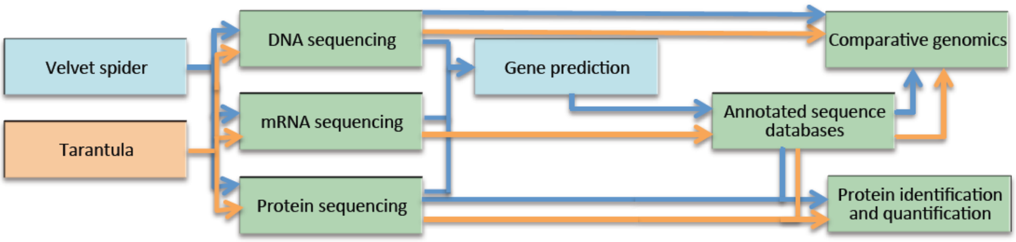
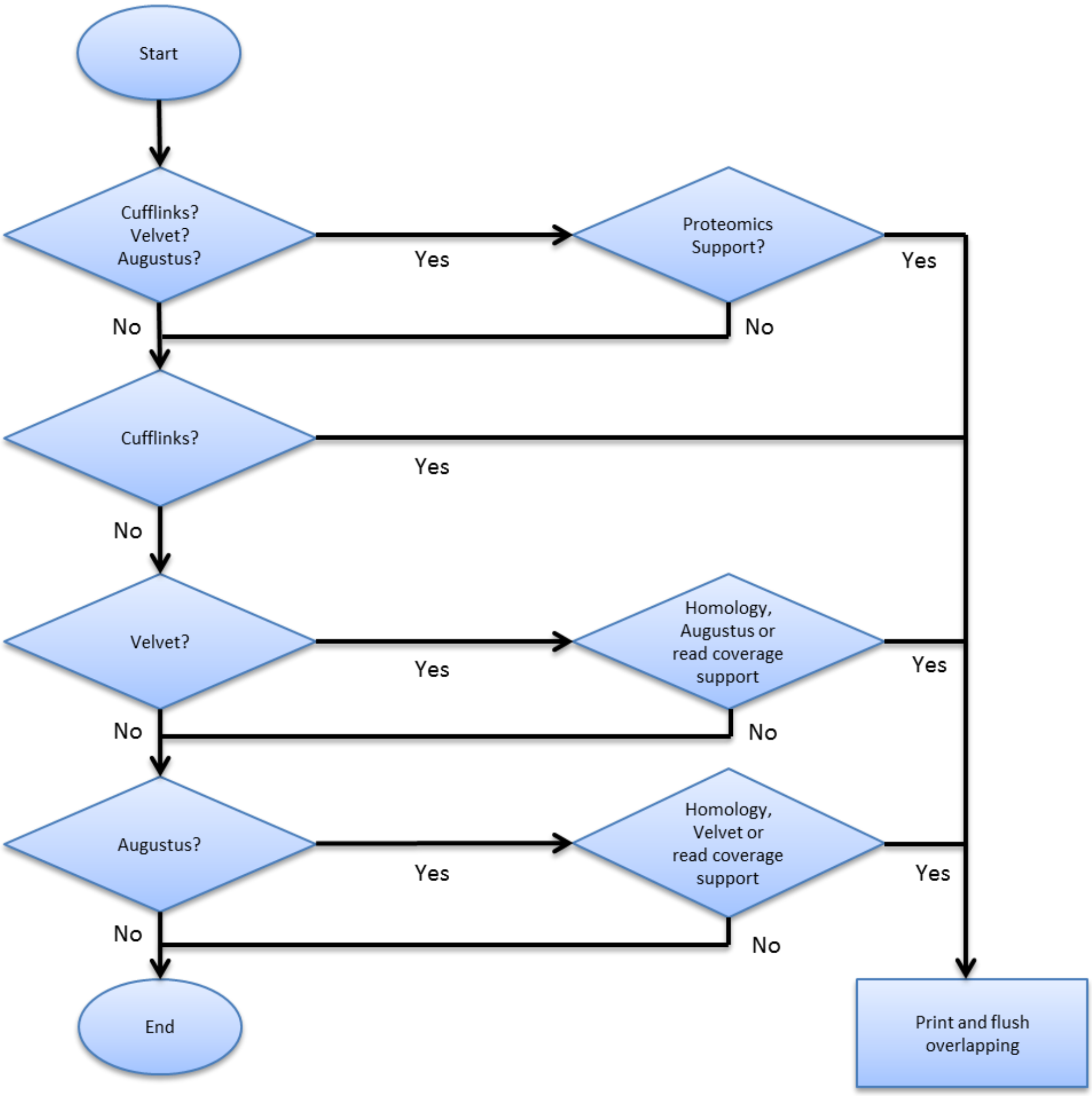
Supplementary Information

Supplementary figures

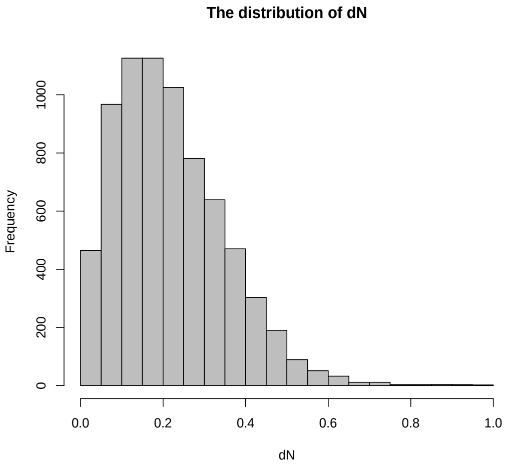


Supplementary Figure 1. Workflow. A schematic overview of the workflow used in this study. The genomes of velvet spider and tarantula were sequenced as well as the transcriptomes of body tissue, venom glands and silk glands. In addition, we performed high throughput shotgun proteomic analyses to provide functional support for genome annotation.

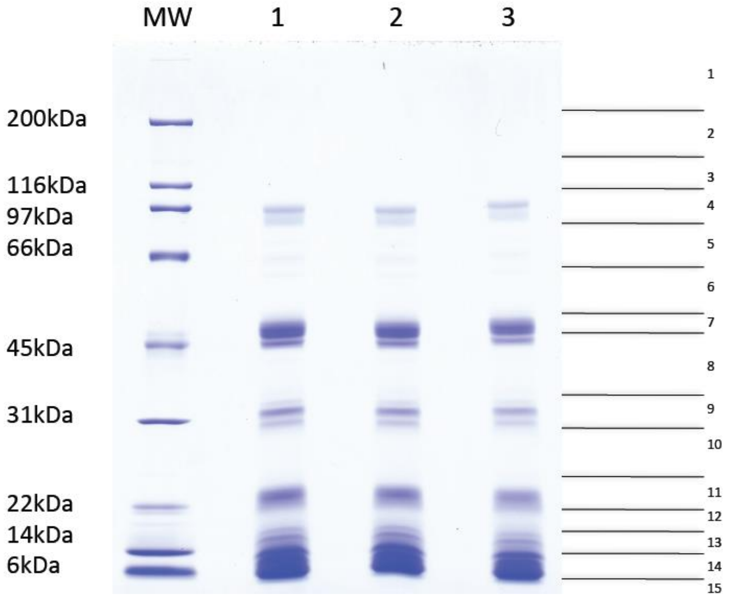


Supplementary Figure 2. Hierarchical gene model selection**.** Gene models were divided into three categories: Cufflinks, Velvet and Augustus. Cufflinks gene models have highest priority followed by Velvet and Augustus. First, all gene models were checked for proteomic support in prioritized order. If proteomics support was found, the gene model was selected and overlapping models with lower priority were filtered out. In the next step, all the Cufflinks gene models were selected, and overlapping models from Velvet and Augustus categories were filtered out. Velvet and Augustus gene models were selected only if they met one of the following criteria:

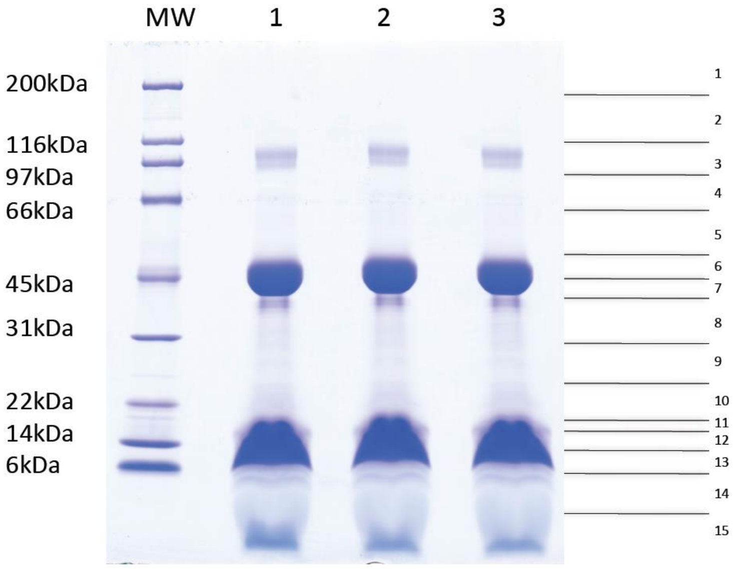
1. has homology to know genes or it contain a known domain.
2. is supported by other evidence.
3. has RNA-seq alignment coverage for at least 80% of the exonic region.



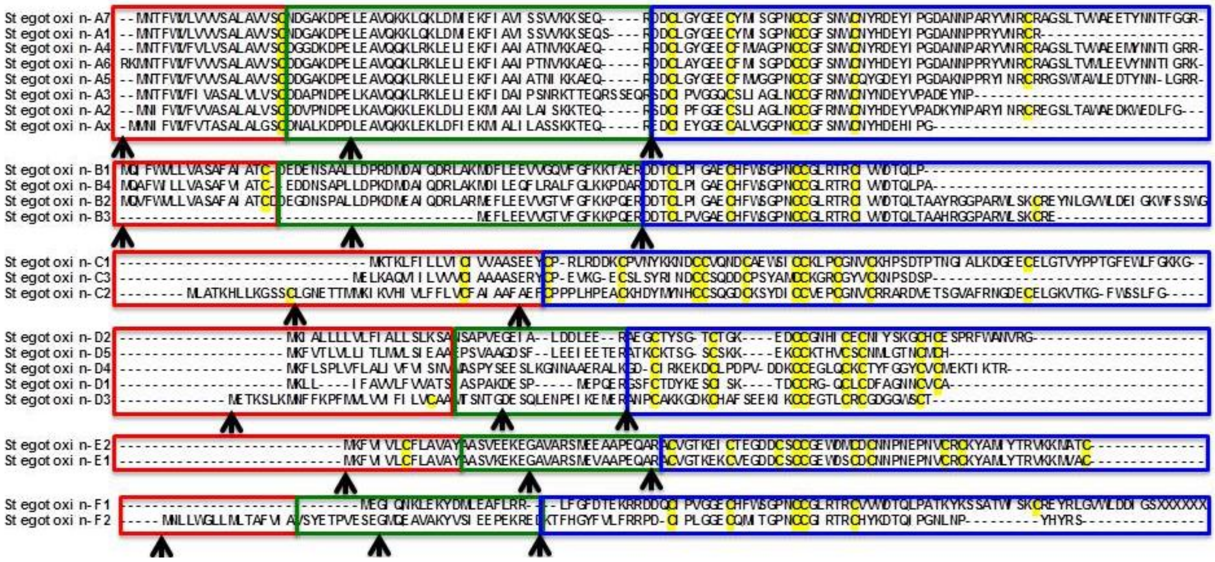
Supplementary Figure 3. The distribution of substitution rate at non-synonymous sites (dN) between velvet spider and tarantula on 8024 orthologous isoforms alignments.dN value between all pairs of orthologs were estimated using the ML method implemented by CODEML.



Supplementary Figure 4. Velvet spider venom. Velvet spider venom from three individuals were subjected to reducing SDS-PAGE (MW: Molecular weight marker). Subsequently the resolved proteins were visualized by Coomassie blue staining. The experiment indicates that the inter-individual variance in venom composition is relatively low. The numbers on the right side of the gel indicates the size and the numbering of the different gel band slices that were excised. The gel slices from all three lanes were digested with trypsin and the resulting peptides micro-purified and finally analyzed by LC-MS/MS.

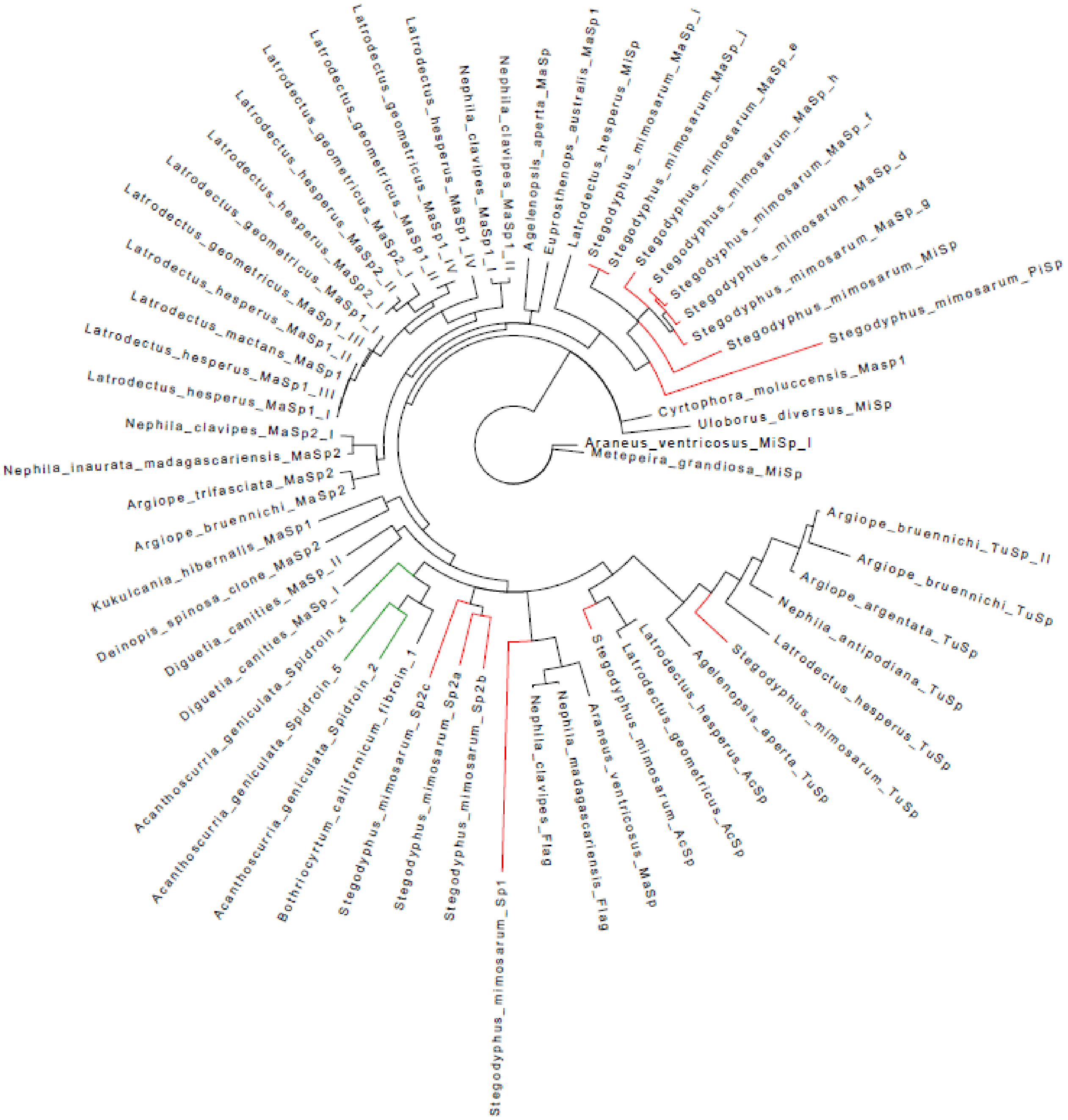


Supplementary Figure 5. Tarantula spider venom. Tarantula venom from three individuals was subjected to reducing SDS-PAGE (MW: Molecular weight marker). Subsequently the resolved proteins were visualized by Coomassie blue staining. The experiment indicates that the inter-individual variance in venom composition is relatively low, and that the venom mainly contains the smaller proteins (app. 5-15 kDa) and one major protein with a molecular weight of app. 45 kDa. The numbers on the right side of the gel indicates the size and the numbering of the different gel band slices that were excised from each lane. The proteins in the 45 (3X15) gel slices were digested with trypsin and the resulting peptides micro-purified and finally analyzed by LC-MS/MS.

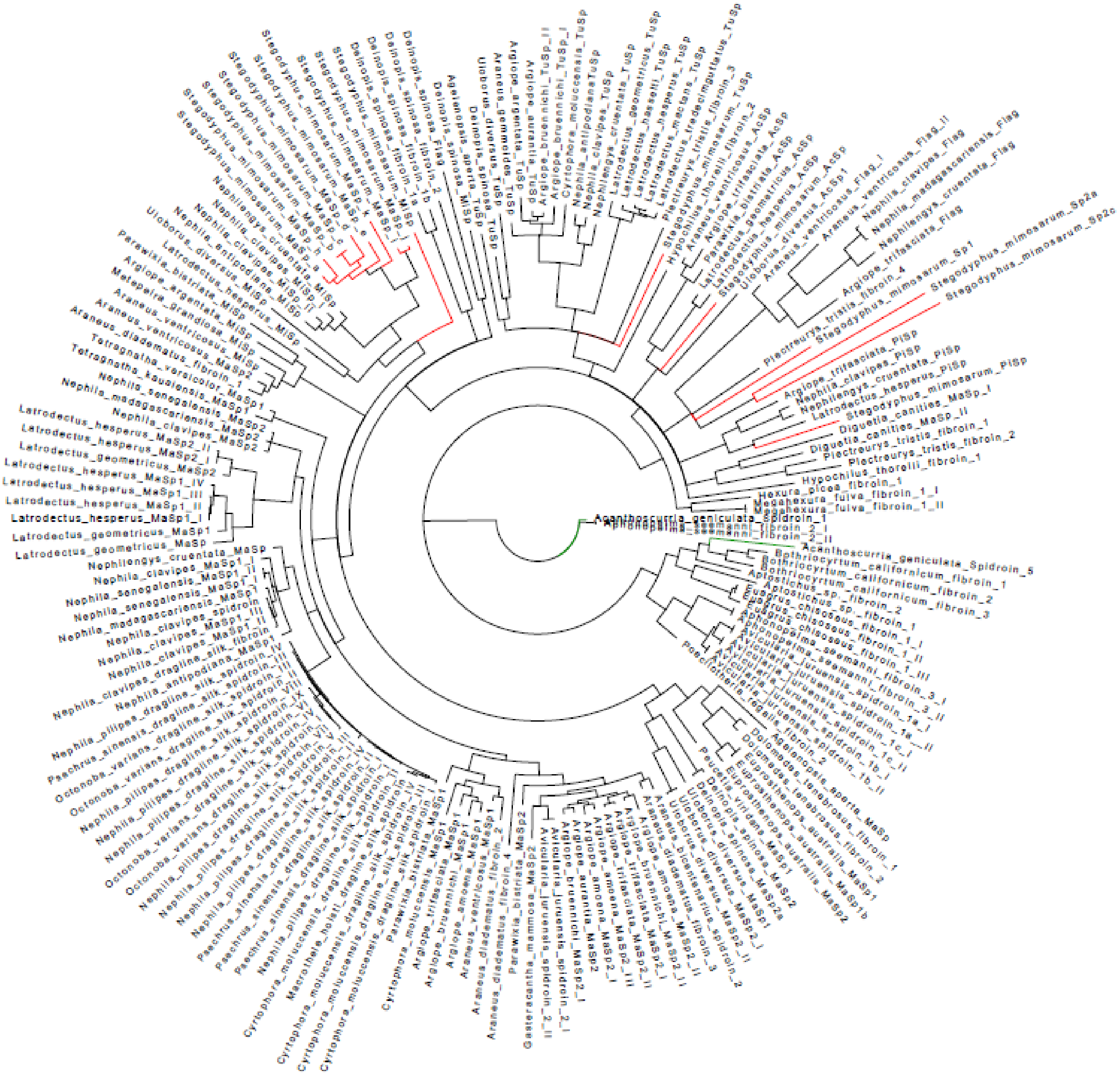


Supplementary Figure 6. Alignment of stegotoxins from the velvet spider. All cysteine-rich protoxins from the velvet spider venom with proteomics support were aligned (except singletons). The red box indicates the signal peptide, the green box indicates the propeptide, and the blue box indicates the mature toxin. The color-coding visualizes that all stegotoxins, except family C, contain a propeptide. The arrows indicate the position of introns at the genomic level and it shows that an intron is positioned between the propeptide and the mature toxin. The cysteine residues are highlighted in yellow, which illustrates the conserved cysteine-rich pattern.

