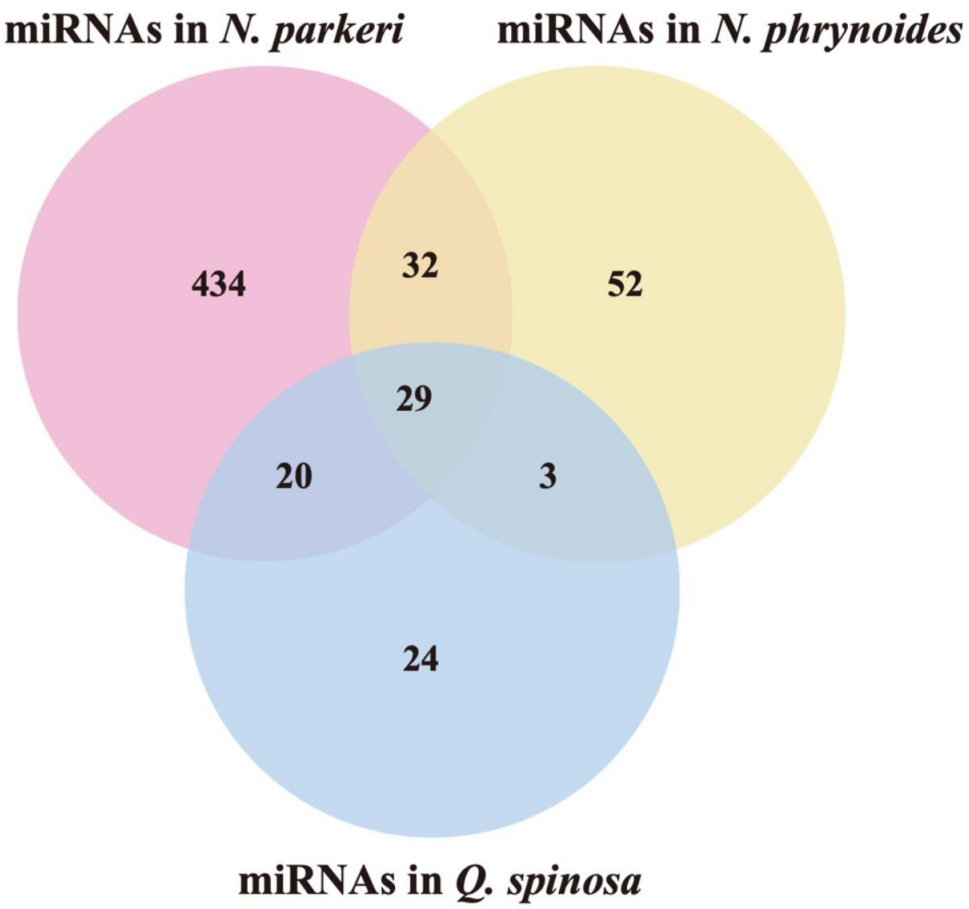

Fig. S9. The Venn diagram of total miRNAs identified in studied frogs.

