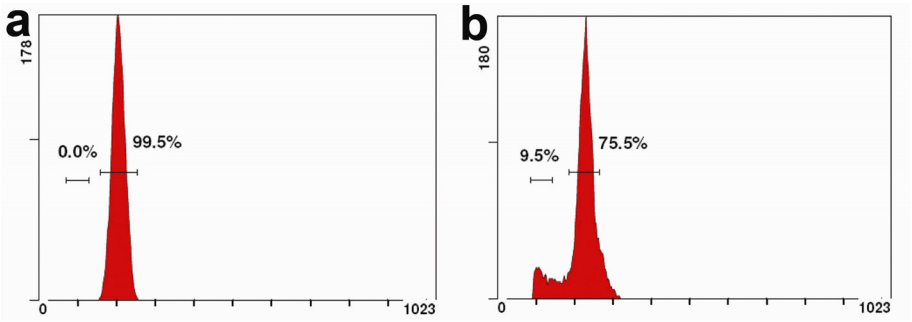
Supplementary Information

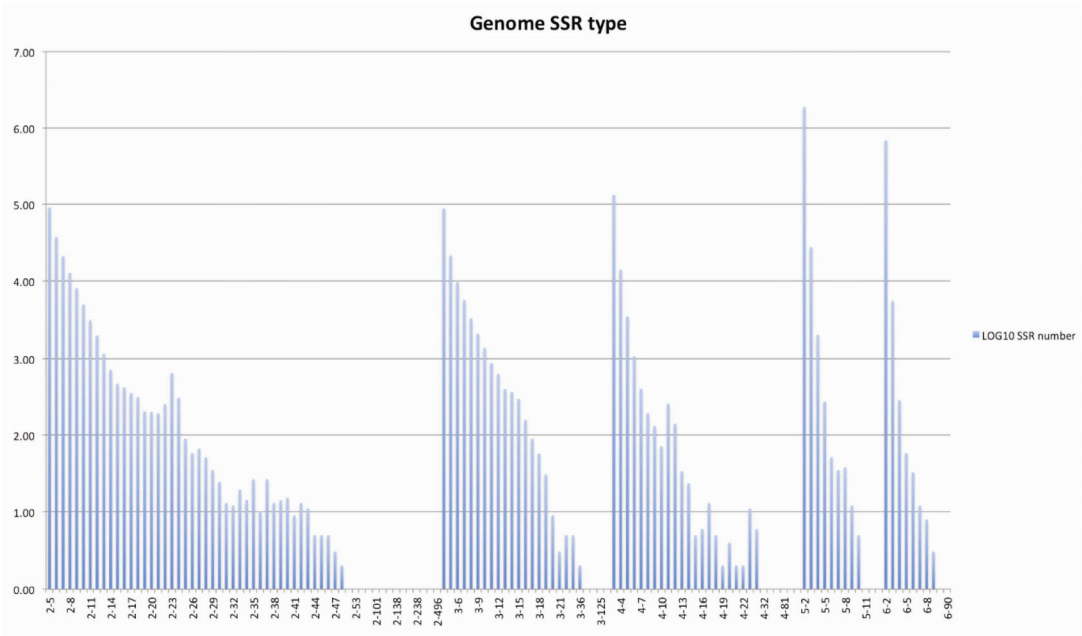
Table of Contents

Supplementary Figures



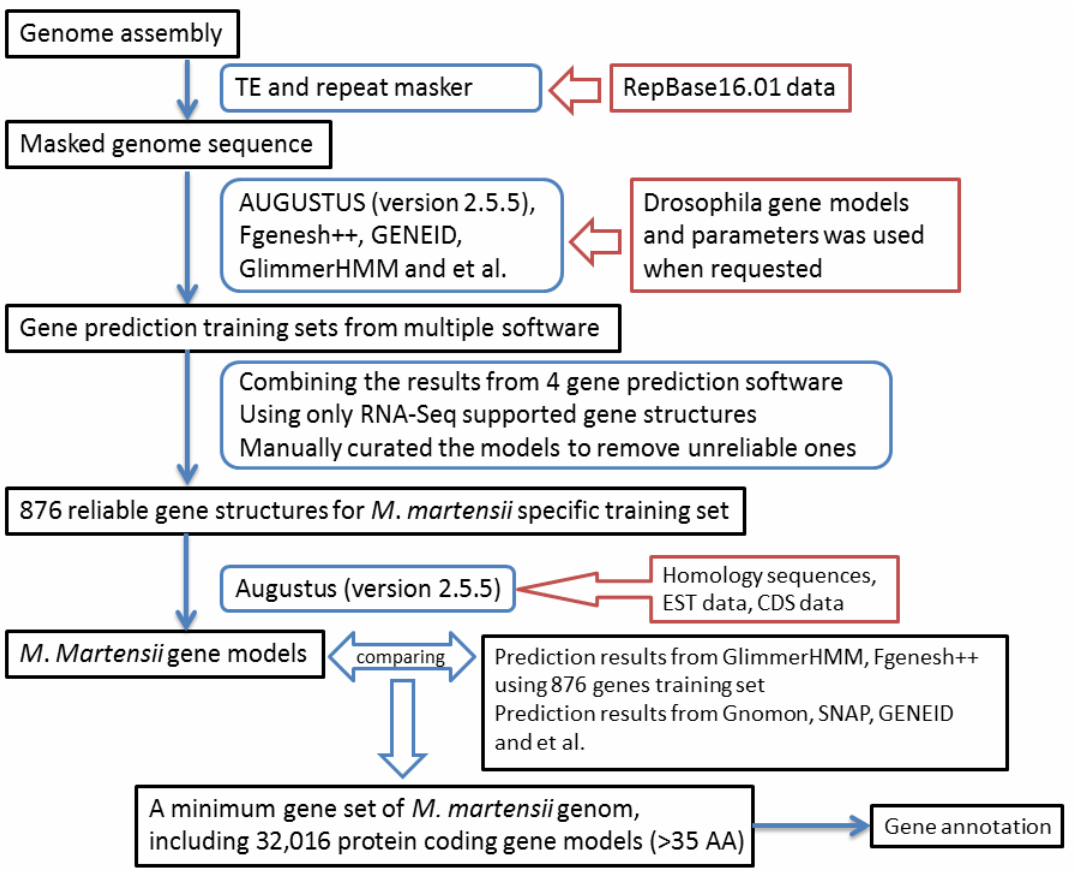
Supplementary Figure S1 Genome size determination of the scorpion *M. martensii* by flow cytometry analysis of fluorescently stained nuclei using chicken erythrocytes as an internal standard.

a, Chicken erythrocytes. b, Scorpion genital gland tissues. X-axis and Y axis show the relative fluorescence and the number of nuclei, respectively.



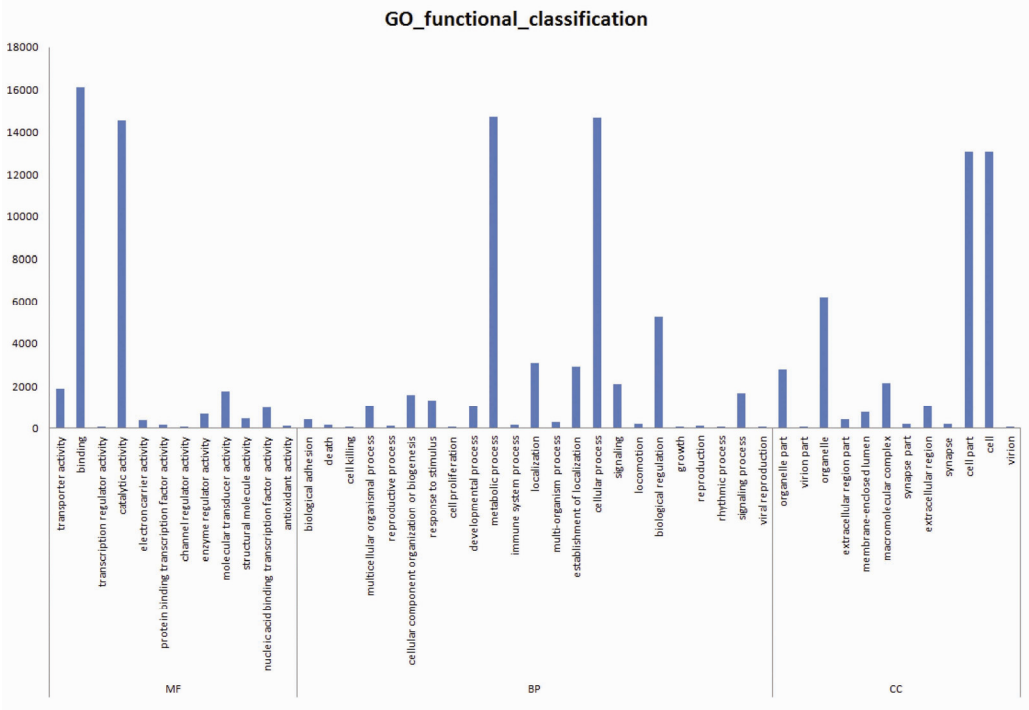
Supplementary Figure S2 Number of different types of SSR from the *M. martensii genome.*

The x-axis represents the type of SSR, that is, '2-6' represents dinucleotide repeats - 6 repeats. The y-axis represents the log10-transformed number of each SSR type.



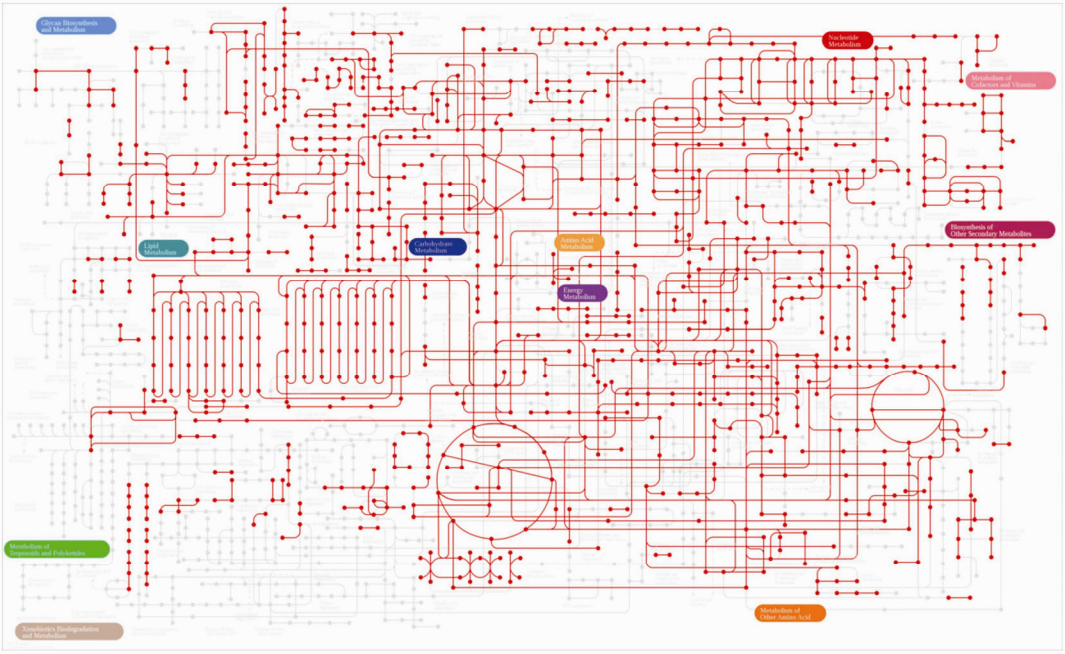
Supplementary Figure S3 Gene prediction pipeline for the *M. martensii* genome.

This figure illustrated the workflow used to predict *M. martensii* gene models. The black boxes represent the major income and outcome. The blue boxes represent the software used in the workflow. The red boxes represent the parameters and dataset required.



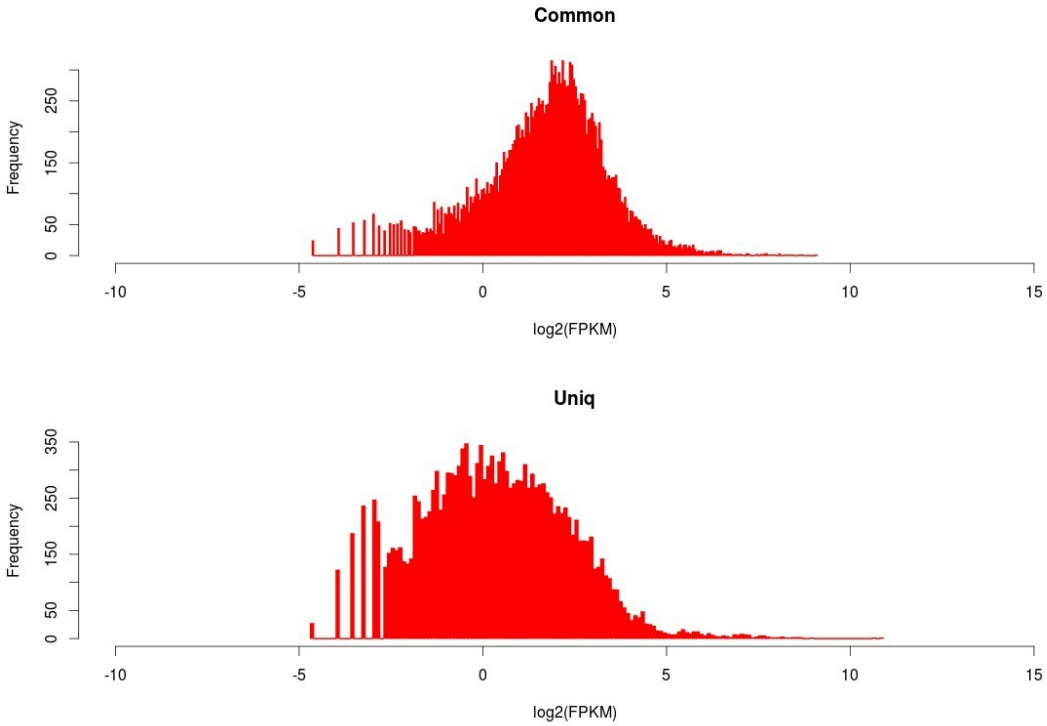
Supplementary Figure S4 Gene Ontology categories of predicted genes.

Gene Ontology (GO) annotation of predicted scorpion genes was conducted using GO annotation of homology sequences from the SwissProt, TrEMBL, and NCBI NR databases. The figure illustrates the number of genes from major GO modules of molecular function (MF), biological process (BP), and cellular component (CC).



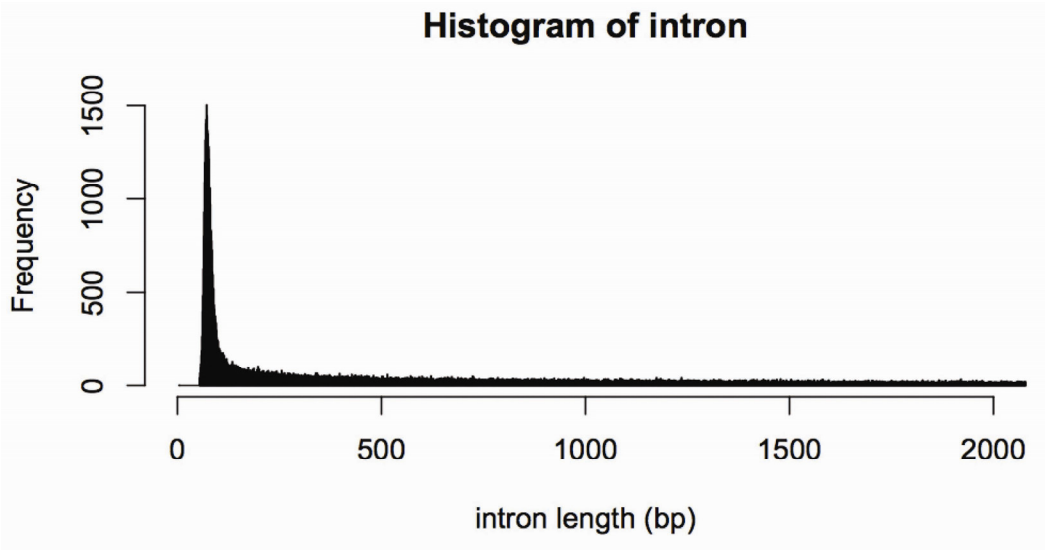
Supplementary Figure S5 KEGG map of predicted genes of *M. martensii*.

The figure was made by Kyoto Encyclopedia of Genes and Genomes (KEGG) Mapper – Color Pathway tools (http://www.genome.jp/kegg). This figure shows the genes involved in the metabolism pathways of *M. martensii*. The read lines represent the enzymes which are coded by genes existed in *M. martensii* genome.



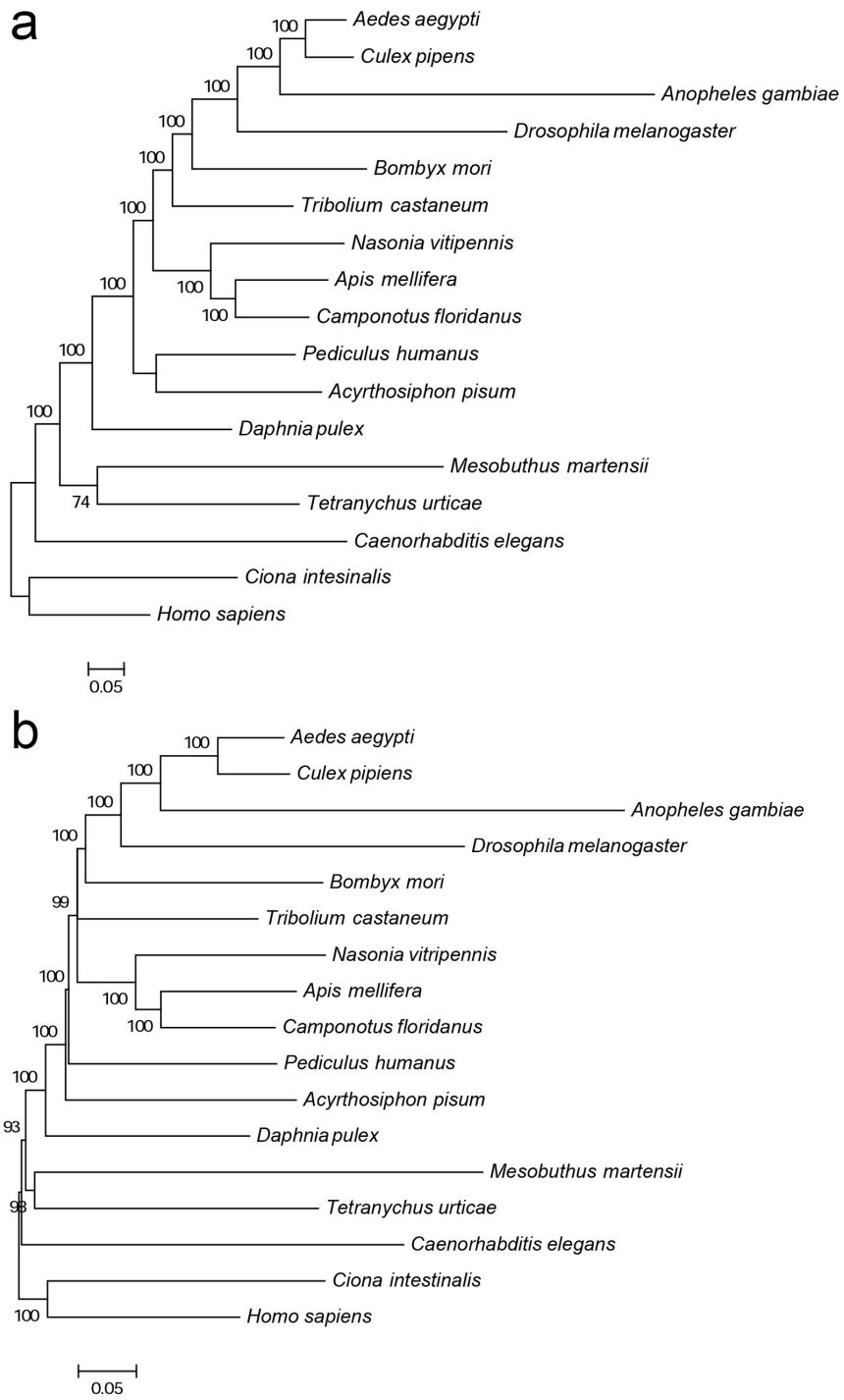
Supplementary Figure S6 Distributions of gene expression.

The distribution of the shared and *Mesobuthus*-specific genes expression levels measured with RPKM. Illumina Solexa Hiseq 2000 sequencing data from mixed tissues were mapped onto predicted gene model with Tophat. The RPKM values were calculated with Cufflinks. The x-axis represents the Log2 transferred RPKM value of each gene. The classification of genes were determined with MCL clustering results.