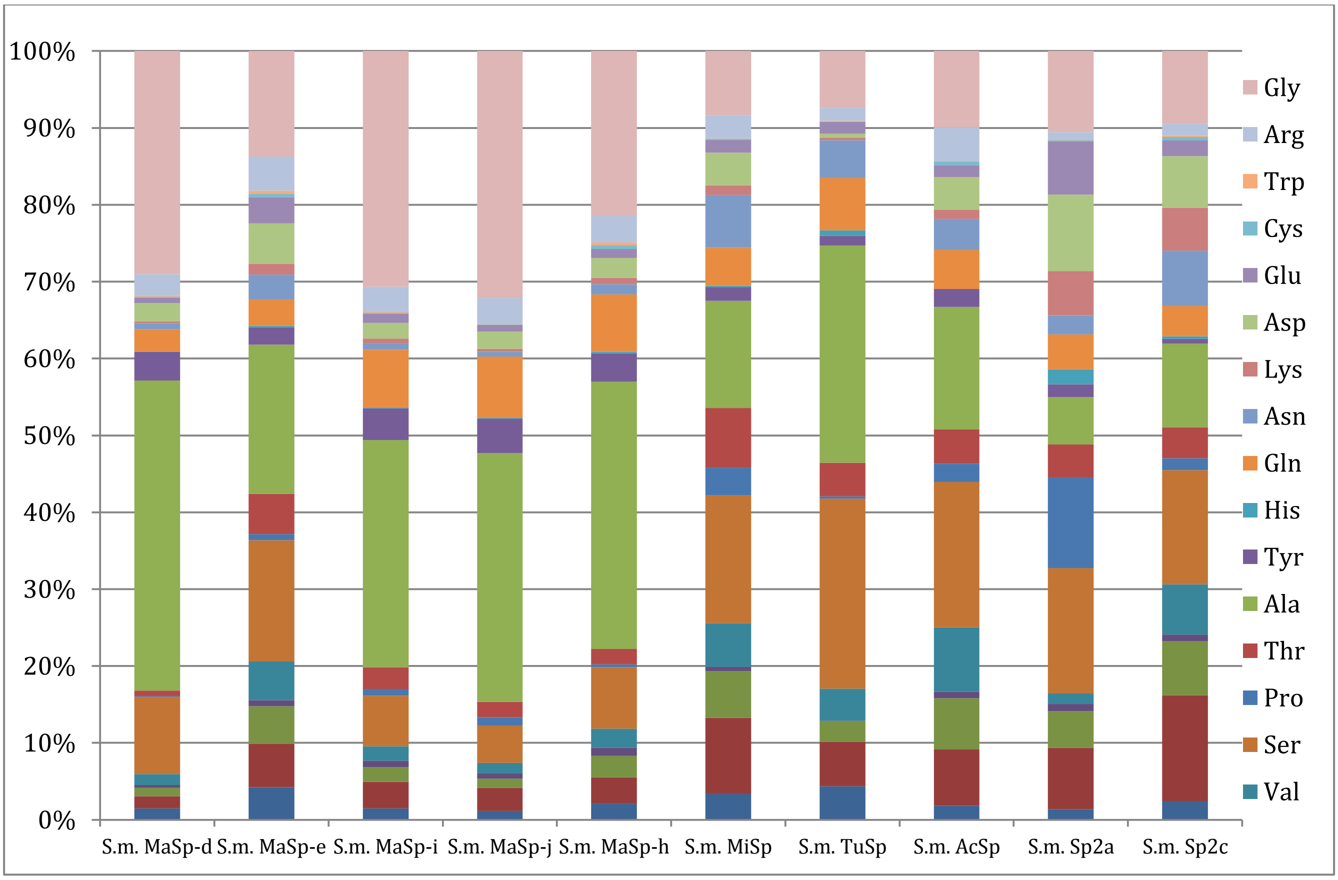
N-terminal domain


Nephila_clavipes_MaSp2_I
Nephila_inaurata_madagascariensis_MaSp2
Argiope_trifasciata_MaSp2
Argiope_bruennichi_MaSp2
Kukulcania_hibernalis_MaSp1
Deinopis_spinosa_clone_MaSp2
Diguetia_canities_MaSp_II
Diguetia_canities_MaSp_I
Acanthoscurria_geniculata_Spidroin_4
Acanthoscurria_geniculata_Spidroin_5
Acanthoscurria_geniculata_Spidroin_2
Bothriocyrtum_californicum_fibroin_1
Stegodyphus_mimosarum_Sp2c
Stegodyphus_mimosarum_Sp2a
Stegodyphus_mimosarum_Sp2b
Stegodyphus_mimosarum_Sp1
Nephila_clavipes_Flag
Nephila_madagascariensis_Flag
Araneus_ventricosus_MaSp
Stegodyphus_mimosarum_AcSp
Latrodectus_geometricus_AcSp
Latrodectus_hesperus_AcSp
Agelenopsis_aperta_TuSp
Stegodyphus_mimosarum_TuSp
Latrodectus_hesperus_TuSp
Nephila_antipodiana_TuSp
Argiope_argentata_TuSp
Argiope_bruennichi_TuSp
Argiope_bruennichi_TuSp_II
Metepeira_grandiosa_MiSp
Araneus_ventricosus_MiSp_I
Uloborus_diversus_MiSp
Cyrtophora_moluccensis_Masp1
Stegodyphus_mimosarum_PiSp
Stegodyphus_mimosarum_MiSp
Stegodyphus_mimosarum_MaSp_g
Stegodyphus_mimosarum_MaSp_d
Stegodyphus_mimosarum_MaSp_f
Stegodyphus_mimosarum_MaSp_h
Stegodyphus_mimosarum_MaSp_e
Stegodyphus_mimosarum_MaSp_j
Stegodyphus_mimosarum_MaSp_i
Latrodectus_hesperus_MiSp
Euprosthenops_australis_MaSp1
Agelenopsis_aperta_MaSp
Nephila_clavipes_MaSp1_II
Nephila_clavipes_MaSp1_I
Latrodectus_hesperus_MaSp1_IV
Latrodectus_geometricus_MaSp1_II
Latrodectus_geometricus_MaSp2_I
Latrodectus_hesperus_MaSp2_II
Latrodectus_hesperus_MaSp2_I
Latrodectus_geometricus_MaSp1_I
Latrodectus_geometricus_MaSp1_III
Latrodectus_hesperus_MaSp1_II
Latrodectus_mactans_MaSp1
Latrodectus_hesperus_MaSp1_III
Latrodectus_hesperus_MaSp1_I

C-terminal domain


Latrodectus_geometricus_MaSp
Latrodectus_geometricus_MaSp1
Latrodectus_hesperus_MaSp1_I
Latrodectus_hesperus_MaSp1_II
Latrodectus_hesperus_MaSp1_III
Latrodectus_hesperus_MaSp1_IV
Latrodectus_geometricus_MaSp2
Latrodectus_hesperus_MaSp2_I
Latrodectus_hesperus_MaSp2_II
Nephila_clavipes_MaSp2
Nephila_madagascariensis_MaSp2
Nephila_senegalensis_MaSp1
Tetragnatha_kauaiensis_MaSp1
Tetragnatha_versicolor_fibroin_1
Araneus_diadematus_MaSp2
Araneus_ventricosus_MiSp
Metepeira_grandiosa_MiSp
Argiope_argentata_MiSp
Parawixia_bistriata_MiSp
Uloborus_diversus_MiSp
Latrodectus_hesperus_MiSp
Nephila_antipodiana_MiSp
Nephila_clavipes_MiSp_II
Nephila_clavipes_MiSp_I
Nephilengys_cruentata_MiSp
Stegodyphus_mimosarum_MaSp_a
Stegodyphus_mimosarum_MaSp_h
Stegodyphus_mimosarum_MaSp_b
Stegodyphus_mimosarum_MaSp_c
Stegodyphus_mimosarum_MaSp_d
Stegodyphus_mimosarum_MaSp_e
Stegodyphus_mimosarum_MaSp_f
Deinopis_spinosa_fibroin_1a
Deinopis_spinosa_fibroin_1b
Deinopis_spinosa_fibroin_2
Deinopis_spinosa_Flag
Deinopis_spinosa_MiSp
Agelenopsis_aperta_TuSp
Deinopis_spinosa_TuSp
Uloborus_diversus_TuSp
Araneus_gemmoides_TuSp
Argiope_argentata_TuSp
Argiope_aurantia_TuSp
Argiope_bruennichi_TuSp_II
Argiope_bruennichi_TuSp_I
Cyrtophora_moluccensis_TuSp
Nephila_antipodianaTuSp
Nephila_clavipes_TuSp
Nephilengys_cruentata_TuSp
Latrodectus_geometricus_TuSp
Latrodectus_hasselti_TuSp
Latrodectus_hesperus_TuSp
Latrodectus_mactans_TuSp
Latrodectus_tredecimguttatus_TuSp
Plectreurys_tristis_fibroin_3
Stegodyphus_mimosarum_TuSp
Hypochilus_thorelli_fibroin_2
Araneus_ventricosus_AcSp
Argiope_trifasciata_AcSp
Parawixia_bistriata_AcSp
Latrodectus_geometricus_AcSp
Latrodectus_hesperus_AcSp
Stegodyphus_mimosarum_AcSp
Uloborus_diversus_AcSp1
Araneus_ventricosus_AcSp
Araneus_ventricosus_Flag_I
Araneus_ventricosus_Flag_II
Nephila_clavipes_Flag
Nephila_madagascariensis_Flag
Nephilengys_cruentata_Flag
Argiope_trifasciata_Flag
Plectreurys_tristis_fibroin_4
Stegodyphus_mimosarum_Sp1
Stegodyphus_mimosarum_Sp2a
Stegodyphus_mimosarum_Sp2c
Argiope_trifasciata_PiSp
Nephila_clavipes_PiSp
Nephilengys_cruentata_PiSp
Latrodectus_hesperus_PiSp
Stegodyphus_mimosarum_PiSp
Diguetia_canities_MaSp_I
Diguetia_canities_MaSp_II
Plectreurys_tristis_fibroin_1
Plectreurys_tristis_fibroin_2
Hypochilus_thorelli_fibroin_1
Hexura_picea_fibroin_1
Megahexura_fulva_fibroin_1_I
Megahexura_fulva_fibroin_1_II
Acanthoscurria_geniculata_Spidroin_1
Aphonopelma_seemanni_fibroin_2_I
Aphonopelma_seemanni_fibroin_2_II
Acanthoscurria_geniculata_Spidroin_5
Bothriocyrtum_californicum_fibroin_1
Bothriocyrtum_californicum_fibroin_2
Bothriocyrtum_californicum_fibroin_3
Aptostichus_sp._fibroin_2
Aptostichus_sp._fibroin_1
Euagrus_chisoseus_fibroin_1_I
Euagrus_chisoseus_fibroin_1_II
Aphonopelma_seemanni_fibroin_1_III
Aphonopelma_seemanni_fibroin_3_I
Avicularia_juruensis_spidroin_3_II
Avicularia_juruensis_spidroin_1a_I
Avicularia_juruensis_spidroin_1a_II
Avicularia_juruensis_spidroin_1c_I
Avicularia_juruensis_spidroin_1c_II
Avicularia_juruensis_spidroin_1b_I
Avicularia_juruensis_spidroin_1b_II
Poecilotheria_regalis_fibroin_2
Agelenopsis_aperta_MaSp
Dolomedes_tenebrosus_fibroin_1
Dolomedes_tenebrosus_fibroin_2
Euprosthenops_australis_MaSp1
Euprosthenops_australis_MaSp1b
Euprosthenops_australis_MaSp2
Peucetia_viridans_MaSp1
Deinopis_spinosa_MaSp2
Deinopis_spinosa_MaSp2a
Uloborus_diversus_MaSp1
Uloborus_diversus_MaSp2_I
Uloborus_diversus_MaSp2_II
Araneus_bicentenarius_spidroin_2
Araneus_diadematus_fibroin_3
Argiope_amoena_MaSp2_II
Argiope_bruennichi_MaSp2_I
Argiope_trifasciata_MaSp2_II
Argiope_trifasciata_MaSp2_III
Argiope_amoena_MaSp2_I
Argiope_aurantia_MaSp2
Argiope_bruennichi_MaSp2_I
Avicularia_juruensis_spidroin_2_I
Avicularia_juruensis_spidroin_2_II
Gasteracantha_mammosa_MaSp2
Parawixia_bistriata_MaSp2
Araneus_diadematus_fibroin_4
Araneus_diadematus_fibroin_2
Araneus_ventricosus_MaSp1
Argiope_amoena_MaSp1
Argiope_bruennichi_MaSp1
Argiope_trifasciata_MaSp1
Cyrtophora_moluccensis_MaSp1
Parawixia_bistriata_MaSp1
Cyrtophora_moluccensis_dragline_silk_spidroin_I
Cyrtophora_moluccensis_dragline_silk_spidroin_II
Cyrtophora_moluccensis_dragline_silk_spidroin_III
Macrothele_holsti_dragline_silk_spidroin
Cyrtophora_moluccensis_dragline_silk_spidroin_IV
Nephila_pilipes_dragline_silk_spidroin_I
Psechrus_sinensis_dragline_silk_spidroin_II
Psechrus_sinensis_dragline_silk_spidroin_I
Nephila_pilipes_dragline_silk_spidroin_III
Nephila_pilipes_dragline_silk_spidroin_V
Octonoba_varians_dragline_silk_spidroin_VII
Octonoba_varians_dragline_silk_spidroin_IV
Nephila_pilipes_dragline_silk_spidroin_VI
Nephila_pilipes_dragline_silk_spidroin_IX
Octonoba_varians_dragline_silk_spidroin_VIII
Octonoba_varians_dragline_silk_spidroin_II
Psechrus_sinensis_dragline_silk_spidroin_III
Nephila_pilipes_dragline_silk_spidroin_IV
Nephila_clavipes_dragline_silk_fibroin
Nephila_antipodiana_MaSp1
Nephila_clavipes_MaSp1_II
Nephila_clavipes_MaSp1_III
Nephila_clavipes_spidroin
Nephila_madagascariensis_MaSp1
Nephila_senegalensis_MaSp1_I
Nephila_senegalensis_MaSp1_II
Nephila_clavipes_MaSp1_I
Nephilengys_cruentata_MaSp

Supplementary Figure 7. Phylogenies of N- and C-terminal domains of the spidroin sequences obtained in this study plus all N- and C- terminal domain sequences found in Genbank. Velvet spider sequences are in red and tarantula (*Acanthoscurria geniculata)* are in green. Roman numbers indicate that there is more than one sequence with same name downloaded from Genbank.



Supplementary Figure 8. Amino acid composition of full length spidroin sequences from the velvet spider.

S.m. MaSp-putative-a

```
GYGGRGYDGGYGGQGAGAGAASAAAAGAGAGQQGQDQGAAAAAAAAAAAAAAAQSYGARGGYGRGAGAAGAAAAGAGARQQGQDQG
AAAAAAAAAGAAAQGYGARGRYGSGASAASAAAAGAGQQARGYDFLADAAALASAAASAFGAG
```

S.m. MaSp-putative-b

```
NNNNNNNNNAAAAAAAAAAAAAAGRGGYGGRGGEGAGGAAGGAAAGAGRGAGGQGDGGAAAAAAAAAAAAAAGG
```

S.m. MaSp-putative-c

```
repeat 1    1 NMLTNDFSLIKILCFLTGYGGRGYGGGYGGDSVAAAASSAAGAGSGAGGEGRDAG---
repeat 2    1 ----MFFSLIKILCFLTGYGGRGYGGGYGGDSGAAAAASAAGAGRGAGGEGRDAGAAA
repeat 3    1 ----MFFSLIKILCFLTGYGDRGYGGGYGGDFGAAVAASAAGAGRGAGGEGRDAG---
repeat 4    1 -CLLMFFSLIKILCFLTGYGGRGYGGGYGGDSGAAAAASAAGAGRGAGGEGRDAGAAA
repeat 5    1 -CLLMFFSLIKILCFLTGYGGRGYGVGYGGDSGAAAAASAAGAGRGAGDEGRDAGAAA
consensus  1    l mfFSLIKILCFLTGYGgRGYGgGYGGDsgAAaAaSAAGAGrGAGgEGRDAGaaa

repeat 1   56 AAAAAAAAAAAGRGGYGGRGGEGAGGAAAGAAAGAAAGAGRGAGGQGDGG----AAAAAA
repeat 2   55 AAAAAAAAAAAGRGGYGGRGGEGAG----GAAAGAAAGAGRGAGGQGDRGAAAAAAAAAA
repeat 3   52 ---AAAAAAAARRGAYGGRGGEGAG----GAAAGAAAGAGRGAG----------------
repeat 4   58 AAAAAAAAAAAGRGGYGGRGGEGAG----GAAAGAAAGAGRGAGGQGDRGAAAAAAAAAA
repeat 5   58 AAAAAAAAAAAGRGGYGGRDGEGAG----CAAAGAAAGAGRGAGGQGDGG----AAAAAA
consensus  61 aaaAAAAAAAAgRGgYGGRgGEGAG    gAAAGAAAGAGRGAGgqgd g    aaaaaa

repeat 1  112 GGNIKGRSP
repeat 2  111 AAAAGG---
repeat 3      ---------
repeat 4  114 AAAAGGM--
repeat 5  110 AAAAGGM--
consensus 121 aaaagg
```

S.m. MaSp-putative-d

```
repeat  1    1 -----XAGAGSGFGGYGQDSGAAAAAAAAAAAAAAAG--QGGYGGRGEAGAGAAS
repeat  2    1 ---AAAAGAGDGSGGYGGDSG--AAAAAAAAAAAAAASGRGGYGGRGGAGAGAAG
repeat  3    1 AAAAAGAGAGSGFGGYEQDSGAAAAAAAAAAAAAAGG--QGRYGGRGGAGAGAAS
repeat  4    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAASGRGGYGGRGGAGAGAAS
repeat  5    1 -AAAAGAGAGSGFGGYGQDSGAAAAAAAAAAAAAAGG--QGGYGGRGGAGAGAAS
repeat  6    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAASGQGGYGGRGGAGAGAAS
repeat  7    1 -AAAAGAGAGSGFGGYGQDSG-AAAAAAAAAAAAAGG--QGGYGGRGGAGAGAAS
repeat  8    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAS-GRGGYGGRGGAGAGAAS
repeat  9    1 -AAAAGAGAGSGFGGYGQDSG-AAAAAAAAAAAAAGG--QGGYGGRGGAGAGAAS
repeat 10    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAASGQGGYGGRGGAGAGAAS
repeat 11    1 -AAAAGAGAGSGFGGYGQDSG-AAAAAAAAAAAAAGG--QGGYGGRGGAGAGAAS
repeat 12    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAASGQGGYGGRGGAGAGAAS
repeat 13    1 -AAAAGAGAGSGFGGYGQDSG---AAAAAAAAAAGGG--QGGYGGRGGAGAGAAS
repeat 14    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAS-GRGGYGGRGGAGAGAAS
repeat 15    1 -AAAAGAGAGSGFGGYGQDSG-AAAAAAAAAAAAAGG--QGGYGGRGGAGAGAAS
repeat 16    1 ---AAAAGAGAGSGGYGRDSGAAAAAAAAAAAAAAAASGQGGYGGRGGAGAGAAS
repeat 17    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAASGRGGYRGRGGAGAGAAS
repeat 18    1 ---AAAAGAGAGSGFGGYGQ-AGAAAAAAAAAAAAGR--QGGYGGRGGSGAGAAS
repeat 19    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAASGQGGYGGRGGAGAGAAS
repeat 20    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAASGRGGYGGRGGAGAGAAS
repeat 21    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAAGSGRGGYGGRGGAGAGAAA
repeat 22    1 ---TAGAGAGSGFGGYGQDSGAAAAAAAAAAAAAAAGGQGGYGGRGGAGAGAAS
repeat 23    1 ---AAAAGAGAGSGGYGGDSG--AAAAAAAAAAAAGG--RVGYGGSGGYGSG---
consensus    1    aaaAGAGaGsGgyggdsg  aAAAAAAAAAAaag gqggYgGrGgaGaGaas
```

S.m. MaSp-putative-f

```
repeat 1   1 GYGGRGYGGGYGAGGAGAAASAAAAGFGAQQQRQDQGAAAASAAAAAAAGQGYGGRG
repeat 2   1 GYGGRGYGGGYGGQGSGAGAASAAAAGAGAGQQGQDQGAAAAAAAAAAAAAQGYAARG
consensus  1 GYGGRGYGGGYGa G GAaAASAAAAG GA QQ QDQGAAAA AAAAAAAgQGYggRG

repeat 1  59 GYEQGSSAAGAAAAGAGSGAGIQRQDQGAAAAAAAAAGGAAGG-
repeat 2  59 GYGRGAGAAGAAAAAAGAGAGQQGQDQGAAAAAAAAAAAAAAGG
consensus 61 GY  G  AAGAAAAgAG GAG Q QDQGAAAAAAAAAggAAgGg
```

MaSp-putative-g

```
GGRGYGGGYEGDSGAAASAAGAGRGAGGEGRDAGAAAAAAAAAAAAAAAGRGGYGGRGGEGAGGAAASAAAGARRGAGGQGDGGAA
AAAAAAAAAAAAG
```