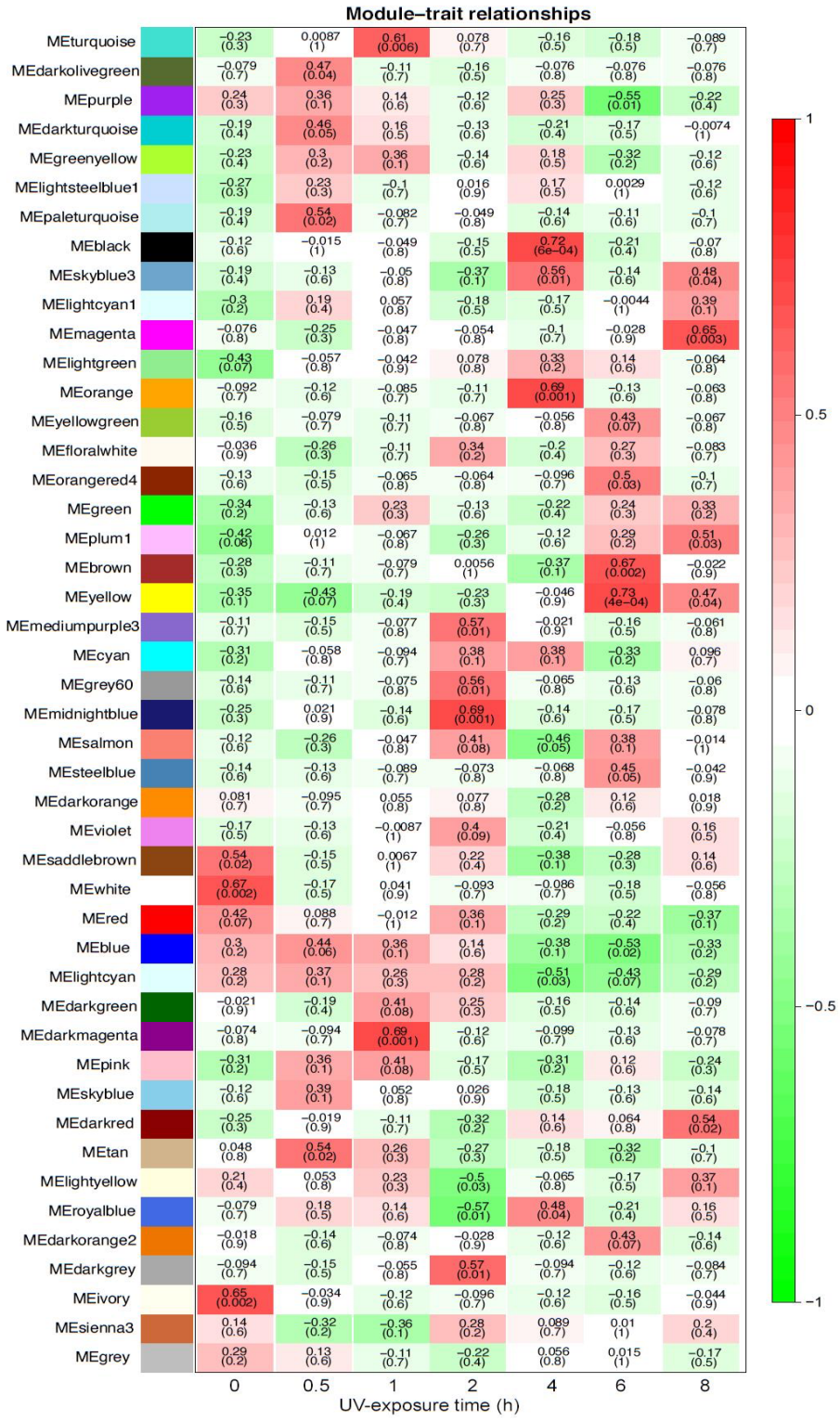

Fig. S10. Module-UV-exposure time relationships of miRNA-mRNAs in *N. parkeri* according to the WGCNA analysis. The color indicates the respective correlation coefficient, as shown in the scale bar.