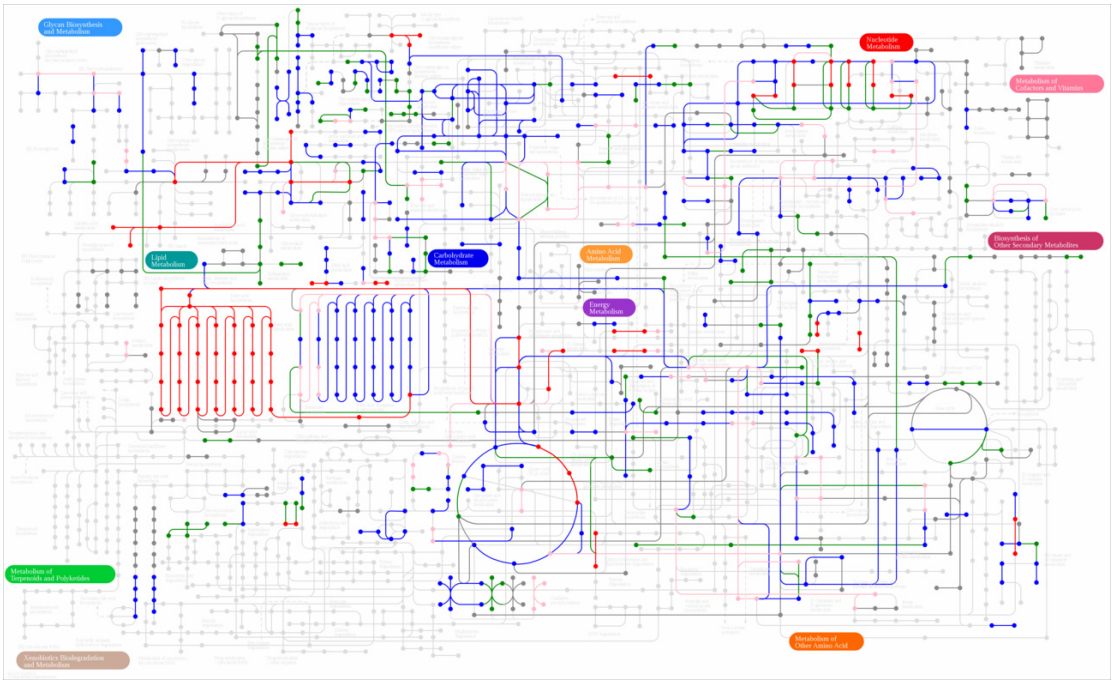
Supplementary Figure S7 Intron length distribution for genes of *M. martensii*.

The average intron size is approximately 2.12 Kb.



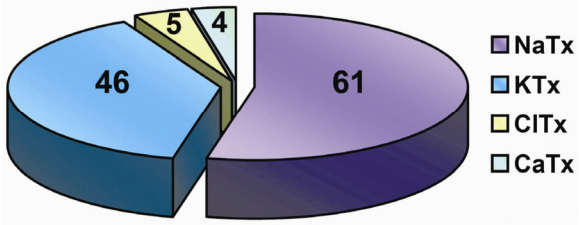
Supplementary Figure S8 Phylogenetic tree of eukaryotic core gene families.

Among all core gene families from eukaryotes, 220 proteins have well-acceptable aligned sequences in the 17 genomes considered. The phylogenetic tree was constructed based on the ML (a) and NJ method (b) in MEGA5[57]. The 500-fold bootstrap testing values (> 70) are shown in the branches. The tree was rooted using *Homo sapiens*, and *Ciona intestinalis* as the most external out-groups.



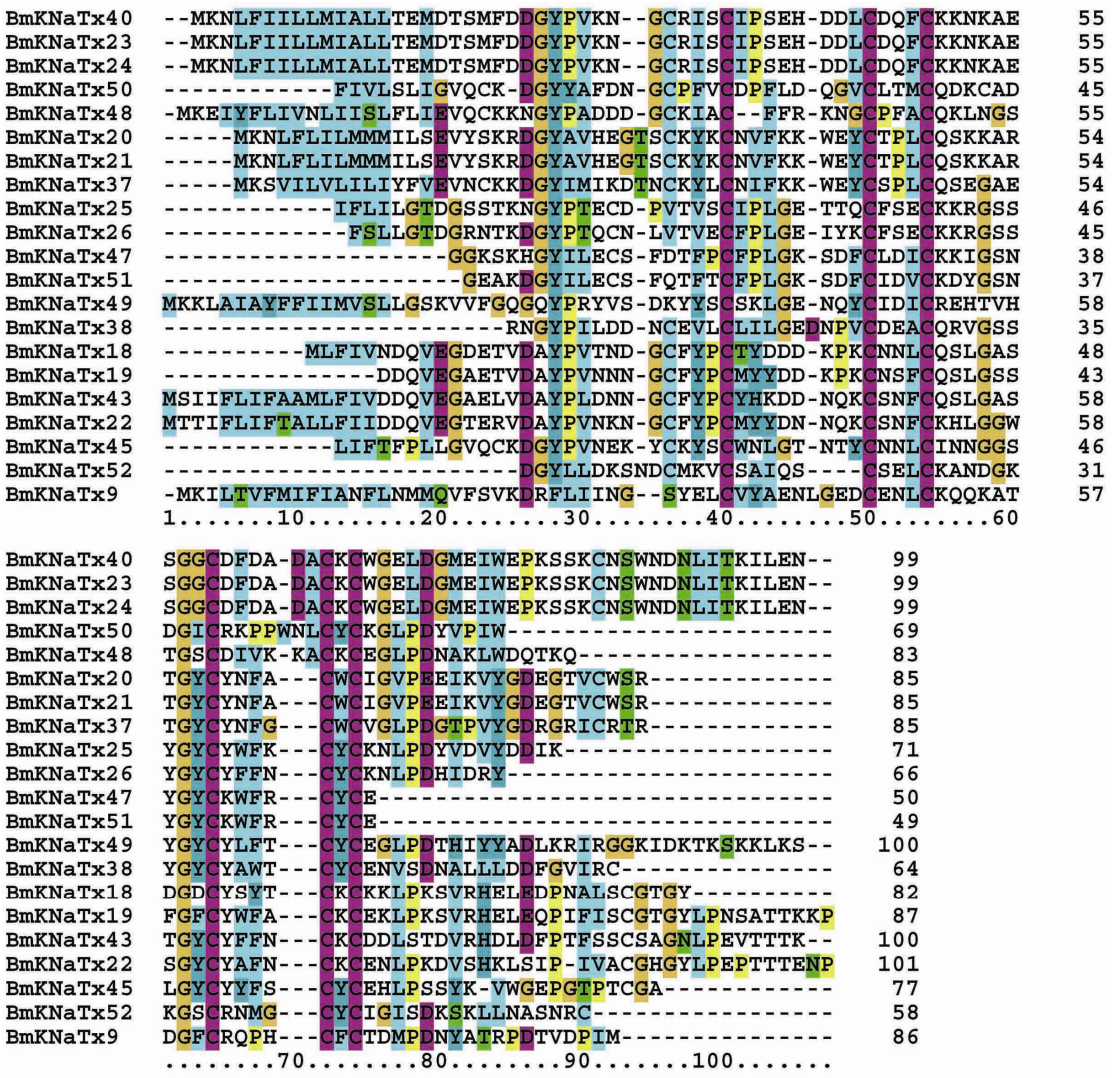
Supplementary Figure S9 KEGG map of expanded gene families from *M. martensii*.

The figure was made by KEGG Mapper – Color Pathway tools (http://www.genome.jp/kegg). This figure shows the genes involved in the metabolism pathways of *M. martensii*. The grey lines represent the KO gene families with single gene; the green lines represent the KO gene families with two genes; the blue lines represent the KO gene families with 3-5 genes; the pink lines represent the KO gene families with 6-10 gene; and the red lines represent the KO gene families with over 10 genes.



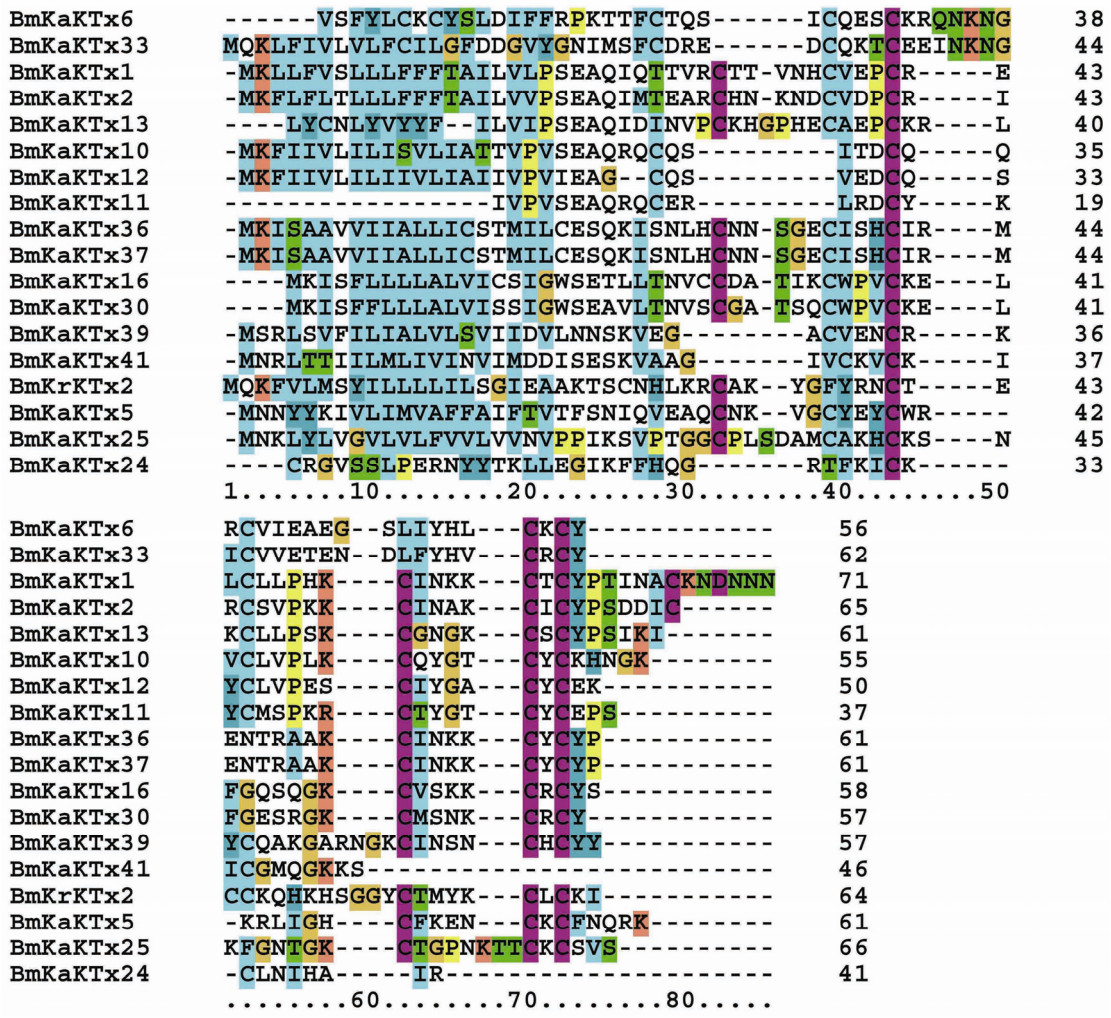
Supplementary Figure S10 Repertoire of venom neurotoxin genes annotated in the *M. martensii* genome.

One hundred and sixteen neurotoxin genes are classified into four types, including 61 NaTx genes, 46 KTx genes, 5 ClTx genes, and 4 CaTx genes.



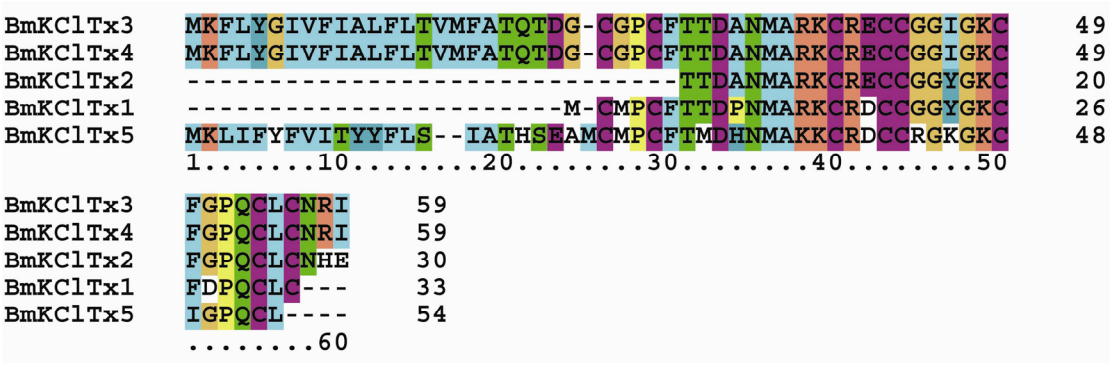
Supplementary Figure S11 Sequences of new NaTx neurotoxins discovered from the *M. martensii* genome.

Twenty-one NaTx neurotoxins (i.e., BmKNaTx9, 18-26, 37, 38, 40, 43, 45, and 47-52) were characterized, and an alignment was performed by ClustalX 1.83. Highly conserved Cys residues are indicated in purple. Color highlighting shows the other conserved amino acid residues. Gaps (-) were introduced to maximize similarities.



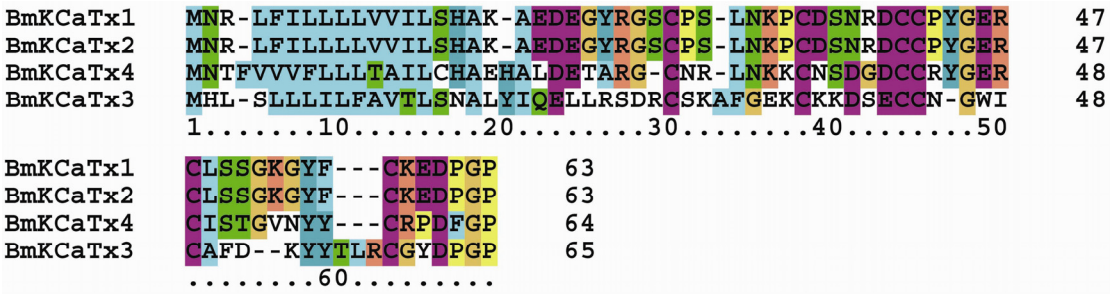
Supplementary Figure S12 Sequences of new KTx neurotoxins discovered from the *M. martensii* genome.

Eighteen KTx neurotoxins (i.e., BmKaKTx1, 2, 5, 6, 10-13, 16, 24, 25, 30, 33, 36, 37, 39, 41, and BmKγKTx2) were, and an alignment was created by ClustalX 1.83. Highly conserved Cys residues are indicated in purple. Color highlighting shows the other conserved amino acid residues. Gaps (-) were introduced to maximize similarities.



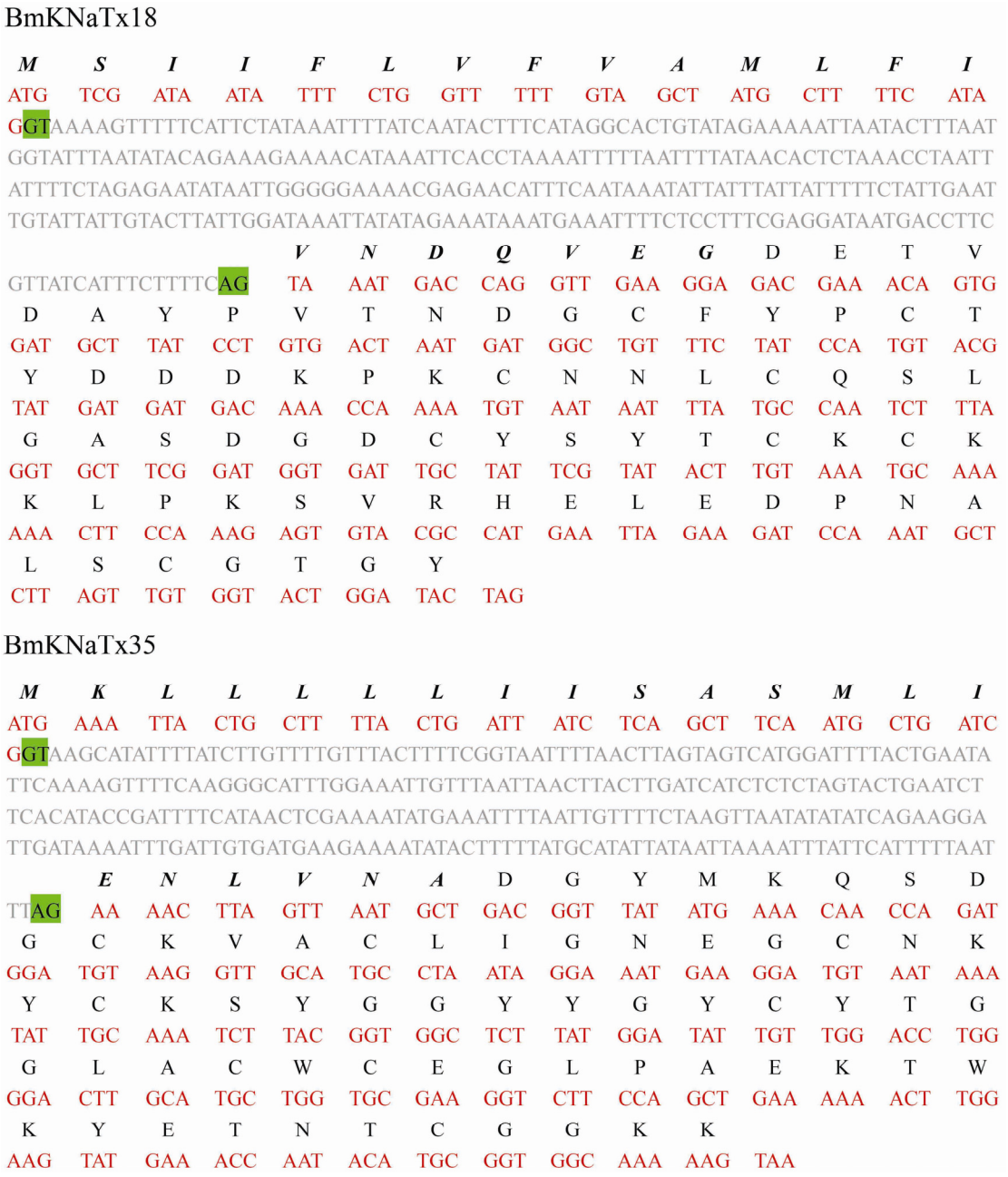
Supplementary Figure S13 Multiple sequence alignments of ClTx neurotoxins characterized from the *M. martensii* genome.

BmKClTx3 and BmKClTx4 had the same deduced amino acid sequences, corresponding to BmKCT (AF135821) and TXCL1 (AF159976), respectively, previously found in *M. martensii*. Three new ClTx neurotoxin genes (i.e., BmKClTx1, BmKClTx2, and BmKClTx5) were characterized. A multiple sequence alignment was created by ClustalX 1.83. Highly conserved Cys residues are indicated in purple. Color highlighting shows the other conserved amino acid residues. Gaps (-) were introduced to maximize similarities.



Supplementary Figure S14 Multiple sequence alignments of CaTx neurotoxins characterized from the *M. martensii* genome.

Three new CaTx neurotoxin genes (i.e., BmKCaTx1, BmKCaTx2, and BmKCaTx3) were characterized; the other CaTx neurotoxin gene, BmKCaTx4, was the same as BmCa1 (DQ206446) previously found in *M. martensii*. A multiple sequence alignment was created by ClustalX 1.83. Highly conserved Cys residues are indicated in purple. Color highlighting shows the other conserved amino acid residues. Gaps (-) were introduced to maximize similarities.

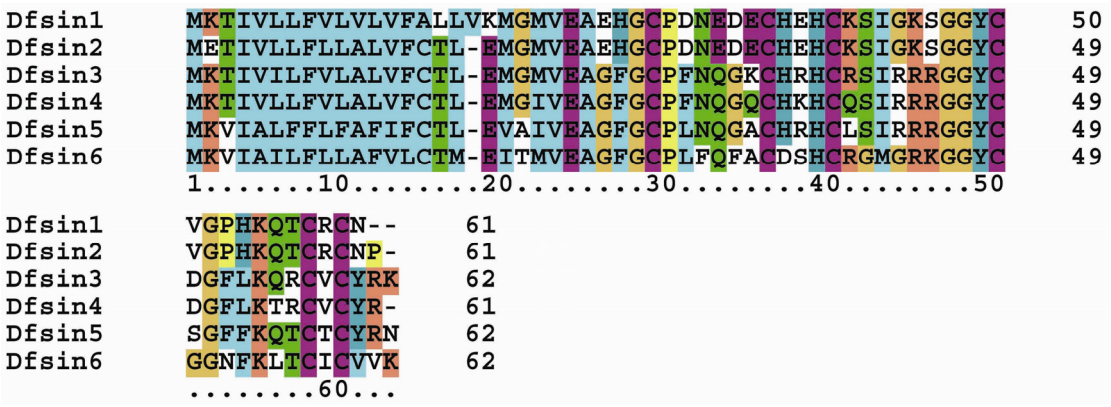


Supplementary Figure S15 Organization and structure of two representative NaTx neurotoxin genes from the scorpion *M. martensii*.

BmKNaTx18 and BmKNaTx35 were selected to show the organization and structure of the NaTx neurotoxin genes in the *M. martensii* genome. The predicted peptide sequences are shown above the nucleotide sequences. The amino acid residues of the signal peptide are in bold face and italics. Exon nucleotide sequences are indicated with red letters, and intron nucleotide sequences are indicated with black letters. Both 5' and 3' consensus splice sites ("GT" and "AG") are shown in green. BmKNaTx18 and BmKNaTx35 neurotoxin genes in *M. martensii* have one phase-I intron with a GT-AG splicing rule located at the end of the signal peptide coding region. Both BmKNaTx18 and BmKNaTx35 have long intron sequences (>400 bp).

Supplementary Figure S16 Organization and structure of two representative KTx neurotoxin genes from the scorpion *M. martensii*.

BmKaKTx3 and BmKaKTx33 were selected to show the organization and structure of the KTx neurotoxin genes in the *M. martensii* genome. The predicted peptide sequences are shown above the nucleotide sequences. The amino acid residues of the signal peptide are in bold face and italics. Exon nucleotide sequences are indicated with red letters, and intron nucleotide sequences are indicated with black letters. Both 5' and 3' consensus splice sites ("GT" and "AG") are shown in green. BmKaKTx3 and BmKaKTx33 neurotoxin genes in *M. martensii* have one phase-I intron with a GT-AG splicing rule located at the end of the signal peptide coding region. Both BmKaKTx3 and BmKaKTx33 have short intron sequences (<200 bp).