S.m. MaSp-putative-h

```
repeat 1   1 RITSLLFIGYGGRGYGGGYGVGAAGAAAASAAGAGAGAGQQRQDQAAAAAATAAAAAAAG
repeat 2   1 KITCLFLLGYGGRGYGGGYGAGGAGAAAASAAGAGAGAGQQRQDQAAAAAAAAAAAAAAG
repeat 3   1 KITCLFLIGYGGRGYGGGYGAGGAGAAAASAAGAGAGAGQQRQDQAAAAAAAAAAAAAAG
repeat 4   1 KITCPYLIGYGGRGYGGGYGAGGAGAAAASAAGAGAGAGQQRQDQAAAAAAAAAAAAAAG
consensus  1 kITclfliGYGGRGYGGGYGaGgAGAAAASAAGAGAGAGQQRQDQAAAAAAaAAAAAAAG


repeat 1  61 QGYGGRGGYGGGAGAAGAAAAAGAGSEAGKQRQDQAAAAAA--------GQGYGGRGGYG
repeat 2  61 QGYGGRGGYGG-------------------------------------------------
repeat 3  61 QGYGSRGGYGRDAGAAGAAAAAGAGSGAGQQRQDQAAAAAAAAAAVAAAGQGYGGRGGSG
repeat 4  61 QGYGSRGGYGR-------------------------------------------------
consensus 61 QGYG RGGYG  agaagaaaaagags ag qrqdqaaaaaa        gqgyggrgg g

repeat 1 113 GGAGAAGAAAAAGAGSGAGQQGQDQGAAAAAAAAAAAAAAAQGYGGRGGYGRGAGAAG-A
repeat 2  72 -GAGAAGAAAAAGAGSGAGQQGQDQGAAAAAAAAAAAAAAAQGYGGRGGYGRGAGAAG-A
repeat 3 121 GGAGAAGAAAAAGAGSGAGQQGQDQG--AAAAAAAAAAAAAQGYGGRGGYGRGAGAAGAA
repeat 4  72 -GAGAAGAAAAAGAGSGAGQQGQDQGAAAAAAAAAAAVSAAQGYGSRSGFES-------S
consensus121 gGAGAAGAAAAAGAGSGAGQQGQDQGaaAAAAAAAAAaaAAQGYGgRgGygrgagaag a

repeat 1 172 AAAAGAGSGAGQQDQGAQAAAAAAAAAAAAAAGGRMVFIYNFDMW
repeat 2 130 AAAAGAGSGAGQQDQGAQAAAAAAAAAAAAAAGGRMVFIYNFDMW
repeat 3 179 AAAAGAGSGAGQQDQGAEAAAAAAAAAAAAAAGGRIRVHLQF---
repeat 4 124 AAASSAG--------------------------------------
consensus181 AAAagAGsgagqqdqgaqaaaaaaaaaaaaaaggrmvfiynfdmw
```

S.m. MaSp-putative-i

```
repeat 1   1 DQGLRGYGQGAGAGAGAAAAAGAGGRGGFGQGQQGYGQGAGAGAAAAAGAGGAGGYDQGV
repeat 2   1 DQGLRGYGQGAGAGAGAAAAAGAGGRGGFGQGQQGYGQGASAGAAAAAGAGGAGGYGQGA
repeat 3   1 DQGLRGYGQGAGASAGAAAAAGAGGRGGFGQGQQGYGQGAGAGAAAAAGAGGAGGYGQGA
repeat 4   1 DLGLRGYGQGAGAGAGAEAVAGAGGGGGFDQGQQGYGQGAGAGAAAAAGAGGAGGYGQSA
repeat 5   1 DQGLRGYGQGAGAGAGAAAAAGAAGQGGFSQGQQGYGQGAGAGAAAAAGARGAGGYGQGA
repeat 6   1 DQGLKGYGQGAGAGAGAAAAAGAGGRGGFGQGQQGYGQGAGAG---AAGAGGAGGYGQGA
repeat 7   1 DQGLRGYGQGAGAGAGAAAAAGAGGRGGFGQGQQGYGQGAGAGAAAAAGAGGTRGYNQGA
consensus  1 DqGLrGYGQGAGAgAGAaAaAGAgGrGGFgQGQQGYGQGAgAGaaaAAGAgGagGYgQga

repeat 1  61 GAGSAAAASAAGLGGLGRGQQGYGQGAGAG--TAAAAGAGGARGPGYG---GGKGAGAAA
repeat 2  61 GAGAAAAAGASGLGGLGSGQQGYGQGAGAAAAAAAAAGAGRAGGPGYGGGQGAGAGAAAG
repeat 3  61 GAGAAAAASASGLGGLGRGQQGYGQGAGAATAAAAAAGAGRTGGLGYGGGQVAGGAAAAA
repeat 4  61 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGARGPGYG---GAGGGAAAA
repeat 5  61 GAGAAAAVSAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGARGPGYGGAEGAGGGAAAA
repeat 6  58 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGARGPGYG---GGQGAGGAA
repeat 7  61 GAGAAAAASAAGAGGYDQGLRGYGQGAGAGGGAAAAAGAGRAGGYDQGVGAGAGAGAAFA
```

```
consensus 61 GAGaAAAasAaGlGGlgrGqqGYGQGAGAgaaaAAAAGAGgarGpgyG g gagggaaaa

repeat 1 116 ASAAGAGGY
repeat 2 121 ASAAGAAGY
repeat 3 121 ARASGAGGS
repeat 4 118 ASVSGAGGS
repeat 5 121 ASASGAGGS
repeat 6 115 ASASGAGGS
repeat 7 121 SSSAATGVD
consensus121 asasgaggs
```

S.m. MaSp-putative-j

```
repeat  1   1 DQGLRGYGQGAGAGAGAAAAAEAGGRGGFGQGQQGYGQGAGAGAAAAAGAGGAGGYGQAA
repeat  2   1 NQGLRGYGQGAGAGAAAAAAAGAGGRGGFGQGQQGYGQGAGAGAAAAAAAGGAGGYDQGA
repeat  3   1 -----GGGQGAGAGAAAAAAAGAGGRGGFGQGQQGYGQGAGAGAAAAAG---AGGYDQGA
repeat  4   1 YQGLRGYGQGAGAG---AAAAGAGERGGFGQGQQGYGQGAGAGAAAAA---AAGGYEQGA
repeat  5   1 DQGLRGYGQ------------GAGGRGGFGQGQQGYGQGA----AAAAGAGGAGGYDQGA
repeat  6   1 DQGLRGYGQGAGAG---AAAAGAGGRGGFGQGQQGYGQGAGAGAAAAAA---AGGYDQGA
repeat  7   1 ---------------------GAGGRVGFGQGQQGYGQGAGAGPAAAAGAGGAGGYDQGA
repeat  8   1 DQGLRGYGQGAGAGAAAAAAAGAGGRGGFGQGQQGYGQGAGAGAAAAAGTGGAGGYDQGA
repeat  9   1 DQGLRGYGQGAGAGASAAAAAGAGGRGGFGQGQQGYGQGAGAGAAAAAGAGGAGVNDQVA
repeat 10   1 DQGLRGYGQGAGAGAGAAAAAGAGGRGGSGQGQQGYGQGAGAGDAAAAGAGGAGGYDQGA
repeat 11   1 DQGLRGYGQGAGAGAGAAAAAGAGGRGGSGQGQQGYGQGADAGDAAAAGAGGAGGYDQGA
repeat 12   1 DQGLRGYGQGAGAGASAAAAAGAGRRGSFGQGQQGYGQGAGAGAAAAAGAGGAGGYDQGA
repeat 13   1 DQGLRGYGQAAGAGASAAAAAGAGGRGGFGQGQQRYGQVAGAGAAAAAGAGGAGRYDQGA
consensus   1 dqglrgygqgagaga aaaaagAGgRggfGQGQQgYGQgAgagaAAAAgaggAGgydQgA

repeat  1  61 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGAKGPGYGGGQGAGAAAAAT
repeat  2  61 GAAAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGPKGPGY-------------
repeat  3  53 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGARGPGYGGGQRAGAGAAAT
repeat  4  55 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGSRGPGYGGGQGAGAGAAAA
repeat  5  45 GAGAAAAGRAAGLGGLGRGQPGYGQGAGAGAATAAAAGAGGARGPGYGGGQGAGAGAAAA
repeat  6  55 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGATGPGYGGGQGAGAGAAAA
repeat  7  40 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGSRGPGYGGGKGAGAGAAAA
repeat  8  61 GAGTAAAASAAGLGGLGRDQQGYGQGAGAGAAAAAAAGAGASRGPGYGRGQGAGAGAAAA
repeat  9  61 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGSRGPGYGGGQGAGAGASGY
repeat 10  61 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGSRGPGYGRGQGAGAGAAAA
repeat 11  61 GAGAAAAASAAGLGGLGRGQQGYGQGAGAGAAAAATAGAGGVRGPGYGGGQDAGAGAAAA
repeat 12  61 GAGAAAAGSAAGLGGLGRGQQGYGQGAGAGAAAAAAAGAGGARGPGYGGGEGAGAGAAAG
repeat 13  61 SAGVSAAASAAGAGGYDRGLRGYGEGAGAGGGAAAAAGAGRAGGYDQGVGAGAGAGAAFA
consensus  61 gAgaaAAasAAGlGGlgRgqqGYGqGAGAGaaaAAaAGAGg rGpgygggqgagagaaaa

repeat  1 121 AAGAGGY--
repeat  2     ---------
repeat  3 113 ASAAG---Y
repeat  4 115 ASAAGAAGY
repeat  5 105 ASAAG---Y
repeat  6 115 AGA------
repeat  7 100 ASAAGAAGY
repeat  8 121 ASAAGAAGY
repeat  9     ---------
repeat 10 121 ASAAGASGY
repeat 11 121 ASAAGAAGY
repeat 12 121 GAGY-----
repeat 13 121 SSSAATGVD
consensus 121 asaag   y
```

S.m. MiSp-putative

```
spacer 1    1 AATGGRQNDFDVTANARGYDSSSRISIQSYKPDTSPNDGYNAGASGAGSPIQQYSANQPN
repeat 1    1 -SDIDVISKFTNSITSSLLSSNDFTSTFRTGLPATTAVNLASSLARSFATQMALDETTI
repeat 2    1 ARNIDVISQFTNTITSSLLSSNEFTSIFGSGLPVTTALNLASNLAQSLAAQIGLDETGI
repeat 3    1 -TDIEVISKFTNTVISSLLSSNDFTSIFGSGLPVTTALNLASNLAQSLATQIGLDEAGI
repeat 4    1 -TDIDVISKFTNAITSTLLSSNDFTSLFKSGLPVTAAVNLASNLALSLANQIGLDQAGI
repeat 5    1 -TDIDVISKFINTISSSLLSSNDFTSIFGSGLPVTSALNLASNLAHSLATQIGLDEAGI
repeat 6    1 -TDIDVISKFTNTISSSLLSSNDFTSIFGSELPVATAFNLASNLAQSLATQIGLDEAGI
repeat 7    1 -TDIGVISKFTNTVSSSLLSSNDFTSIFGSGLPVTTAFNLASNLAQSLATQIGLDEAGI
spacer 2    1 GLDSAVNPRVTRPGFITAGSPASGATGAAIFESDLIEQVNVSPSAVSDTDLGNGAINSA
consensus   1 -tdIdVISkFtNtitSsLLSSNdFTSiFgsgLPvttAlNLASnLAqSlAtQigLDeagI

spacer 1   61 DAGAGTRYSQASSDVSYSPVNAQSFSQNVGINAFGSTGNLGFGDRRYGQDSTAVASANAP
repeat 1   59 NTLLPLVSQYVSEISSSADVSAYANAISRAVGDALASTGNVPPVLTASLAPADTQPIANL
repeat 2   60 NYLLSLLRQYISAIDLSADASAYANAVSRAIGNALASAGNLSPALASSLASADTQPIANL
repeat 3  59 NSVLSLLNQYISAIDPSADASTYANALSLAIGNTLASAGSLSPALASSLASVDAQPIANF
repeat 4  59 NSLLSLLSQYISTIDSSADASAYANALSLAIVNTLANAESLSPTLASYLASADTQPFANF
repeat 5  59 NSLLSLLSQYISAIDSSADASAYANALSLAIGSTLASAGSLSPALASSLASADTQRIANF
repeat 6  59 NSLLSLLSQYISAIGSSGDASAYANALSLAIGNTLASAGSLSPALASSLASANAQAIANF
repeat 7  59 NSLLSLLSQYISAIGSSADASAYANALSLAIGNTLAKAGNLSPILTASLASADAQPIANL
spacer 2  60 ATSPSIEVGGGSVSP
consensus 61 NslLsLlsQYiSaIdsSaDaSaYANAlSlAigntLAsagslsPaLassLAsadtQpiANf

spacer 1 121 YYSDYSQQGDNRQIYPMDTDRAAGDATANDTPSLSGDTSFYDVSFATRNRNNGGNDREIL
repeat 1 119 VNSVTSTNVNEQQSNLVRRGGGSTLRNIAAKQVQENRQNLGSVQKVVETRIQPSLSRFPL
repeat 2 120 VNSVTSSTLIAQQPKLIGSGGAS-----------------------TVGTKPEPSLSGFPG
repeat 3 119 VSSVASKTLNAQQPNLVRSGGTSTFRNVPFKQVQRKRGNLGSAQSAVGTKLQPSLSGFPG
repeat 4 119 VSSVTSRTLNAQQPKIVRNGRTSSFNNIPFTQVEGNRGNLGSVQSAVGTGLQPSLSRYPG
repeat 5 119 VSSVTSRTLNTQQPNLVRSGVASTFRNAPLAAVQGNRGNLGSVQIAVGTRFQPSLSGFPG
repeat 6 119 VRSVTSRTLNVQQPILVGSGGVSTFRNVPLTEVQGNRGNLGSVQRAVGTRLQPSLSRFPG
repeat 7 119 VNSVTSSTLIAQQPNLIRRGGIS-------------------------------------
spacer 2
consensus121 VsSVtSrtlnaQQpnlvrsGggStfrnip  qvqgnrgnlgsvq aVgTrlqPSLS fPg


spacer 1 181 SS
repeat 1 179 PGA-SAAANGGSGGAQATDITS-
repeat 2 158 QSGASAAATAAAGGGQGAGTSST
repeat 3 179 QSA-SAAASAAAGGGQGVGTSST
repeat 4 179 QGA-IAAASAAAGGAQVAGTSNS
repeat 5 179 QSA-SAATSTAAGGAQGAGTYGT
repeat 6 179 QDA-TSAASTAAGGAQVAGTSST
repeat 7     -----------------------
spacer 2
consensus181 qsa saAasaaaGGaQgagtsst
```

S.m. AcSp-putative

```
repeat  1   1 -YGTPSAAVTPSGIISEVTNNLASALLRSNVFQRVFNNRVPSSISTRIASELAQSIISKL
repeat  2   1 -YGLPSAVNVPSGVISNVANNLVTALLRSNVFQRAFNSRVPSSVVNRIAVALAQSIASSL
repeat  3   1 DYGALSGGAVPSAVISDVANNLASALLRSNIFQRSFNARISASVANRIAAALAQSIASSF
repeat  4   1 DYGAPSSGAVPSSLISDVANNIASALLRSNIFQRAFNARVSTSVANRIAVALTQSIASTF
repeat  5   1 EYGAPAPGVAPSGVISDVANNLASALLRSNIFQRAFNARVSSSVANRISAALAQTIASSL
repeat  6   1 EYGAPSSGAVPSALISDLANNLASALLRSNVFQRAFNARNSSAVTNRIAAALAQSIVSSL
repeat  7   1 DYGAPSAGAAPS-VVSDVANTLASGLLTSNAFQRAFNSRISSSVANRIAAALAQSVASSM
repeat  8   1 DYGASSTGAAQSAVVSDVANKIASALLRSNLFQRVFNTRISSSVASRIATTLAQTTASSL
repeat  9   1 DYSVPSVPAVPTGIISDVANNIASALLRSNIFQRAFNARVSTSVANRIAAALAQTIASSL
repeat 10   1 DYGAPTSGAVPSGIISDVANNLASALLRSNVFKRAFNVRVSSNVANRIAGALAQSIASSL
repeat 11   1 DYGEPSAAVLPSSIISDVANNLASALLRSNIFQRAFNARISSSVANRIAAALAQSISSSM
repeat 12   1 DYGALTSGAVPSGVISDVANNLASALLRSNVFQRAFNVRVSSSVANRIAGVLAQSIASSL
repeat 13   1 EYGEPSAAVLPSGIISDVGNNLASALLRSNVFQRAFNARISSSVVSRIATALTQSISSSM
repeat 14   1 DYGAPSGVVVPSGIISDVANNLASALLRSNVFQRAFNARISSSIANKVAAALTQTLASSL
repeat 15   1 DFSVASGGALPSDVISDVANNLASALLRSNVFQRSFNPRVSSSVTNRIAAALAQSICTSL
repeat 16   1 DYGAPSSGGVPSGLISDIASNLASALVRSKIFQRAFNARVSSSIANRIASALTQSIASSL
repeat 17   1 DYGAPSGSVVPSGLISEVASKLASALLRSNIFQLAFNARVSSSVASRIAAVLVQSIASSL
repeat 18   1 DYPAPSGAAVPSRVISEVANKLASALLRSSVFQRAFNTRVSSSVANRIASALAQSIASSL
repeat 19   1 DYGAPSGGAVPSGVISDVANNLASALLRSNIFQRAFNGRVSSSVANRIGAALAQSIASTL
repeat 20   1 ------AAAVPSGVISDVVNNLASALLRSNSFQRAFNARVSSSVANRIVVALSQSIASNL
consensus   1 dygapsagavpsgviSdvannlasaLlrSnvFqraFNaRvsssvanriaaaLaQsiassl

repeat  1  60 QLDYTTASKCRNSIIQAVSGIRSGSDTRVYAQAIASVLTSELATTGRLNASNASVIGSSI
repeat  2  60 QLDYGTASKCRNAITQALAGVRSGSDTRAYAVAIASAVSGQLAAVGRLNSSNASSIGSSL
repeat  3  61 QLDYATASKCRNAIMQALSSVRSGSDTRTYATAIATTLASQLAAAGRLNTSNASGIGTTL
repeat  4  61 QLDYGTASKCRNAVMQALSSVRSGSDTRVYALAIASALAAQLAAAGRLNASNASSIGSSL
repeat  5  61 QLDYATAVKCRNAIMQAISGVRSGSDTRAYALAIASALAAQLGNAGRLNASNASGIGSSL
repeat  6  61 QLDYGTASKFRNAITQALSSVRSGSNTRVYALAIASALAAQLAAAGRLNASNASSIGSSL
repeat  7  60 QLDYGTASKCRNAIMQALSSVRSGSDTRVYALTIASSLATQLANAGVLNASNMSSIGSSL
repeat  8  61 QLDYGTASKCRNAIMQALSGVRTGSDTRVYALAIASALAAQLAASGRLNASNASGIGSSV
repeat  9  61 QLDNATAAKCRNAIMQALSGVRSGSDTRIYALAIASALAGQLAAAGRLNASNASGIGSSL
repeat 10  61 QLDFGTASKCRNAITQALSGVRSGSDTRVYALAISSALTAQLAAAGRLNASNASGIGSSL
repeat 11  61 QLDYATASKCRNAIMQALSGVRSGSDTRVYALAIASALVAQLAAAGRLNASNASGIGSSV
repeat 12  61 QLDYGTASKCRNAIMQALSGVRSGSDTRVYALAIASALTTQLAAAGRLNESNASGIGSSL
repeat 13  61 QLDYATASKCRNAIMQALSGVRSGSDTRTYALAIASALVAQLAAAGRLNASNASGIGSSL
repeat 14  61 QLDYSTAAKCRNAIMQALSAVRSGSDTRVYALAIASSLVAQLAAAGRLDASNASGIGSSL
repeat 15  61 QLDYRTASKCRSAIMQALSSVRTGSDTRVYALAIASALAAQLAASGRLNASNASSIGSSL
repeat 16  61 QLDNTTASKCRIAVTQALSSVRSGSDTRAYALSIASALARQLAAVGRLNSSNASSIGSSL
repeat 17  61 QLDYGTASKCRNAIMQALSGVRTGSETRAYALTVASALATQLAGAGRLNASNASDIGSSV
repeat 18  61 QLDYATASKCRNAIMQALSGVRSGSDTSVYALAIASALAGQLAATGRLSASNASGIGSSL
repeat 19  61 QLDYGTAAKCRNAIMQALSGVRSGSDTRVYALAIASAVVAQLAAAGRLNTSNASGIGSSL
repeat 20  55 QLDYGTASKLRNAVVQALSGVRSGSDTRVYAVTIASSLAAQLANAGLLKASNASSIGSSL
consensus  61 QLDygTAsKcRnaimQAlsgvRsGSdTrvYAlaiasalaaqLaaaGrLnaSNaSgIGssl

repeat  1 120 LSGILQGAYSAARQAGLDLSGIDVTSDISSSLSAYSSSSAAPQTVAETQQLTAVISD---
repeat  2 120 LSSVVQGAYSAARQAGIDVSGVDVSSDISSSISAYGTGPAVAFDTAITPQIPESISD---
repeat  3 121 LSGVLQGAYSGARQAGVDVSGVDVSTDISSSVSAYAGGPAAGQVPAMSAQYAEGISD---
repeat  4 121 LSGVVQGAYSGARQAGVDVSGVDVSSDISSSISAYGAGSAAGQDIVAAQQFTEGISD---
repeat  5 121 LSGVVQGAYSGARQAGVDMSGVDVSSDISSSISAYSAGPTAGQVPAVTQQFSEGISG---
repeat  6 121 LSGVVQGAYSGARQAGVDVSGVDVSSDISSSISAYGAGSAAGQDVVAAQQFTEGISD---
repeat  7 120 LSSVVQGAYSGARQAGIDVSGIDVSSDISSSISAYGGSRTGGQETGISTQFPGGISS---
repeat  8 121 ISGVVQGTYSGASQAGVDVSGVDVSSDISSSISAYGRGSAVGQDIAGPQKITESISD---
repeat  9 121 LSGVVQGAYSGARQTGIDVSGVDVSSDISSSISAFAAGSTAGQDVASAQLFTESMAD---
repeat 10 121 LSGVVQGTYSGAKQAGVDVSGVDVSSDISSSVSAYGAGPTGAQESDVSSLLPDGISD---
repeat 11 121 LPGVVQGAYSGARQAGVDVTGVDLSSDISSSISAFGGSSIGGQGIAAAPQFAESISD---
repeat 12 121 LSGVVQGTYSGAKQAGVDVSGVDVSSDISSSVSAYGAGPTGAQESVVSSLLPEGITD---
repeat 13 121 LSGVVQGAYSGARQAGVDVTGVDVSSDISSSISAFGGSSTGGQGIAAAQQFAESISD---
repeat 14 121 LSGVVQGAYSGARQAGVDVSGIDVSTDISSSISTYGAGSPAGQDIAATSQFTAGISD---
repeat 15 121 LSGVVQGAYSGARQAGVDVSGVDVSTDISSSISAYGAGSTAAQDISAAAQFTGGVSD---
repeat 16 121 LSGVVQGAYSGARQAGVDVSGVDVSSDISSSVSAYGAGRTVSSETDVTSLLTEGISD---
repeat 17 121 LSAAVQGAYSGASQAGVDVSGVDVSSDISSSISAYGAGPTGGSETGLTSLLAQGISD---
repeat 18 121 LSGVVQGSYSGARQAGVDLSGVDVSSDISSSLSAYGADSSAGQDIAPSQPFTEGISD---
repeat 19 121 LSGVVQGTYSGAKQAGVDVSGVDVSSDISSSISAYGAGPMG----EVSSLLAGGISD---
repeat 20 115 LSSIVQGSYSGARQAGVDVSGIDIRSDISTSASAYSSSASSIQTSSVSLPLPEGVSQGLS
consensus 121 lsgvvQGaYSgArQaGvDvsGvDvssDISsSiSaygagstagqdvait qftegisd---

repeat  1 177 -VSKDFQGGYEPISKTGP-----
repeat  2 177 -ISQGISGVSEGISGPS------
repeat  3 178 -ISQSILGVPEGITSPG------
repeat  4 178 -ISQGISAITAGVAGPR------
repeat  5 178 -ISQDISALPEGVASPG------
repeat  6 178 -ISQGISEVSQSSPGTG------
repeat  7 177 -ISQGISGASQGIAGPG------
repeat  8 178 -ISQAVSGVSEGIAGLG------
repeat  9 178 -ISQGVSGVSAGFSGPG------
repeat 10 178 -ISQGLSAITGKVTGPT------
repeat 11 178 -ISQGVS---ADISGPS------
repeat 12 178 -ISQGFSAITGKVTGPA------
repeat 13 178 -ISQGVSGVSEAIAGSG------
repeat 14 178 -VLQGVSGVSEVMTGPG------
repeat 15 178 -ISQGVSGVSEGIAGPG------
repeat 16 178 -ISQGVSGISGGISGPG------
repeat 17 178 -VSQKISAISDGVTVPG------
repeat 18 178 -ISQGVSGASEGISGSG------
repeat 19 174 -ISQGLSAVTEGVTGPGADYGAP
repeat 20 175 ETSRGVLGVSEGISESAFDFGGP
consensus 181  isqgisgvsegisgpg------
```