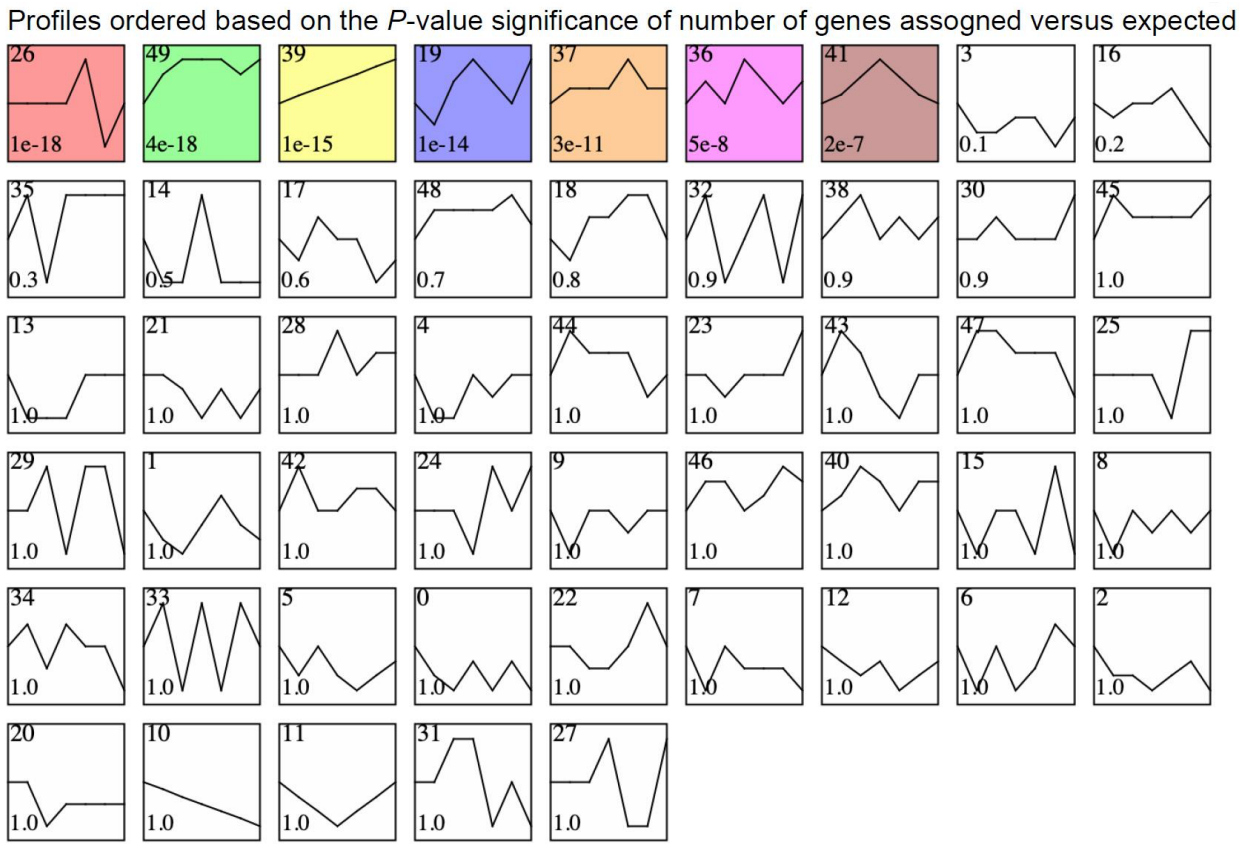

Fig. S11. Clustering of miRNA expression levels in *N. parkeri* by STEM. Profiles that can be significantly ($P < 0.001$) clustered are shown in color and ordered by significance. Other profiles are shown in white. The profile number is at the top-left corner and the *P*-value of significance is at the bottom left corner. The miRNAs in Profile #39 which also showed inverse expression patterns relative to their predicted mRNAs were considered as co-expressed miRNAs in gradually-decreased group.

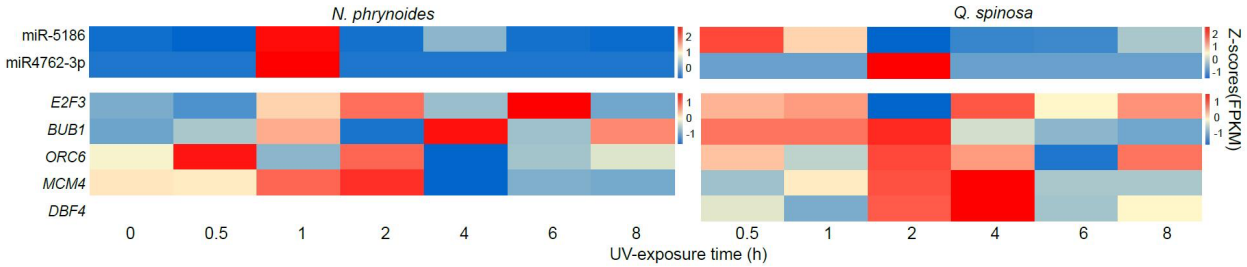

Fig. S12. Expression heatmaps of orthologous miRNAs (upper panels) and mRNAs (lower panels) in *N. phrynoides* and *Q. spinosa*. The color represents an average of expression level (FPKM) normalized by Z-scores.