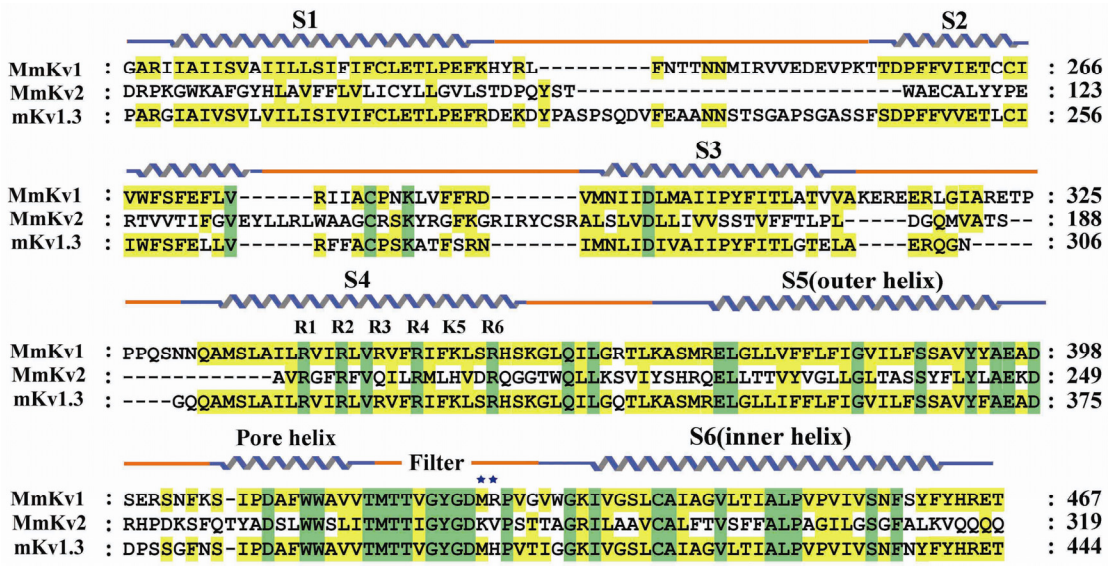


Supplementary Figure S17 Multiple sequence alignments of new defensins characterized from the *M. martensii* genome.

Six defensins (BmKDfsin1-6) were characterized. A multiple sequence alignment was created by ClustalX 1.83 software. Highly conserved Cys residues are indicated in purple. Color highlighting shows the other conserved amino acid residues. Gaps (-) were introduced to maximize similarities.

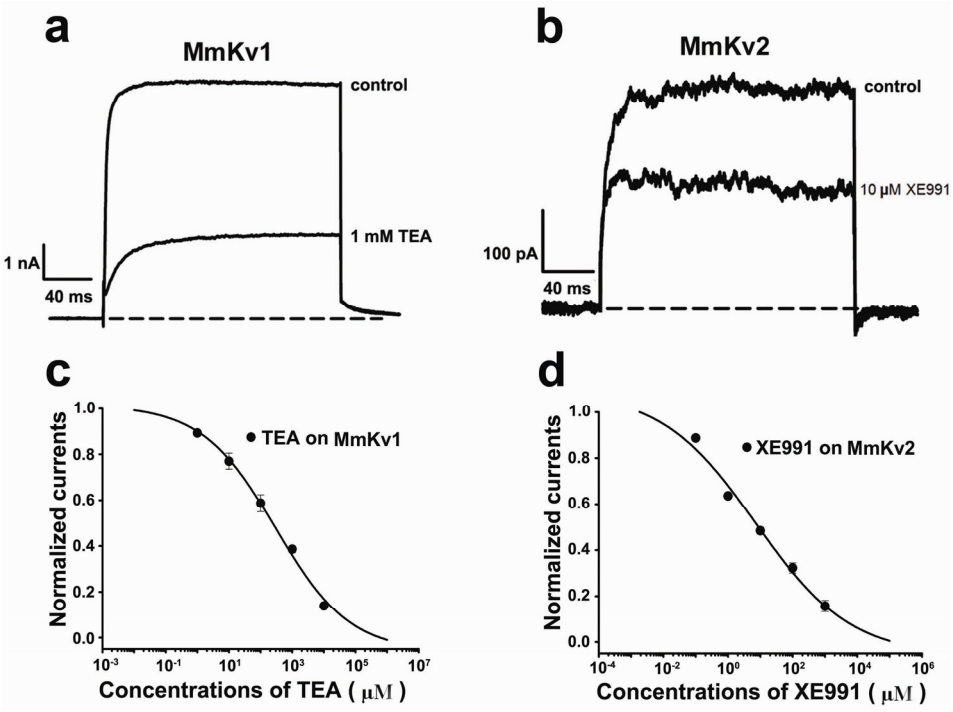
Supplementary Figure S18 Organization and structure of two representative defensin genes discovered in the scorpion *M. martensii*.

The predicted peptide sequences are shown above the nucleotide sequences. The amino acid residues of the signal peptide are in bold face and italics. Exon nucleotide sequences are indicated with red letters, and intron nucleotide sequences are indicated with black letters. Both 5' and 3' consensus splice sites (“GT” and “AG”) are shown in green. Two defensins in *M. martensii* had one phase-I intron with a GT-AG splicing rule located at the end of the signal peptide coding region. BmKDfsin1 has longer intron sequences (>400 bp) than BmKDfsin2 (<200 bp).



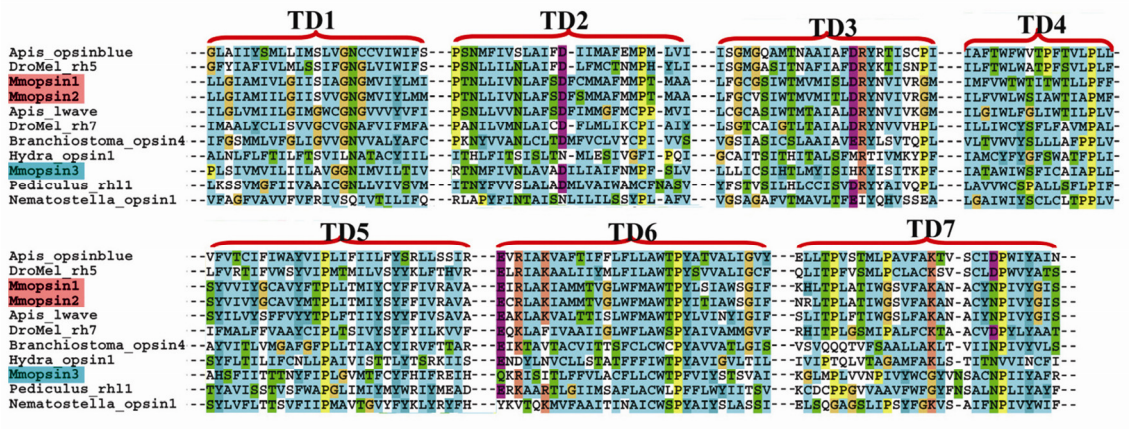
Supplementary Figure S19 Sequence alignments of scorpion potassium channel and mouse Kv1.3 (GI, 1345815) transmembrane regions.

Secondary structure elements are indicated above the sequences. The conserved positively charged residues in the voltage sensor of S4 segment are highlighted, and the characteristic residues of ion selectivity filter *"TVGYGD"* are also indicated.



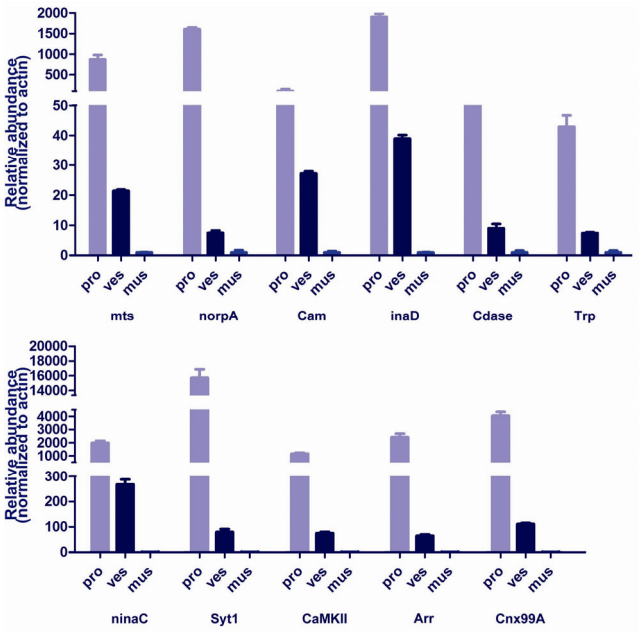
Supplementary Figure S20 Identification of functional scorpion K^+ channels.

a, Inhibitory activity of TEA to MmKv1 K^+ channel. 1 mM TEA (tetraethyammonium) blocked 59.4% of MmKv1 currents. b, Blocking effect of XE991 to MmKv2 K^+ channel. 10 μM XE991 (10,10-bis(4-pyridinylmethyl)-9(10H)-anthracenone) blocked 51.6% of MmKv2 currents. c, Dose dependence of TEA on the scorpion MmKv1 K^+ channel. The curve was fitted using the Hill equation, IC_{50} = 300.8 ± 91.3 μM. d, Dose dependence of XE991 on the scorpion MmKv2 K^+ channel. The curve was fitted using the Hill equation, IC_{50} = 7.3 ± 4.6 μM.



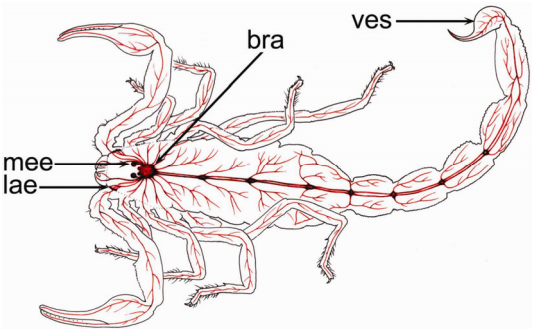
Supplementary Figure S21 Multiple sequence alignments of opsins' transmembrane regions.

Seven transmembrane regions (TD1-7) of opsin proteins are highly homologous. The identity and detailed information of each sequence for mutiple alignments are listed in Supplementary Tables S17 and S20.



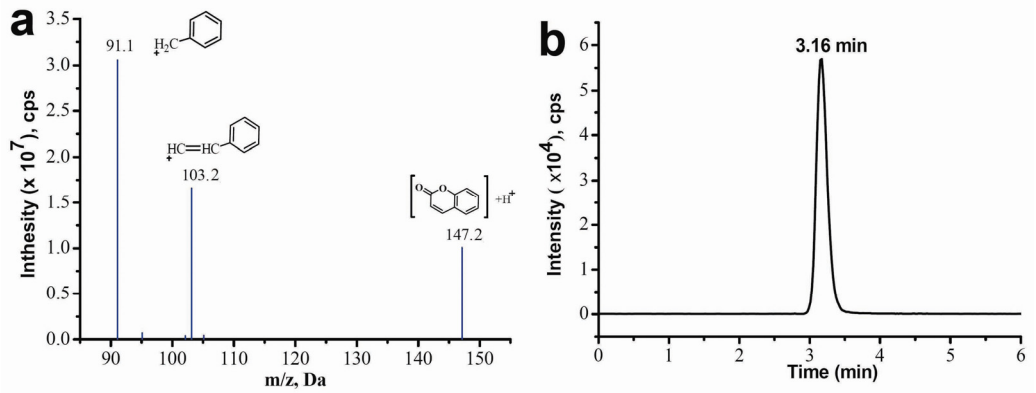
Supplementary Figure S22 Quantitative expression analysis of phototransduction pathway genes in the tail (vesicle) of the scorpion *M. martensii*.

pro, prosoma; ves, vesicle or venom gland; mus, muscle from chelas and metasomal segments I-V. These experimental data indicate that 11 phototransduction pathway genes (i.e., mts, norpA, Cam, inaD, Cdase, Trp, ninaC, Syt1, CaMKII, Arr, and Cnx99A) have the same expression patterns among the different tissues of *M. martensii*: prosoma > vesicle > muscle. Data are expressed as the mean ± standard deviation from three replicates.



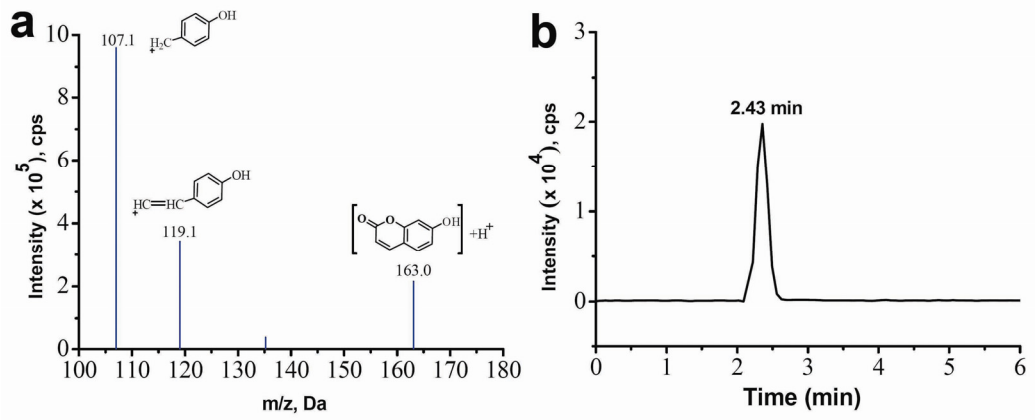
Supplementary Figure S23 Nervous system of scorpions.

Nerve fibers are indicated with red lines, while neural ganglion and organ (brain) are highlighted with black. lae, lateral eyes; mee, median eyes; bra, brain; ves, vesicle.



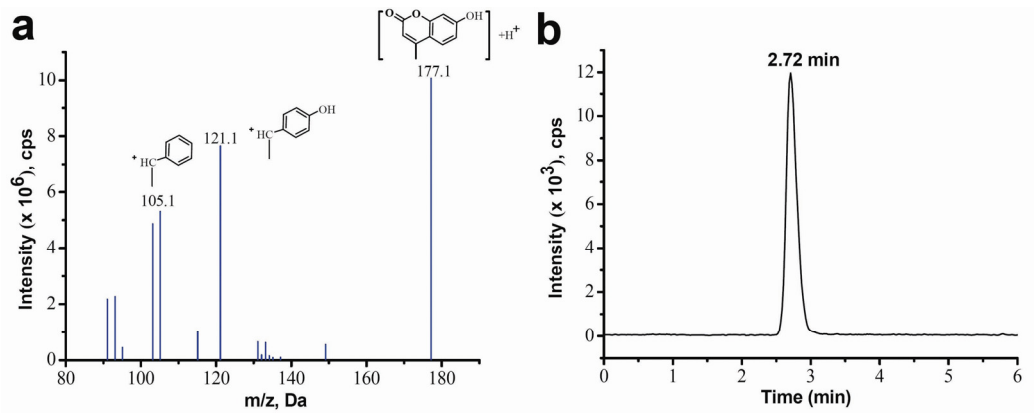
Supplementary Figure S24 Product ion spectrum and LC-ESI-MS/MS (MRM mode) of coumarin standard.

a, Product ion spectrum of coumarin standard. b, LC-ESI-MS/MS (MRM mode) chromatogram of coumarin standard. The LC-ESI-MS/MS conditions are described in the experimental section. Coumarin was detected by monitoring *m/z* 147.2/103.2.



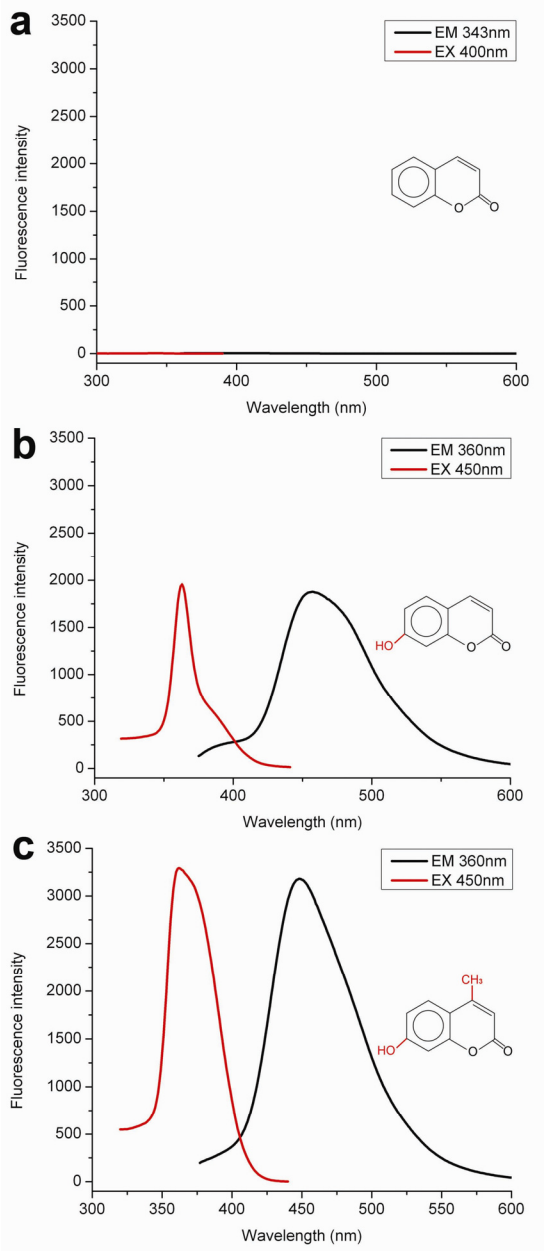
Supplementary Figure S25 Product ion spectrum and LC-ESI-MS/MS of 7-hydroxy-coumarin standard.

a, Product ion spectrum of 7-hydroxy-coumarin standard. b, LC-ESI-MS/MS of 7-hydroxy-coumarin standard. The LC-ESI-MS/MS conditions are described in the experimental section. 7-hydroxy-coumarin was detected by monitoring *m/z* 163.0/107.1.



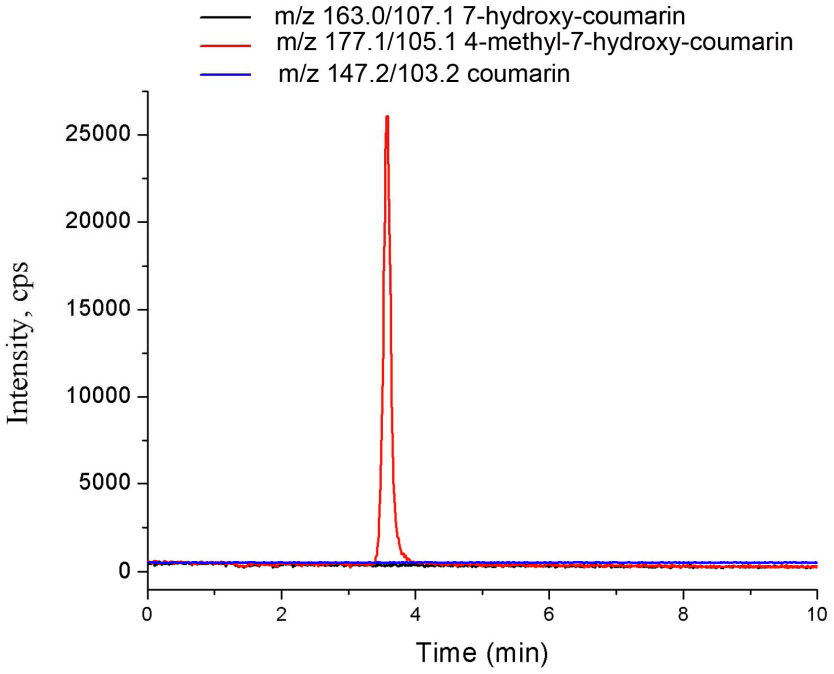
Supplementary Figure S26 Product ion spectrum and LC-ESI-MS/MS (MRM mode) chromatogram of 4-methyl-7-hydroxy-coumarin standard.

a, Product ion spectrum of 4-methyl-7-hydroxy-coumarin standard. b, LC-ESI-MS/MS chromatogram of 4-methyl-7-hydroxy-coumarin standard. The LC-ESI-MS/MS conditions are described in the experimental section. 4-methyl-7-hydroxy-coumarin was detected by monitoring *m/z* 177.1/105.1.