S.m. TuSp-putative

```
repeat  1   1 ---------------------------------SSNNISSRAEDSASAFARSSATSLASS
repeat  2   1 SAFAQSASQAASQAGSRSATTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
repeat  3   1 SAFAQSASQAASQAGSRSTTTTTSISQAASQETSSSSASSTAEASASAFAQSSASSLSSS
repeat  4   1 SAFAQSASQAASQAGSRSTTTTTSISQAASQETSSSSASSTAEASASAFAQSSASSLASS
repeat  5   1 SAFAQSASQAASQAGSRSTTTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
repeat  6   1 SAFAQSASQAASQAGSRSTTTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
repeat  7   1 SAFAQSASQAASQAGSRSTTTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
repeat  8   1 SAFAQSASQAASQAGSRSTTTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
repeat  9   1 SAFAQSASQAASQAGSRSTTTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
repeat 10   1 SAFAQSASQAASQAGSRSTTTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
repeat 11   1 SAFAQSASQAASQAGSRSATTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
repeat 12   1 SAFAQSASQAASQAGSRSATTTTSISQAASQETSSSSASSRAEASASAFAQSSASSLASS
consensus   1 safaqsasqaasqagsrstttttsisqaasqetSSssaSSrAEaSASAFAqSSAsSLaSS

repeat  1  28 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat  2  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat  3  61 SSFARAFSSASSAEAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat  4  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat  5  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat  6  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat  7  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat  8  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat  9  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat 10  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGATANAYA
repeat 11  61 SSFARAFSSASSAAAAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
repeat 12  61 SSFARAFSSASSAETAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGASANAYA
consensus  61 SSFARAFSSASSAaaAGSIAYQGGLLAAQNLGIGNAVGLANALSQAVSSVGVGAsANAYA

repeat  1  88 NAVANTVGHFWAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat  2 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat  3 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat  4 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat  5 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat  6 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat  7 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat  8 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat  9 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat 10 121 NAVANTVGHFLAGQGNLTQGNASRLASAFSNAFASS----AASVAAASSAFSQSAVAAQS
repeat 11 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
repeat 12 121 NAVANTVGHFLAGQGILTQGNASGLASAFSNAFASSAASAAASVAAASSAFSQSAAAAQS
consensus 121 NAVANTVGHFlAGQGiLTQGNASgLASAFSNAFASSaasaAASVAAASSAFSQSAaAAQS

repeat  1 148 AS
repeat  2 181 AS
repeat  3 181 AS
repeat  4 181 AS
repeat  5 181 AS
repeat  6 181 AS
repeat  7 181 AS
repeat  8 181 AS
repeat  9 181 AS
repeat 10 177 AS
repeat 11 181 AS
repeat 12 181 AS
consensus 181 AS
```

S.m. PiSp-putative

```
repeat 1   1 AISSGEISVTDVIYFASQDLAQKYGLSQDFVQSILSQSLSEYGTGSSAEEITQALATASS
repeat 2   1 AISTGQLSVQNVISVASQVLANSFGISQSSAQSILSQALSNFGRGSSAQAVATALASASS
repeat 3   1 AISTGQLSVQNVISVASQVLANSFGISQRSAQSILSQALSNFGTGSSAQAVATALASASS
repeat 4   1 AISTGQLSVQNVISVASQVLANSFGISQSSAQSILSQALSNFGRGSSAQAVATALASASS
repeat 5   1 AISTGQLSVQNVISVASQVLANSFGISQSSAQSILSQALSNFGRGSSAQAVATALASASS
repeat 6   1 AISTGQLSVQNVISVASQVLANSFGISQSSAQSILSQALSNFGRGSSAQAVATALASASS
repeat 7   1 AISTGQLSVQNVISVASQVLANSFGISQSSAQSILSQALSNFGRGSSAQAVATALASASS
repeat ?    1 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
consensus   1 aistgqlsvqnvisvasqvlansfgisqssaqsilsqalsnfgrgssaqavatalasass

repeat 1   61 EILVQSGAVTAGQEESVGQSVGSILSSALQQLLSQISRPAPAPAPRPLPAPRPAPFIAQQ
repeat 2   61 QVLVQTGAVTAGQEQSVGQSFGSILLSALQQLLSQISRPAPAPAPRPLPAPRSAPFIAQQ
repeat 3   61 QVLVQTGAVTAGQEQSVGQSFGSILLSALQQLLSQISRPAPAPAPRPLPAPRPAPFIAQQ
repeat 4   61 QVLVQTGAVTAGQEQSVGQSFGSILLSALQQLLSQISRPAPAPAPRPLPAPRPAPFIAQQ
repeat 5   61 QVLVQTGAVTAGQEQSVGQSFGSILLSALQQLLSQISRPAPAPAPRPLPAPRPAPFIAQQ
repeat 6   61 QVLVQTGAVTAGQEQSVGQSFGSILLSALQQLLSQISRPAPAPAPRPLPAPRPAPFIAQQ
repeat 7   61 QVLVQTGAVAAGQEQSVGQSFGSILLSTLQQLLSQISRPAPAFAPRPLPAPRPAPFIAQQ
repeat ?   61 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNPAPAPAPRPLPAPRPAPFIAQQ
consensus 61 qvlvqtgavtagqeqsvgqsfgsillsalqqllsqisrPAPApAPRPLPAPRpAPFIAQQ

repeat 1 121 TQQAASLSSASSASSSTSTSQAVQTSSASQFTAASSQTSASVSVSSQALQSAIISKIASS
repeat 2 121 TQQAASLSSASSAASSTSTSEAVQTSSASQFTAASSQTSASVSVSSQALQSAIISNIASS
repeat 3 121 TQQAASLSSASSAASSTSTSQAVQTSSASQFTAASSQTSASVSVSSQALQSAIISNIASS
repeat 4 121 TQQAASLSSASSAASSTSTSQAVQTSSASQFTAASSQTSASVSVSSQALQSAIISNIASS
repeat 5 121 TQQAASLSSASSAASSTSTSQAVQTSSASQFTAASSQTSASVSVSSQALQSAIISNIASS
repeat 6 121 TQQAASLSSASSAASSTSTSQAVQTSSASQFTAASSQTSASVSVSSQALQSAIISNIASS
repeat 7 121 TQQNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
repeat ? 121 TQQAASLSSASSAASN--------------------------------------------
consensus121 TQQaaslssassaasststsqavqtssasqftaassqtsasvsvssqalqsaiisniass

repeat 1 181 SALN
repeat 2 181 SALN
repeat 3 181 SALN
repeat 4 181 SALN
repeat 5 181 SALN
repeat 6 181 SALN
repeat 7 181 NNNN
repeat ?     ----
consensus181 saln
```

S.m. Sp2a

```
repeat 1    1 DGIGRKQSPDSDAGATKPPNSEAKAIVPLKALDKLSQLYPDTESEGTGDGSPSGSPKSPT
repeat 2    1 DGIGKKQIPDSDAGSPKSPDSKAKEIAPLLALEMLSQLYPDTESEGTGDGSPSGSPKSPT
repeat 3    1 DGIGKKQSPDSDAGSPKSPDSEAKEIAPLLALEKLSQLYPDTESEGTGDGSPSGFPKSPT
repeat 4    1 DGIGKKQIPDSDAGSPKSPDSKAKEIAPLLALEMLSQLYPDTESEGTGDGSPSGSPKSPT
repeat 5    1 DGIGKKQSPDSDAGSPKSPDSEAKEIAPLLALEKFSQLYPDTESEGTGDGSPSGSPKSPT
repeat 6    1 DGIGKKQSPDSDAGSPKSPDSEAKEIAPLLALEKLSQLYPDTESEGTGDGSPSGFPKSPT
consensus   1 DGIGkKQsPDSDAGspKsPdSeAKeIaPLlALeklSQLYPDTESEGTGDGSPSGsPKSPT

repeat 1   61 GPGAPEHSGSSDGEPTNSDKGGKQPDDASSSPRHGSDGKIPDKDTSALLLVDIDIATLLP
repeat 2   61 GPGDPEHSGSSDGEPTNSDEGGKQPDDASSSPGHRT--MVX---TSALLLEDIDIATLLP
repeat 3   61 GPGAPEHSGYSDGEPTNSDEGGKQPDDASSSPGHGSDGKIPDKDTSALLLVDIDIATLLP
repeat 4   61 GPGDPEHSGSSDGEPTNSDEGGKQPDDASSSPGHRSDGKIPDKDTSALLLVDIDIATLLP
repeat 5   61 GPGAPEHSGSSDGEPTNSDEGGKQPDDASSSPGHGSDGKIPDKDTSALLLVDIDIATLLP
repeat 6   61 GPGAPEHSGYSDGEPTNSDEGGKQPDDASSSPGHGSDGKIPDKDTSALLLVDIDIATLLP
consensus 61 GPGaPEHSGsSDGEPTNSDeGGKQPDDASSSPgHgsdgkipdkdTSALLLvDIDIATLLP

repeat 1 121 SSQPGEGPSDDSLGGSEGPTGPDNASPSQPSSAAPSGELPDSATIQSLYDLLSKLPISLP
repeat 2 116 SSQHGEGPSDDSLGGSESPTGPDNASSSQPSSAAPSGELPDSATIQSLYDLLSKLPIPLP
repeat 3 121 SSQHGEGPSDDSLGGSESPTGPDNASSSQPSSAAPSGELPDSATIQSLYDLLSKLPIPLP
repeat 4 121 SSQPGEGPSDDSLGGSESPTGPENASLFQPSSAAPSGQLPDSATIQSLYDLLSKLPIPLP
repeat 5 121 SSQPGEGPSDDSLGGSESPTGPDNASPSQPSSAAPSGELPDSATIQSLYDLLSKLPIPLP
repeat 6 121 SSQPGEGPSDDSLGGSEGPTGPDNASPSQPSSAAPSGELPDSATIQSLYDLLSKLPISLP
consensus121 SSQpGEGPSDDSLGGSEsPTGPdNASpsQPSSAAPSGeLPDSATIQSLYDLLSKLPIpLP

repeat 1 181 DQGAPSDQNKGPSGQDAGTPESGMSPEDKGPYGGSDGESPESGDQSDTLSKEPELVSLIS
repeat 2 176 DRGAPNDKNRGPSGQDAGTPEGGMSPEDKGPYGGSDGESPESGDQIDTLSKEPELVSLIS
repeat 3 181 DQGAPNDQNRGPSGQDAGTPESGMSPEDKGPYGGSDGESPESGDQIDTLSKEPELVSLIS
repeat 4 181 DQGAPNDQNRGPSGQDAGTPESGMSPEDKAPYGGSDGESPESGDQIDTLSKEPELVSLIS
repeat 5 181 DQGAPNDQNRGPSGQDAGTPESGMSPEDKAPYGGSDGESPESGDQIDTLSKEPELVSLIS
repeat 6 181 DQGAPSDQNKGPSGQDAGTPESGMSPEDKGPYG---------------------------
consensus181 DqGAPnDqNrGPSGQDAGTPEsGMSPEDKgPYGgsdgespesgdqidtlskepelvslis

repeat 1 241 NLLDDIPS
repeat 2 236 NLLDDIPS
repeat 3 241 NLLDDIPS
repeat 4 241 NLLDDIPS
repeat 5 241 NLLDDIPS
repeat 6     --------
consensus241 nllddips
```

S.m. Sp2c

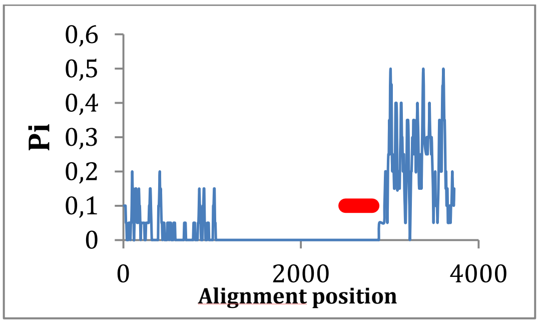
Repeat type 1

```
repeat 1    1 NILQTQGLLNKSLDSIITQTIEGILQGLGQALNINIDIKKALDLAAQVKVDAGVGLNANT
repeat 2    1 NILQSQGLLNVNLNTLLTQATECTLLGLSQALNINIDIKKALNLAGQVKVDIGIGANTDI
consensus   1 NILQtQGLLN  L siiTQ  E  L GL QALNINIDIKKAL LAaQVKVD GvG N

repeat 1   61 GAAVGADIGAALGADVGVGLGSDVGLGLDANVGLDANANNEASANAGISTNLGLGFSPSA
repeat 2   61 SGSLRADAEVGVGADVPVGLGTDVDTRFEADVGIGLDASLGTGANADLNANLGLGLLQSA
consensus  61  a v AD   alGADV VGLGsDV    dA VGl   A     ANA i  NLGLG   SA

repeat 1 121 DVGLGVGFNAPNTGLKLKNLLGLKLKATGVLDILASKKPSKSDIANISKLICRFLANKFQ
repeat 2 121 DKGLGVGLNIPNFGLKLTNLLALKLKATGVFNVLAKKTPSQSDFLNISKLISRLLANKFQ
consensus121 D GLGVG N PN GLKL NLLgLKLKATGV  iLA K PS SD  NISKLI R LANKFQ

repeat 1 181 IQLNASMIKLLYGSLIKLNARARPEDFGNVLAAVI
repeat 2 181 IQLNASLVKLFYGSLIKLNGRAKPEDFANVLAAST
consensus181 IQLNASmiKL YGSLIKLNaRArPEDFgNVLAA
```

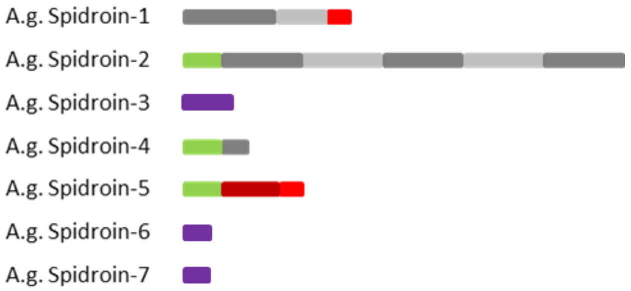
Repeat type 2

```
repeat 1    1 ASVGVDTNLGLDLSPSTGIGQQTGLNVPNLGLKLTNLLGLKLKATGMLNILATKTPSRSH
repeat 2    1 ANADVNTNLGLGLSPSTGMRLGVGLNVPDLGLKLKNLLVLKLKAAGVLNILATKAPSRSD
consensus   1 A   V TNLGL LSPSTGi    GLNVP LGLKL NLL LKLKA GmLNILATK PSRS

repeat 1   61 IVNISKSVSKLLATKFQIQFNGSMIKLFYNSLAKLDATRKPDDFANVLAAVTINILQSQG
repeat 2   61 IVNISKSICRLLANKFQIKLDVSMIKLLYGSLAKFDATAKPDDFANVLAAVTMNILQSQG
consensus  61 IVNISKSv kLLA KFQI    SMIKL Y SLAK DAT KPDDFANVLAAVTiNILQSQG

repeat 1 121 LLNINLDSLLSQATECILLGLREALNINVDIKKALDLAAQVKVDVGADVGLAGNVAIGVG
repeat 2 121 LLNINLDTLLTQATECILLGLGQALNIDIDIKSALDLAAKMKVDAGAGVDVDVGVGLGAD
consensus121 LLNINLDsLLsQATECILLGL  ALNI vDIK ALDLAA vKVD GA V l   VaiG

repeat 1 181 ADAGVDANTNLGIQEGPNIDAS---------
repeat 2 181 IEAGVGLGAKAGIGLDAGVGIDADANLGIQV
consensus181  dAGV      GI     i   adanlgiqv
```

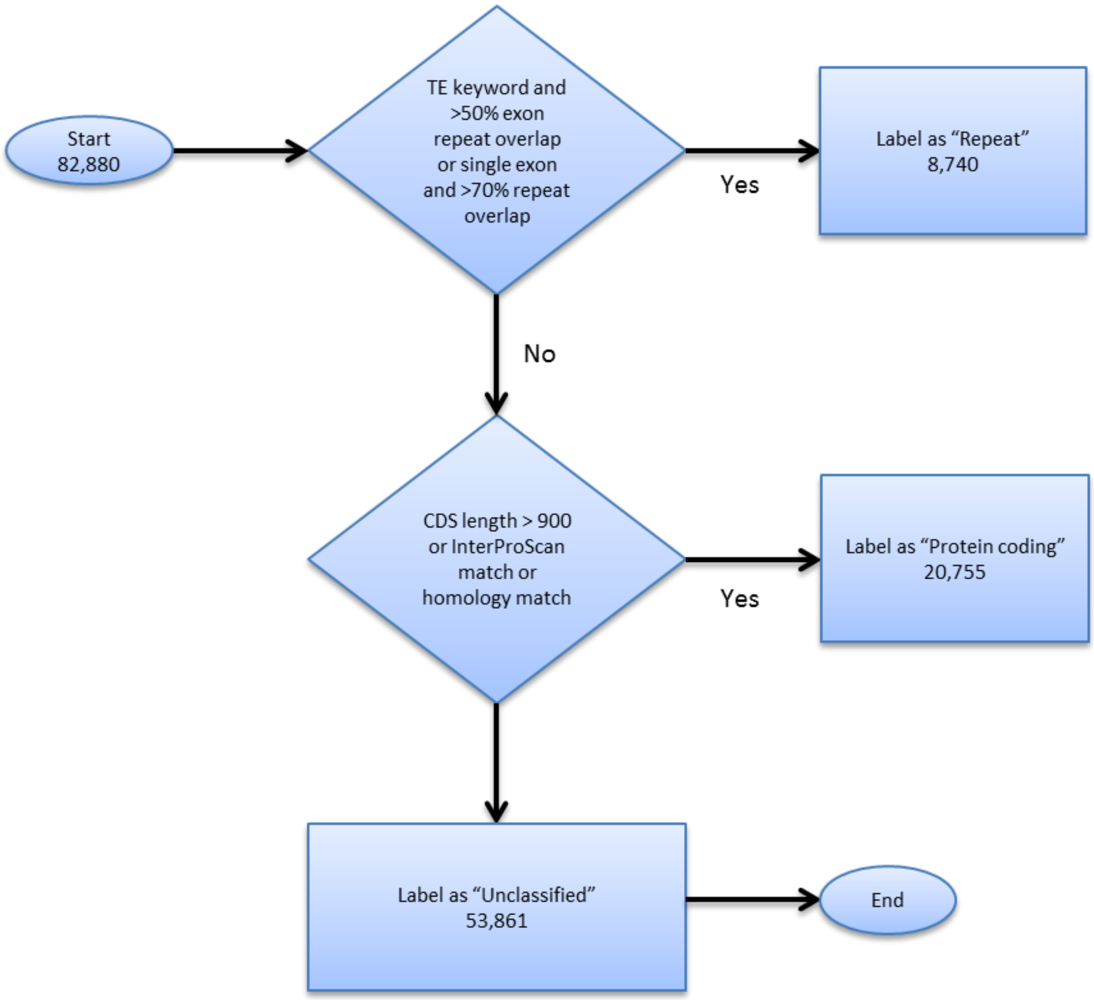
Supplementary Figure 9. Shows alignments of velvet spider silk gene repeats for each gene separately. Below each alignment line is given the consensus sequences. Left of each sequence is given the first amino acid position in the given line. *S. m. MiSp-putative* also includes two spacer sequences that flank the repetitive core region. For *S.m. MaSp-putative-a*, *S.m. MaSp-putative-b* and *S.m. MaSp-putative-g* only one or a partial repeat was sequenced, and therefore no alignment is presented. The figure was produced using BoxShade (http://www.ch.embnet.org/software/BOX_form.html).