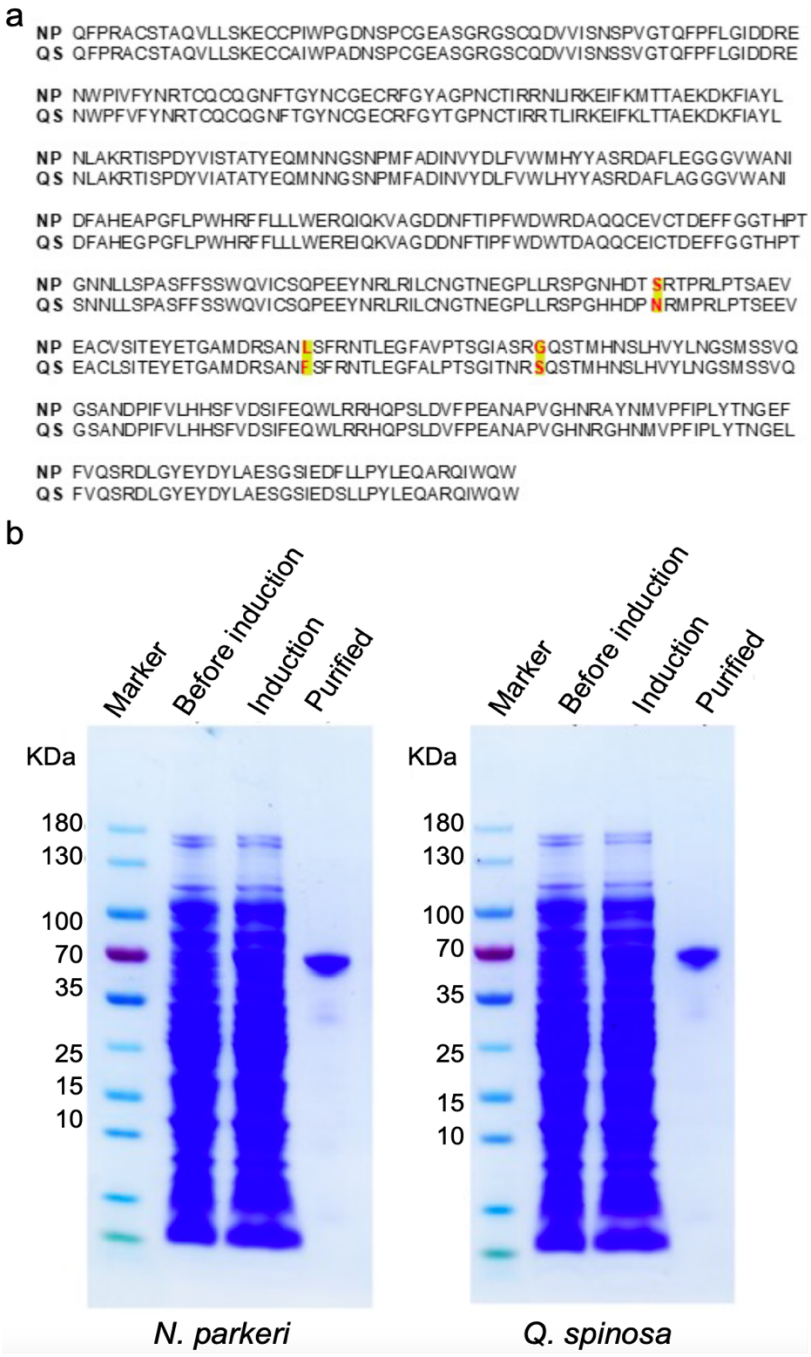

Fig. S13. Amino acids sequences (**a**) and expression *in vitro* of tyrosinase (TYR) (**b**) in *N. parkeri* and *Q. spinosa*. **a,** Sequences aligned here were used to synthesize TYR and then to identify *in vitro* enzymological kinetic analyses. Amino acids replacements that are unique to *N. parkeri* are marked with red font and yellow highlighting. These *N. parkeri* unique replacements were found to be under positive selection. **b**, Synthetic TYR was expressed in BL21, purified with Ni-NTA agarose columns, and analyzed by SDS-PAGE. Lane 1: protein marker; lane 2: negative control; lane 3: protein expression after induction; and lane 4: purified protein.