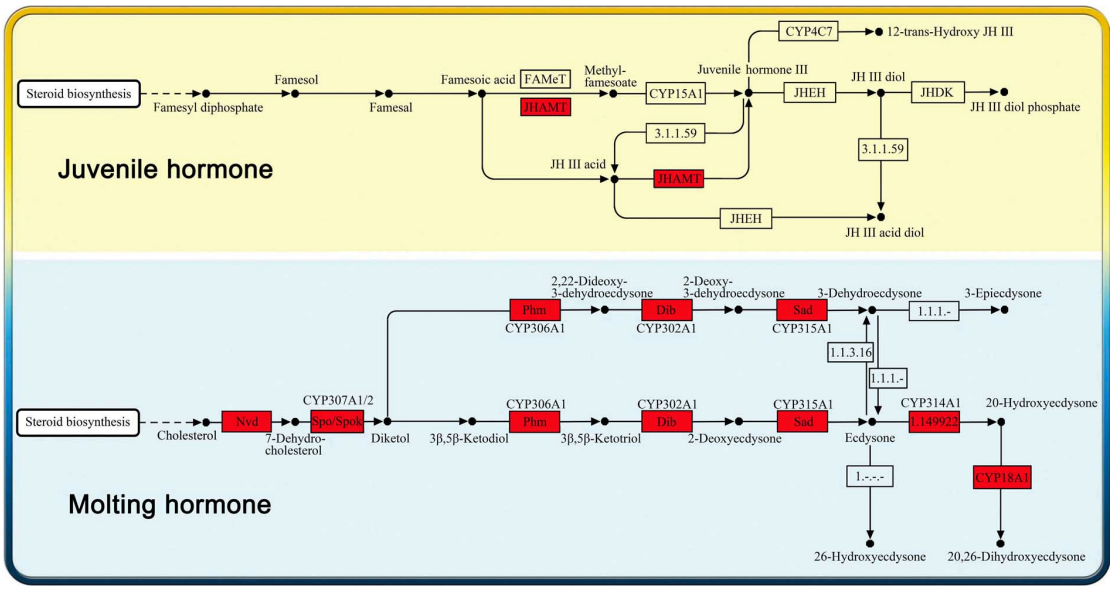
Supplementary Figure S27 Fluorescence spectra of coumarin and its derivative standards.

a, Fluorescence spectra of 1 mg/ml coumarin standard resolved in ethanol. b, Fluorescence spectra of 1 mg/ml 7-hydroxy-coumarin standard resolved in ethanol. c, Fluorescence spectra of 1 mg/ml 4-methyl-7-hydroxy-coumarin standard resolved in ethanol. Excitation spectra of coumarin, 7-hydroxy-coumarin, and 4-methyl-7-hydroxy-coumarin were obtained by monitoring the emission of light at 400, 450, and 450 nm, respectively. Emission spectra of coumarin, 7-hydroxy-coumarin, and 4-methyl-7-hydroxy-coumarin were obtained by monitoring emission flowing excitation at 343, 360, and 360 nm, respectively. All fluorescence spectra of coumarin and its derivatives were recorded under the same instrument parameters using the F-4600 FL Spectrophotometer (Hitachi, Japan). The excitation and emission slits were 5 nm and 5 nm, respectively. X-axis shows the light wavelength (nm), and Y-axis represents fluorescence intensity. X- and Y-axes have the same values for all three fluorescence spectra.



Supplementary Figure S28 Detection of coumarin, 7-hydroxy-coumarin, and 4-methyl-7-hydroxy-coumarin in the ethanol solution of 4-methyl-7-hydroxy-coumarin standard (stored for more than one year).

The analysis was performed using LC-ESI-MS/MS (MRM mode) by monitoring a transition pair of m/z 147.2/103.2 for coumarin, m/z 163.0/107.1 for 7-hydroxy-coumarin and m/z 177.1/105.1 for 4-methyl-7-hydroxy-coumarin. Only 4-methyl-7-hydroxy-coumarin is observed (red line), and neither 7-hydroxy-coumarin (black line) nor coumarin (blue line) are found in the stored solution of 4-methyl-7-hydroxy-coumarin standard.



Supplementary Figure S29 KEGG maps of the biosynthetic pathways for juvenile hormone and molting hormone in the scorpion *M. martensii*.

Annotations for the KEGG pathways were performed using KAAS from KEGG. The biosynthetic pathway of juvenile hormone involves Jhamt, but not CYP15A1. *M. martensii* has the complete biosynthetic and metabolic pathways for molting hormone: Nvd, CYP307A1/2, CYP306A1, CYP302A1, CYP315A1, CYP314A1, and CYP18A1.

Supplementary Tables

Supplementary Table S1 C-value and genome size of *M. martensii* determined by flow cytometry.

Samples	C-value (pg)	Average C-value (pg)	Average Genome Size (Mb)
1	1.4	1.35 ± 0.04	1323.73 ± 39.12
2	1.36		
3	1.33		
4	1.37		
5	1.37		
6	1.39		
7	1.35		
8	1.28		
9	1.37		
10	1.28		
11	1.37		

Supplementary Table S2 Statistics of raw sequence data from different Illumina libraries.

Lib Type	Library size	Read number	Read Length	Total Size
Illumina-PE	180 bp	540,074,026	100bp x2	108 Gb
	300 bp	857,374,353	100bp x2	171 Gb
	420 bp	114,738,294	100bp x2	23 Gb
Illumina-MP	5 Kb	90,917,717	75bp x2	14 Gb
	10 Kb	27,999,437	75bp x2	4 Gb
454-FLX	700 bp	3,691,676	～450 bp	1.3 Gb

Supplementary Table S3 Statistics of genome assembly v1.0 for *M. martensii*.

Scaffold	Scaffold Number	Contigs in Scaffolds	Total Size (bp)	Gaps Percentage (%)
>1000 Kb	27	1,745	33,171,018	3.92%
500Kb ~ 1000 Kb	253	8,519	167,014,543	3.62%
200Kb ~ 500 Kb	1,135	18,811	346,568,078	4.07%
100Kb ~ 200 Kb	1,410	12,257	202,225,120	4.64%
50Kb ~ 100 Kb	1,497	7,697	109,051,448	5.49%
20Kb ~ 50 Kb	1,819	5,775	60,029,970	8.04%
10Kb ~ 20 Kb	1,419	3,201	20,339,418	16.48%
5Kb ~ 10 Kb	1,938	3,491	13,938,965	31.24%
2Kb ~ 5 Kb	1,909	2,578	6,583,288	21.52%
1Kb ~ 2 Kb	3,391	3,669	4,560,614	0.66%

Supplementary Table S4 Sample of simple repeats in the *M. martensii* genome.

Contig	Motif Length	Motif	Repeats	Start	End
NODE_888	6	tttgtt	2	159	170
NODE_730	5	atctt	2	61	70
NODE_6843	6	tcaaaa	2	858	869
NODE_64	5	caaaa	2	77	86
NODE_521	3	aat	4	72	83
NODE_467	5	tagtt	2	88	97
NODE_3834	5	taaat	2	75	84
NODE_3716	5	ctaat	2	26	35
NODE_3650	5	ttcaa	2	53	62
NODE_3514	2	ac	5	83	92
NODE_3370	5	ccatg	2	5	14
NODE_3245	6	ttattt	2	25	36
NODE_3152	5	gcctc	2	1	10
NODE_299	5	taact	2	55	64
NODE_2970	6	acaaac	2	84	95
NODE_2879	5	ttctg	2	7	16
NODE_270	5	gaaga	2	66	75
NODE_2397	6	taaaaa	2	79	90
NODE_2156227	6	gtttta	2	50	61
NODE_2154331	5	aatta	2	2,679	2,688
NODE_2152134	6	ccgttt	2	1,572	1,583
NODE_2142885	5	taaat	2	2,253	2,262
NODE_2083608	3	att	5	22,676	22,690
NODE_2083608	5	taagg	2	576	585
NODE_1984	5	ttttg	2	48	57
NODE_1634	5	caaaa	2	51	60
NODE_154	3	gag	5	88	102
NODE_134	5	cagtt	2	428	437
NODE_1244	2	ta	5	34	43
NODE_1189	6	gggtct	2	55	66

The full table of simple repeat sequences in *M. martensii* genome could be found in our website: http://www.genoportal.org/mm/simple-repeat-sequence-v1.0.txt.zip

Supplementary Table S5 Samples of transposable elements in the *M. martensii* genome.

TE	Gene Number	Size (bp)	Proportion
Gypsy	1,908	2,906,178	9.95%
hAT	1,637	2,606,179	8.92%
Mariner	1,408	2,553,107	8.74%
CR1	1,313	2,008,692	6.88%
Copia	854	1,319,627	4.52%
L1	708	723,407	2.48%
RTE	576	1,039,713	3.56%
SMAR	560	1,023,636	3.50%
ERE2_EH_1p	473	869,314	2.98%
P-x	457	564,393	1.93%
BEL-x	406	410,036	1.40%
Poseidon	382	737,556	2.52%
Penelope	360	663,764	2.27%
Helitron	321	351,348	1.20%
LOA	243	571,859	1.96%
Polinton	243	289,705	0.99%
Harbinger	236	302,213	1.03%
Dong	218	470,859	1.61%
LINE	205	321,823	1.10%
Sola	200	262,000	0.90%
MuDr	195	249,466	0.85%
BEL	186	273,634	0.94%
EnSpm	178	242,024	0.83%
TRAS	153	367,732	1.26%
L1HS	150	20,562	0.07%
POLB	149	168,731	0.58%
Piggy	147	150,022	0.51%
Ginger	145	241,314	0.83%
L1PREC2	143	15,240	0.05%
LIN	141	229,900	0.79%
ERV	130	200,089	0.68%
R1	130	239,830	0.82%
Crack	125	175,560	0.60%
Transib	119	129,447	0.44%
Tc1	110	169,040	0.58%
CRMAR	99	200,662	0.69%
BLACKJACKp	95	157,559	0.54%
Tx	93	131,967	0.45%
I-x	90	144,127	0.49%
L1PB2c	88	10,856	0.04%

L2	88	131,831	0.45%
TX1	82	110,164	0.38%
I-2	75	134,677	0.46%
Chapaev-x	74	75,288	0.26%
FAMAR1	69	134,999	0.46%
Mari2	65	117,162	0.40%
I-3	63	94,099	0.32%
Hoana	62	87,403	0.30%
PENEL1_Nvi	62	120,046	0.41%
Proto	61	89,898	0.31%
DIRS	59	83,607	0.29%
DESMAR1	57	87,179	0.30%
INT	55	66,781	0.23%
Kolobok	55	69,712	0.24%
BovB	53	103,362	0.35%
RT	53	104,592	0.36%
Chapaev	52	64,513	0.22%
Crack-x	51	88,073	0.30%
IS4EU	51	65,396	0.22%
Rehavkus	50	66,299	0.23%
VENSMAR1	47	95,342	0.33%
NAVIMAR	46	89,592	0.31%
PY	46	44,345	0.15%
ATCOPIA	45	60,117	0.21%
L1MB	45	5,922	0.02%
Galileo	44	61,083	0.21%
SHALINE	44	68,379	0.23%
ISL2EU	43	47,540	0.16%
Tc3	42	81,262	0.28%
transposase	41	77,697	0.27%
MARIAM	36	62,608	0.21%
PERERE	36	65,751	0.23%
SHACOP	36	51,941	0.18%
Ag	35	65,331	0.22%
ATP-x_p	34	32,261	0.11%
numbus	34	37,552	0.13%
Galluhop	33	77,921	0.27%
R2	33	49,811	0.17%
Neptune	32	48,370	0.17%
Nimb	32	60,170	0.21%
Howilli	31	47,568	0.16%
PX	31	32,073	0.11%
Ingi	30	57,313	0.20%

AgaPx	29	27,719	0.09%
Kiri	29	40,957	0.14%
QUETZAL	29	42,933	0.15%
Baggins-x	28	56,893	0.19%
L2B	28	29,007	0.10%
R4	28	62,850	0.22%
I-1	27	57,495	0.20%
Ogre	27	33,599	0.12%
COP	26	39,462	0.14%
AeHerves	25	28,127	0.10%
KBOC_DB_1p	25	31,784	0.11%
HARB-1_AP_1p	23	25,094	0.09%
Loner	22	32,322	0.11%
Zator	22	27,692	0.09%
Jockey	21	23,947	0.08%
PRO-1	21	15,085	0.05%
Academ	19	22,188	0.08%
MacERV	19	16,324	0.06%
Ambal	18	24,104	0.08%
ATP-x	18	24,765	0.08%
I-4	18	30,188	0.10%
Perere_Smed	18	34,827	0.12%
Vingi	18	37,248	0.13%
CAULIV	17	27,877	0.10%
L1MC	17	13,236	0.05%
Proto2-1	17	11,478	0.04%
ATHAT	16	20,848	0.07%
CYPSM	16	26,733	0.09%
HEL	16	23,000	0.08%
AeBuster	15	25,744	0.09%
HARB	15	16,857	0.06%
Lian-Aa1	15	27,620	0.09%
PNL	15	20,258	0.07%
REX	15	20,484	0.07%
TRE	15	26,190	0.09%
AMARI	14	25,935	0.09%
Chimpo-I	14	8,301	0.03%
GYPS	14	20,789	0.07%
HELIBAT1	14	8,289	0.03%
ITmD37D	14	15,566	0.05%
RETRO	14	22,077	0.08%
AVMAR	13	34,091	0.12%
Kiritsubo	13	19,896	0.07%

Randl	13	23,299	0.08%
ATRPA	12	14,911	0.05%
I_Ele	12	16,270	0.06%
L1B	12	4,663	0.02%
HAL1-1_Cja_1p	11	2,639	0.01%
HELMET	11	12,837	0.04%
Kibi	11	15,748	0.05%
Myotis_hAT1	11	16,659	0.06%
R5	11	12,938	0.04%
Waldo	11	23,739	0.08%

Supplementary Table S6 Single nucleotide polymorphism in the diploid

Dataset	**Data size(fq)**	**Covered genome loci**	**Heterozygous loci (80%)**	**Heterozygous percentage (80%)**
Genome Seq L1	33G	913,483,460	3,153,071	0.35%
Genome Seq L2	19G	675,503,758	2,092,611	0.31%
Genome Seq L7	21G	908,031,513	2,843,783	0.31%

Supplementary Table S7 Statistics of gene models of *M. martensii* minimum protein coding genes.

Gene Models	**Num of Exons Per Model**	**Average Exon Length Per Model**	**Total Exon Length Per Model**
32,016	3.9	198.12 bp	676.6 bp

Supplementary Table S8 Summary of potential alternatively spliced genes in *M. martensii*.

Alternatively Spliced Transcripts	**Gene Number**
2	3,737
3	1,536
4	508
5	204
>5	154
Total	6,139

Supplementary Table S9 Species used for comparative genomics analysis in the present study.

Tax Name	Resource	Citation or GenBank Number
Acyrthosiphon pisum	NCBI	BUILD.1.1
Aedes aegypti	Vectorbase	AaegL1.3, April 2012
Anopheles gambiae	NCBI	BUILD.2.1
Apis mellifera	NCBI	BUILD.4.1
Bombyx mori	Silkworm Genome Database	SilkDB V2.0
Caenorhabditis elegans	NCBI	BUILD.8.1
Camponotus floridanus	Camponotus Floridanus Genome Project	Cflo_3.3
Ciona intestinalis	NCBI	update 16 September 2008
Culex pipiens	BROAD INSTITUTE	Jun 21, 2012
Daphnia pulex	wFleaBase	Daphnia pulex Genes 2010
Drosophila melanogaster	FlyBase	r5.38
Homo sapiens	NCBI	BUILD.37.2
Nasonia vitripennis	NCBI	NCBI BUILD.1.1
Pediculus humanus	Vectorbase	PhumU1.2, July 2009
Tetranychus urticae	Bogas	Nov 15 2012
Tribolium castaneum	NCBI	NCBI BUILD.1.1

Supplementary Table S10 The number of expanded gene family and large gene family.

Species	**Expand Family***	**Large Family#**	**All Gene Families**
Aedes aegypti	388	5	6,728
Drosophila melanogaster	999	3	6,972
Tribolium castaneum	37	0	5,660
Camponotus floridanus	80	19	6,381
Acyrthosiphon pisum	371	10	6,309
Daphnia pulex	269	78	7,831
Mesobuthus martensii	496	69	5,947
Tetranychus urticae	172	25	5,382
Caenorhabditis elegans	210	56	6,282
Ciona intestinalis	126	7	6,126
Homo sapiens	1421	31	8,225

Expand Family*: the families with z-score>2

Large Family*: the families with >20 numbers

(More details in Supplementary Note 3)

Supplementary Table S11 Ten annotated expanded gene family in the *M. martensii* genome.

ID	Gene number	Z score	Function	Three Representative Genes
F8	157	2.58	Down syndrome cell adhesion molecule	MMa20200,MMa46644,MMa30953
F40	89	2.72	Choline dehydrogenase	MMa51732,MMa10041,MMa33113
F29	87	2.45	ATP-binding cassette	MMa40312,MMa15313,MMa39839
F24	71	2.25	Myosin	MMa55824,MMa39812,MMa00137
F46	65	2.30	Dynein heavy chain	MMa26304,MMa27825,MMa43716
F35	59	2.26	MFS transporter, OCT family	MMa45571,MMa52554,MMa52089
F53	59	2.65	glutamate receptor	MMa38974,MMa22238,MMa51194
F44	56	2.57	low density lipoprotein-related protein	MMa51846,MMa42212,MMa41894
F76	34	2.64	26S proteasome regulatory subunit	MMa27265,MMa17809,MMa23037
F150	28	2.49	aldehyde dehydrogenase	MMa30843,MMa12877,MMa14909

Supplementary Table S12 Neurotoxin genes characterized from the *M. martensii* genome.

Name	*M. martensii* genome ID	Homologous gene (GenBank ID)	E-values
BmKNaTx1	MMa13618	O61705.1	4.00E-37
BmKNaTx2	MMa13616	O61705.1	3.00E-36
BmKNaTx3	MMa13619	P45698.1	4.00E-37
BmKNaTx4	MMa17865	AAM49598.1	2.00E-33
BmKNaTx5	MMa17864	AAM49598.1	1.00E-35
BmKNaTx6	MMa17854	AAM49598.1	1.00E-36
BmKNaTx7	MMa17853	Q9UAC8.1	9.00E-31
BmKNaTx8	MMa17863	Q9UAC9.1	3.00E-30
BmKNaTx9	MMa21232	ABJ09780.1	1.00E-38
BmKNaTx10	MMa26033	ADT64283.1	2.00E-20
BmKNaTx11	MMa29117	Q9BKJ1.1	2.00E-35
BmKNaTx12	MMa29118	Q9BKJ1.1	1.00E-09
BmKNaTx13	MMa32665	P01497.2	8.00E-17
BmKNaTx14	MMa32664	P01497.2	8.00E-17
BmKNaTx15	MMa32661	O77091.1	8.00E-39
BmKNaTx16	MMa32660	O77091.1	8.00E-39
BmKNaTx17	MMa32663	O61668.1	2.00E-36
BmKNaTx18	MMa35303	P0CI57.1	3.00E-05
BmKNaTx19	MMa35304	C9X4J9.1	3.00E-06
BmKNaTx20	MMa35396	ADK12684.1	2.00E-25
BmKNaTx21	MMa35393	ADK12684.1	2.00E-25
BmKNaTx22	MMa37864	P68725.1	2.00E-05
BmKNaTx23	MMa43956	ABR21048.1	9.00E-38
BmKNaTx24	MMa43955	ABR21048.1	9.00E-38
BmKNaTx25*	MMa44056	P0CI56.1	3.00E-07
BmKNaTx26	MMa44055	P0CI56.1	8.00E-06
BmKNaTx27	MMa47002	P0C5I5.1	2.00E-06
BmKNaTx28*	MMa46992	P68726.1	2.00E-12
BmKNaTx29	MMa53032	P24336.1	5.00E-16
BmKNaTx30	MMa53030	Q86M31.1	5.00E-21
BmKNaTx31	MMa55372	AAF31295.1	5.00E-37
BmKNaTx32	MMa00206	AAM49598.1	2.00E-18
BmKNaTx33	MMa00705	ADT82854.1	3.00E-20
BmKNaTx34	MMa01249	AAK53809.1	1.00E-37
BmKNaTx35	MMa04432	Q26292.1	7.00E-29
BmKNaTx36	MMa04555	O77091.1	5.00E-36

BmKNaTx37	MMa10609	P0CI55.1	1.00E-21
BmKNaTx38*	MMa35093	1PE4	9.00E-13
BmKNaTx39	MMa12629	Q9GQW3.1	4.00E-38
BmKNaTx40	MMa13212	ABR21048.1	3.00E-41
BmKNaTx41	MMa19599	O61668.1	1.00E-18
BmKNaTx42	MMa20191	P59853.1	2.00E-20
BmKNaTx43	MMa37863	P0C292.1	5.00E-06
BmKNaTx44	MMa23370	AAF29465.1	5.00E-36
BmKNaTx45*	MMa24896	ADY17426.1	6.00E-08
BmKNaTx46	MMa34630	AAK53809.1	2.00E-37
BmKNaTx47*	MMa31503	Q4LCT0.1	5.00E-06
BmKNaTx48	MMa41588	ADY17426.1	7.00E-08
BmKNaTx49	MMa37070	Q4LCT0.1	5.00E-10
BmKNaTx50	MMa45084	P0CI56.1	3.00E-06
BmKNaTx51*	MMa44674	Q4LCT0.1	1.00E-05
BmKNaTx52	MMa47685	Q1I165.1	7.00E-05
BmKNaTx53	MMa51877	AAK94769.1	2.00E-20
BmKNaTx54	MMa55680	P45697.2	1.00E-25
BmKNaTx55	MMa00001	P59853.1	1.00E-12
BmKNaTx56	MMa02192	Q9BKJ1.1	8.00E-23
BmKNaTx57	MMa02242	P59853.1	3.00E-14
BmKNaTx58	MMa02691	AAF29465.1	5.00E-36
BmKNaTx59	MMa05947	AAP33620.1	2.00E-14
BmKNaTx60	MMa37863	Q9Y1U3.2	1.00E-10
BmKNaTx61	MMa14634	Q9BKJ1.1	5.00E-11
BmKaKTx1*	MMa16285	ACJ23153.1	3.00E-10
BmKaKTx2	MMa16284	ACJ23153.1	2.00E-07
BmKaKTx3	MMa21259	3E8Y	2.00E-09
BmKaKTx4	MMa21265	Q9TVX3.1	2.00E-23
BmKaKTx5	MMa21260	Q9TVX3.1	2.00E-05
BmKaKTx6	MMa21379	Q8MUB1.1	8.00E-15
BmKaKTx7	MMa21378	Q8MUB1.1	4.00E-12
BmKaKTx8	MMa26036	Q5F1N4.1	7.00E-30
BmKaKTx9	MMa35053	AAV59462.1	2.00E-14
BmKaKTx10	MMa35044	AAV59462.1	2.00E-08
BmKaKTx11	MMa35054	AAV59462.1	7.00E-08
BmKaKTx12	MMa35043	AAV59462.1	4.00E-05
BmKaKTx13	MMa35048	ACJ23153.1	3.00E-16
BmKaKTx14	MMa43624	Q9NII6.1	1.00E-21
BmKaKTx15	MMa43622	Q9NII5.1	8.00E-22
BmKaKTx16	MMa43630	Q9NII5.1	3.00E-13

BmKaKTx17	MMa43621	Q8MQL0.1	2.00E-16
BmKaKTx18	MMa43631	Q9NII7.1	3.00E-22
BmKaKTx19	MMa01676	Q9U8D1.1	2.00E-21
BmKaKTx20	MMa57009	Q8MUB1.1	4.00E-13
BmKaKTx21	MMa00460	Q9BKB4.1	5.00E-15
BmKaKTx22	MMa04118	Q9NBG9.1	9.00E-18
BmKaKTx23	MMa05343	A7KJJ7.1	1.00E-15
BmKaKTx24	MMa06076	AAC72985.1	6.00E-05
BmKaKTx25	MMa31503	ABW90713.1	8.00E-13
BmKaKTx26	MMa09960	Q86BX0.1	7.00E-16
BmKaKTx27	MMa16352	B5KF99.1	2.00E-18
BmKaKTx28	MMa23444	Q9BKB4.1	1.00E-18
BmKaKTx29	MMa25600	AAP33621.1	6.00E-21
BmKaKTx30	MMa26924	Q9NII5.1	3.00E-15
BmKaKTx31	MMa34377	AAK61820.1	2.00E-18
BmKaKTx32	MMa34333	AAM91031.1	4.00E-17
BmKaKTx33	MMa34788	Q8MUB1.1	4.00E-09
BmKaKTx34	MMa52078	Q8I0L5.1	5.00E-16
BmKaKTx35	MMa54042	AAM91031.1	1.00E-12
BmKaKTx36	MMa12594	ACJ23147.1	3.00E-12
BmKaKTx37	MMa20479	ACJ23147.1	3.00E-12
BmKaKTx38	MMa29598	|AAC72985.1	6.00E-05
BmKaKTx39	MMa36174	A7KJJ7.1	1.00E-07
BmKaKTx40	MMa41975	AAC72985.1	2.00E-05
BmKaKTx41	MMa42180	Q95P89.1	4.00E-08
BmKbKTx1	MMa43637	Q9N661.1	1.00E-34
BmKbKTx2	MMa35531	Q9NJC6.1	1.00E-23
BmKrKTx1	MMa43623	P59938.1	1.00E-21
BmKrKTx2	MMa54155	Q86QT3.1	7.00E-06
BmKKTx1	MMa56512	ABN04116.1	3.00E-45
BmKClTx1	MMa37070	P15222.2	2.00E-10
BmKClTx2	MMa44674	Q9BJW4.1	7.00E-09
BmKClTx3	MMa00982	Q9UAD0.1	1.00E-09
BmKClTx4	MMa06682	Q9UAD0.1	6.00E-09
BmKClTx5	MMa08179	ABR21063.1	5.00E-11
BmKCaTx1	MMa15573	B8XH22.1	9.00E-11
BmKCaTx2	MMa48745	B8XH22.1	1.00E-11
BmKCaTx3	MMa56389	B8XH22.1	3.00E-05
BmKCaTx4	MMa18822	Q8I6X9.2	5.00E-15

*These toxin genes were not detected to be expressed in the venom gland using RNA-Seq data from this study.