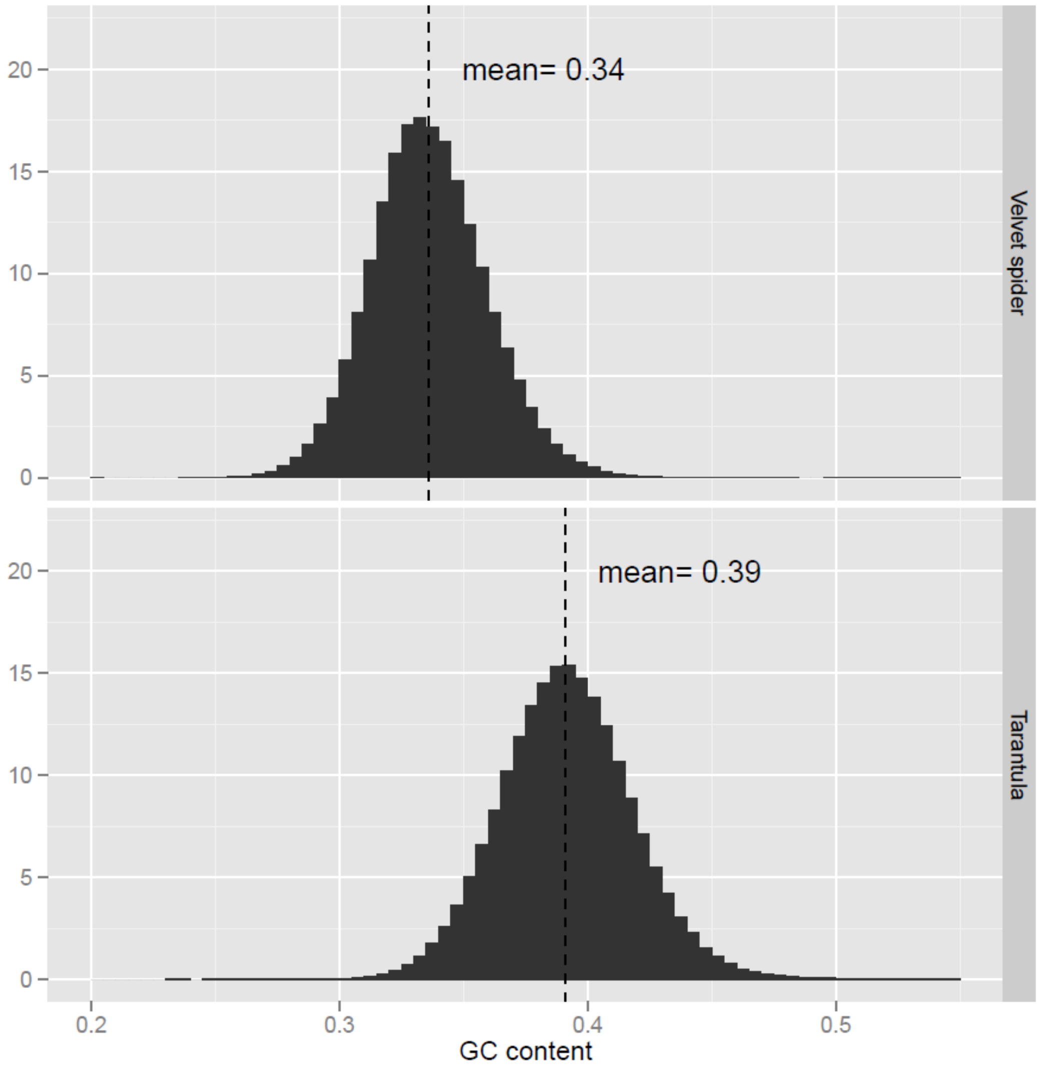
Supplementary Figure 10. Shows the diversity of the aligned C-terminal domains and the flanking regions of S.m. MaSp-putative-b and S.m. MaSp- putative-c. The position of the C-terminal domain is indicated by the red box. The sequences downstream to the C- terminal domains do not share evolutionary history, and the pi values estimated for this region do not reflect true divergence.



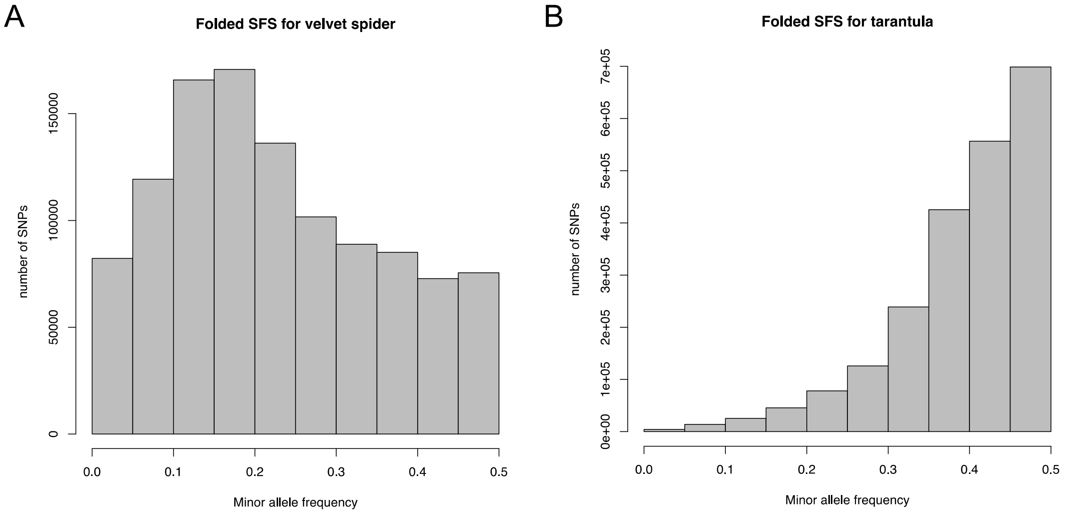
Supplementary Figure 11. Schematic overview of spidroins identified in the tarantula. A total of 12 transcripts were found to show similarity to published spidroin sequences by blast. A.g. Spidroin 1 and 2 are put together from 2 and 5 transcripts, respectively. PCR verified that the five transcripts from A.g. Spidroin 2 most likely come from same locus. Colors indicate: green N-terminal domains, red C-terminal domains, light and dark grey repetitive domains, brown internal non-repetitive domain and purple potential fragments of spidroin genes likely from the repetitive core region. Only A.g. Spidroin 5 is complete.


Start
82,880
TE keyword and >50% exon repeat overlap or single exon and >70% repeat overlap
Yes
Label as "Repeat"
8,740
No
CDS length > 900 or InterProScan match or homology match
Yes
Label as "Protein coding"
20,755
Label as "Unclassified"
53,861
End

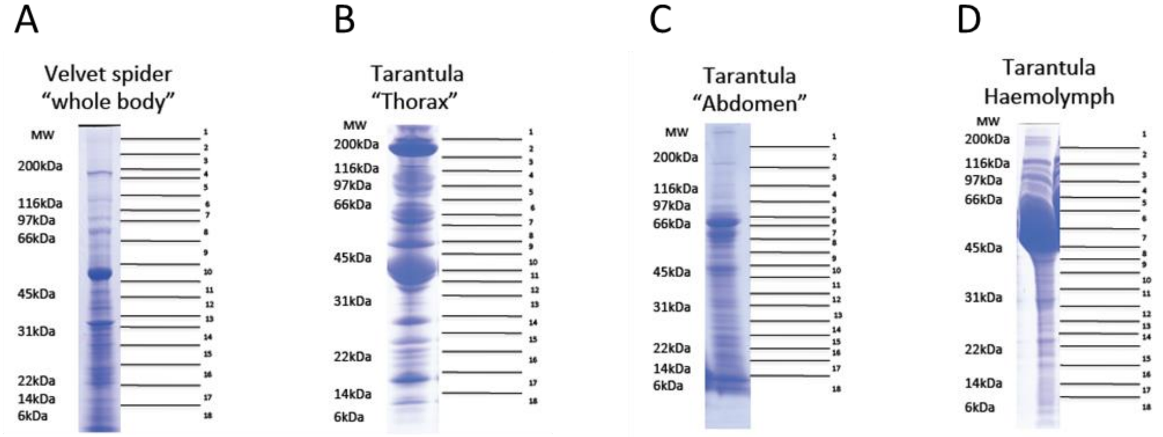
Supplementary Figure 12. Gene model classification. To identify gene models corresponding to transposable elements (TEs), RepeatScout[1] was used to build a library of repeat elements. Gene models were categorized as repeats if their annotation contained TE-related keywords and more than 50% of the gene model exons overlapped with repeat masked regions or if they were single-exon genes with more than 70% repeat masked overlap. Next, we categorized gene models as protein coding if they had a coding sequence (CDS) length of more than 900 bp, contained conserved protein domains or had homology to known proteins. The remaining gene models were designated "Unclassified".



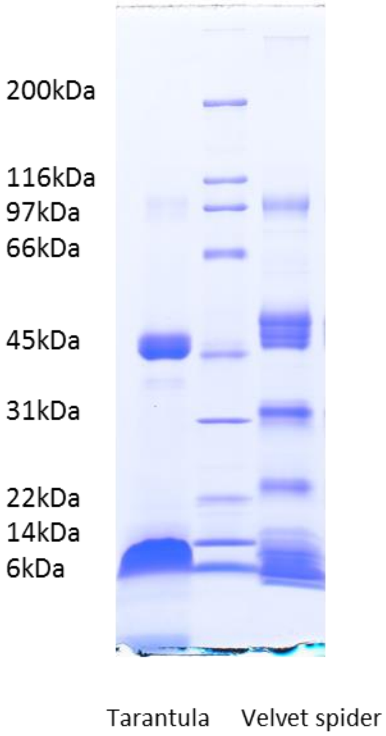
Supplementary Figure 13. GC content in the two species.



Supplementary Figure 14. Folded site frequency spectrum for velvet spider (A) and tarantula (B).



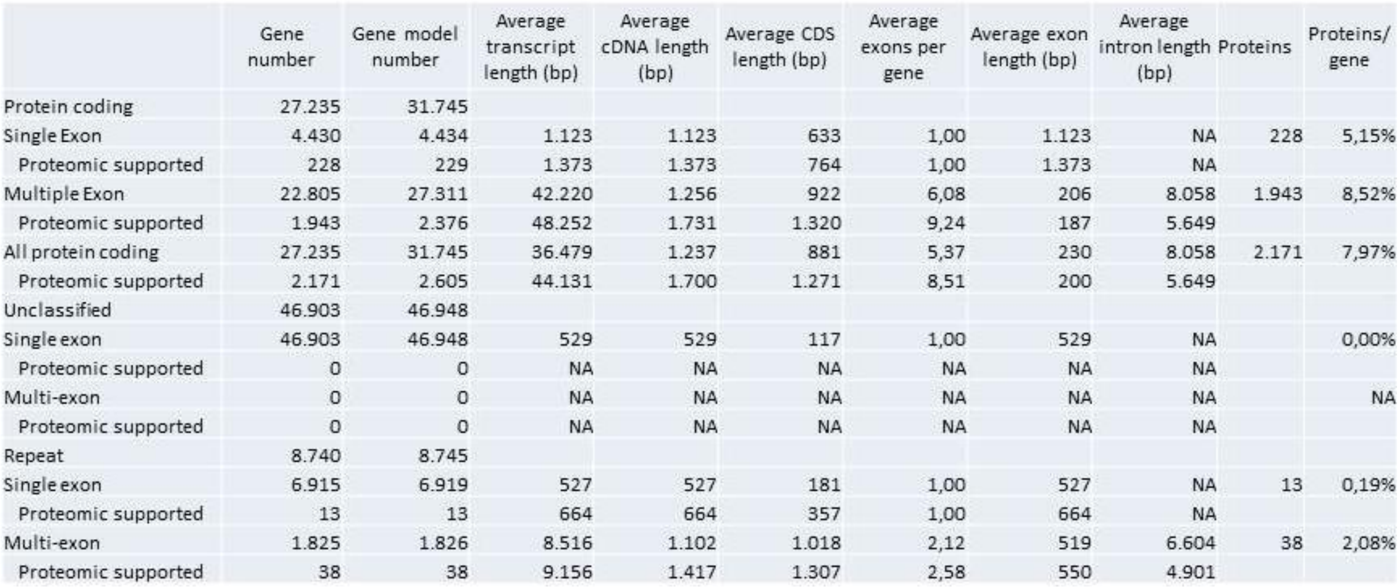
Supplementary Figure 15. Coomassie blue-stained gels of samples for proteomics. The proteins were extracted from the different tissues and resolved by reducing SDS-PAGE and subsequently visualized by staining (MW: molecular weight marker). The gel lanes were then cut into 18 gel pieces as indicated at the right. The proteins in the different gel slices were in-gel digested with trypsin and the resulting tryptic peptides micro-purified prior to LC-MS/MS analyses. **A.** Analysis of intact velvet spider. **B.** Analysis of thorax from a tarantula. **C.** Analysis of the tarantula abdomen sample, and **D.** Analysis of tarantula haemolymph.



Supplementary Figure 16. Coomassie-blue stained SDS-gel of venom from tarantula and velvet spider, respectively. The middle lane is weight marker with corresponding values to the left.

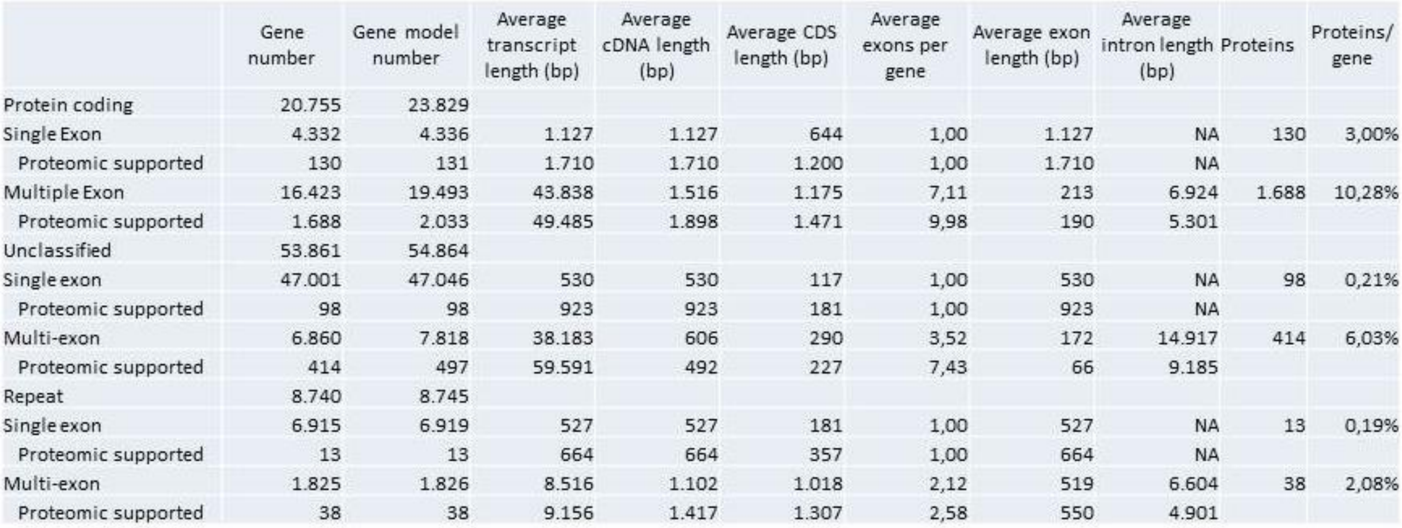
Supplementary tables

Supplementary Table 1. Proteomics-supported final gene model classification



	Gene number	Gene model number	Average transcript length (bp)	Average cDNA length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)	Proteins	Proteins/ gene
Protein coding	27.235	31.745								
Single Exon	4.430	4.434	1.123	1.123	633	1,00	1.123	NA	228	5,15%
Proteomic supported	228	229	1.373	1.373	764	1,00	1.373	NA		
Multiple Exon	22.805	27.311	42.220	1.256	922	6,08	206	8.058	1.943	8,52%
Proteomic supported	1.943	2.376	48.252	1.731	1.320	9,24	187	5.649		
All protein coding	27.235	31.745	36.479	1.237	881	5,37	230	8.058	2.171	7,97%
Proteomic supported	2.171	2.605	44.131	1.700	1.271	8,51	200	5.649		
Unclassified	46.903	46.948								
Single exon	46.903	46.948	529	529	117	1,00	529	NA		0,00%
Proteomic supported	0	0	NA	NA	NA	NA	NA	NA		
Multi-exon	0	0	NA	NA	NA	NA	NA	NA		NA
Proteomic supported	0	0	NA	NA	NA	NA	NA	NA		
Repeat	8.740	8.745								
Single exon	6.915	6.919	527	527	181	1,00	527	NA	13	0,19%
Proteomic supported	13	13	664	664	357	1,00	664	NA		
Multi-exon	1.825	1.826	8.516	1.102	1.018	2,12	519	6.604	38	2,08%
Proteomic supported	38	38	9.156	1.417	1.307	2,58	550	4.901		

Supplementary Table 2. Initial gene model classification



	Gene number	Gene model number	Average transcript length (bp)	Average cDNA length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)	Proteins	Proteins/ gene
Protein coding	20.755	23.829								
Single Exon	4.332	4.336	1.127	1.127	644	1,00	1.127	NA	130	3,00%
Proteomic supported	130	131	1.710	1.710	1.200	1,00	1.710	NA		
Multiple Exon	16.423	19.493	43.838	1.516	1.175	7,11	213	6.924	1.688	10,28%
Proteomic supported	1.688	2.033	49.485	1.898	1.471	9,98	190	5.301		
Unclassified	53.861	54.864								
Single exon	47.001	47.046	530	530	117	1,00	530	NA	98	0,21%
Proteomic supported	98	98	923	923	181	1,00	923	NA		
Multi-exon	6.860	7.818	38.183	606	290	3,52	172	14.917	414	6,03%
Proteomic supported	414	497	59.591	492	227	7,43	66	9.185		
Repeat	8.740	8.745								
Single exon	6.915	6.919	527	527	181	1,00	527	NA	13	0,19%
Proteomic supported	13	13	664	664	357	1,00	664	NA		
Multi-exon	1.825	1.826	8.516	1.102	1.018	2,12	519	6.604	38	2,08%
Proteomic supported	38	38	9.156	1.417	1.307	2,58	550	4.901		

Supplementary Table 3. Overview of the performed LC-MS/MS analyses.

Tissue samples	Tarantula – thorax – 1 gel-lane	18
	Tarantula – abdomen – 1 gel-lane	18
	Tarantula – hemolymph – 1 gel-lane	18
	Velvet spider – “whole body” – 1 gel-lane	18
Venom samples	Tarantula – 3 gel-lanes	45
	Velvet spider – 3 gel-lanes	45
	Tarantula - in-solution digest, 1 sample with 4 replica	4
	Velvet spider – in-solution digest, 1 sample with 8 replica	8
Silk samples	Tarantula –3 samples with 3 replica of each	9
	Velvet spider – whole web, 3 samples with 3 replica of each	9
	Velvet spider – dragline, 1 sample	1
	Velvet spider – egg-case, 1 sample	1
Total number of LC-MS/MS analyses:		**194**

Supplementary Table 4. Repeatmasker results for the tarantula genome (the entire assembly, 3562063354 bp excluding N/X runs), all % are including N/X runs (which should be ignored).