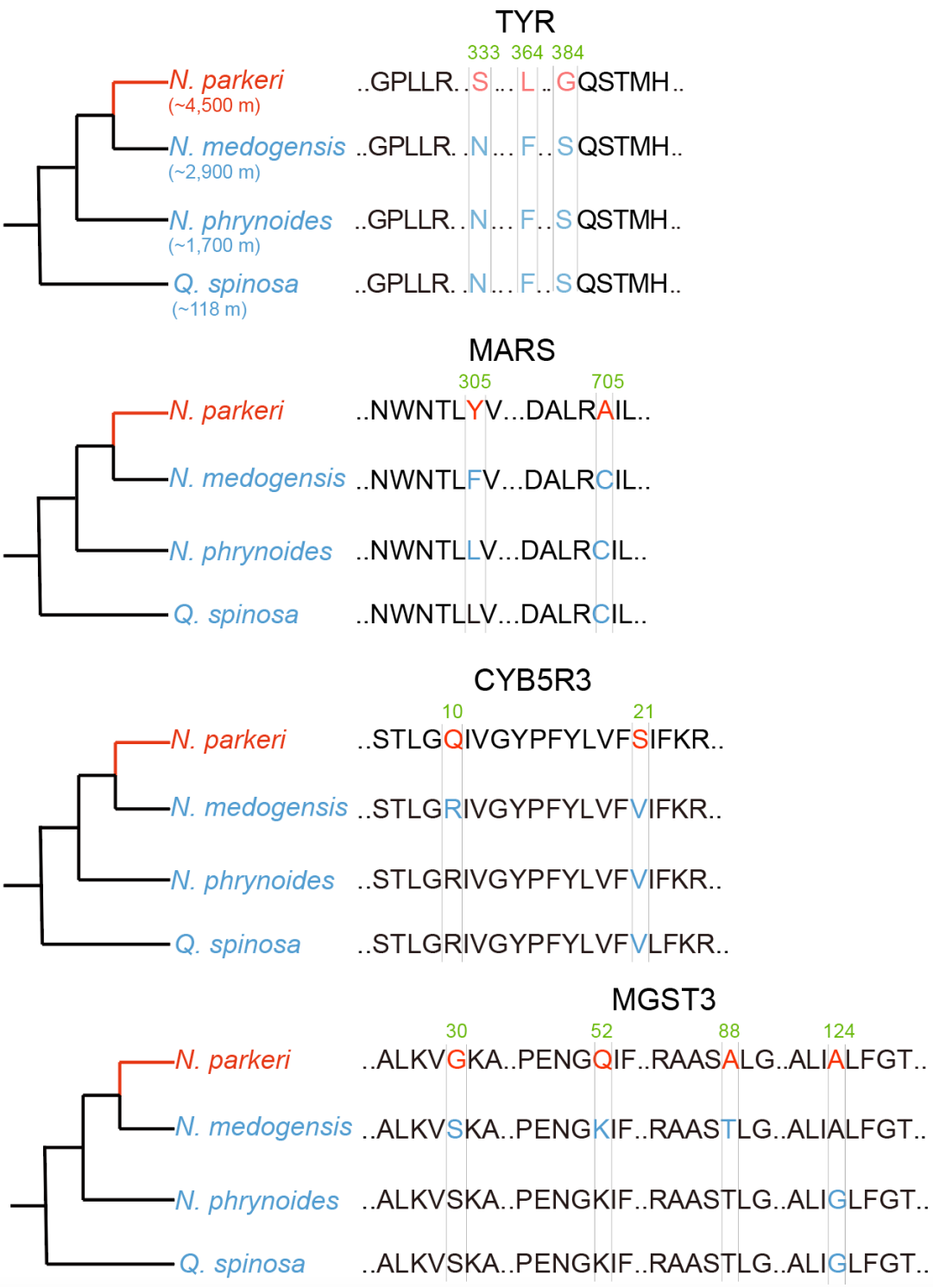

Fig. S14. Phylogenetic relationships and examples of schematic alignment of amino acid (AA) replacements in *N. parkeri* and lower-elevation relatives. AA replacements unique to *N. parkeri* are marked with red font. The number above the alignments refers to the reference sequence of *N. parkeri.* Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; I, Ile; K, Lys; L, Leu; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; Y, Tyr.

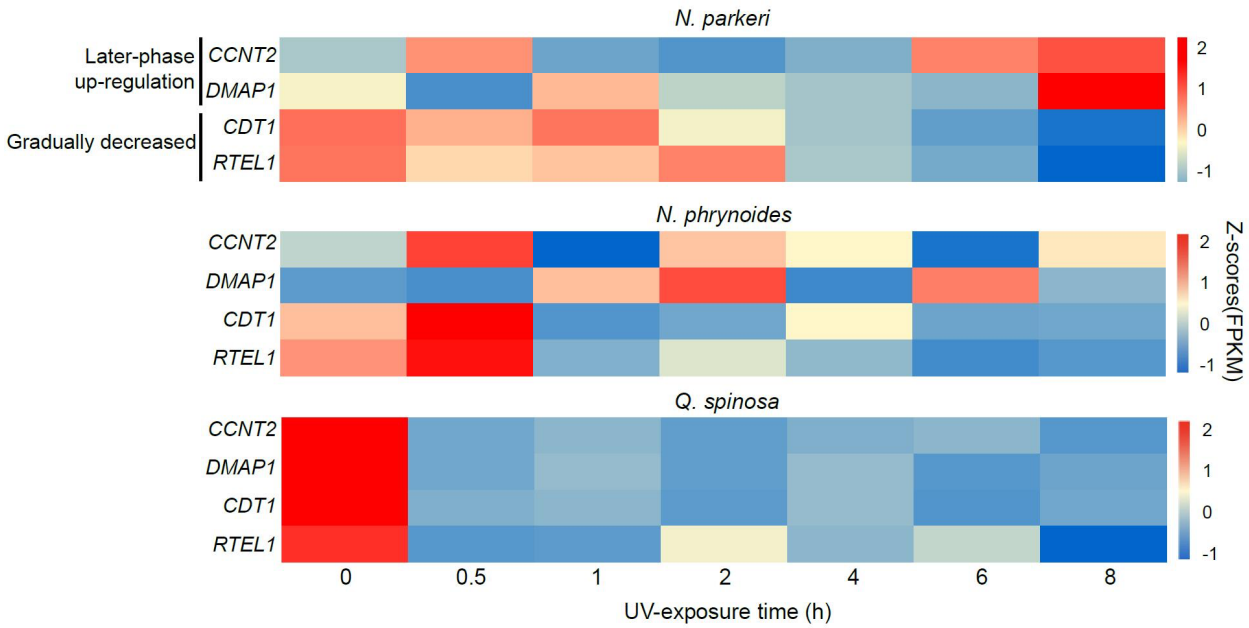

Fig. S15. Expression changes of temporal expression genes in *N. parkeri* and corresponding orthologs in lower-elevation relatives (*N. phrynoides* and *Q.spinosa*) throughout UV exposure. The color represents an average of expression level (FPKM) at a time point normalized by Z-scores.