Supplementary Table S13 Neurotoxin and defensin gene clusters summarized from the *M. martensii* genome.

Serial Number	Scaffold ID	Toxin Gene
1	NODE_5541725	BmKNaTx1
		BmKNaTx2
		BmKNaTx3
2	NODE_5965609	BmKaKTx1
		BmKaKTx2
3	NODE_6184654	BmKNaTx4
		BmKNaTx5
		BmKNaTx6
		BmKNaTx7
		BmKNaTx8
4	NODE_6517216	BmKNaTx9
		BmKaKTx3
		BmKaKTx4
		BmKaKTx5
5	NODE_6524412	BmKaKTx6
		BmKaKTx7
6	NODE_6814096	BmKNaTx10
		BmKaKTx8
7	NODE_6975215	BmKNaTx11
		BmKNaTx12
8	NODE_7134751	BmKNaTx13
		BmKNaTx14
		BmKNaTx15
		BmKNaTx16
		BmKNaTx17
9	NODE_7241642	BmKaKTx9
		BmKaKTx10
		BmKaKTx11
		BmKaKTx12
		BmKaKTx13
10	NODE_7255846	BmKNaTx18
		BmKNaTx19
11	NODE_7263488	BmKNaTx20
		BmKNaTx21
12	NODE_7365295	BmKNaTx43
		BmKNaTx22
13	NODE_7554576	BmKaKTx14
		BmKaKTx15
		BmKaKTx16

		BmKaKTx17
		BmKaKTx18
		BmKbKTx1
		BmKrKTx1
14	NODE_7565769	BmKNaTx23
		BmKNaTx24
15	NODE_7568079	BmKNaTx25
		BmKNaTx26
16	NODE_7649271	BmKNaTx27
		BmKNaTx28
17	NODE_7818526	BmKNaTx29
		BmKNaTx30
18	NODE_7421102	BmKDfsin3
		BmKDfsin4
19	NODE_4521545	BmKDfsin1
		BmKDfsin2

Supplementary Table S14 Six new defensin genes discovered from the *M. martensii* genome.

Name	Gene ID	Homologous gene (GenBank No.)	E-values
BmKDfsin1	MMa09283	P56686.1	6.00E-05
BmKDfsin2	MMa09284	P56686.1	6.00E-05
BmKDfsin3	MMa39356	P41965.1	9.00E-11
BmKDfsin4	MMa39352	P41965.1	9.00E-10
BmKDfsin5	MMa36626	P41965.1	3.00E-14
BmKDfsin6	MMa52629	P41965.1	2.00E-05

Supplementary Table S15 Death of the cockroach *B. dubia* injected with fresh venom from *M. martensii*.

Venom Dose (μl/individual)	Animal Number	Survival Number	Death Number	Death Rate
1/2	10	0	10	100%
1/4	10	2	8	80%
1/8	10	4	6	60%
1/16	10	5	5	50%
1/32	10	8	2	20%
1/64	10	9	1	10%
1/128	10	10	0	0
0	10	10	0	0

Supplementary Table S16 Death of the scorpion *M. martensii* injected with its own fresh venom.

Venom Dose (μl/individual)	Animal Number	Survival Number	Death Number	Death Rate
10	10	10	0	0
5	10	10	0	0
0	10	10	0	0

Supplementary Table S17 Genes involved in the phototransduction pathway in the *M. martensii* genome.

Gene name (Drosophila)	*M. martensii* genome ID	E-values	Scores
Arr	MMa03492	3.00E-81	298
Cac	MMa39395	4.00E-105	380
	MMa18173	7.00E-57	219
	MMa23612	9.00E-44	176
	MMa42091	2.00E-39	161
	MMa06161	7.00E-39	160
	MMa03418	1.00E-30	132
	MMa17617	4.00E-29	127
CaMKII	MMa41427	8.00E-32	134
	MMa41654	6.00E-27	118
	MMa45215	1.00E-31	134
	MMa42996	5.00E-25	112
	MMa01898	3.00E-22	102
	MMa18645	4.00E-18	89
	MMa18646	4.00E-12	69.3
	MMa21302	9.00E-30	127
	MMa22962	3.00E-28	122
Cam(CALM)	MMa28976	1.00E-36	148
	MMa28325	2.00E-33	137
	MMa31790	4.00E-53	202
	MMa31244	3.00E-13	70.5
	MMa33967	7.00E-31	129
	MMa36464	7.00E-51	195
	MMa36827	1.00E-81	297
	MMa41518	4.00E-23	103
	MMa44822	4.00E-29	123
	MMa48681	7.00E-36	145
	MMa50924	4.00E-23	103
	MMa49931	7.00E-13	69.3
	MMa52175	4.00E-54	206
	MMa52883	1.00E-27	118
	MMa03550	2.00E-34	140
	MMa08456	7.00E-61	228
	MMa09585	1.00E-50	194
	MMa09473	4.00E-49	189
	MMa13555	4.00E-34	139

	MMa15813	4.00E-34	139
	MMa23541	3.00E-29	124
	MMa23239	4.00E-63	236
Cnx99A	MMa12354	7.00E-72	268
	MMa11165	5.00E-53	205
	MMa47989	1.00E-52	204
CDase	MMa34926	2.00E-175	612
DopR	MMa39583	4.00E-102	368
	MMa11918	5.00E-92	334
	MMa34740	2.00E-61	233
	MMa08722	1.00E-51	200
	MMa08030	4.00E-51	198
	MMa30406	7.00E-51	197
	MMa25619	1.00E-49	194
	MMa00895	2.00E-47	186
	MMa22488	2.00E-46	183
	MMa11827	3.00E-46	182
	MMa41467	9.00E-46	181
	MMa34743	4.00E-43	172
	MMa17020	6.00E-41	165
	MMa18652	2.00E-40	163
	MMa31446	7.00E-39	158
Galpha49B	MMa19058	1.00E-86	316
	MMa54844	1.00E-82	302
InaC	MMa55320	2.00E-153	539
	MMa31916	8.00E-96	347
InaE	MMa16629	8.00E-56	215
Mts	MMa06470	3.00E-122	434
	MMa20540	1.00E-80	296
	MMa13913	9.00E-71	263
	MMa13914	5.00E-67	250
	MMa16565	2.00E-66	249
	MMa51972	2.00E-63	239
	MMa25229	2.00E-53	206
ninaC	MMa16182	2.00E-103	374
norpA	MMa54976	0	679
	MMa20228	3.00E-106	382
Rh*	MMa00264	3.00E-109	391
	MMa40170	2.00E-105	378
	MMa17758	3.00E-23	105
Shab	MMa48920	2.00E-87	320

	MMa05372	3.00E-86	316
	MMa32844	7.00E-70	262
	MMa03794	8.00E-70	262
	MMa14698	4.00E-67	253
	MMa14694	2.00E-66	251
	MMa14695	5.00E-66	249
	MMa14697	9.00E-64	241
	MMa10705	2.00E-63	241
	MMa34046	3.00E-62	237
	MMa21085	3.00E-61	233
Syt1	MMa29228	2.00E-137	485
Trp	MMa29466	0	788
	MMa45115	7.00E-123	438
	MMa56220	7.00E-91	332
	MMa35182	7.00E-80	295
	MMa18482	4.00E-78	290
	MMa08818	3.00E-53	207
E	MMa54427	3.00E-80	296

*The Rhs annotated from the *M. martensii* genome were Mmopsin1, Mmopsin2 and Mmopsin3, corresponding to the gene ID numbers of MMa00264, MMa40170and MMa17758, respectively.

Supplementary Table S18 Phototransduction pathway related-genes expressed in the tail of *M. martensii*.

Gene name (Drosophila)	*M. martensii* tail transcriptome ID	E-values	Scores
Arr1	comp133_c0_seq1	2.00E-98	355
Arr2	comp32991_c0_seq1	2.00E-108	389
Cac	comp9808_c0_seq2	0	866
	comp9808_c0_seq1	0	866
	comp5255_c0_seq3	0	721
	comp5255_c0_seq7	0	721
	comp5255_c0_seq4	0	721
	comp5255_c0_seq8	0	720
	comp5255_c0_seq5	1.00E-138	491
	comp5255_c0_seq1	1.00E-138	491
	comp5255_c0_seq6	3.00E-138	490
	comp5255_c0_seq2	3.00E-138	490
CaMKII	comp1793_c0_seq1	0	725
	comp1793_c0_seq2	0	718
	comp516_c1_seq9	2.00E-178	622
	comp516_c1_seq7	2.00E-178	622
	comp516_c1_seq1	2.00E-178	622
	comp516_c1_seq15	4.00E-83	305
	comp516_c1_seq10	4.00E-83	305
	comp516_c1_seq3	4.00E-83	305
	comp516_c1_seq17	6.00E-82	301
	comp516_c1_seq12	6.00E-82	301
	comp516_c1_seq5	6.00E-82	301
	comp3230_c0_seq1	3.00E-64	242
	comp3786_c0_seq2	1.00E-62	237
	comp3786_c0_seq1	1.00E-62	237
	comp7128_c0_seq1	2.00E-61	233
	comp516_c1_seq16	3.00E-57	219
	comp516_c1_seq16	6.00E-17	85.5
	comp516_c1_seq11	3.00E-57	219
	comp516_c1_seq4	3.00E-57	219
	comp516_c1_seq18	4.00E-57	219
	comp516_c1_seq13	4.00E-57	219
	comp516_c1_seq6	4.00E-57	219
	comp9047_c0_seq1	2.00E-55	213
	comp2178_c0_seq1	4.00E-54	209
	comp3464_c0_seq2	5.00E-54	208
	comp3464_c0_seq1	5.00E-54	208
	comp7335_c0_seq4	6.00E-52	201

	comp7335_c0_seq3	7.00E-52	201
	comp4522_c0_seq2	1.00E-51	201
Cam(CALM)	comp242_c0_seq1	2.00E-81	297
	comp146_c0_seq1	2.00E-59	224
CDase	comp926_c0_seq1	0	640
Cnx99A	comp357_c0_seq1	4.00E-149	525
	comp357_c0_seq3	2.00E-115	412
	comp357_c0_seq2	6.00E-89	325
	comp357_c0_seq4	4.00E-55	212
DopR	comp5875_c0_seq1	8.00E-119	423
	comp32630_c0_seq1	3.00E-60	229
	comp9821_c0_seq1	3.00E-52	202
	comp16092_c0_seq1	4.00E-50	195
Galpha49B	comp3854_c0_seq1	0	648
	comp3571_c0_seq1	0	633
	comp3854_c0_seq2	3.00E-176	614
	comp6597_c0_seq1	4.00E-98	354
	comp3277_c0_seq1	1.00E-91	332
	comp10678_c0_seq1	4.00E-85	311
	comp13043_c0_seq1	2.00E-76	282
	comp2252_c0_seq1	4.00E-76	281
	comp6597_c0_seq2	3.00E-72	268
	comp2252_c0_seq3	7.00E-56	214
inactivation no afterpotential C	comp10430_c0_seq5	0	657
	comp4273_c0_seq1	2.00E-169	352
	comp10430_c0_seq6	5.00E-139	491
	comp10430_c0_seq7	1.00E-99	360
	comp15883_c0_seq1	2.00E-91	333
	comp4034_c0_seq1	4.00E-90	328
	comp1885_c0_seq3	4.00E-89	325
	comp1885_c0_seq2	6.00E-88	321
	comp1885_c0_seq1	1.00E-79	294
	comp3156_c0_seq1	3.00E-77	286
InaD	comp15375_c0_seq1	3.00E-74	276
	comp15375_c0_seq2	3.00E-74	276
	comp599_c1_seq2	1.00E-53	207
	comp599_c1_seq1	1.00E-53	207
InaE	comp16066_c0_seq4	0	749
	comp16066_c0_seq6	0	739
	comp16066_c0_seq11	0	737
	comp16066_c0_seq1	0	737
	comp16066_c0_seq2	0	728
	comp16066_c0_seq9	0	726

	comp16066_c0_seq7	0	705
	comp16066_c0_seq8	0	696
	comp16066_c0_seq3	0	695
	comp16066_c0_seq12	0	695
	comp16066_c0_seq5	0	686
	comp16066_c0_seq10	0	684
	comp14188_c0_seq1	3.00E-79	292
	comp14188_c0_seq2	2.00E-76	283
Laza	comp2647_c0_seq1	2.00E-32	135
	comp10578_c0_seq1	3.00E-28	122
Mts	comp1645_c0_seq1	7.00E-179	622
	comp9662_c0_seq1	3.00E-143	504
	comp1645_c0_seq2	4.00E-141	497
	comp2026_c0_seq1	2.00E-101	365
	comp567_c0_seq1	3.00E-79	291
	comp16991_c0_seq1	2.00E-69	259
	comp1143_c0_seq1	5.00E-68	254
	comp4777_c0_seq1	1.00E-57	219
	comp4454_c1_seq1	2.00E-56	215
	comp12732_c0_seq1	2.00E-55	213
	comp8732_c0_seq1	5.00E-55	211
NinaC	comp8925_c0_seq2	2.00E-64	244
	comp8925_c0_seq1	2.00E-64	244
	comp4741_c0_seq1	2.00E-60	231
	comp4508_c0_seq1	1.00E-58	225
	comp13759_c0_seq1	2.00E-58	224
	comp5601_c0_seq1	8.00E-57	219
	comp1481_c0_seq2	3.00E-53	207
	comp7775_c0_seq1	4.00E-53	206
	comp7375_c0_seq1	7.00E-53	205
	comp9720_c0_seq7	2.00E-52	204
	comp9720_c0_seq4	2.00E-52	204
	comp9720_c0_seq3	2.00E-52	204
	comp9720_c0_seq2	2.00E-52	204
	comp9720_c0_seq1	2.00E-52	204
norpA	comp7707_c0_seq4	0	696
	comp7707_c0_seq2	0	693
	comp7707_c0_seq6	0	683
	comp7707_c0_seq8	0	682
	comp7707_c0_seq5	0	681
	comp7707_c0_seq7	0	679
	comp7707_c0_seq3	0	352
	comp7707_c0_seq1	0	352

	comp11183_c0_seq1	5.00E-107	385
	comp6205_c0_seq2	1.00E-102	371
	comp6205_c0_seq1	1.00E-102	371
Rh	comp16187_c0_seq1	1.00E-23	107
Shab	comp13950_c0_seq1	0	756
	comp13950_c0_seq2	0	756
	comp13950_c0_seq9	0	755
	comp13950_c0_seq8	0	741
	comp8211_c0_seq1	1.00E-124	444
	comp13950_c0_seq4	4.00E-123	439
	comp13950_c0_seq3	4.00E-123	439
	comp13950_c0_seq11	5.00E-123	438
	comp13950_c0_seq10	2.00E-118	423
	comp13260_c0_seq2	1.00E-87	320
	comp13260_c0_seq1	1.00E-87	320
	comp15083_c0_seq2	1.00E-70	264
	comp15083_c0_seq1	1.00E-70	264
	comp8840_c1_seq2	2.00E-70	263
	comp8840_c1_seq1	2.00E-70	263
	comp15083_c0_seq4	8.00E-70	262
	comp15083_c0_seq3	8.00E-70	262
	comp2113_c0_seq4	2.00E-61	234
	comp2113_c0_seq1	2.00E-61	234
	comp2113_c0_seq5	2.00E-61	234
	comp2113_c0_seq2	2.00E-61	234
Syt1	comp12565_c0_seq1	2.00E-137	485
	comp12565_c0_seq2	3.00E-67	252
	comp16717_c0_seq1	8.00E-65	244
	comp16717_c0_seq2	9.00E-65	244
	comp11967_c0_seq1	3.00E-56	215
Trpl	comp18136_c0_seq1	4.00E-75	279

Supplementary Table S19 Primers used for qPCR to quantitatively detect the expression of phototransduction pathway genes.

Genes	Directions	Sequences
ninaC	Forward	5’-AGCAATTTTCCTAAGGCTAC-3’
	Reverse	5’-AAACATTTTGGATATCATCG-3’
Mts	Forward	5’-TTAGTAGATGGTCAGATTT-3’
	Reverse	5’-CACATAGGACCCTCATGA-3’
Syt1	Forward	5’-TTCAACTTTAAGGTACCGTT-3’
	Reverse	5’-CTAGCTTGTTTTCCTGTCC-3’
CaMKII-PD	Forward	5’-GCTGGGGCTTATGATTAT-3’
	Reverse	5’-GTGTAGAGGCAACACGCT-3’
CaMKII-PH	Forward	5’-TGGATATGTCAAAGGGA-3’
	Reverse	5’-CCCCTTGAGTTTACGTC-3’
norpA	Forward	5’-ACTTCAAGCTGGTTATCGC-3’
	Reverse	5’-TGCATCCATTAAATCTCCT-3’
Cam(CALM)	Forward	5’-GACTGAAGAACAAATTGCTG-3’
	Reverse	5’-GTCAATGGTTCCATTACCA-3’
InaD	Forward	5’-ATATTTACAAGAAGAAACTC-3’
	Reverse	5’-GAGGACCTTTTCTTGTT-3’
Arr	Forward	5’-AAAGCCTCATAAAAGGAATT-3’
	Reverse	5’-ATGATGGTACAACTCTTTATCC-3’
Cdase	Forward	5’-CAATGGAACTCTTCACTACT-3’
	Reverse	5’-ATAACACGGTAAGTATCATT-3’
Cnx99A	Forward	5’-ATTTGTGCTCCAATATGAAG-3’
	Reverse	5’-TAGTTTCTGCTCAGTACCAC-3’
Trp	Forward	5’-CATCAGGATGTTTATCTT-3’
	Reverse	5’-GAACCTTTTCTGTTTGTA-3’
Rh	Forward	5’-ATCCTATAGTTTATTGGTGT-3’
	Reverse	5’-ACCTCCTGTACTATGTATTG-3’
β-actin	Forward	5’-GGTATAGTGACAAATTGGGATG-3’
	Reverse	5’-TTGCCTTAGGATTCAGTGGG-3’

Supplementary Table S20 Sequences used for opsin phylogenetic and light-wavelength bias analyses.