		Count	Length	%
SINEs:		558915	76503680 bp	1.35 %
	ALUs	0	0 bp	0.00 %
	MIRs	0	0 bp	0.00 %
LINEs:		1072136	260480958 bp	4.59 %
	LINE1	5947	1157289 bp	0.02 %
	LINE2	115915	29288607 bp	0.52 %
	L3/CR1	65625	19022251 bp	0.34 %
LTR elements:		53436	21829186 bp	0.38 %
	ERVL	0	0 bp	0.00 %
	ERVL-MaLRs	0	0 bp	0.00 %
	ERV_classI	0	0 bp	0.00 %
	ERV_classII	3791	745898 bp	0.01 %
DNA elements:		3815184	1021971808 bp	18.00 %
	hAT-Charlie	48460	12889325 bp	0.23 %
	TcMar-Tigger	44971	13023408 bp	0.23 %
Unclassified:		4916608	630743359 bp	11.11 %
Total interspersed repeats:			2011528991 bp	35.44 %
Small RNA:		113193	17790495 bp	0.31 %
Satellites:		3435	1471274 bp	0.03 %
Simple repeats:		153010	7565684 bp	0.13 %
Low complexity:		412111	9969528 bp	0.18 %

Supplementary Table 5. Repeatmasker results for the velvet spider genome (the entire assembly)

	Count	Length (bp)	%
DNA elements	1,941,556	396,267,039	13.76
LINEs	510,765	104,996,763	3.64
LTRs	999,082	201,755,688	7.00
Other	273	18,930	0.00
SINEs	10,787	211,503	0.01
Simple_repeats	17,424	2,624,997	0.09
Unclassified	4,085,830	847,865,628	29.43
Total interspersed repeats:		1,551,876,153	53.87

Supplementary Table 6. Functional support for silk genes in the velvet spider.

Silk gene	Transcriptome support	Proteome support-whole web silk	Proteome support-dragline silk	Proteome support-egg case silk
S.m. MaSp-putative-d	X	X	X	
S.m. MaSp-putative-e	X	X	X	
S.m. MaSp-putative-i	X	X	X	X
S.m. MaSp-putative-j	X	X	X	X
S.m. MaSp-putative-h	X	X	X	X
*S.m. MaSp-putative-b**	X	X		
*S.m. MaSp-putative-a**	X	X		
*S.m. MaSp-putative-c**	X	X		
S.m. MaSp-putative-f	X	X	X	X
S.m. MaSp-putative-g	X	X	X	X
S.m. MaSp-putative-k				
S.m. MiSp-putative	X	X	X	
S.m. AcSp-putative	X	X	X	X
S.m. TuSp-putative				X
S.m. PySp-putative	X	X	X	X
S.m. Sp2a	X	X		
S.m. Sp2b	X	X	X	X
S.m. Sp1	X	X		X
S.m. Sp2c	X		X	

* These were only identified by one unique peptide. However, the MS/MS spectra were manually inspected and uninterrupted y- or b-ion serials were present. In addition, these spidroins were also identified in 6 of the 11 velvet spider silk LC-MS/MS analyses, which supports their presence.

Supplementary Table 7. List of spidroin sequences found in the velvet spider.

Spidroin	Sequence	Length bp	# exons	# repeats
S.m MaSp-putative-d	Complete	8328	2	28
S.m. MaSp-putative-e	Complete	22270	3	1*
S.m. MaSp-putative-i	Complete	4021	1	7
S.m. MaSp-putative-j	Complete	5918	1	13
S.m. MaSp-putative-h	Complete	11544	5	4
S.m. MaSp-putative-b	C-term	-	-	-
S.m. MaSp-putative-a	C-term	-	-	-
S.m. MaSp-putative-c	C-term	-	-	-
S.m. MaSp-putative-f	N-term	-	-	-
S.m. MaSp-putative-g	N-term	-	-	-
S.m. MaSp-putative-k	C-term	-	-	-
S.m. MiSp-putative	Complete	5531	1	7
S.m. TuSp-putative	Complete	7388	1	12
S.m. AcSp-putative	Complete	12650	1	20
S.m. PiSp-putative	Complete**	5037**	1**	7**
S.m. Sp2a	Complete	5717	1	6
S.m. Sp2b	N-term	-	-	-
S.m. Sp1	Complete	1092	1	-
S.m. Sp2c	Complete	11286	2	4

* a single internal exon with repeat-like amino acid composition.

** S.m. PiSp consists of an N- and C- terminal domain on different scaffolds, but PCR verified that they belong to same locus. A part of the repetitive core sequence is missing, and the length, exon- and repeat numbers are therefore not certain.

Supplementary Table 8. Accession numbers for sequences used in the phylogenetic analyses of spidroins. All sequences in Genbank were used except if identical sequences were uploaded.

N-terminal domain
Araneus_ventricosus_MiSp gb|JX513956.1|
Argiope_bruennichi_TuSp_I dbj|AB242145.1|
Argiope_bruennichi_TuSp_II dbj|AB242144.1|
Argiope_bruennichi_MaSp2 gb|JX112872.1|
Latrodectus_hesperus_AcSp gb|JX978171.1|
Latrodectus_hesperus_MaSp1_I gb|EF595246.1|
Agelenopsis_aperta_TuSp gb|HM752576.1|
Agelenopsis_aperta_MaSp gb|HM752573.1|
Araneus_ventricosus_MaSp gb|AY945306.1|
Argiope_argentata_TuSp gb|HM752577.1|
Argiope_trifasciata_MaSp2 gb|DQ059136.1|DQ059136S1
Cyrtophora_moluccensis_Masp1 gb|KF032719.1|
Deinopis_spinosa_MaSp2 gb|HM752568.1|
Diguetia_canities_MaSp_I gb|HM752566.1|
Diguetia_canities_MaSp_II gb|HM752564.1|
Euprosthenops_australis_MaSp1 emb|AM259067.1|
Kukulcania_hibernalis_MaSp1 gb|HM752563.1|
Latrodectus_hesperus_MaSp2_I gb|EF595248.1|
Latrodectus_hesperus_TuSp gb|DQ379383.1|
Latrodectus_hesperus_MaSp2_II gb|DQ379382.1|
Latrodectus_hesperus_MiSp gb|HM752570.1|
Latrodectus_hesperus_MaSp1_II gb|EU177665.1|
Latrodectus_hesperus_MaSp1_III gb|EU177663.1|
Latrodectus_hesperus_MaSp1_IV gb|EU177650.1|
Metepeira_grandiosa_MiSp gb|HM752575.1|
Latrodectus_geometricus_AcSp gb|JX978180.1|
Latrodectus_geometricus_MaSp1_I gb|EU177669.1|
Latrodectus_geometricus_MaSp1_II gb|DQ059133.1|DQ059133S1
Latrodectus_geometricus_MaSp1_III gb|EU177667.1|
Latrodectus_geometricus_MaSp1_IV gb|EU177660.1|
Latrodectus_geometricus_MaSp2 gb|EU177657.1|
Latrodectus_mactans_MaSp1 gb|HM752779.1|
Uloborus_diversus_MiSp gb|HM752574.1|
Nephila_inaurata_madagascariensis_MaSp2 gb|DQ059135.1|
Nephila_madagascariensis_Flag gb|AF218623.1|AF218623S1
Nephila_clavipes_Flag gb|AF027972.1|AF027972
Nephila_clavipes_MaSp2 gb|EU599243.1|
Nephila_clavipes_MaSp1_I gb|EU599242.1|
Nephila_clavipes_MaSp1_II gb|EU599241.1|
Nephila_antipodiana_TuSp gb|EU730637.1|
Bothriocyrtum_californicum_fibroin_1 gb|HM752562.1|

C-terminal domain
Agelenopsis_aperta_MaSp gb|AY566305.1|
Agelenopsis_aperta_TuSp gb|HM752572.1|
Aphonopelma_seemanni_fibroin_2 gb|JX102558.1|
Aphonopelma_seemanni_fibroin_3 gb|JX102559.1|
Aptostichus_sp._fibroin_1 gb|EU117160.1|
Aptostichus_sp._fibroin_2 gb|EU117161.1|
Araneus_bicentenarius_spidroin_2 gb|U20328.1|ABU20328
Araneus_diadematus_fibroin_1 gb|U47853.1|ADU47853
Araneus_diadematus_fibroin_2 gb|U47854.1|ADU47854
Araneus_diadematus_fibroin_3 gb|U47855.1|ADU47855
Araneus_diadematus_fibroin_4 gb|U47856.1|ADU47856
Araneus_gemmoides_TuSp gb|AY855101.1|
Araneus_ventricosus_AcSp gb|HQ008714.1|
Araneus_ventricosus_Flag_I gb|EF025541.1|
Araneus_ventricosus_Flag_II b|AY587193.1|
Araneus_ventricosus_MaSp1 gb|JN857964.2|
Araneus_ventricosus_MaSp2 gb|AY177203.1|
Araneus_ventricosus_MiSp gb|JX513956.1|
Argiope_amoena_MaSp1 gb|AY263390.1|
Argiope_amoena_MaSp2_I gb|AY365021.1|
Argiope_amoena_MaSp2_II gb|AY365020.1|
Argiope_amoena_MaSp2_III b|AY365018.1|
Argiope_argentata_MiSp gb|JQ713004.1|
Argiope_argentata_TuSp gb|AY953084.1|
Argiope_aurantia_MaSp2 gb|AF350263.1|AF350263
Argiope_aurantia_TuSp gb|AY855099.1|
Argiope_bruennichi_MaSp1 gb|JX112871.1|
Argiope_bruennichi_MaSp2_I gb|JX112872.1|
Argiope_bruennichi_MaSp2_II gb|JX202781.1|
Argiope_bruennichi_TuSp_I dbj|AB242145.1|
Argiope_bruennichi_TuSp_II dbj|AB242144.1|
Argiope_trifasciata_AcSp gb|AY426339.1|
Argiope_trifasciata_Flag gb|AF350264.1|AF350264
Argiope_trifasciata_MaSp1 gb|AF350266.1|AF350266
Argiope_trifasciata_MaSp2_I gb|AF350267.1|AF350267
Argiope_trifasciata_MaSp2_II gb|DQ059137.1|DQ059136S2
Argiope_trifasciata_PiSp gb|GQ980328.1|
Avicularia_juruensis_spidroin_1a gb|EU652181.1|
Avicularia_juruensis_spidroin_1b gb|EU652182.1|
Avicularia_juruensis_spidroin_1c gb|EU652183.1|
Avicularia_juruensis_spidroin_2 gb|EU652184.1|
Bothriocyrtum_californicum_fibroin_1 gb|EU117162.1|
Bothriocyrtum_californicum_fibroin_2 gb|EU117163.1|
Bothriocyrtum_californicum_fibroin_3 gb|EU117164.1|
Cyrtophora_moluccensis_dragline_silk_spidroin_I gb|AY666063.1|
Cyrtophora_moluccensis_dragline_silk_spidroin_II gb|AY666061.1|

Cyrtophora_moluccensis_dragline_silk_spidroin_III gb|AY666060.1|
Cyrtophora_moluccensis_dragline_silk_spidroin_IV gb|AY666062.1|
Cyrtophora_moluccensis_MaSp1 gb|KF032719.1|
Cyrtophora_moluccensis_TuSp gb|AY953083.1|
Deinopis_spinosa_fibroin_1a gb|DQ399326.1|
Deinopis_spinosa_fibroin_1b gb|DQ399327.1|
Deinopis_spinosa_fibroin_2 gb|DQ399323.1|
Deinopis_spinosa_Flag gb|DQ399325.1|
Deinopis_spinosa_MaSp2 gb|DQ399329.1|
Deinopis_spinosa_MaSp2a gb|DQ399328.1|
Deinopis_spinosa_MiSp gb|DQ399324.1|
Deinopis_spinosa_TuSp gb|AY953073.1|
Diguetia_canities_MaSp_I gb|HM752567.1|
Diguetia_canities_MaSp_II gb|HM752565.1|
Dolomedes_tenebrosus_fibroin_1 gb|AF350269.1|AF350269
Dolomedes_tenebrosus_fibroin_2 gb|AF350270.1|AF350270
Euagrus_chisoseus_fibroin_1 gb|EU117165.1|
Euprosthenops_australis_MaSp1 emb|AM490183.1|
Euprosthenops_australis_MaSp1b emb|AM490191.1|
Euprosthenops_australis_MaSp2 emb|AM490169.1|
Gasteracantha_mammosa_MaSp2 gb|AF350272.1|
Hexura_picea_fibroin_1 gb|JX102565.1|
Hypochilus_thorelli_fibroin_1 gb|JX102555.1|
Hypochilus_thorelli_fibroin_2 gb|JX102556.1|
Latrodectus_geometricus_AcSp gb|JX978181.1|
Latrodectus_geometricus_MaSp gb|DQ059134.1|DQ059133S2
Latrodectus_geometricus_MaSp1 gb|AF350273.1|AF350273
Latrodectus_geometricus_MaSp2 b|AF350275.1|AF350275
Latrodectus_geometricus_TuSp gb|AY953079.1|
Latrodectus_hasselti_TuSp gb|AY953080.1|
Latrodectus_hesperus_AcSp gb|EU025854.1|
Latrodectus_hesperus_MaSp1_I gb|EU177650.1|
Latrodectus_hesperus_MaSp1_II gb|EU177648.1|
Latrodectus_hesperus_MaSp1_III gb|EU177655.1|
Latrodectus_hesperus_MaSp1_IV gb|EU177653.1|
Latrodectus_hesperus_MaSp2_I gb|EF595245.1|
Latrodectus_hesperus_MaSp2_II gb|DQ409058.1|
Latrodectus_hesperus_MiSp gb|EU394445.1|
Latrodectus_hesperus_PiSp gb|HQ005791.1|
Latrodectus_hesperus_TuSp gb|AY953070.1|
Latrodectus_mactans_TuSp gb|AY953077.1|
Latrodectus_tredecimguttatus_TuSp gb|AY953078.1|
Macrothele_holsti_dragline_silk_spidroin gb|AY666068.1|
Megahexura_fulva_fibroin_1 gb|JX102566.1|
Metepeira_grandiosa_MiSp gb|HM752569.1|
Nephila_antipodiana_MaSp1 gb|DQ338461.1|
Nephila_antipodiana_MiSp gb|DQ338462.1|
Nephila_antipodiana_TuSp gb|DQ089048.1|
Nephila_clavipes_dragline_silk_fibroin gb|M37137.2|NEPDSF
Nephila_clavipes_Flag gb|AF027973.1|AF027973

Nephila_clavipes_MaSp1_I gb|AY654292.1|
Nephila_clavipes_MaSp1_II gb|EU617338.1|
Nephila_clavipes_MaSp1_III gb|AY654289.1|
Nephila_clavipes_MaSp2 gb|AY654297.1|
Nephila_clavipes_MiSp_I gb|AF027736.1|AF027736
Nephila_clavipes_MiSp_II gb|AF027735.1|AF027735
Nephila_clavipes_PiSp gb|HM020705.1|
Nephila_clavipes_spidroin gb|U20329.1|NCU20329
Nephila_clavipes_TuSp gb|AY855102.1|
Nephila_madagascariensis_Flag gb|AF218624.1|AF218623S2
Nephila_madagascariensis_MaSp1 gb|AF350277.1|AF350277
Nephila_madagascariensis_MaSp2 gb|AF350278.1|AF350278
Nephila_pilipes_dragline_silk_spidroin_I gb|AY666077.1|
Nephila_pilipes_dragline_silk_spidroin_II gb|AY666075.1|
Nephila_pilipes_dragline_silk_spidroin_III b|AY666076.1|
Nephila_pilipes_dragline_silk_spidroin_IV gb|AY666073.1|
Nephila_pilipes_dragline_silk_spidroin_V gb|AY666055.1|
Nephila_pilipes_dragline_silk_spidroin_VI gb|AY666053.1|
Nephila_pilipes_dragline_silk_spidroin_VII gb|AY666059.1|
Nephila_pilipes_dragline_silk_spidroin_VIII gb|AY666050.1|
Nephila_pilipes_dragline_silk_spidroin_IX gb|AY666049.1|
Nephila_senegalensis_MaSp1 gb|AF350279.1|AF350279
Nephila_senegalensis_MaSp2 gb|AF350280.1|AF350280
Nephilengys_cruentata_Flag gb|EF638444.1|
Nephilengys_cruentata_MaSp gb|EF638446.1|
Nephilengys_cruentata_MiSp gb|EF638447.1|
Nephilengys_cruentata_PiSp gb|GU062417.1|
Nephilengys_cruentata_TuSp gb|EF638445.1|
Octonoba_varians_dragline_silk_spidroin_I gb|AY666059.1|
Octonoba_varians_dragline_silk_spidroin_II gb|AY666076.1|
Octonoba_varians_dragline_silk_spidroin_III gb|AY666057.1|
Octonoba_varians_dragline_silk_spidroin_IV gb|AY666059.1|
Parawixia_bistriata_AcSp gb|GQ275356.1|
Parawixia_bistriata_MaSp1 gb|GQ275359.1|
Parawixia_bistriata_MaSp2 gb|GQ275360.1|
Parawixia_bistriata_MiSp gb|GQ275358.1|
Peucetia_viridans_MaSp1 gb|GU306168.1|
Plectreurys_tristis_fibroin_1 gb|AF350281.1|AF350281
Plectreurys_tristis_fibroin_2 gb|AF350282.1|AF350282
Plectreurys_tristis_fibroin_3 gb|AF350283.1|AF350283
Plectreurys_tristis_fibroin_4 gb|AF350284.1|AF350284
Poecilotheria_regalis_fibroin_2 gb|JX102561.1|
Psechrus_sinensis_dragline_silk_spidroin_I gb|AY666067.1|
Psechrus_sinensis_dragline_silk_spidroin_II gb|AY666066.1|
Psechrus_sinensis_dragline_silk_spidroin_III gb|AY666065.1|
Psechrus_sinensis_dragline_silk_spidroin_IV gb|AY666064.1|
Tetragnatha_kauaiensis_MaSp1 gb|AF350285.1|AF350285
Tetragnatha_versicolor_MaSp1 gb|AF350286.1|AF350286
Uloborus_diversus_AcSp1 gb|DQ399333.1|
Uloborus_diversus_MaSp1 gb|DQ399331.1|

Uloborus_diversus_MaSp2_I gb|DQ399334.1|
Uloborus_diversus_MaSp2_II gb|DQ399335.1|
Uloborus_diversus_MiSp gb|DQ399332.1|
Uloborus_diversus_TuSp gb|AY953072.1|

Supplementary Table 9. Functional support for silk genes in the tarantula. * A.g. Spidroin-1 and A.g. Spidroin-2 consists of two transcripts and 5 transcripts, respectively, that we merged based on highly similar repeat sequences and PCR verification. There is protein support for both A.g. Spidroin-1 transcripts.

Putative spidroins	Proteome support
*A.g. Spidroin-1**	X
*A.g. Spidroin-2**	X
A.g. Spidroin-3	
A.g. Spidroin-4	X
A.g. Spidroin-5	X
A.g. Spidroin-6	
A.g. Spidroin-7	X

Supplementary Table 10. Summary statistics of tarantula raw data, assuming the genome size is 6.0G

Pair-end libraries	Insert Size	Total Data(G)	Reads Length	Sequence coverage (X)
Solexa Reads	250bp	48.11	150_150	8,02
	500bp	42.85	150_150	7,14
	2kb	52.26	49_49	8,71
	5kb	35.70	49_49	5,95
	10kb	24.15	49_49	4,03
	20kb	25.41	49_49	4,24
Total		228.47		38,08

Supplementary Table 11. Summary statistics of velvet spider raw data, assuming the genome size is 3.0G

Pair-end libraries	Insert Size	Total Data(G)	Reads Length	Sequence coverage (X)
Solexa Reads	250bp	145.99	150_150	48.66
	500bp	80.78	150_150	26.93
	2kb	68.99	49_49	23.00
	5kb	33.86	49_49	11.29
	10kb	17.55	49_49	5.85
	20kb	6.80	49_49	2.27
Total		353.99		118.00

Supplementary Table 12. Transcriptome sequencing data statistics.

Species	Tissues	Insert Size	Reads Length	Raw Data(G)
Tarantula	Whole body	150-250	101	33.65
	Venom gland	200	90	7.01
	Opistosomal gland	200	90	7.28
Velvet spider	Whole body	200	90	7.07
	Venom gland	200	90	7.2

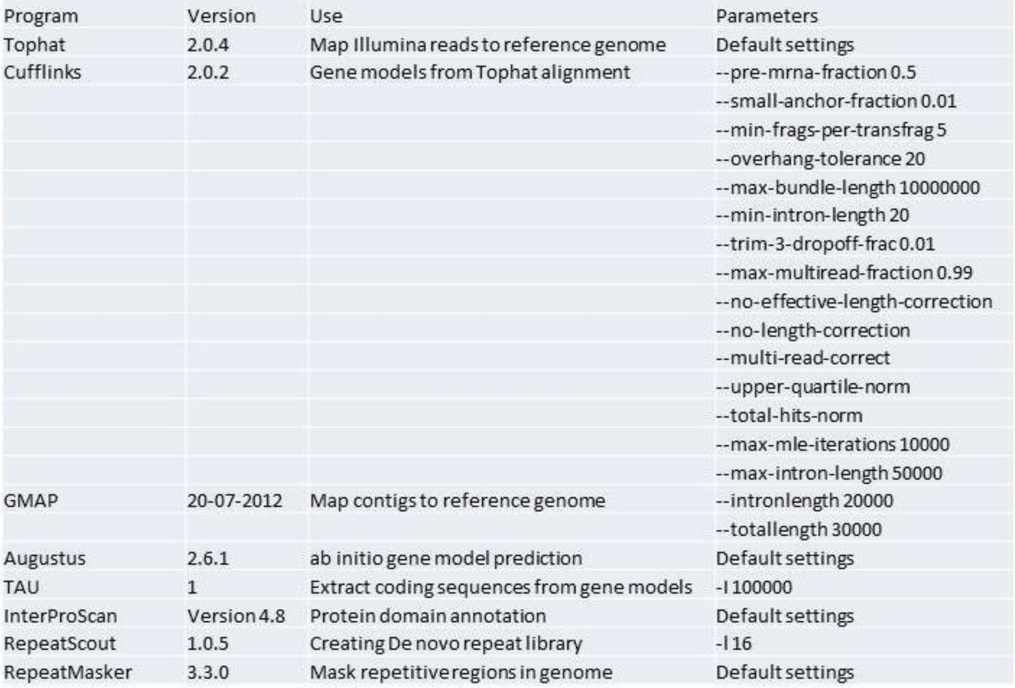
Supplementary Table 13. Statistics of the assembled sequence length in the tarantula.