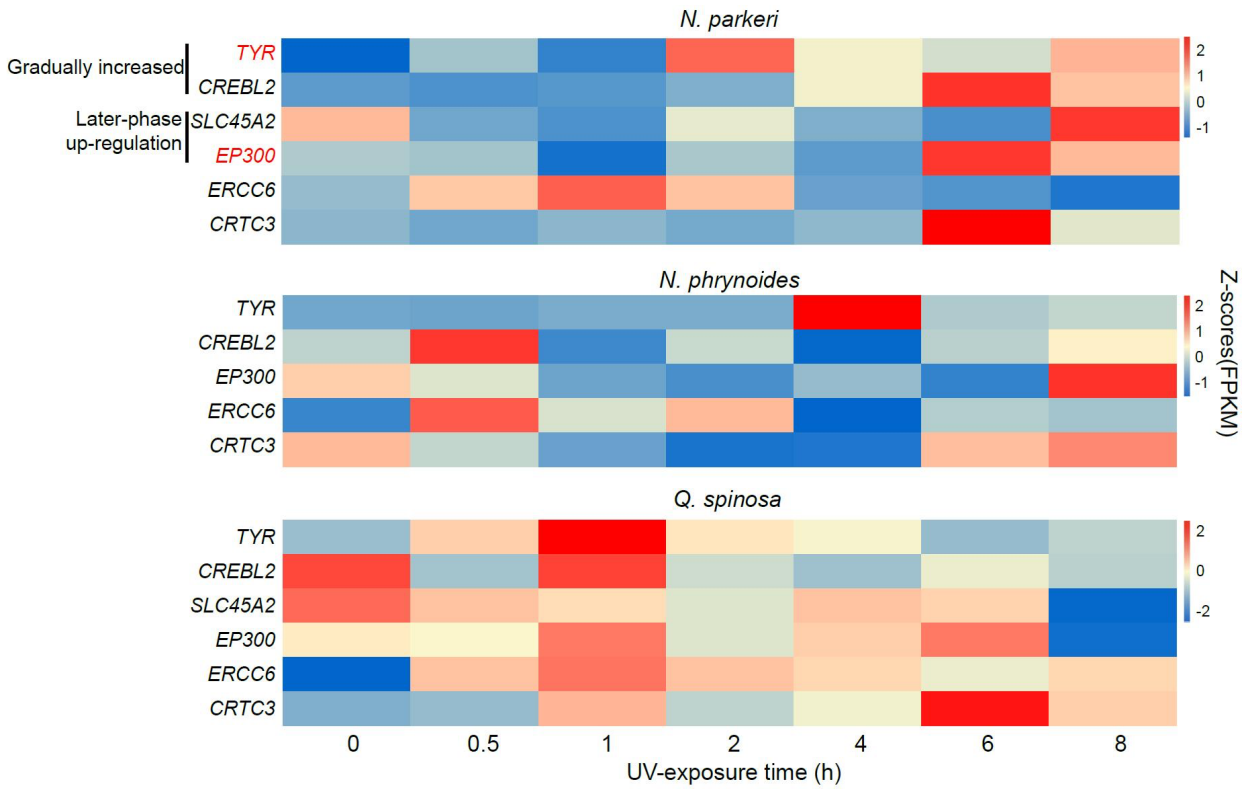

Fig. S16. Expression changes of PSGs and temporal expression genes in *N. parkeri*, and corresponding orthologs in lower-elevation relatives (*N. phrynoides* and *Q.spinosa*) throughout UV exposure. PSGs in *N. parkeri* are in red font. The color of the scare bar represents an average expression level (FPKM) at a time point normalized by Z-scores.

Table S1. Specimen information, collection sites, and use of samples (Y = yes; N = no).

	N. parkeri	***N. phrynoides***	***Q. spinosa***
Sampling location	Mozhugongka, Tibet, China (4,478 m)	Jingdong, Yunnan Province, China (1,760 m)	Huangshan, Anhui Province, China (176 m)
Histological sections	Y	Y	Y
Identification of $ABTS^{+}$ free radical-scavenging activity	Y	N	Y
Metabolite identification	Y	N	Y
Genome sequencing	Y	N	N
Time-course UV exposure	Y	Y	Y
RNA-seq	Y	Y	Y
miRNA-seq	Y	Y	Y
Tyrosinase synthesis and activity identification	Y	N	Y

Table S2. Statistics of contig-level genome assemble of *N. parkeri*. N90: the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 90% of the assembly, and so on for N80 to N50.