Species Name	Gene Name	Accession Number
Anopheles gambiae	Anopheles_cil	XP_312503
Apis mellifera	Apis_pteropsin	NP_001035057
Bos Taurus	Bos_rhodopsin	62460472
Bos Taurus	Bos_RGR	NP_786969
Carassius auratus	Carassius_cone	P32310
Ciona intestinalis	Ciona_cil	BAB68391
Gallus gallus	Gallus_pineal	P51475
Gallus gallus	Gallus_melanopsin	NP_001038118
Homo sapiens	Homo_peropsins	NP_006574
Macaca mulatta	Macaca_ensephalopsin	XP_001094239
Loligo forbesi	Loligo_rh	P24603
Mizuhopecten yessoensis	Mizuhopecten_rh	O15973
Mizuhopecten yessoensis	Mizuhopecten_Go	O15974
Mus musculus	Mus_neuropsin	NP_861418
Mus musculus	Mus_RGR	AAC69836
Papilio glaucus	Papilio_rh	AAD29445
Petromyzon marinus	Petromyzon_pineal	Q98980
Platynereis dumerilii	Platynereis_rh	CAC86665
Platynereis dumerilii	Platynereis_cil	AAV63834
Petromyzon marinus	Petromyzon_P-opsin	AAC41240
Schistocerca gregaria	Schistocerca_rh	Q94741
Takifugu rubripes	Takifugu_TMT	NP_001027778
Uta stansburiana	Uta_parietopsin	AAZ79904
Branchiostoma	Branchiostoma_melanopsin	Q4R1I4
Pediculus humanus corporis	Pediculus_rhl1	XP_002422997
	Pediculus_rhl2	XP_002426976
	Pediculus_rhl3	XP_002423973
	Pediculus_rhl4	XP_002430734
Apis mellifera	Apis_opsinblue	NP_001011606
	Apis_opsinuv	NP_001011605
Siproeta stelenes	Siproeta_lwave	gb|AAU93396
Mouse	Mouse_mwave	O35599
	Mouse_swave	P51491
Poecilia formosa	Poecilia_lwave	AEQ53949
Apis mellifera	Apis_lwave	NP_001071293
Hydra vulgaris	Hydra_opsin1	BAD67148
	Hydra_opsin2	BAD67147

	Hydra_opsin3	BAD67146
	Hydra_opsin4	BAD67145
	Hydra_opsin5	BAD67144
	Hydra_opsin6	BAD67143
	Hydra_opsin7	BAD67142
	Hydra_opsin8	BAD67141
Hydra magnipapillata	Hydra_opsin9	XP_002160448
	Hydra_opsin10	XP_002163358
	Hydra_opsin11	XP_002163327
	Hydra_opsin12	XP_002163209
	Hydra_opsin13	XP_002157157
Nematostella vectensis	Nematostella_opsin1	FAA00408
	Nematostella_opsin2	FAA00400
	Nematostella_opsin3	FAA00395
	Nematostella_opsin4	FAA00413
	Nematostella_opsin5	FAA00411
	Nematostella_opsin6	FAA00412
	Nematostella_opsin7	FAA00410
	Nematostella_opsin8	FAA00409
	Nematostella_opsin9	FAA00407
	Nematostella_opsin10	FAA00406
	Nematostella_opsin11	FAA00405
	Nematostella_opsin12	FAA00403
	Nematostella_opsin13	FAA00402
	Nematostella_opsin14	FAA00404
	Nematostella_opsin15	FAA00401
	Nematostella_opsin16	FAA00399
	Nematostella_opsin17	FAA00398
	Nematostella_opsin18	FAA00396
	Nematostella_opsin19	FAA00394
	Nematostella_opsin20	FAA00397
	Nematostella_opsin21	FAA00393
	Nematostella_opsin22	FAA00392
	Nematostella_opsin23	FAA00391
	Nematostella_opsin24	FAA00390
	Nematostella_opsin25	FAA00389
	Nematostella_opsin26	FAA00388
	Nematostella_opsin27	FAA00387
	Nematostella_opsin28	FAA00386
	Nematostella_opsin29	FAA00385
	Nematostella_opsin30	FAA00384
	Nematostella_opsin31	FAA00383

Branchiostoma belcheri	Branchiostoma_opsin1	BAC76019
	Branchiostoma_opsin2	BAC76020
	Branchiostoma_opsin3	BAC76021
	Branchiostoma_opsin4	BAC76024
	Branchiostoma_opsin5	BAC76023
	Branchiostoma_opsin6	BAC76022
Drosophila melanogaster	DroMel_rh1	*
	DroMel_rh6	*
	DroMel_rh2	*
	DroMel_rh3	*
	DroMel_rh4	*
	DroMel_rh5	*
	DroMel_rh7	*
Drosophila simulans	DroSim_rh7	*
Drosophila sechellia	DroSec_rh7	*
Drosophila yakuba	DroYak_rh7	*
Drosophila erecta	DroEre_rh7	*
Drosophila ananassae	DroAna_rh7	*
Drosophila pseudoobscura	DroPse_rh7	*
Drosophila persimilis	DroPer_rh7	*
Drosophila willistoni	DroWil_rh7	*
Drosophila virilis	DroVir_rh7	*
Drosophila mojavensis	DroMoj_rh7	*
Drosophila grimshawi	DroGri_rh7	*

* The opsins of *Drosophila melanogaster*, *Drosophila simulans*, *Drosophila sechellia*, *Drosophila yakuba*, *Drosophila erecta*, *Drosophila ananassae*, *Drosophila pseudoobscura*, *Drosophila persimilis*, *Drosophila willistoni*, *Drosophila virilis*, *Drosophila mojavensis*, *and Drosophila grimshawi* are obtained from http://genomewiki.ucsc.edu/index.php/Main_Page.

Supplementary Table S21 CYP genes encoding P450 enzymes annotated from the genome of *M. martensii*.

***M. martensii* gene ID**	**Annotation**
CYP2 clan	
MMa00296	CYP18A1
MMa09921	CYP18A1
MMa00297	CYP2D
MMa36879	CYP2A6
MMa00298	CYP2R1
MMa09922	CYP2R1
MMa20055	CYP2R1
MMa40517	CYP2R1
MMa29763	CYP306A1
MMa30372	CYP306A1
MMa53042	CYP306A1
MMa43925	CYP307A
CYP3 clan	
MMa00594	CYP3A
MMa01230	CYP3A
MMa01349	CYP3A
MMa01701	CYP3A
MMa02709	CYP3A
MMa02961	CYP3A
MMa03581	CYP3A
MMa03628	CYP3A
MMa03629	CYP3A
MMa06752	CYP3A
MMa08586	CYP3A
MMa08649	CYP3A
MMa08689	CYP3A
MMa10169	CYP3A
MMa13538	CYP3A
MMa13956	CYP3A
MMa14354	CYP3A
MMa14736	CYP3A
MMa15771	CYP3A
MMa16541	CYP3A
MMa19201	CYP3A
MMa19203	CYP3A
MMa20086	CYP3A
MMa21635	CYP3A

MMa21817	CYP3A
MMa22909	CYP3A
MMa22910	CYP3A
MMa22911	CYP3A
MMa25314	CYP3A
MMa25720	CYP3A
MMa29197	CYP3A
MMa29558	CYP3A
MMa29559	CYP3A
MMa29560	CYP3A
MMa30715	CYP3A
MMa32122	CYP3A
MMa32127	CYP3A
MMa32960	CYP3A
MMa33836	CYP3A
MMa34060	CYP3A
MMa34689	CYP3A
MMa35723	CYP3A
MMa36734	CYP3A
MMa38228	CYP3A
MMa38829	CYP3A
MMa39295	CYP3A
MMa39427	CYP3A
MMa40200	CYP3A
MMa42528	CYP3A
MMa42964	CYP3A
MMa43104	CYP3A
MMa43293	CYP3A
MMa44743	CYP3A
MMa45585	CYP3A
MMa45606	CYP3A
MMa47409	CYP3A
MMa47411	CYP3A
MMa49266	CYP3A
MMa49346	CYP3A
MMa50803	CYP3A
MMa51809	CYP3A
MMa51951	CYP3A
MMa52068	CYP3A
MMa52292	CYP3A
MMa53921	CYP3A

MMa54567	CYP3A
MMa55239	CYP3A
MMa29389	CYP5A
MMa52069	CYP5A
MMa16508	CYP6
MMa41083	CYP6
MMa43991	CYP6
MMa46905	CYP6
MMa02971	CYP6G1
MMa22623	CYP9
MMa29198	CYP9
CYP4clan	
MMa29608	CYP4
MMa29609	CYP4
MMa37178	CYP4
MMa47974	CYP4
MMa47975	CYP4
MMa56901	CYP4
MMa56977	CYP4
MMa51869	CYP4
MMa56069	CYP4
MMa01910	CYP4
MMa02347	CYP4
MMa03517	CYP4
MMa10063	CYP4
MMa15900	CYP4
MMa19194	CYP4
MMa23015	CYP4
MMa25020	CYP4
MMa04333	CYP4B
MMa26868	CYP4V
MMa26869	CYP4V
MMa30075	CYP4V
MMa28734	CYP4V
MMa36428	CYP4V
MMa37047	CYP4V
MMa37048	CYP4V
MMa37179	CYP4V
MMa38799	CYP4V
MMa41395	CYP4V
MMa42755	CYP4V

MMa42253	CYP4V
MMa40786	CYP4V
MMa44661	CYP4V
MMa44199	CYP4V
MMa45665	CYP4V
MMa48244	CYP4V
MMa51741	CYP4V
MMa11223	CYP4V
MMa00502	CYP4V
MMa56082	CYP4V
MMa56611	CYP4V
MMa56261	CYP4V
MMa56263	CYP4V
MMa03089	CYP4V
MMa03484	CYP4V
MMa06754	CYP4V
MMa06371	CYP4V
MMa10391	CYP4V
MMa15094	CYP4V
MMa14283	CYP4V
MMa14284	CYP4V
MMa14285	CYP4V
MMa14287	CYP4V
MMa14288	CYP4V
MMa14289	CYP4V
MMa15899	CYP4V
MMa17295	CYP4V
MMa21472	CYP4V
MMa22660	CYP4V
MMa25021	CYP4V
MMa17917	CYP4V
MMa56739	CYP4X
mitochondrial CYP clan	
MMa38874	CYP12
MMa12181	CYP12
MMa51548	CYP24A
MMa54014	CYP27A
MMa38873	DIB
MMa42737	DIB
MMa46762	DIB
MMa46763	DIB

MMa55958	DIB
MMa25315	SAD, CYP315A1
MMa21388	SHD, CYP314A1

Supplementary Table S22 Comparison of CYP gene numbers in insects, crustacean, mite, and scorpion. CYP gene number of *M. martensii* is derived from this study, and CYP gene numbers of insects, crustacean, and mite are from Grbic et al. (2011).

Species	**CYP2 Clan**	**CYP3 Clan**	**CYP4 Clan**	**Mitochondrial CYP Clan**	**Total**
Insecta					
Drosophila melanogaster	7	36	32	11	88
Anopheles gambiae	10	40	46	9	105
Aedes aegypti	12	82	57	9	160
Bombyx mori	7	30	36	12	85
Apis mellifera	8	28	4	6	46
Nasonia vitripennis	7	48	30	7	92
Tribolium castaneum	8	72	45	9	134
Acyrthosiphon pisum	10	23	23	8	64
Pediculus humanus	8	11	9	8	36
Crustacea					
Daphnia pulex	20	12	37	6	75
Acari					
Tetranychus urticae	48	10	23	5	86
Scorpiones					
Mesobuthus martensii	12	76	61	11	160

Supplementary Table S23 CYP gene clusters summarized from the *M. martensii* genome.

Numbers	Scaffold ID	Strand	Gene ID	Gene Annotation
1	NODE_1109744	+	MMa00296	CYP18A1
		+	MMa00297	CYP2D
		-	MMa00298	CYP2R1
2	NODE_2308936	-	MMa02961	CYP3A
		-	MMa02971	CYP6G1
3	NODE_2478777	-	MMa03628	CYP3A
		-	MMa03629	CYP3A
4	NODE_3760419	+	MMa06752	CYP3A
		+	MMa06754	CYP4V
5	NODE_4719697	+	MMa09921	CYP18A1
		+	MMa09922	CYP2R1
6	NODE_568659	-	MMa14283	CYP4V
		-	MMa14284	CYP4V
		-	MMa14285	CYP4V
		-	MMa14287	CYP4V
		-	MMa14288	CYP4V
		-	MMa14289	CYP4V
7	NODE_5919781	+	MMa15899	CYP4V
		+	MMa15900	CYP4
8	NODE_6306247	-	MMa19201	CYP3A
		-	MMa19203	CYP3A
9	NODE_6629391	-	MMa22909	CYP3A
		-	MMa22910	CYP3A
		-	MMa22911	CYP3A
10	NODE_6740423	-	MMa25020	CYP4
		-	MMa25021	CYP4V
11	NODE_6772116	-	MMa25314	CYP3A
		+	MMa25315	SAD, CYP315A1
12	NODE_6871636	+	MMa26868	CYP4V
		-	MMa26869	CYP4V
13	NODE_6978841	-	MMa29197	CYP3A
		-	MMa29198	CYP9
14	NODE_6988257	-	MMa29558	CYP3A
		-	MMa29559	CYP3A
		+	MMa29560	CYP3A
15	NODE_6991193	-	MMa29608	CYP4
		-	MMa29609	CYP4
16	NODE_7120533	-	MMa32122	CYP3A
		-	MMa32127	CYP3A
17	NODE_7344930	+	MMa37047	CYP4V

		+	MMa37048	CYP4V
18	NODE_7349596	+	MMa37178	CYP4
		+	MMa37179	CYP4V
19	NODE_7398863	-	MMa38873	DIB
		-	MMa38874	CYP12
20	NODE_7643922	-	MMa46762	DIB
		-	MMa46763	DIB
21	NODE_7658622	+	MMa47409	CYP3A
		+	MMa47411	CYP3A
22	NODE_7681785	-	MMa47974	CYP4
		-	MMa47975	CYP4
23	NODE_7790040	-	MMa52068	CYP3A
		-	MMa52069	TBXAS1, CYP5A
24	NODE_8036453	+	MMa56069	CYP4
		-	MMa56082	CYP4V
25	NODE_8045593	+	MMa56261	CYP4V
		+	MMa56263	CYP4V

Supplementary Table S24 Genes involved in juvenile hormone and molting hormone biosynthesis pathways in *M. martensii*.

Gene name (Drosophila)	*M. martensii* gene ID
JHAMT	MMa48581
	MMa07889
	MMa14956
Nvd	MMa30051
	MMa33903
	MMa33904
	MMa40424
	MMa45132
	MMa54826
	MMa01559
	MMa19962
	MMa22754
	MMa23199
spo, spok, CYP307A	MMa43925
CYP306A1 (ecdysteroid 25-hydroxylase)	MMa29763
	MMa30372
	MMa53042
Dib	MMa38873
	MMa42737
	MMa46762
	MMa46763
	MMa55958
SAD, CYP315A1	MMa25315
ecdysone oxidase	MMa36642
SHD, CYP314A1	MMa21388
CYP18A1	MMa00296
	MMa09921

Supplementary Notes

Supplementary Note 1. Genome assembly, gene features and transcriptome analysis

Animals

M. martensii individuals were collected from the Fu-niu Mountains, Xichuan County (33.13-33.17°N, 111.48-111.52°E), Henan Province, China. They were observed and manipulated under a Motic K700 stereoscopic microscope. *M. martensii* has the following characteristics: (1) moderate size, with males reaching 53 mm and females 60 mm in total length; (2) general coloration is yellowish to reddish-yellow, with some darker brownish zones on tergites; (3) metasomal segment V with a conspicuous dark spot covering all faces of the segment; (4) carinae and granulations moderately to weakly marked on carapace, tergites, and metasomal segments; (5) pectinal teeth count of 21-26 in males and 17-22 in females; and (6) trichobothrial pattern of orthobothriotaxic type A (beta)[58]. Remarkable sexual dimorphisms allowed the identification of *M. martensii* males easily. First, males have more globular, bigger chela and shorter fingers than those of females. Second, chela fingers on adult males are scalloped, whereas they are nearly straight in females. Third, the mesosome is round in females, whereas it is generally wizened and flat in males. Finally, male pectinal teeth and metasomes are bigger than those of females. Moreover, we analyzed the hemispermatophores of males before extracting nucleic acid. All animal studies were approved by the Institutional Animal Care and Use Committee at Wuhan University.

DNA and RNA preparation

Previously, it was reported that the scorpion *M. martensii* had an XY sex chromosome system (http://www.scdwzz.com.cn/index_en.aspx)[59]. An adult *M. martensii* male was selected for extraction of DNA samples from muscle tissues of the pedipalp and metasoma (to minimize microbial contamination) for genomic sequencing. The prepared sample was then quickly ground in liquid nitrogen, and DNA extraction was performed using TIANamp Genomic DNA Kit DP304-2 (TIANGEN, China) according to the manufacturer's protocol. The quality and quantity of DNA sample was examined by ultraviolet (UV) absorbance and gel electrophoreses.

We prepared total RNA from a single male scorpion for whole transcriptome sequencing. Briefly, an adult *M. martensii* male was washed three times with 95% ethanol to reduce microbial contamination from its body surface. After ethanol volatilization, the individual was then quickly ground into a fine powder in liquid nitrogen. Total RNA was prepared using the TRIZOL Reagent (Invitrogen, Carlsbad, CA, USA). RNA quantitation was performed by UV absorbance and its quality was further confirmed by gel electrophoreses. Meanwhile, total RNA from the venom

gland of the scorpion *M. martensii* was isolated for transcriptome study. Thirty scorpion individuals (15 males and 15 females) were used for dissection of venom glands (telson). The dissected venom gland tissues were then quickly ground in liquid nitrogen. Subsequently, the preparation and characterization of the total RNA sample were performed as described above.

Flow cytometry analysis

We determined the genome size of *M. martensii* by flow cytometry analysis of fluorescently stained nuclei using chicken erythrocytes as an internal standard[60,61]. The haploid size of *M. martensii* was estimated to be 1.35 ± 0.04 pg, corresponding to 1,323.73 ± 39.12 Mb (Supplementary Table S1). We dissected genital gland tissues of an adult *M. martensii* male individual to prepare a sample for flow cytometry. As shown in Supplementary Fig. S1, the flow cytometry results for *M. martensii* show two peaks. The peak value of the diploid somatic cell DNA is twice the value of sperm cell DNA, which further confirms that the genome size of *M. martensii* determined by flow cytometry is accurate.

Genome shotgun sequencing and assembly

We sequenced the genome of an adult *M. martensii* male using a whole-genome shotgun approach. For sequencing using the Illumina Genome Analyzer or Hiseq2000, paired-end (PE) libraries with insert sizes of 180 bp, 300 bp, and 420 bp were constructed separately, and mate-pair (MP) libraries with circular DNA sizes of 5 Kb and 10 Kb were constructed. The preparation of the paired-end library and sequencing with the Illumina GAIIx/Hiseq2000 sequencing was performed according to standard Illumina protocols (Illumina, San Diego, CA, USA). For the Roche GS-FLX platform, the library was constructed and sequenced using standard protocols(454 Life Sciences, Roche, Brandford, CT, USA). Raw sequence data from each library were summarized in Supplementary Table S2. In addition, the genome was sequenced using a Roche 454 GS-FLX instrument with an average read length of 450 bp. The 454 reads were later used in the gap-filling step.

The Illumina data were first processed to filter out low-quality reads. We eliminated the low-quality bases using the PERL script of fastq_qualitytrim_window.pl from (http://genepool.bio.ed.ac.uk/). The quality threshold and window size were set as 20 and 5, respectively. Data were then screened against microbial, plasmid, and organelle sequences, and removed from the subsequent assembly if found to be contaminants. The remaining sequence reads were assembled using Velvet[33] (version: 1.1.04). The assembling was performed on a computing server, I950R-GP (Dawning Information Industry, Tianjin, China) with eight 2.67-GHz Intel XEONE7-8837 (96 cores) and 1024 G of memory. The operating system is Ubuntu 10.04 LTS. The parameters for Velvet are: kmer = 49, cov_cutoff = 1, min_contig_lgth = 100, min_pair_count = 2, shortMatePaired = yes, scaffolding = yes. The generated scaffolds (with a total size of 1,015.13 Mb) were gap-filled with GapCloser[34] (version: 1.12) using short sequence reads from GAIIx/Hiseq2000 and long reads from Roche 454-FLX. The parameters for GapCloser were: max_rd_len = 1000, avg_ins = 180, reverse_seq = 0, -k = 25,

asm_flags = 3. The resulting genome assembly (v1.0) had a final scaffold N50 length of 223,560 bp and a contig N50 length of 43,135 bp. The largest scaffold had a size of 1.85 Mb. The assembled genome is in good agreement with the experimentally determined genome size, and 95.6% of the genome was composed of a non-gap region (Supplementary Table S3).

The overall GC content in *M. martensii* is 29.6%, which is similar to its close neighbor, *T. urticae*, whose overall GC content is 32.3% (http://www.ncbi.nlm.nih.gov/genome/?term=Tetranychus+urticae)[14]. However, they are both lower than other more distant arthropod neighbors, e.g. *D. pulex* (overall GC content 40.8%; genome version: JGI v.1.0, July 5, 2007)[13], and *D. melanogaster* (overall GC content 42.2%; http://www.ncbi.nlm.nih.gov/genome/47). Whether the lower overall GC content represents a phenomenon special to arachnid remains to be further studied.

Analysis of microsatellite DNA and transposable elements (TEs)

We identified simple sequence repeats, low complexity DNA sequences, and satellites from the *M. martensii* genome using SSRIT[35] and RepeatMasker[36] (version 3.2.9). Default parameters were used for screening. The sizes of simple sequence repeats (Supplementary Tables S4) and low complexity sequences were 34,111,238 bp and 32,106,783 bp, respectively (Supplementary Fig. S2).

Since there is no chelicerate-specific TE library, we constructed a putative *M. martensii* specific TE library. Homologs from RepBase16.01[38], which does not contain chelicerate-specific TEs, were used to query the genome with tblastn (E-value of 1E-5). We used relaxed parameters to collect protein domains associated with TEs. Sequences were extended from TE-domains as far as possible to obtain full-length TEs. As estimated by the collected putative TE sequences, the total size of TEs from *M. martensii* approached 35 Mb. The most abundant TE families were Mariner, hAT, and Gypsy (Supplementary Table S5, http://www.genoportal.org/mm/transposable-element-lib-v1.0.fasta).

To investigate the percentage of raw sequence data mapped to TEs, we sampled ~40G (400,000,000 x 100bp) reads as queries to blast the library of *M. martensii* specific transposable element sequences we constructed earlier. Any reads having >50 bp region with 90% identity toward a TE element were counted as TE-reads. Out of the sampled raw sequence reads, roughly 13.0% reads were mapped to TE sequences.