	Contig		Scaffold	
	Size(bp)	Number	Size(bp)	Number
N90	118	15168132	646	794718
N80	139	11458794	1748	192171
N70	164	8305732	14325	69268
N60	205	5705116	30019	42072
N50	277	3696252	47827	26834
Longest	15869		2755643	
Total Size	4737985631		5787464414	
Total Number (>100bp)		19550163		2432292
Total Number (>2kb)		91877		174012

Supplementary Table 14. Statistics of the assembled sequence length in the velvet spider.

	Contig		Scaffold	
	Size(bp)	Number	Size(bp)	Number
N90	3,260	180,442	68,248	7,328
N80	7,055	123,914	175,692	4,843
N70	10,362	91,013	267,727	3,527
N60	13,682	67,267	356,755	2,600
N50	17,272	48,813	456,729	1,886
Longest	160,587		4,549,793	
Total Size	2,835,815,719		2,880,654,633	
Total Number(>100bp)		681,210		1,232,544
Total Number(>2kb)		204,058		14,958

Supplementary Table 15. Program parameters.



Program	Version	Use	Parameters
Tophat	2.0.4	Map Illumina reads to reference genome	Default settings
Cufflinks	2.0.2	Gene models from Tophat alignment	--pre-mrna-fraction 0.5
			--small-anchor-fraction 0.01
			--min-frags-per-transfrag 5
			--overhang-tolerance 20
			--max-bundle-length 10000000
			--min-intron-length 20
			--trim-3-dropoff-frac 0.01
			--max-multiread-fraction 0.99
			--no-effective-length-correction
			--no-length-correction
			--multi-read-correct
			--upper-quartile-norm
			--total-hits-norm
			--max-mle-iterations 10000
			--max-intron-length 50000
GMAP	20-07-2012	Map contigs to reference genome	--intronlength 20000
			--totallength 30000
Augustus	2.6.1	ab initio gene model prediction	Default settings
TAU	1	Extract coding sequences from gene models	-l 100000
InterProScan	Version 4.8	Protein domain annotation	Default settings
RepeatScout	1.0.5	Creating De novo repeat library	-l 16
RepeatMasker	3.3.0	Mask repetitive regions in genome	Default settings

Supplementary Table 16. Overview of de novo transcriptome assembly.

Species	Tissue	Size (Mbp)	Number of transcripts	Min length (bp)	Average length (bp)	Max length (bp)	N50 length (bp)	# transcripts annotated by NCBI nr	% transcripts annotated by NCBI nr
Tarantula	Opistosomal gland	11.15	22,480	113	495	6,306	564	14,134	62.9%
	Venom gland	4.20	9,619	107	436	12,112	437	5,579	58.0%
	The whole body	70.81	84,299	100	840	24,532	1,515	38,711	45.9%
Velvet spider	Venom gland	0.53	1,269	105	416	5,857	388	768	60.5%
	The whole body	11.20	25,892	106	432	9,887	435	16,164	62.4%

Supplementary Table 17. The criteria for assigning the best description of nr database annotation. The right column shows how the regular expression used in Perl language.

Uninformative description **(case ignoring)**	**Regular expression in Perl**
'hypothetical protein'	m/hypothetical protein/i
'novel protein [' or 'novel protein ('	m/novel protein [\[\(]/i
'unnamed protein product'	m/unnamed protein product/i
'predicted protein'	m/predicted protein/i
Starting with 4-10 characters including numbers, alphabet and dots followed by '['. e.g. 'GJ10650 [Drosophila virilis]'	m/^ [\w\.]{4,10} \[/i
Starting with Uncharacterized protein plus 4-10 characters including numbers, alphabet and dots followed by '['. e.g. ' Uncharacterized protein K03H1.5 [Harpegnathos saltator]'	m/^ Uncharacterized protein [\w\.]{4,10} \[/i

Supplementary Notes

Supplementary Note 1. Biology of the two spider species

Tarantula (Acanthoscurria geniculata, Araneae, Mygalomorphae, Theraphosidae)

The *Acanthoscurria geniculata* spider is found in the northern part of Brazil and is a terrestrial tarantula. It usually prefers to hide in pre-existing holes in the ground, owing to the fact that it's a wandering species. If provoked, instead of biting, *Acanthoscurria geniculata* quite often performs urticating hairs-shooting behavior, as a deterrent against predators. The Brazilian white-knee tarantula can reach a leg span of 20 cm and a body length of 8-9 cm for the females, while males are usually smaller. The male is characterized by the presence of tibial spurs on the dorsal part of the anterior legs, differently from most Theraphosidae genera, in which the spurs are usually located in the ventral part of the tibiae. These spurs are used by males during mating to handle the female, in order to manage its aggressiveness and lift its frontal legs to reach the epyginium with the pedipalps. Males reach adulthood around 1 year and 6 months, and they are fertile usually until 6/8 months later, while females become adult after 3-4 years. After mating, it takes around 3 months for the female to produce a cocoon, which normally hatches after 4-6 weeks. It's a very prolific species, and up to 2000 spiderlings can hatch from a single egg sac, that measure at the time of birth 5-8 mm[2]. The *Acanthoscurria geniculata* used in this study originated from a captive bred stock and were obtained from commercial dealers. Upon purchase they were kept in individual terraria, containing Exoterra plantation soil (made from compressed coconut husk fibres), a shelter and branches of wood within the animal care facility at Department of Biosciences (Aarhus University). The daily light:dark cycle was 14:10h, temperature was 27-29^{o}C and air humidity around 80%. The tarantulas were fed on cockroaches on a weekly basis and increased body mass during captivity.

Velvet spider, (Stegodyphus mimosarum, Araneae, Araneomorphae, Eresidae)

Stegodyphus mimosarum lives in the southern and eastern part of Africa[3,4]. It is a social species, a trait that has evolved three times independently in the *Stegodyphus* genus. Social behavior is characterized by colony living. Colonies are founded by single mated females, and individual spiders typically stay in the colony throughout their lifetime, with very little dispersal among colonies. The number of spiders in newly founded colonies quickly rises, and mature colonies often have as many as 300 to 500 individuals. Since colonies are founded by single mated females inbreeding among colony members is extreme. Furthermore, the sex ratio is female biased (about 8:1), there is reproductive skew among females, and populations often undergo boom-and-bust dynamics, where whole populations quickly die out while new ones arise. All these factors mean that the effective population size is expected to be very low, as is the level of genetic variation. See Lubin and Bilde (2007)[5] for more details.

The females of the velvet spider reach a body length of ~1.5cm. Males are smaller. For this study a single colony was collected in South Africa (GPS position: 29° 39' 16.46" S, 30° 27' 35.55" E) and kept in the lab until extraction of DNA and RNA, dissection of venom glands, sampling of silk, and milking of venom. Spiders were fed twice per week with *Calliphora* flies and sprayed with water once per month. The colony consisted of about 300 individuals. Only females were used in all analyses.

Supplementary Note 2. Venomics

The proteinaceous part of venom has traditionally been divided in the proteins below 10 kDa, which contains the protoxins, and the other proteins (mainly enzymes) with molecular weights above 10 kDa. Gel-based separation of proteins and in-gel trypsin-digestion is a very sensitive proteomics method and well suited for identification of proteins above 10 kDa, and the present study is the most comprehensive analysis of spider venom proteins performed. However, the approach is not optimal for the identification of the smaller protoxins, since they only contain a few tryptic peptides suitable for the mass spectrometer and the protoxins co-migrate on the gel. Consequently, we decided to use a bioinformatics approach employing the genomics, transcriptomics, and proteomics data, generated in the present study, to identify cysteine-rich protoxins.

Identification of protoxin encoding genes in the velvet spider

To secure that all toxin sequences are actually annotated as toxins, all non-annotated sequences were, after the annotation step based on BLASTP against NCBInr, compared with Araneomorphae toxin sequences extracted from the Arachnoserver (http://www.arachnoserver.org/mainMenu.html). Only cysteine-rich peptide toxins (based on number of cysteines and molecular weight) from the Arachnoserver were used for this comparison. After finalizing all entries that did not fulfill the following criteria were removed from the annotated protein sequence database: i) the protein should contain the word “toxin” or “non-annotated”, ii) the molecular weight should be between 4.000 and 25.000 dalton, and iii) the sequence should contain more than 4 cysteine residues among the C-terminal 80 amino acid residues. The two last criteria are based on the known primary structure of these cysteine-rich protoxins in other spider species[6]. Using this approach we reduced the database to app. 200 sequences of which 54 sequences were annotated as toxins.

In order to evaluate whether the non-annotated sequences are actually toxin-coding, we did a multiple alignment of the sequences and tried to cluster the sequences using ClustalW (http://www.ebi.ac.uk/Tools/msa/clustalw2/). However, the sequences were too distantly related for this exercise, and instead we looked for transcriptomics support in our velvet spider venom gland transcriptome, but only 5 of the non-annotated sequences were present in the transcriptome.

In contrast, 28 of the 54 toxin-annotated sequences were present in the transcriptome. Taken together, these different lines of evidence suggest that the non-annotated sequences should not be regarded as cysteine-rich spider peptide toxins and these sequences were removed from the database. Then we evaluated whether the remaining sequences had proteomics support. LC-MS/MS generated data of both the in-gel trypsin-digested venom and the in-solution trypsin-digested venom were used to query a database containing the 54 toxin-annotated sequences. The Mascot search parameters for these analyses and the criteria for protein identification are described in Methods and Supplementary Methods. Using these criteria 26 cysteine-rich protoxins were identified. No toxins were identified in the gel-based samples, which were not present in the in-solution-based approach. As previously mentioned, it underlines that the gel-based approach is not optimal for the identification of protoxins.

All 54 toxin sequences in the database were then subjected to SignalP (http://www.cbs.dtu.dk/services/SignalP/) for signal peptide prediction. Proteins with no predicted signal peptide, with no transcriptomics support, and with no proteomics support were removed from the list of toxins reducing the list to 51 velvet spider protoxins (Supplementary Data 9). Afterwards the genomic localization of the 51 genes was examined and the number of introns, the scaffold number, and the number of toxins belonging to the same cluster were reported (Fig. 3d). The 26 sequences with proteomics support were further characterized using ClusterW. Based on the manual inspection of the generated guided tree and evaluation of the alignments with focus on cysteine pattern, 9 toxin families were identified (Supplementary Fig. 6). These were named A, B, C, D, E, F, G, H, and I, with A containing the highest number of sequences, and G, H, and I being singletons. We named the 26 toxins Stegotoxin-XY, where “X” represents the capital letter representing the toxin family and “Y” being a number added to unambiguously identify the different sequences. These data shows that similar toxins, that was grouped together based on alignments, where all located on the same scaffold (Fig. 3d).

In order to obtain an estimate of the presence of sequences with toxin-homology in the genome, we performed a BLASTX search of the genome against the identified mature toxin peptides from velvet spider. These identified genome sequences may be predicted genes, which are already annotated as toxins, pseudogenes or sequences, which have not been identified during the gene finding stage. This approach identified 252 sequences with potential to encode toxin peptides. The venom mass-spectrometry analyses were used to query the database containing the 252 sequences. 37 sequences showed proteomics support, showing that most sequences had been previously identified using the method outlined before. In addition, some of the sequences among the 252 sequences might represent exons from the same gene, which will affect both the total number sequences and the number of sequences with proteomics support. This result shows that the conservative gene prediction pipeline has probably not exhaustively identified toxin-coding genes, but the coding potential for additional protoxins seems limited. As the evidence for the additionally sequences is low, and since these sequences do not necessarily represent true, intact and / or full-length protein coding genes, we did not pursue this approach further.

Identification of transcripts encoding protoxins in tarantula

The approach to identify tarantula transcript encoding for cysteine-rich peptide toxins was similar to the described approach for the velvet spider. The tarantula sequences, based on the merged transcriptomes, were annotated based on i) BLASTP against both NCBInr and ii) filtered mygalomorphae toxin-sequences extracted from the Arachnoserver, as described in Methods and Supplementary Methods. To remove the non-relevant sequences from the database we used the same criteria as described for the velvet spider, except that in addition to annotation as "toxin", one sequence annotated, as "HWTX-XVa2" protein was also included. Requirements for molecular weight and cysteines were similar. Using this approach 78 tarantula toxins were identified. Then we looked for proteomics support using the described criteria and identified 8 toxins in the venom. The 8 toxin sequences were analysed using ClustalW and grouped into 6 families (named A-F) with two families containing two sequences and the remaining four sequences being singletons. We named these 8 toxins Genicutoxin-XY, where "X" is a capital letter from A to F representing the family, and "Y" is a number added to unambiguously identify the different sequences.

We estimated the presence of sequences with toxin-homology in the genome, as described for the velvet spider. This approach identified 18 sequences with potential to encode toxin peptides. The venom mass-spectrometry analyses were used to query the database containing the 18 sequences, but none of them showed proteomics support, and we did not pursue this approach further.

Supplementary Note 3. Silkomics

Identification of the silk genes

The velvet spider

A selected set of spider N- and C-terminus terminal domain sequences spanning the major phylogenetic groups described in[7] were blasted to the genome using tblastx with a cutoff e-value of 0.01. Accession numbers can be found in Supplementary Table 6 for N- and C terminal domains respectively. Twelve complete spidroin sequences were identified, while incomplete sequences with either an N- or a C-terminal domain were found 7 times. In six of the complete spidroins, the sequences between the N- and C-terminal domains were complete open reading frames with easily identifiable repeats. However, in four of them exon-intron structures were identified. Exons were identified in two ways; 1) using mRNA sequence data when possible (many mRNA sequences are too fragmented to cover the complete transcript), and 2) searching for repeated amino acid sequences by translating the nucleotide sequence in all three frames and assuming that these are coding. Of the incomplete spidroin sequences all but one (*S.m MaSp-k*) were identified next to scaffold ends. All of them have an orientation suggesting that the complete spidroin sequences span two scaffolds. An N- and a C-terminal domain blasted to piriform sequences previously published, suggesting they belong to the same locus (see below). We did PCR with primers designed to anneal to the repeat sequence and the C-terminal domain, which verified that the two sequences identified belong to the same locus. All identified putative spidroin sequences are listed in Supplementary Table 7.

The tarantula

The same strategy as used for the velvet spider was tried, but no N- or C- terminal domains were found, suggesting that the genome assembly is too fragmented. In the transcriptome database 12 sequences were identified that show similarity to spidroin sequences identified by blast. There are two groups of sequences (of 5 and 2 sequences) that are highly similar. A PCR with primers designed in the N-terminal domain and the repeats of *A.g. Spidroin-2* gave several bands. Sequencing of those bands demonstrated that these five transcripts likely belong to same locus. It was not possible to amplify *A.g. Spidroin-1* using the same strategy. On the assumption that highly similar transcripts are repeats from same locus the twelve transcript sequences represent 7 distinct loci structured like shown in Supplementary Fig. 11. We note that the only evidence of *A.g. Spidroin-3,-6* and *-7* being spidroins are the similarity to published sequences by blast. They all blast to repetitive core region of other *Mygalomorph* species.

Functional grouping of spidroins

The spidroin sequences were grouped to previously published sequences (major ampullate, minor ampullate, aciniform, tubiliform and piriform) by blast and phylogenetic analyses. Both N- and C-terminal domain sequences were blasted using tblastx to the NCBI non-redundant database. Phylogenies were constructed for both N- and C- terminal domain sequences by aligning the sequences obtained in this study to all previously published sequences (see Supplementary Table 8 for Genbank accession numbers) using Muscle[8]. The best fitting substitution model for sequence evolution was estimated using jModelTest 2.1.1[9] with 11 substitution schemes. Model selection was computed using the Akaike information criterion (AIC). The phylogenies for both C-term and N-term sequences were constructed using the Bayesian method implemented in MrBayes 3.2[10]. According to the results from jModelTest we applied the General Time Reversible-model of sequence evolution with a Gamma distribution for the rate variation among sites and Invariable sites (GTR + G + I) for both C-term and N-term sequences. Two chains were run for four million generations, with a sampling frequency of 1000 and a burn-in of 500,000. The program Tracer v1.5.0 [11] was used to check for convergence of the model likelihood and parameters between the two runs. The resulting trees were visualized and graphically edited with FigTree v1.4.0[12] (see Supplementary Fig. 7).

The resolution of especially deep splits is quite low due to relatively high divergence among sequences. The phylogenetic analyses of both N- and C- terminal domains do not reveal a monophyletic group of the minor ampullate sequences, and the putative minor ampullate sequence of the velvet spider group closely with major ampullate sequences. The amino acid composition of the minor ampullate sequence differs from the major ampullate sequences, with major ampullate loci having high alanine and glycine content compared to the putative minor ampullate sequence from the velvet spider (Supplementary Fig 8). The ensemble repeats characterized in published minor ampullate spidroins are not found in the putative minor ampullate sequence from the velvet spider. However, the repetitive regions of the major ampullate spidroins have the characteristic poly A runs, GGX and GA motifs, which lacks in the minor ampullate repetitive region (Supplementary Fig. 9). For those reasons we exclude the putative minor ampullate spidroin sequence even though it groups phylogenetically with the major ampullate sequences.

Functional support of spidroins

Functionality of the identified spidroin sequences was ascertained by both transcriptional and proteomic support.

The velvet spider

The spidroin sequences were blasted against the whole body transcriptome sequence database which returned identical RNA sequences for all putative spidroin sequences, except for *S.m. TuSp* and *S.m. MaSp-k* (Supplementary Table 6). The individual used for transcriptome sequencing was sub-adult, and tubiliform (egg case silk) transcripts were therefore not expected to be present.

Three samples of silk were used for mass spectrometry analyses; 1) dragline silk, 2) egg case silk, and 3) whole web silk. Proteome support was found for all identified putative spidroin sequences, except for *S.m. MaSp-k*. This incomplete sequence consisting of a C–terminal domain is located in the middle region of a large scaffold, but no N-terminal domain or repeat-like sequences could be identified in the proximity. *S.m. MaSp-k* is based on this evidence most likely not functional. Evidence of functional support is summarized in Supplementary Table 6. For more details regarding method and quantification see Methods, Supplementary Methods and Supplementary Data 7 and 8.

Tarantula

All spidroin sequences identified from the tarantula come from transcriptome sequencing directly giving transcriptional support. In addition proteome support was obtained for 5 of the 7 hypothesized genes (see Supplementary Table 9) by mass spectrometry analyses of burrow lining silk. We note that other silk types like sperm web and egg case silk are produced by tarantulas, which was not analyzed in this study. For more details regarding method and quantification see Methods, Supplementary Methods and Supplementary Data 7 and 8.

Major ampullate evolution

A phylogenetic tree of the major ampullate C-terminal domain sequences (Fig. 3b) was constructed by neighbor-joining using Mega5[13],to study the molecular evolution of the major ampullate genes.

Gene conversion

The C-terminal domain and about 1500bp upstream sequence of two of the identified major ampullate sequences, *S.m. MaSp-b* and *S.m. MaSp-c*, are almost identical. Upstream from this the similarity is of same magnitude as for sequences from different spidroin loci, and downstream the sequences do not align (Supplementary Fig. 10). The most plausible explanation for this result is a gene conversion event that occurred recently.

We tested if the inferred gene conversion was real or due to mis-assembly. A primer common to the two loci was designed in the identical C-terminal domain region, and two primers were designed downstream to the C- terminal domain where the sequences are highly divergent. PCR and Sanger sequencing showed that the identical region is present in both loci.

The diversity profile of the C- terminal domains of *S.m. MaSp-b* and *S.m. MaSp-c* and the flanking regions were constructed using a sliding window approach in DNAsp 5[14] with window length 25 and step size 5. Pi values were uncorrected.

Gene duplication

Phylogenetically the two major ampullate loci *S.m. MaSp-I* and *S.m. MaSp-j* cluster closely (Fig. 4b, Supplementary Fig. 7), both in N- and C- terminal domains. Also the repeats are very similar. These results suggest a whole gene duplication event. The time since this duplication event was estimated based on divergence. The synonymous difference between the C- terminal domains of S.m. MaSp-i and S.m. MaSp-j was estimated in DNAsp5[14] to be 0.1356. Based on a molecular clock assumption, each sequence has since the duplication diverged by π_S=0.0677. Based on the mutation rates suggested by Mattila et al (2012)[15] the duplication event is estimated to have occurred around 10 mya.

Repeat evolution of S.m. MaSp-i and S.m. MaSp-j

The repeats of *S.m. MaSP-i* and *S.m. MaSp-j* are highly similar and easily alignable, so repeat evolution since the duplication event was investigated by a phylogenetic analysis. The repeat sequences were aligned using Muscle[8]. The best fitting substitution model for sequence evolution was estimated using jModelTest 2.1.1[9] with 11 substitution schemes. Model selection was computed using the Akaike information criterion (AIC). The phylogeny was constructed using the Bayesian method implemented in MrBayes 3.2[10]. According to the results from jModelTest we applied the General Time Reversible model of sequence evolution with a Gamma distribution for the rate variation among sites (GTR + G). We run two chains for four million generation, with a sampling frequency of 1000 and a burn-in of 500’000. Convergence of the model likelihood and parameters between the two runs were checked with Tracer v1.5.0[11]. The resulting tree was visualized and graphically edited in FigTree v1.4.0[12].