	Origin Falcon		Polish		Pilon	
	Size (bp)	**Number**	**Size (bp)**	**Number**	**Size (bp)**	**Number**
N90	183,302	1,563	184,350	1,565	184,362	1,565
N80	635,982	854	640,635	855	640,378	855
N70	1,181,135	575	1,187,113	576	1,186,991	576
N60	1,733,635	406	1,745,301	406	1,744,822	406
N50	2,323,300	282	2,337,137	283	2,336,739	283
Longest	13,630,581	----	13,716,666	----	13,710,411	----
Total Size	2,456,683,934	----	2,473,585,870	----	2,472,697,895	----
>=100bp	----	7,134	----	7,134	----	7,134
>=2kb	----	6,884	----	6,887	----	6,886

Table S3. Statistics of *N. parkeri* genome assessment with BUSCO.

Number	Percentage (%)	Type
2272	87.9	Complete BUSCOs (C)
2146	83	Complete and single-copy BUSCOs (S)
126	4.9	Complete and duplicated BUSCOs (D)
145	5.6	Fragmented BUSCOs (F)
169	6.5	Missing BUSCOs (M)
2586	----	Total BUSCO groups searched

Table S4. Statistics of chromosome-level genome assemble of *N. parkeri*.

	Scaffold	
	Size(bp)	**Number**
N90	107,000,516	11
N80	116,390,974	9
N70	126,942,726	7
N60	150,785,030	6
N50	268,565,388	5
Longest	407,978,197	----
Total Size	2,475,211,768	----
Total Number(>=100bp)	----	1,869
Total Number(>=2kb)	----	1,732

Table S5. Comparison of genome assemble and annotation among anuran amphibians.

Species	Assembled genome size (Gb)	Contig N50 (bp)	Scaffold N50 (bp)	Repetitive sequences (%)	Protein coding genes
Nanorana parkeri	2.47	2,336,739	268,565,388	58.9	22,884
Leptobrachium leishanense	3.5	1,931,129	394,693,044	77.1	23,420
Leptobrachium ailaonicum	3.58	821,125	412,424,790	69.48	26,227
Rana catesbeiana	5.8	5,239	68,964	66.9	22,000
Rhinella marina	2.6	167,489	167,489	86.6	58,302
Oophaga pumilio	5.5	3,942	59,503	62.9	17,051
Xenopus tropicalis	1.5	71,041	135,134,832	51.9	27,047
Xenopus laevis	2.7	19,713	136,570,856	38.3	37,385

Table S6. Statistics of gene annotation in *N. parkeri* genome.

		Number	Percent (%)
Total		22,884	
Annotated	InterPro	21,140	92.38
	GO	11,768	51.42
	KEGG	19,005	83.05
	Swissprot	21,378	93.42
	TrEMBL	22,619	92
	NR	22,679	99.1
Unannotated		96	0.42

Table S7. Statistics of gene prediction.

Gene set		Number	Average transcript length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
de novo	*AUGUSTUS*	21,608	52,280	1,338	7.7	173	7,575
	H. sapiens	19,133	39,177	1,661	9.4	177	4,462
	G. gorilla	20,479	33,465	1,501	8.6	173	4,187
Homolog	*M. mulatta*	20,538	34,652	1,499	8.7	172	4,300
	M. musculus	17,756	38,852	1,571	9	174	4,664
	R. roxellana	20,279	34,597	1,585	8.9	177	4,165
Final set		22,884	31,131	1,382	7.78	177	4,381

Table S8. Statistics of annotated transposable elements (TEs).

Type	Repbase TEs length (bp)	% in genome	TE proteins length (bp)	% in genome	De novo length (bp)	% in genome	Combined TEs	
							Length (Bp)	% in genome
DNA	94,358,202	3.82	15,323,977	0.61	530,152,426	21.44	574,368,176	23.23
LINE	85,162,083	3.44	127,854,008	5.17	570,544,992	23.07	622,326,934	25.17
SINE	8,690,955	0.35	0	0	20,450,549	0.83	28,042,113	1.13
LTR	40,483,305	1.64	24,854,865	1.01	351,387,174	14.21	366,123,006	14.81
Other	14,470	0.0006	0	0	9,698	0.0004	24,049	0.001
Unknown	0	0	0	0	4,154,629	0.17	4,154,629	0.17
Total	220,290,016	8.91	168,006,648	6.79	1,320,029,475	53.38	1,364,981,392	55.2

Table S9. Statistics of non-coding RNA prediction.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		220	94.66	20,826	0.000842
tRNA		19,627	75.12	1,474,328	0.059624
rRNA	rRNA	1,509	137.07	206,838	0.008365
	18S	122	293.8	35,843	0.00145
	28S	284	196.34	55,760	0.002255
	5.8S	23	105.52	2,427	0.000098
	5S	1,080	104.45	112,808	0.004562
snRNA	snRNA	330	128.66	42,457	0.001717
	CD-box	41	104.61	4,289	0.000173
	HACA-box	25	143.16	3,579	0.000145
	splicing	258	128.64	33,190	0.001342

Table S10. Statistics of the transcriptome assemblies of *N. phrynoides* and *Q. spinosa*.