Single nucleotide polymorphism (SNP) and INDEL in the diploid

Bowtie2[39], a BWT-based aligner, was used to align genome sequencing reads to the assembly. Totally 73G filtering fastq files from three different sequence read data sets were used for alignment with at 'local' condition and other parameters being default. Then genotype calling was performed by Samtools pileup with default setting. These sites with sequence depth >4 were accepted in SNP/INDEL calling. A heterozygous genotype for a base would be called if the frequency of the non-consensus allele were between 20% and 80%[40,41]. As displayed in Supplementary Table S6, all three data sets showed a similar level of single nucleotide polymorphism (~3.24 per 1000 bp).

Gene modeling and annotation

Our gene prediction pipeline is composed of the following steps as displayed in Supplementary Fig. S3. The pipeline was developed based on the published protocol in John K. Colbourne's work[13].

(1) TEs and other sequence repeats in the assembly were masked before gene modeling. We used homologs from RepBase16.01 to construct a *M. martensii* specific TE database as described below. The masked genome sequence was used in the following gene perdition.

(2) The gene models of *M. martensii* genome were predicted using several methods: AUGUSTUS (version 2.5.5)[42,62], Fgenesh++[43], GENEID (version 1.2)[44,63], SNAP[45], GlimmerHMM[46] and Gnomon. Since no gene pattern information was available for scorpions, in the first round of gene prediction we constructed a gene prediction training set. Several self-trained gene model prediction methods, such as Fgenesh++ and GENEID, were used to product gene models respectively. Several training set relied gene prediction software, such as AUGUSTUS and GlimmerHMM, were used to predict gene model with parameters from drosophila gene models.

(3) The prediction results from these four softwares were compared, and only identical gene structures from all four prediction results were considered as potential reliable gene models. By manually curating the gene structures from training set with transcriptome sequencing data alignment results (described below), 876 reliable gene structures were recognized as reliable *M. martensii* gene prediction for training set. Using AUGUSTUS retraining program (optimize_augustus, and etraining model), optimized *M. martensii* specific parameters for gene prediction were obtained.

(4) Next, we used a comprehensive gene prediction pipeline combining ab initio modeling, homology based modeling and EST based modeling. The gene model structures were mainly from the AUGUSTUS prediction results using contrasted training set (from step 3) and transcriptome data as hints file, which is a file with extrinsic evidence about the location and structure of genes, see documentation of AUGUSTUS. Using this training set models, we also retained GlimmerHMM parameters and conducted the second round of gene perdition with GlimmerHMM. And we conducted the prediction with Gnomon and SNAP, using transcriptome data as evidence. These results were compared with AUGUSTUS prediction. Those gene structures have evidence from less than two software were removed.

(5) We also compared the predicted gene with protein homology from public databases (NCBI, Swiss-Prot, Pfam and et al.), as additional but not required evidence. In addition, these gene models were also manually checked. Gene models without clear open reading frame (ORF) were removed.

Eventually, we presented a minimum gene set of *M. martensii* genome. This refers to version 1.0 gene models (http://www.genoportal.org/mm/gene-models-v1.0.gff), including 32,016 protein coding gene models (>35 AA). They are grouped into 5,947 distinct gene families and 3,211 orphan genes using OrthoMCL (more details in Supplementary Note 2). The script used in gene prediction pipeline will be provided upon request.

Software list:
(1) AUGUSTUS - http://augustus.gobics.de/
(2) GlimmerHMM - http://ccb.jhu.edu/software/glimmerhmm/
(3) Fgenesh++ - http://www.molquest.com/help/2.3/programs/Fgenesh++/about.html
(4) GENEID - http://genome.crg.es/software/geneid/index.html
(5) Gnomon - https://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml
(6) SNAP - http://korflab.ucdavis.edu/software.html

For 32,016 protein-coding genes in *M. martensii*, they have a median gene size of approximately 6.7 Kb (open reading frame plus included introns). On average, each gene has 3.9 exons, similar to that of *T. urticae* and *D. melanogaster* (Supplementary Table S7). The average GC content in the genes was 42.7%. The average intron size was approximately 2.12 Kb (Supplementary Fig. S7). Approximately 65% of the introns have lengths less than 2,000 bp.

'Pseudogenes' are defined as dysfunctional genes that have evolved to lose their protein-coding ability or to be no longer expressed in the cells due to mutations. The process for estimating pseudogenes in *M. martensii* is: (1) Gene models were collected from gene prediction results. The predicted gene models were translated into protein sequences with 3 possible reading frames with transeq program from EMBOSS package. (2) We identified their homolog sequences from public database: Uni-Prot database with BLAST (E value < 1E-5). These gene models were aligned to their homologues in all 3 different frames. (3) Those genes with two different frames aligned to one single known protein were identified as genes with frameshift. We identified and counted genes with either premature stop codons or indels that cause frameshifts. (4) For the rest genes without known homolog from other species, those belonging to gene clusters were aligned to their homologues within their corresponding clusters. Using similar protocol as described above, genes with premature stop codons or frameshifts were identified. In the *M. martensii* genome, 15,825 gene models are homologous to proteins from other sources, and within them 526 genes were found to have a premature inframe stop codon and 26 ones with frameshifts, which are considered pseudogenes. For the 16,191 genes without homology to known proteins, pseudogenes were estimated as predicted CDS containing premature stop codon, in which 784 were found. And 55 genes were identified to have clear frameshift compared with their paralogs. The combining 1,391 (~4.34%) genes were considered possible pseudogenes in *M. martensii*, which is in line to studies on other arthropods. The pseudogene list was provided in our web site (http://www.genoportal.org/mm/pseudogene-list.txt).

tRNAs were identified using the program t-RNAscan SE. The RNAmmer program was used to predict rRNAs. These rRNAs, as well as small nuclear and nucleolar RNAs, were included in the tRNA/rRNA gene file (http://www.genoportal.org/mm/noncoding-RNA.xls).

Gene annotation was performed using a BLAST search against the NCBI nr, Swiss-Prot, and TrEMBL databases using default parameters (setting E value of 1E-5). The CDD and PFAM databases were used for functional domain annotation. Gene

ontology analysis[64] was performed using the annotation results from the NR, Swiss-Prot, and TrEMBL databases (Supplementary Fig. S4). The metabolic pathway was constructed based on the KEGG database using KAAS provided by the KEGG database (Supplementary Fig. S5)[49].

Of the total number of 32,016 protein-coding genes, nearly 50% were annotated (http://www.genoportal.org/mm/gene-annotation-v1.0.xlsx). The percentages of the predicted genes annotated to the NCBI nr, Swiss-Prot, TREMBL, CDD, and PFAM databases were 40.34%, 31.88%, 42.90%, 24.87%, and 37.82%, respectively.

To compare the protein coding genes between *M. martensii* and *T. urticae*, the same annotation pipeline was perform on genes from *T. urticae* genome. In total, 18,961 (59.2%) show similarity to proteins from non-*Mesobuthus* species, while the proportion in *T. urticae* was 60.8%.

Evidence for predicted gene models

A minimum gene set of 32,016 prtoein-coding genes (excluding Transposase-like genes) was found in the *M. martensii* genome, which is the most so far among sequenced arthropods. The protein-coding gene models for *M. martensii* was assessed and supported by the following evidence.

To provide experimental support for the gene models and to facilitate gene annotation, we isolated mRNA from mixed tissues and venom glands to be sequenced using Illumina Hiseq2000. Library construction and sequencing were performed according to Illumina's standard protocols. Approximately 27,665,490 and 22,607,723 reads (2 × 100 bp) were obtained from the mixed tissues and venom glands, respectively. We also isolated and sequenced mRNA from mixed whole-body samples using the 454 GS-FLX platform. Sample preparation and sequencing were performed using standard protocols from Roche 454. A total of 1,169,644 reads were generated with an average length 410 bp.

Illumina Solexa reads were mapped to the predicted gene set with Tophat using default parameters. The mapped reads of each gene model were counted from output files (sam format). The expression level of genes is measured using RPKM (http://www.genoportal.org/mm/RNA-expressioin-RPKM-v1.0.txt). To investigate the expression of gene model, we filtered out these genes with < 2 reads mapped onto coding region. Out of 32,016 gene models, 29,837 (93.2%) predicted genes were broadly supported by RNA-seq data. The coding regions were supported by at least 2 reads. Junctions between two exons may be covered by RNA-Seq reads, but were not required. When the shared and *Mesobuthus*-specific genes are counted separately, both have the support: 97.4% vs 88.4%. Also, the distributions of gene expression between the two groups show no significant difference (Supplementary Fig. S6). These data strongly indicate that *Mesobuthus*-specific genes are no more "artifact" than those genes shared by the species.

We downloaded the newly published transcriptome data from *C. noxius*[15], a scorpion species from the deferent genus of *M. martensii*, and compared them to the predicted gene models of *M. martensii*.. The sequencing reads were obtained from SRA database provided from the reference[15]. We aligned the read with predicted gene

models with BLAST program (E value < 1E-5). The regions of gene models covered by more than one reads were summarized. The validated gene models were defined as these genes with >40% of its region covered by *C. noxius* transcriptome sequencing reads. The result indicated that 14,785 (46.2%) gene models were found to have homologs in *C. noxius*. When the shared and *Mesobuthus*-specific genes from *M. martensii* were accounted separately, their rates were ~65% and ~25%, respectively.

To demonstrate the completeness of *M. martensii* assembly and gene modeling, 457 of the 458 (> 99%) core eukaryotic genes CEGMA (Core Eukaryotic Genes Mapping Approach: http://korflab.ucdavis.edu/datasets/cegma/)[16] were identified in the gene models for *M. martensii*. 96% of core eukaryotic genes were found in its transcriptome data. The BLAST search program (blastp) was used to align *M. martensii* protein with core eukaryotic genes from CEGMA by setting the cutoff for E-value < 1E-5.

To study the gene duplication events in *M. martensii*, we estimated the synonymous distances (Ks values) among all paralog pairs (more details in Supplementary Note 3). The possibility of erroneous assembly of alleles might result in the artificial gene paralogs. According to Aphid Genomics Consortium[65], the divergence values for allelic variants are generally very low, thus the cutoff of Ks (<0.01) could be considered to filter out potential allelic variant from true paralogs. The possibility of erroneous assembly of alleles from the same locus as separate genes was estimated to be smaller than 1%.

Analysis of putative alternatively spliced genes

To investigate alternatively spliced genes in *M. martensii*, we used the RNA-seq data to detect potential alternative patterns for the predicted gene models. The raw RNA-seq reads were trimmed from 5' end to 3' end, by setting the window size as 5 bp and mean quality value of each window over 20. Reads containing 'N' were removed. The clean data were mapped onto the *M. martensii* genome (v1.0) using TopHat[66] (version 1.3.1). The mapping was performed with paired-end data, using the predicted gene structures as reference gene models by setting '-G' parameter. Alternatively spliced genes and the transcript products were predicted using Cufflinks[67] (version 2.0.1). The accepted bam file was used by Cufflinks (with default parameters) to detect both known and novel transcript structures. The reported transcript structures (gtf format) were then compared with predicted gene structures using Cuffcompare. As a result, 6,139 gene models from the predicted gene sets were detected to have at least two transcripts of differently spliced form, supported by RNA-seq data (Supplementary Table S8).

Supplementary Note 2. Comparative genomics and evolution

Dataset collection and phylogenetic tree construction

To study the comparative genomics *M. martensii* with *T. urticae*, *D. melanogaster, Acyrthosiphon pisum, Apis mellifera, Bombyx mori, Camponotus floridanus,*

Anopheles gambiae, Aedes aegypti, Culex pipiens, Pediculus humanus, *Tribolium castaneum*, *Nasonia vitripennis*, *Daphnia pulex*, *Caenorhabditis elegans*, *Ciona intestinalis*, and *Homo sapiens*, we collected the genome dataset by downloading their genomic sequences and predicted gene/protein data from NCBI (ftp://ftp.ncbi.nih.gov/) (Supplementary Table S9). Orthologous gene families were identified using OrthoMCL[53] with default parameters, with the –I parameter set to 3.0, for all proteins of the various species (17 species, including *M. martensii*).

To construct the phylogenetic tree for these 17 species, we collected the eukaryotic core gene set from 17 genomes, and found that out of 457 core genes, 220 were conserved among all the species (http://www.genoportal.org/mm/common-gene-family.txt). These 220 orthologs present in each of the 17 species were aligned using ClustalX[68] (version 1.83) and concatenated. A phylogeny was built with MEGA5 using neighbor-joining (NJ) and maximum likelihood (ML) methods (500 replicas for the bootstrapping test). The two types of trees had identical topologies (Supplementary Fig. S8).

The result clearly shows that while *M. martensii* is morphologically conserved, its gene sequences diverged at a rate comparable to (or even faster than) those of the sampled insect species, a pattern similar to that observed with the gene family gain/loss analysis.

Orthologous clustering and comparative genomics

Considering that *D. pulex* and *T. urticae* are the two sequenced species most closely related to *M. martensii*, we compared the gene sets of three species to identify homologous genes (all-against-all blastp, and E-value cutoff of 1E-5). *M. martensii* contains 14,298 genes that have homologs in the *D. pulex* genome, and 4,241 pairs among this set were identified as both best hit (BBH) gene pairs. The numbers of homologous genes and BBH genes between *M. martensii* and *T. urticae* are similar with those between *M. martensii* and *D. pulex*.

The gene families for the 11 species (*Aedes aegypti*, *Drosophila melanogaster*, *Tribolium castaneum*, *Camponotus floridanus*, *Acyrthosiphon pisum*, *Daphnia pulex*, *Mesobuthus martensii*, *Tetranychus urticae*, *Caenorhabditis elegans*, *Ciona intestinalis*, and *Homo sapiens*), were identified using OrthoMCL[53] clustering. All sequences were processed with orthomclFilterFasta to eliminate poor-quality sequences. The filter is based on length and percent stop codons according to the orthoMCL standard protocol. An all-versus-all BLAST search was performed with threshold setting at 1E-5 for the E-value. In total, 32,016 gene families were identified in all species, whereas 5,947 gene families and 3,211 orphan genes were identified in the proteome genome of *M. martensii* using default parameters for MCL clustering (I = 3.0) (http://www.genoportal.org/mm/gene-family-v1.0.txt).

The definition of gene family classes are as the followings:

(A) Lineage-specific gene families: these *M. martensii* families that contains no gene from other species were defined as *Mesobuthus*/Scorpiones lineage specific families.

(B) Shared gene families: these *M. martensii* families that contains homology genes from other species were defined as shared families.

(C) Gain of a gene family: For a species, there exists at least one member of the gene family, but none in any other species studied.
(D) Loss of a gene family: For a species, member(s) of a gene family exist in both the neighboring branch and the out-group, but none exits in itself. According to this standards, the loss of a gene family was defined as the process of gene family lost in the period of evolution from the most recent common ancestor (MRCA).
The events of gene family gain and loss for extant species were identified according to the definition above. The gene family list of their most recent common ancestor can be inferred as (1) consisting of the common gene families shared by its two direct descendants, (2) plus the gene families shared by one of the direct descendants and the out-group. The gene family lists of all ancestor nodes in the phylogenetic tree were inferred with their direct descendants through the linage, starting from the extant species. As a result, the MRCA of *A. aegypti* and *D. melanogaster* has 6,595 gene families; the MRCA of *A. aegypti, D. melanogaster* and *T. castaneum* has 6,604 gene families; the MRCA of *Aedes aegypti*, *D. melanogaster*, *T. castaneum* and *C. floridanus* has 6,865 gene families; the common ancestor of insects (CAoI) has 6,802 gene families, the number of gene families of MRCA of insecta and *D. pulex* is 6,811. The MRCA of *M. martensii* and *T. urtice* has 5,842 gene families. Out of these 5,842 gene families, 1,302 ones were lost through the evolution, and other 1,407 ones were gained in *M. martensii* (Fig.1b). The MRCA of arthropoda has 6,750 gene families.
From the above process, the gene family list of any node (including ancestors and extant species) was constructed. In addition, the gene family gain and loss from an ancestor node to a descendant node can be inferred according to the definition (C) and (D) by comparing the gene family list of the two related nodes. A gain of gene family through the path is the gene family absent from the ancestor but present in the descendant. A loss of gene family through the path is the gene family absent from the descendant but present in the ancestor. Over the last ~400 million years, the Mesobuthus lineage has a gene family turn-over of 2,709 (1,407+1,302). We computed the gene family turn-over value for each insect lineage from the common ancestor of insects (CAoI). The gene family turn-over from CAoI to *Aedes aegypti* is 2,398; the gene family turn-over from CAoI to Drosophila melanogaster is 2,698; the gene family turn-over from CAoI to *Tribolium castaneum* is 1,759; the gene family turn-over from CAoI to *C. floridanus* is 1,502; the gene family turn-over from CAoI to *A. pium* is 2,047. These results clearly show that comparing to these sampled insect lineages, the Mesobuthus lineage has a greater gene family turn-over despite its apparent conservation in morphological evolution.

Supplementary Note 3 Expansion of shared and lineage-specific gene families

Expansion of *M. martensii* gene families

Gene family expansions in the *M. martensii* genome were studied based on the homologous gene families of predicted genes. The gene families existing in at least three species besides *M. martensii* were considered. The z-score of each gene family was calculated by: (the gene number for each family - the average gene number of the family from all species) / the standard divergence of gene numbers of the family from all species. The families with z-scores ≥ 2 were extracted, which represents significantly expanded gene families in *M. martensii*. As a result, 496 gene families in *M. martensii* were identified as significantly expanded gene families. In contrast, *T. urticae* has only 172 gene families with a z-score ≥ 2.

The numbers of expanded gene families and large gene families in other species were listed in Supplementary Table S10. The most significantly expanded gene families were acetyl-CoA acyltransferase, DnaJ homolog subfamily C member 11, phospholipase like protein family, histone-lysine N-methyltransferase SETMAR, abhydrolase domain-containing protein 11, and et al. (Supplementary Table S11). Besides the significant expanded families, there exist 90 largely expanded gene families in *M. martensii*, which contain over 20 numbers in each family and exist in less than three species including *M. martensii*. The function of most of these large gene families in *M. martensii* remains unknown. The annotation of expanded gene families and large gene families could be found in supplementary files (http://www.genoportal.org/mm/expanded&large-family-annotatino-v1.0.xlsx).

From the comparison, we conclude that while generally Daphnia and Mesobuthus have similar number of genes and their distributions of gene family size are quite similar as well, Mesobuthus has on average a larger gene family size than Daphnia and higher number of families with 100+ members. This confirms a general trend (that can also be drawn from Fig. 1b of this text): Daphnia has more expanded families that on average are less expanded, and Mesobuthus has fewer families that on average are more expanded. We looked closely at those most expanded families in Daphnia and Mesobuthus. We found there is no overlapping of most expanded families between the two. While the expanded families in Daphnia concentrate in protocadherin family, aldehyde dehydrogenase family, ATP-binding cassette sub-family, long chain fatty acid CoA ligase, coronin family, flavin-containing monooxygenase, calcium release-activated calcium channel protein, and clathrin heavy chain protein family, they in Mesobuthus concentrate in transposase, histone-lysine N-methyltransferase, phospholipase like protein family, ryanodine receptor, ubiquitin carboxyl-terminal hydrolase, FKBP12-rapamycin complex-associated protein family, E3 ubiquitin-protein ligase, and calcitonin receptor-like protein family (Supplementary Fig. S9).

Synonymous mutation of gene paralog in *M. martensii*

Paralogy genes from *M. martensii* were identified with BLAST program, using default parameters, setting the cutoff of E-value as 1E-5. For pairwise alignment results, there gene pairs with identity > 70% and alignment length > 70 amino acid were considered as potential gene paralogs. The evolution divergence (number of synonymous substitutions per synonymous site or Ks) between paralogs was

estimated with codeml program from PAML package[50], using maximum-likelihood method. According to John K. Colbourne's protocol[13], the Ks values were corrected by multiply the ratio between the alignment length and the length of the alignment minus the conversion tract. The synonymous rates of gene families in *M. martensii* were estimated by using the average Ks of gene paralogs in each family.
The same protocol was performed on the gene set from *T. urticae* and *D. pulex*. To investigate the differential duplicate rates among these three species, the number of single duplicated paralogs within the youngest cohort (Ks < 0.01) was calculated. The gene birth rate was estimated by: number of youngest single duplicate gene pairs / (the number of single copy genes + the number of single duplicate gene pairs)[13]. The gene birth rate of *M. martensii* genome (0.076) is 68% of that in *D. pulex* genome (0.113), and is 7 times of that in *T. urticae* genome (0.011).

Supplementary Note 4. Genetic diversity of venom neurotoxins and their receptors

Biodiversity of neurotoxin genes

All reported scorpion neurotoxins were retrieved and collected from the GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and EMBL (http://www.ebi.ac.uk/embl/) sequence databases. A total of 490 scorpion neurotoxins were obtained from both databases. With these sequences, a scorpion neurotoxin query set was constructed and used in a tblastn search against the *M. martensii* genome. E-values less than 10^{-4} were used as a cut-off for significant hits of candidate neurotoxin genes. Each hit was further verified manually, based on the similarity of their sequences, cysteine patterns, and gene structural organizations. Neurotoxin homologs in *M. martensii* were classified into four families, including NaTx (toxins for sodium channels), KTx (toxins for potassium channels), ClTx (toxins for chloride channels), and CaTx (toxins for ryanodine receptors), A total of 116 neurotoxin genes were found in *M. martensii*, including 61 NaTx, 46 KTx, 5 ClTx, and 4 CaTx. They were named following the traditional nomenclature, and annotated on the *M. martensii* genome (Supplementary Table S12 and Supplementary Fig. S10). Forty-five are novel neurotoxins (21 NaTx, 18 KTx, 3 ClTx, and 3 CaTx) that were newly discovered and characterized in *M. martensii* (Supplementary Figs. S11 to S14). A similar approach was used to analyze the transcriptome of scorpion tails. Out of 116 neurotonxins genes found in the *M. martensii* genome, 109 (94%) were confirmed to be expressed in its venomous gland.