Pseudo-functionalization

As mentioned above the *S.m. MaSp-k* C- terminal domain sequence does not seem to be functional based on no transcriptional or proteomics support. However, the sequence still has an open reading frame suggesting that pseudo-functionalization happened recently, since a locus with no function is expected to lose its open reading frame relatively fast either due to point mutations leading to a stop codon or insertions/deletions of bases not a multiple of 3. Assuming that only point mutations will ruin the open reading frame, we estimate the maximum age of the pseudo-functionalization. The open region frame of the C-terminal region of *S.m. MaSp-k* is 318 bp long. If a mutation rate of 1E-8 per site per year is assumed, 3.18E-6 mutations in this region are expected per year. 954 different mutations are possible in this region, 3 per site. Forty two of these will lead to a stop mutation in the sequence as it is currently. Therefore, in average it will take about 7 million years for a stop codon to occur. This estimate of the maximum time since pseudogenization of the *S.m. MaSp-k* locus is conservative, since it does not consider the possibility that an insertion or deletion ruins the open reading frame. Further, the fact that we don’t find a closely related C- terminal domain sequence or a non-functional N- terminal domain sequences suggest that pseudo-functionalization occurred by a deletion event, and not an unequal recombination event.

Silk related protein

We identified a protein in all three types of silk that was not identified in the search for spidroins. This protein has an N-terminal domain not similar to the spidroin followed by a highly repetitive domain. The repetitive domain does not consist of a single conserved repeat type like the spidroins, but several in different lengths. We therefore hypothesize that this protein is not a spidroin. However, the repetitive region has a high proportion of glycine (45%) and alanine (26%) similar to spidroins. Very short repeats of GA are abundant in this protein, interfered mostly by single leucines and valines instead of the alanines. The protein was detected in all three web types (Supplementary Data 1). The nucleotide and amino acid sequence of this protein can be found in Supplementary Data 5 and 6 named ‘CUFF.83830.1_Ste Silk-related protein’.

Supplementary Methods

DNA extraction of the tarantula. DNA for both short and long insert libraries (250 bp, 500 bp, 2,000 bp, 5,000 bp, 10,000 bp, 20,000 bp) was extracted using same protocol for all libraries. Hemolymph was removed with a syringe before dissecting out soft tissue from the abdomen. About 1 gram of soft tissue was snap frozen in fluid nitrogen and grinded to powder before adding 10 ml extraction buffer (10mM Tris pH 8, 100mM EDTA, 0.02 mg RNase/ml buffer, 0.5% SDS). After incubation at 37°C for 1 h, 50 µl proteinase K (20mg/ml) was added and the sample was incubated in a 50°C water bath for 3 hours. The sample was equilibrated to room temperature before 10 ml of phenol was added. After mixing gently for 10 min, the sample was centrifuged for 15 min at 3000 rpm. The viscous aqueous phase was transferred to a new tube using a wide-pore glass pipette. Phenol extraction was repeated two times. Two ml ammonium acetate (10M) was added and the sample was mixed gently. After adding 2 volumes of ethanol at room temperature, DNA was collected using a bended pipette tip and air dried for about 10 min and dissolved in TE buffer.

DNA extraction of the velvet spider. DNA for short insert libraries (250 bp, 500 bp, 2,000 bp, 5,000 bp) was extracted from whole bodies (a single spider for each library). 350 ml CTAB were added to each sample and squashed 30 seconds using a TissueLyser. Five µl proteinase K (20mg/ml) was added before incubating the samples at 60°C for 1 h. 350 µl phenol was added, and the sample was centrifuged 2 min at 13,000 rpm. The upper phase was transferred to a new tube and 1 µl RNase was added before incubation the sample 15 min at room temperature. One volume of Chlorophorm/Isoamylalcohol (24:1) was added, and the sample was mixed gently and centrifuged for 2 min at 13,000rpm. The upper phase was transferred to a new tube, and 1 volume of Isopropanol was added. The sample was mixed gently and put at -20°C overnight. Next, the sample was centrifuged 20 min at 13,000 rpm. The supernatant was removed, and the pellet was washed by adding 100 µl 70 % ethanol, followed by 2 min centrifugation at 13,000 rpm. The supernatant was removed and the pellet air dried for 15 min. The DNA was dissolved in 50 µl TE buffer. DNA for long insert libraries (10,000 bp, 20,000 bp) was extracted using the same protocol as used for *A. geniculata*, except that tissue was pooled from 100 spiders from the same colony and DNA was dissolved in distilled water.

Gland dissections for RNA sequencing. Venom glands and an opistosomal gland were dissected out from a single individual of tarantula. The spider was anaesthetized by putting it in a chamber with carbon dioxide for about 10 min until no more movements were observed. The spider was fixed

with needles and venom glands were dissected out. Before freezing at -80° the glands were quickly washed in a Ringer solution[16] to remove tissues and cells not from the venom glands.
Venom glands were dissected out from about 50 velvet spider individuals giving a total of about 100 glands. The glands were washed in the same Ringer solution as above and deposited on a glass plate placed on an ice block from a -80° freezer before they were put in an Eppendorf tube at -80°.

RNA library construction. All RNA libraries except the Tarantula whole body, were constructed using the Illumina mRNA-Seq Prep Kit. Briefly, oligo(dT) magnetic beads were used to purify polyA containing mRNA molecules. The mRNA was further fragmented and randomly primed during the first strand synthesis by reverse transcription. This procedure was followed by second-strand synthesis with DNA polymerase I to create double-stranded cDNA fragments. The double stranded cDNA was subjected to end repair by Klenow and T4 DNA polymerases and A-tailed by Klenow lacking exonuclease activity. Ligation to Illumina Paired-End Sequencing adapters, size selection by gel electrophoresis and then PCR amplification completed library preparation. Similarly, transcripts from the total RNA sample, originating from the whole bodies of tarantulas, were purified, broken in the presence of Zn^{2+}, and double-stranded cDNA synthesis was performed using random primers and RNaseH. After end repair and purification, the fragments were ligated with bar-coded paired-end adapters, and fragments with insert sizes of approximately 150-250 bp were isolated from an agarose gel and split in three. These were all amplified by PCR to generate DNA colonies template library and the libraries were then purified. Two of the libraries were normalized using two different normalization protocols. Library quality of all 3 samples was then assessed by a titration-run (1 x 50 bp) on an Illumina HiSeq 2000 instrument.

Genome assemblies. The following list of filters was used:

- Remove reads where Ns or polyA structures constitutes more than 20 percent of the read length for the large insert-size library data and 25 percent for the short ones.
- Remove low quality reads with quality score < 8 for >30 bases (large insert-size libraries) and >50 bases (short insert-size libraries).
- Remove reads with adapter contamination. Reads that aligns with >10bp to the adapter sequence with at most 2 mismatches were removed.
- Remove small insert size reads if the paired reads overlap more than 10 bp.
- Remove PCR duplicates defined as sets of paired reads with identical mapping positions.
- Trimming the first 3bp and last 4bp of paired end reads from the short insert-size library data.

Repeat-masking of the velvet spider. We searched the genome for tandem repeats with the help of software named Tandem Repeats Finder[17] (TRF). Transposable elements (TEs) were identified in the genome using a combination of homology-based and de novo approaches. Homology-based approach involves commonly used databases of known repetitive (for example repbase[18]), while de novo prediction approach generates a library of repetitive sequence.

1) Homolog based prediction

We use the known repbase[18] (composition of many TEs) to find the repeat. TEs in the genome assembly were identified both at the DNA and protein level. RepeatMasker[19] was applied for DNA-level identification using a custom library (a combination of Repbase, plant repeat database and our genome de novo TE library). At the protein level, RepeatProteinMask, updated software in the RepeatMasker package, was used to perform WuBlastX against the TE protein database.

2) De novo prediction

Firstly, we use two denovo softwares LTR_FINDER[20] and RepeatModeler to build de novo repeat library in base of genome. The softwares mentioned predict repeats in different ways: 1)full length LTR(Long terminal repeat retrotransposons) has typical structure and contain a ~18bp sequence complemented to the 3' tail of some tRNA, LTR_FINDER search the whole genome for the LTR typical structure; 2) At the heart of RepeatModeler are two de novo repeat finding programs(RECON[21] & RepeatScout[22]) which employ complementary computational methods for identifying repeat element boundaries and family relationships from sequences.

Then we filtered contamination and multicopy genes in the library. We classified this library and used it as the input library of RepeatMasker[19] and finally ran the software again to find homolog repeats in the genome.

Gene family evolution. Treefam's methodology was used.

1) BlastP was used on all the protein sequences against a database containing a protein dataset of all the species under E-value 1E-7, and conjoined fragmental alignments for each gene pairs by Solar. We assigned a connection (edge) between two nodes (genes) if more than 1/3 of the region aligned to both genes. An Hscore that ranged from 0 to 100 was used to weigh the similarity (edge). For two genes G1 and G2, the Hscore was defined as score (G1G2) / max(score(G1G1), score(G2G2)), the score here is the BLAST raw score.

2). Extracting gene families, i.e. clustering by Hcluster_sg. We used the average distance for the hierarchical clustering algorithm, requiring the minimum edge weight (Hscore) to be larger than 5, and the minimum edge density (total number of edges /theoretical number of edges) to be larger than 1/3. The clustering for a gene family would also stop if it already had one or more of the outgroup genes.

In CAFE, a random birth and death model was proposed to study gene gain and loss in gene families across a user-specified phylogenetic tree. A global parameter λ, which described both the gene birth (λ) and death (μ = -λ) rates across all branches of the tree for all gene families, was estimated using maximum likelihood. A graphical model can be used to calculate the most likely family size in the ancestral species, and this for each family CAFE calculates these so-called Viterbi assignments, and a comparison of these estimated sizes at all parent and descendant nodes allows one to infer the direction and size of change in gene family sizes along each branch. For each of the gene families in the data file, CAFE computes a *p*-value associated with the gene family sizes in the extant species given our model of gene family evolution. Branches with low *p*-values represent unusually large changes, either contractions or expansions. Families with conditional p-values less than the threshold (0.05) were considered to have an accelerated rate of expansion and contraction[23].

Collection of tissue samples for proteomics. Tarantulas were euthanized by incubation at -80°C for 30 min before the hemolymph was extracted with a needle and snap-frozen in liquid nitrogen. The organs were dissected from both the thorax and the abdomen, and snap-frozen separately in liquid nitrogen. Subsequently, both tissue samples and hemolymph were lyophilized overnight. Afterwards the thorax tissue was homogenized in liquid nitrogen using a mortar. Finally all samples were stored at -80°C. To obtain the "whole body" samples from the velvet spider, three individuals were dried overnight in a desiccator and homogenized in liquid nitrogen using a mortar. The tissue samples were subjected to SDS-PAGE as described in the Method section of the main paper, see also Supplementary Fig. 15.

Collection of venom samples for proteomics. Three tarantula spiders were anaesthetized by placing them in a chamber with carbon dioxide for about 10 min until no more movements were observed. The spiders were then placed on their back and immobilized. Plastic tubes were placed on the fang and electrical stimulation was applied to facilitate the release of venom. The venom was collected, snap-frozen in liquid nitrogen, and stored at -80°. The venom samples from the three individuals were not pooled. From the velvet spider, venom was extracted from six individuals. The spiders were anaesthetized in carbon dioxide for about 2 min or until no movements were observed. The spiders were immobilized and electrical stimulation was applied to facilitate the release of venom. Small droplets of venom on the tip of the fangs were collected by glass capillaries, and stored at -80°. The venom from the six individuals was not pooled. The venom samples were analyzed, as described in the Method section of the main paper, see also Supplementary Figs. 4 and 5.

Collection of silk samples for proteomics. Tarantula silk from three spiders kept in a vivarium was collected. The silk type was surface lining silk that is used to stabilize burrows, for prey sensing and to walk on. The silk was not pooled. In order to obtain velvet spider "whole web" silk three sets of 10 velvet spiders were placed into three clean boxes to build clean webs without prey remnants. The three spider webs were subsequently collected. The web is a mixture of more than one web type, and is referred to as 'whole web' in the text. Velvet spider dragline silk was collected by letting a spider attach a dragline to a piece of filter paper, and allowing it to walk. The end of the dragline was grabbed by forceps and attached to a reel. The reel was lifted, hanging the spider by its dragline, which was spun to the reel, as the spider extended it. A velvet spider egg case was taken from a colony, and the eggs removed. As the egg sac is attached to the nest part of the colony and is often moved and reoriented, it cannot be ruled out, that silk types other egg case silk, could be present in this sample. After collection, all silk samples were dissolved and treated with trypsin as described in the "In-solution treatment of silk and venom" paragraph in the Method section of the main paper.

LC-MS/MS analyses. In total, 194 LC-MS/MS analyses were performed, as outlined in Supplementary Table 3. These analyses were performed on a TripleTOF 5600 mass spectrometer (AB Sciex) coupled in-line with an EASY-nLC II system (Thermo Scientific). The trypsin digested samples were dissolved in 0.1% formic acid, injected, trapped and desalted isocratically on a ReproSil-Pur C18-AQ column (5 µm, 2 cm × 100 µm I.D; Thermo Scientific) after which the peptides were eluted from the trap column and separated on a home packed analytical ReproSil-Pur C18-AQ 3 µm capillary column (16 cm × 75 µm I.D) connected in-line to the mass spectrometer at 250 nL/min using a 50 min gradient from 5 % to 35 % phase B (0.1 % formic acid and 90 % acetonitrile). An Information dependent acquisition method was employed to automatically run experiments acquiring up to 50 MS/MS spectra per cycle using 2.8 s cycle times or up to 25 MS/MS spectra per cycle using 1.6 s cycle times both with an exclusion window of 6 s.

Settings and criteria for LC-MS/MS-based protein identification. The collected MS files were converted to Mascot generic format (MGF) using the AB SCIEX MS Data Converter beta 1.3 (AB SCIEX) and the "proteinpilot MGF" parameters. Mascot 2.3.02 (Matrix Science) was used to, based on the generated peak lists, identify proteins in the produced spider-protein databases[13]. For in-gel trypsin digested samples, propionamide was set as a fixed modification in the search parameters and one missed trypsin cleavage site was allowed. For the in-solution trypsin digested silk samples, a combination of CNBr and Trypsin cleavage was applied in the search parameters, and for the in-solution trypsin digested venom samples only trypsin was selected as enzyme. Both for in-solution trypsin digested silk and venom samples one missed cleavage was allowed and carbamidomethyl was used as a fixed modification in the search parameters. Oxidation of methionine residues was entered as variable modification for all searches. The mass accuracy of the precursor and product ions were set between 15 and 30 ppm and 0.2 Da, respectively, and the instrument setting was specified as ESI-QUAD-TOF. The significance threshold (p) was set at 0.01 and an ion score cut-off, set to the ion score homology value indicated by Mascot, was applied to all searches. Protein identifications were only excepted if they were based on two unique peptides. However, three exceptions from this rule were applied. The exceptions are related to the identification of velvet spider spidroins (see Supplementary Table 6), and to the identification of protoxins and to the identification of proteins used for gene-prediction support, as described below.

The LC-MS/MS data obtained from the in-solution trypsin digested venom samples were used to query the databases containing cysteine-rich protoxin sequences (see Supplementary Note 2 and Supplementary Data 9). Semi-trypsin was selected as enzyme, but apart from this, the search parameters were the same as in all other searches, as described above. But, in contrast to the main-part of the other identified proteins in this study, we accepted identifications based on only one peptide. The criterion for identification was changed to account for the small size of the mature cysteine rich peptide toxins. The ion score cut-off was set to 30 in these searches. Only toxins identified in at least three of the four (tarantula) or three of the eight (velvet spider) technical replica analyses were accepted. If the identification was based on only one peptide, the MS/MS spectrum was manually inspected and only accepted if an uninterrupted y- or b-ion series was present. One protein identified as "same sets" in the Mascot output from the velvet spider analyses was not included in the final number of identified toxins. The full list of identified toxins and the peptides used for the identifications are included in the Supplementary Data 1 and 2.

To support the gene prediction pipeline in relation to the velvet spider, MS files were searched against six frame translations of the transcriptomic derived sequences and the different gene prediction models (see Methods) of the genomic derived sequences with the same search criteria as above. For gene prediction, all protein hits identified based on at least one peptide with an ion score above the ion score homology value, indicated by Mascot, was accepted.

Settings and criteria for extracted ion-chromatography (XIC)-based protein quantification. The MS data was processed using the default settings from the ABSciex_5600.opt file except that the MS/MS Peak Picking "Same as MS Peak Picking" was deselected and "Fit method" was set to "Single Peak". After peak picking all scans, a Mascot search was performed using the same settings as for protein identification, as described above, except that the default average [MD] quantitation protocol was selected using a significance threshold at 0.01, number of peptides used for quantitation was 3, matched rho was 0.7, XIC threshold was 0.1 and isolated precursor threshold was set at 0.5.

In the quantification of tarantula silk and velvet spider whole web silk, three biological samples with three technical replicates, were analyzed. The quantification of the proteins, in these analyses, was only reported if they were detected by three quantifiable peptides in at least two of the three technical replicas for all three biological replicas. If a peptide is shared between two proteins, the intensity of the peptide is, if the peptide is among the top three most intense peptides for both proteins, used for quantification of both proteins, when the Distiller software is used. It should be taken into consideration that this potentially could influence the quantification and abundance ranking of the silk proteins, since some of the spidroin sequences contain identical trypsin/CNBr-generated peptides. Furthermore, in the pellet, present after the combined CNBr-, acid-, and trypsin-treatment of silk, proteins are likely to be present, and certain proteins, e.g. certain spidroins, might be more difficult to solubilize and digest, than other proteins. Therefore, the results of the quantitative analyses of silk should only been seen as a rough indication of the actual amount of the different silk proteins.

In the analyses of venom, one sample (from each species) was generated based on pooling of three individuals. Eight technical replicas were performed on the pooled velvet spider sample, and four technical replicas were obtained for the tarantula venom sample. Proteins were only accepted if they could be quantified (based on three quantifiable peptides) in at least three of the LC-MS/MS analyses. The characterization of the quantified proteins in tarantula venom revealed two sequences representing the N- and C-terminal of the same protein (the venom hyaluronidase). These sequences were merged and the three most intense peptides after merging of the sequences were manually identified and included in the calculations of the relative abundance

Supplementary references

1 Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, I351-I358, doi:10.1093/bioinformatics/bti1018 (2005).

2 Caratozzolo, S. I ragni giganti. (2000).

3 Kraus, O. & Kraus, M. in *Verhandlungen des Naturwissenschaftlichen Vereins in Hamburg (NF) 30* (ed Otto Kraus) 151-254 (Verlag Paul Parey, 1988).

4 Majer, M., Svenning, J. C. & Bilde, T. Habitat productivity constrains the distribution of social spiders across continents - case study of the genus Stegodyphus. *Frontiers in zoology* **10**, 9, doi:10.1186/1742-9994-10-9 (2013).

5 Lubin, Y. & Bilde, T. in *Advances in the Study of Behavior, Vol 37* Vol. 37 *Advances in the Study of Behavior* (eds H. J. Brockmann *et al.*) 83-145 (Elsevier Academic Press Inc, 2007).

6 Tang, X. *et al.* Molecular diversification of peptide toxins from the tarantula *Haplopelma hainanum* (*Ornithoctonus hainana*) venom based on transcriptomic, peptidomic, and genomic analyses. *J Proteome Res* **9**, 2550-2564, doi:10.1021/pr1000016 (2010).

7 Garb, J. E., Ayoub, N. A. & Hayashi, C. Y. Untangling spider silk evolution with spidroin terminal domains. *BMC Evol. Biol.* **10**, doi:10.1186/1471-2148-10-243 (2010).

8 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics* **5**, 1-19, doi:10.1186/1471-2105-5-113 (2004).

9 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772-772 (2012).

10 MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space (Systematic Biology 61 (3): 539-542, 2012).

11 Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer (2007).

12 Rambaut, A. *FigTree, http://tree.bio.ed.ac.uk/software/figtree/*, 2007).

13 Tamura, K. *et al.* MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* **28**, 2731-2739, doi:10.1093/molbev/msr121 (2011).

14 Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451-1452, doi:10.1093/bioinformatics/btp187 (2009).

15 Mattila, T. M., Bechsgaard, J. S., Hansen, T. T., Schierup, M. H. & Bilde, T. Orthologous genes identified by transcriptome sequencing in the spider genus Stegodyphus. *BMC Genomics* **13**, doi:10.1186/1471-2164-13-70 (2012).

16 Muller, H. M. IONIC CONCENTRATIONS, OSMOLARITY AND PH OF THE HEMOLYMPH OF THE COMMON HOUSESPIDER TEGENARIA-ATRICA KOCH,CL (AGELENIDAE, ARACHNIDA). *Comparative Biochemistry and Physiology a-Physiology* **87**, 433-437, doi:10.1016/0300-9629(87)90148-4 (1987).

17 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580 (1999).

18 Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-467, doi:10.1159/000084979 (2005).

19 Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 10, doi:10.1002/0471250953.bi0410s25 (2009).

20 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265-268, doi:10.1093/nar/gkm286 (2007).

21 Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* **12**, 1269-1276, doi:10.1101/gr.88502 (2002).

22 Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**, i351-358, doi:10.1093/bioinformatics/bti1018 (2005).

23 Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome research* **15**, 1153-1160, doi:10.1101/gr.3567505 (2005).