	N. phrynoides	***Q. spinosa***
Number of seqs	178,551	81,931
Mean length (bp)	973.96	1259.67
Longest (bp)	26,386	90,220
N90	402	498
N70	731	1,218
N50	1,471	2,071

Table S11. Relationships of co-expressed miRNA-mRNAs in gradually-decreased and later-phase temporal groups in *N. parkeri*.

Gradually decreased group			
miRNA	**mRNA**	**Pearson Correlation**	***P*-value**
gga-miR-456-3p	*C1orf112*	-0.8191004	0.02418845
	ATAD2B	-0.7916125	0.03391129
hsa-miR-4711-5p	*CHAF1B*	-0.7801322	0.03852019
	MSANTD3	-0.8699893	0.01090092
	NFIA	-0.8444652	0.01682064
	CDCA7	-0.8504237	0.01530709
	UBAP1L	-0.9033303	0.00529432
	SLC8A1	-0.7823912	0.03758749
	BCAT1	-0.7940028	0.0329926
	ACAD8	-0.8179993	0.02454278
hsa-miR-4762-3p	*B3GALT5*	-0.7964512	0.03206616
	CRP	-0.8493524	0.01557322
	DBF4	-0.763739	0.04566992
gga-miR-6587-5p	*LRRFIP2*	-0.7713126	0.04228331
	CDT1	-0.7570061	0.04880151
hsa-miR-5186	*ORC6*	-0.8105978	0.0270001
	MCM4	-0.7696406	0.0430186
	RNASEH2C	-0.8340297	0.01966894
hsa-miR-1288-5p	*ORC6*	-0.7679027	0.04379029
hsa-miR-1288-3p	*PAPPA*	-0.7971506	0.03180424
	CRP	-0.7832334	0.037243
	GINS4	-0.765581	0.04483301
	THUMPD1	-0.835378	0.01928667
	ORC6	-0.8230241	0.02294939
	HSD17B12	-0.7582626	0.04820844
	SLC24A2	-0.9020196	0.00547163
	IQCH	-0.8417394	0.01754025
	TMEM144	-0.8672699	0.01146243
	SPECC1	-0.8187286	0.02430777
hsa-miR-744-3p	*IFRD1*	-0.7802601	0.03846707
	MSANTD3	-0.8639703	0.01216559
	SYNC	-0.7652601	0.0449782
	SETMAR	-0.8711499	0.01066621
	E2F3	-0.8299849	0.02084133
	PSMC3IP	-0.917823	0.0035558
	CDCA8	-0.8563841	0.013874
	C1orf112	-0.8987079	0.00593486
	BUB1	-0.7790372	0.03897688
gga-miR-128-3p	*PAPPA*	-0.7679027	0.04379029
	HCN4	-0.8688884	0.01112629
	E2F3	-0.7955461	0.03240694
gga-miR-148a-3p	*HMGA2*	-0.9049274	0.00508284
	E2F7	-0.8630111	0.01237452
	E2F3	-0.7628172	0.04609191
gga-miR-31-5p	*DGKI*	-0.8344002	0.01956348
gga-miR-425-5p	*B3GALT5*	-0.8144547	0.02570314
Later-phase upregulation group			
miRNA	**MM**	***P*-value**	
hsa-miR-3617-3p	-0.67	0.0016	
hsa-miR-4440	-0.6	0.006	
hsa-miR-6847-5p	-0.63	0.004	

Dataset S1. Photographs and statistics of histological skin sections in studied species.
Dataset S2. Metabolomics identification.
Dataset S3. Orthologous genes across studied species.
Dataset S4. Genes of early-phase up-regulation and associated functional enrichments.
Dataset S5. Genes of later-phase up-regulation and associated functional enrichments.
Dataset S6. Genes that gradually increased in expression and associated functional enrichments.
Dataset S7. Genes that gradually decreased in expression and associated functional enrichments.
Dataset S8. Conservative and novel miRNAs predicted in studied frogs.
Dataset S9. Normalized expression levels of miRNAs in studied frogs.
Dataset S10. List of PSGs and associated functional enrichments.