Organization and structure of neurotoxin genes

By comparison of neurotoxins' cDNAs to their corresponding genomic sequences, the genomic organization and structure of the neurotoxin genes were clarified.
The neurotoxin genes contained one, two, or none intron. The form with one intron and two exons represents the majority of neurotoxin genes: the first exon spans the 5'

untranslated region and the stretch encoding the first two-thirds of the signal peptide, whereas the second exon encodes the last third of the signal peptide, the mature neurotoxin, and the 3' untranslated region. As the two main components of scorpion venoms, although the NaTx neurotoxins share similar genomic organization and structure with the KTx neurotoxins, the NaTx and KTx neurotoxins show some differences on the genomic level. The intron sizes of NaTx neurotoxin genes were generally greater than 400 bp in length, significantly longer than those of KTx neurotoxin genes (<200 bp) (Supplementary Figs. S15 and S16). The scaffold location analysis of neurotoxin genes indicated that they exist in cluster on the *M. martensii* genome. The clustered neurotoxin genes were summarized in Supplementary Table S13.

New defensin genes and their structural organization

Previously, 14 defensin peptides were isolated and identified from the hemolymph and venom glands of scorpions[69-71]. These peptide sequences were used for a tblastn search against the *M. martensii* genome. E-values less than 10^{-4} were taken to represent significant hits of similarity during the sequence search. As a result, we discovered six new defensin genes (i.e., BmKDfsin1, BmKDfsin2, BmKDfsin3, BmKDfsin4, BmKDfsin5, and BmKDfsin6) in the *M. martensii* genome (Supplementary Table S14 and Supplementary Fig. S17). Further sequence analysis showed that all six defensin genes in *M. martensii* had one phase-I intron with a GT-AG splicing rule located at the end of signal peptide coding region, which shared the same genomic organization and structure as the NaTx and KTx neurotoxins (Supplementary Fig. S18). Interestingly, the intron sizes of the six defensin genes were 398 bp, 147 bp, 131 bp, 128 bp, 416 bp, and 104 bp, respectively. Two defensin genes (i.e., BmKDfsin1 and BmKDfsin5) from *M. martensii* have similar intron lengths to those of the NaTx neurotoxin genes (Supplementary Figs. S15 and S18), whereas the remaining four defensin genes (i.e., BmKDfsin2, BmKDfsin3, BmKDfsin4, and BmKDfsin6) had similar intron lengths as those of the KTx neurotoxin genes (Supplementary Figs. S16 and S18). The scaffold location analysis of defensin genes indicated that at least four of the six defensin genes cluster on the scorpion genome (Supplementary Table S13).

Origin and evolution of neurotoxin genes

One hundred and sixteen genes encoding putative scorpion neurotoxins were characterized in the genome of *M. martensii*. We searched the NCBI NR database for homologs of scorpion neurotoxin genes, and none were found from species other than scorpions. These results suggest that these neurotoxin genes evolved independently within scorpions. These genes were grouped into four families through functional annotations. The most diversified families are NaTx and KTx neurotoxins, with 61 and 46 genes in the genome, respectively. Among the 116 neurotoxin genes, the NaTx and KTx groups account for more than 92%, representing the two main active components of the scorpion venom arsenal for defense.

To investigate the origin and evolution of neurotoxin genes in *M. martensii*, 54 NaTx, 41 KTx, and 5 ClTx near full-length neurotoxin genes were selected together with 6 defensin genes to perform phylogenetic analysis. We attempted to build a phylogenetic tree to infer their phylogeny using the NJ, ML, or UPGMA method. These efforts were hindered by the difficulty in producing informative multiple sequence alignments of the neurotoxin genes due to the short sequence lengths of the molecules and the large divergence at each amino acid position. Alternatively, hierarchical clustering[17,72] was used to group the related neurotoxin genes. Each pair of neurotoxin genes was compared and an alignment score was used to represent the similarity between the pair of sequences. Hierarchical clustering was performed based on the pairwise similarity matrix of the neurotoxin genes using the stats package from R (version 2.11.1, http://www.r-project.org/).

Resistance of *M. martensii* to its own venom

The resistance of the scorpion *Androctonus australis* to its own venom suggested that the venom from *A. australis* was pharmacologically inactive on the voltage-sensitive K^+ and Na^+ channels of its own muscle and nerve cells[18]. Here, an acute poisoning experiment was carried out with the scorpion *M. martensii* and the cockroach *Blaptica dubia*. Scorpion and cockroach individuals fed in the lab for 1 month to allow adaptation to the experimental environment. All animals were healthy adults. The venom of *M. martensii* was obtained by electrical stimulation of adult individuals (average venom production of an adult scorpion: 1-2 μl) and immediately kept in a 4°C icebox. Eighty cockroaches were randomly selected into eight groups (10 per group). Each group was injected with water (control) or with one of seven doses of fresh venom: 1/2 μl, 1/4 μl, 1/8 μl, 1/16 μl, 1/32 μl, 1/64 μl, 1/128 μl. Meanwhile, 30 scorpions were randomly selected into three groups (10 per group). Each group was injected with water (control), or 10 μl or 5 μl fresh venom. Animals were singly fed and observed for 24 hours after ventral subcutaneous injection of the venom. The number and course of animal death were recorded. Statistical analysis of the data was carried out using Microsoft Excel version 2007 to determine median lethal dose (LD_{50}). Measurement data were expressed as the mean ± standard deviation (SD), and the Student's *t* test was used for two-group comparisons. The LD_{50} value of fresh venom injected into *B. dubia* was 0.055 μl/g, 100 times lower than that of *M. martensii* (LD_{50} > 5.56 μl/g). The experimental results demonstrated the resistance of *M. martensii* to its own venom (Supplementary Tables S15 and S16).

Cloning and sequencing of the scorpion K^+ ion channels

Based on the annotated *M. martensii* genome, two K^+ ion channel gene homologs (MmKv1 and MmKv2) were identified and found to be expressed (Supplementary Fig. S19). To investigate the electrophysiological function of the scorpion K^+ ion channels, we used reverse transcription-polymerase chain reaction (RT-PCR) to clone full-length cDNAs encoding MmKv1 and MmKv2. Total RNA was prepared from *M. martensii* using the method described in S2.1.2. RNA quantitation was performed by UV absorbance, and quality was further confirmed by gel electrophoreses. First-strand

cDNA was synthesized from *M. martensii* total RNA using random hexamers (AP Biotech, Germany) and SuperScriptII reverse transcriptase according to the manufacturer's protocol (Invitrogen). The cDNAs of MmKv1 and MmKv2 were amplified by PCR. The PCR products were cloned into the pTZ57R/T TA cloning vector (Fermentas, USA). *E. coli* were transformed with the cloning vector and allowed to grow. DNA was extracted from the bacterial cultures and isolated PCR inserts were sequenced in both directions with an ABI 3100 sequencer (Applied Biosystems, Foster City, CA, USA).

Expression and electrophysiological recordings

The cDNAs encoding MmKv1 and MmKv2 were subcloned into the *Xho*I/*Bam*HI sites of the pIRES2-EGFP vector (Clontech, USA) and transiently transfected into HEK293 cells (CCTCC, China) using the FuGene Transfection Reagent (Roche Diagnostics, Switzerland). The ion channel currents were measured 1-3 days after transfection. To measure the Kv channel currents, the internal pipette and external solutions were prepared according to the previous procedure[51,52]. The voltage-gated K^+ channel currents were elicited by 200-ms depolarizing voltage steps from the holding potential of -80 mV to +50 mV. The formation of functional K^+ channels in MmKv1 or MmKv2 transfected cells was verified by blocking their currents with TEA (Sigma, USA), XE-991 (Tocris Bioscience, USA), or scorpion toxin ChTX (Alomone, Jerusalem, Israel) (Supplementary Fig. S20). As a comparison, the mouse K^+ channel mKv1.3 was transiently transfected into HEK293 cells and its pharmacological activity was detected by the addition of *M. martensii* venom or ChTX.

Supplementary Note 5. Genetic basis for photosensor function in the tail

Genes involved in phototransduction in the genome

Sequence similarity searches against predicted proteins of the *M. martensii* genome were performed by blastp using 25 *Drosophila* genes described to be involved in visual signal transduction (http://www.flybase.org) as "bait." E-values less than 10^{-4} and scores greater than 200 were considered significant hits. Each hit gene from *M. martensii* was manually checked against the NCBI GenBank database (Supplementary Table S17). Homologs of *Drosophila* light signal transduction gene were found in the genome of *M. martensii.* Particularly, three opsin-homologous genes (Mmopsin1, Mmopsin2, and Mmopsin3) were identified according to the NCBI domain database (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

Expression of phototransduction genes in scorpion tails

Although scorpions have median and lateral eyes [73,74], their eyes may not adequately function in the generation of visual images. The limited functions of scorpion eyes

may be complemented by light sensitivity in the tail Anatomically, a ventral nerve cord runs from the central nervous system in the brain to the venom apparatus in the tail (Supplementary Fig. S23). Subsequently, we performed tblastn searches of the scorpion tail transcriptome using 25 *Drosophila* genes involved in visual signal transduction (http://www.flybase.org). E-values less than 1E-4 and scores greater than 200 were taken to represent significant hits of similarity. Most of the visual signal transduction pathway genes were found to be expressed in the scorpion tails (Supplementary Table S18).

Quantitative expression analysis of phototransduction pathway genes

Total RNA samples for quantitative PCR (qPCR) verification were prepared by dissecting three separate tissues, that is, the prosoma (head with eyes), vesicle (tail or telson), and muscle, as described in S2.1.2. First-strand cDNA was synthesized from 1 μg total RNA using random hexamers (AP Biotech) and SuperScriptII reverse transcriptase according to the manufacturer's protocol (Invitrogen). The qPCR primers were designed for 14 important phototransduction genes, according to their sequences obtained from genome and transcriptome analyses (Supplementary Table S19), and were synthesized at the Tsingke Biotechnology Limited Company (Wuhan, China). The qPCR assays were performed using SYBR ReadyMix (Sigma) and an ABI 7500HT real-time PCR system. Combined with the tail transcriptome analysis of light signal transduction pathway genes, the qPCR results confirmed that the scorpion tail tissues express all checked genes involved in the phototransduction pathway (Supplementary Fig. S22).

Evolution and light-wavelength bias of opsins

Opsin proteins are essential molecules in mediating the light signal transduction, which inputs a light signal to initiate phototransduction[75,76]. Based on the annotated genome and tail transcriptome from *M martensii*, three opsin homologs (Mmopsin1, Mmopsin2, and Mmopsin3) were found in *M martensii*. Mmopsin1 and Mmopsin2 have high homology to eye-related rhabdopsins and were only expressed in the prosoma of scorpion, whereas Mmopsin3, with low similarity to eye-related rhabdopsins, was highly expressed in both scorpion eyes and tails. To further study Mmopsin3, we retrieved from GenBank the sequences of two melanopsins from *Drosophila* (DroMel_rh5, and DroMel_rh7), two ospins from bee (Apis_opsinblue, and Apis_lwave), one opsin from *Hydra* (Hydra_opsin1), one opsin from *Pediculus* (Pediculus_rhl1), one opsin from *Nematostella* (Nematostella_opsin1), and one opsin from *Branchiostoma* (Branchiostoma_opsin4). Multiple sequence alignments of the opsins were carried out using ClustalX 1.83 and adjusted manually, and the result showed that Mmopsin3 shared seven conserved transmembrane domains with the other opsin proteins (Supplementary Fig. S21).

To understand the evolutionary relationship of the three opsins from scorpion, we performed a phylogenetic analysis using opsin proteins from inferior invertebrates and insects, which were downloaded from the NCBI (http://www.ncbi.nlm.nih.gov) and UCSC Genome Browser Wiki sites

(http://genomewiki.ucsc.edu/index.php/Main_Page, Supplementary Table S20). One hundred and four opsin protein sequences were used to construct the phylogenetic tree following the methods described in S2.1.2.

In order to study the spectrum bias of the three opsins from the scorpion *M. martensii*, we extracted sequences for 47 opsins with known light-wave length of sensitivity from the UCSC Genome Browser Wiki site (http://genomewiki.ucsc.edu/index.php/Main_Page). A phylogeny tree was then constructed for them. The light-wave length bias analysis indicated that Mmopsin3 is sensitive to short waves (UV to blue), but both Mmopsin1 and Mmopsin2 are the members of the long-wave-sensitive opsin family. The spectrum sensitivity of Mmopsin3 agrees well with previously documented scorpion behavioral results under different wave-lengths. It was previously shown that scorpions with blocked eyes were more likely to react to wave-length at 395 nm (long UV) than to at 505 nm (cyan-green)[26]. The tails of *Urodacus* (a scorpion species) had the spectral sensitivity in the range from 400 nm to 480 nm[8].

Supplementary Note 6. P450 genes in detoxification, fluorescence and hormone biosynthesis

CYP genes encoding P450 enzymes

Cytochrome P450 genes comprise a large and diverse superfamily found in most classes of organisms. After the *M. martensii* genome was annotated as described in Supplementary Note 1, we searched for the homologs of CYP genes encoding P450 enzymes and found 160 homologues of CYP genes in *M. martensii*. They belong to the CYP 2, 3, 4, and mitochondrial clans, similar to the four clans found in other sequenced protostome genomes [29,77]. All 160 CYP genes were systematically classified according to their P450 characteristics (Supplementary Table S21). The CYP 2, 3, 4, and mitochondrial clans in *M. martensii* had 12, 76, 61, and 11 members, respectively. Thus, the CYP 3 and 4 clans were significantly expanded in *M. martensii* (Supplementary Table S22). The scaffold location analysis of CYP genes indicated that they exist in cluster on the *M. martensii* genome. The clustered CYP genes were summarized in Supplementary Table S23.

Coumarin and its derivatives

The extraction of fluorescent compounds from *M. martensii* was performed according to Forest et al.[10]. Briefly, 100 adult *M. martensii* individuals were immersed in 300 ml 75% ethanol (v:v) solution and kept under room temperature for more than 12 months. The fluorescence from the extract solution was observed under UV light, which indicated that the fluorescent substance was soluble in 75% ethanol.

The chemical standards, coumarin, 7-hydroxy-coumarin, and 4-methyl-7-hydroxy-coumarin were purchased from Sigma (USA) as the reference control. Reverse-phase high-performance liquid chromatography (HPLC) with a C18

column (10 mm × 250 mm, 5 μm, DaLian Elite, China) was performed to collect fractions containing coumarin, 7-hydroxy-coumarin, or 4-methyl-7-hydroxy-coumarin. Buffer A was 0.1% trifluoroacetic acid in water, and buffer B was 80% methanol in water. The flow rate was 5ml/min. Fractions were collected every 0.5 min throughout the gradient elution. The fluorescence of each HPLC fraction was measured using a fluorimeter at 350 nm. The HPLC fractions were lyophilized and resolved in methanol.

Analysis of coumarin, 7-hydroxy-coumarin, and 4-methyl-7-hydroxy-coumarin was performed on the HPLC-electrospray ionization-tandem mass spectrometry (ESI-MS/MS) system consisting of an AB 3200 QTRAP liquid chromatography (LC)-MS/MS (Applied Biosystems) with an ESI source (Turbo Ionspray) and a Shimadzu LC-20 AD HPLC (Tokyo, Japan) system. Data acquisition and processing were performed using Analyst 1.5 software (Applied Biosystems). The HPLC separation was performed on a Shim-pack VP-ODS (150 mm × 2.0 mm i.d., 5 mm) column from Shimadzu. A mixture of 0.1% formic acid and 70% methanol in water (v/v) was used as the mobile phase with isocratic elution. The flow rate was set to 0.2 ml/min. The analysis was performed in multiple reaction monitoring (MRM) mode by monitoring a transition pair of *m/z* 147.2 (molecular ion)/103.2 (fragment ion) for coumarin, *m/z* 163.0/107.1 for 7-hydroxy-coumarin, and *m/z* 177.1/105.1 for 4-methyl-7-hydroxy-coumarin, and a scan time of 200 ms for each pair. The optimized ESI conditions were as follows: curtain gas flow, 40 psi; collision activated dissociation gas flow, medium; Turbo Ionspray voltage, + 5,500 V; turboprobe temperature, 450°C; nebulizer gas flow, 50 psi; heater gas flow, 65 psi; declustering potential voltage, 20 V; entrance potential voltage, 4 V; collision energy, 20 V; and collision cell exit potential, 5 V.

To address the question that 4-methyl-7-hydroxy-coumarin may degrade into 7-hydroxy-coumarin or coumarin, we analyzed the ethanol solution of the chemical standard for 4-methyl-7-hydroxy-coumarin which was prepared in May 2012. As shown in the figure (Supplementary Fig. S28), we detected neither 7-hydroxy-coumarin nor coumarin in the ethanol solution of the 7-hydroxy-4-methylcoumarin standard (stored for more than one year) on September 5, 2013. The analysis was performed by LC-ESI-MS/MS in multiple reaction monitoring (MRM) mode by monitoring a transition pair of *m/z* 147.2/103.2 for courmain, *m/z* 163.0/107.1 for 7-hydroxycourmain and *m/z* 177.1/105.1 for 4-methyl-7-hydroxy-coumarin. These data support the presence of all three compounds in the cuticles.

CYP genes involved in juvenile and molting hormone biosynthesis

In most insect species, hormones regulate critical physiological processes such as metamorphosis and reproduction, in which ecdysone initiates molting or ecdysis, and juvenile hormone controls major developmental transitions from egg to larva, larva to pupa, and pupa to adult. There are very few reports of the hormonal regulation of development and reproduction in arachnids[78].

We performed KEGG mapping using the sequence data from the *M. martensii* genome. The results of KEGG mapping showed that *M. martensii* has key

biosynthesis-related genes for both juvenile and molting hormone biosynthesis: Jhamt, Nvd, CYP307A1/2 (Spo/Spok), CYP306A1 (Phm), CYP302A1 (Dib), CYP315A1 (Sad), CYP314A1, and CYP18A1 (Supplementary Table S24 and Supplementary Fig. S29). We further identified CYP306A1 and CYP18A1 manually by performing blastp searches of the predicted proteins from *M. martensii* using the *Drosophila* CYP306A1 and CYP18A1 genes http://www.flybase.org). For the biosynthesis pathway of juvenile hormone, *M. martensii* had the Jhamt homolog but not the ortholog of CYP15A1, which was consistent with that of the spider mite *T. urticae*, the only other sequenced chelicerate. The result suggests that *M. martensii* has only methyl farnesoate (MF), but not juvenile hormone III. On the other hand, *M. martensii* has the genes of complete biosynthetic-metabolic pathway for molting hormone ecdysone 20E: Nvd, CYP307A1/2, CYP306A1, CYP302A1, and CYP315A1, whereas CYP314A1 and CYP18A1 function to degrade the molting hormone (Supplementary Fig. S29). These results shed light on the biosynthesis and metabolism of scorpions' hormones.

Supplementary References

57 Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution* **17**, 368-376 (1981).

58 Qi, J. X., Zhu, M. S. & Lourenço, W. R. Redescription of *Mesobuthus martensii martensii* (Karsch, 1879) (scorpiones: buthidae) from China. *Revista Ibérica de Aracnología* **10**, 137-144 (2004).

59 Li, W. S. The meiosis and karyotype of male Mesobuthus martensii. *Sichuan Journal of Zoology* **10**, 18-19 (1991).

60 Vindelov, L. L., Christensen, I. J. & Nissen, N. I. Standardization of high-resolution flow cytometric DNA analysis by the simultaneous use of chicken and trout red blood cells as internal reference standards. *Cytometry* **3**, 328-331 (1983).

61 Arumuganathan, K. & Earle, E. D. Nuclear DNA content of some important plant species *Plant Molecular Biology Reporter* **9**, 208-218 (1991).

62 Stanke, M., Schoffmann, 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 (2006).

63 Blanco, E., Parra, G. & Guigó, R. *Using geneid to Identify Genes*. (John Wiley & Sons Inc., 2002).

64 Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic acids research* **32**, D258-261 (2004).

65 Genome sequence of the pea aphid *Acyrthosiphon pisum*. *PLoS biology* **8**, e1000313 (2010).

66 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)* **25**, 1105-1111 (2009).

67 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578 (2012).

68 Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* **Chapter 2**, Unit 2 3 (2002).

69 Ehret-Sabatier, L. *et al.* Characterization of novel cysteine-rich antimicrobial peptides from scorpion blood. *The Journal of biological chemistry* **271**, 29537-29544 (1996).

70 Rodriguez de la Vega, R. C. *et al.* Antimicrobial peptide induction in the haemolymph of the Mexican scorpion Centruroides limpidus limpidus in response to septic injury. *Cell Mol Life Sci* **61**, 1507-1519 (2004).

71 D'Suze, G., Schwartz, E. F., Garcia-Gomez, B. I., Sevcik, C. & Possani, L. D. Molecular cloning and nucleotide sequence analysis of genes from a cDNA library of the scorpion Tityus discrepans. *Biochimie* **91**, 1010-1019 (2009).

72 Glazko, G. V. & Mushegian, A. R. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. *Genome Biol* **5**, R32 (2004).
73 Hjelle, J. T. *Anatomy and morphology*. (Standford University Press, 1990).
74 Root, T. M. *Neurobiology*. (Standford University Press, 1990).
75 Porter, M. L. *et al.* Shedding new light on opsin evolution. *Proc Biol Sci* **279**, 3-14 (2011).
76 Fain, G. L., Hardie, R. & Laughlin, S. B. Phototransduction and the evolution of photoreceptors. *Curr Biol* **20**, 114-124 (2010).
77 Nelson, D. R. Cytochrome P450 and the individuality of species. *Arch Biochem Biophys* **369**, 1-10 (1999).
78 Rees, H. H. Hormonal control of tick development and reproduction. *Parasitology* **129 Suppl**, S127-143 (2004).