OPEN ACCESS Freely available online

PLoS one

# Blueprint for a High-Performance Biomaterial: Full-Length Spider Dragline Silk Genes

**Nadia A. Ayoub*, Jessica E. Garb, Robin M. Tinghitella, Matthew A. Collin, Cheryl Y. Hayashi**

Department of Biology, University of California Riverside, Riverside, California, United States of America

**Spider dragline (major ampullate) silk outperforms virtually all other natural and manmade materials in terms of tensile strength and toughness. For this reason, the mass-production of artificial spider silks through transgenic technologies has been a major goal of biomimetics research. Although all known arthropod silk proteins are extremely large (>200 kiloDaltons), recombinant spider silks have been designed from short and incomplete cDNAs, the only available sequences. Here we describe the first full-length spider silk gene sequences and their flanking regions. These genes encode the MaSp1 and MaSp2 proteins that compose the black widow's high-performance dragline silk. Each gene includes a single enormous exon (>9000 base pairs) that translates into a highly repetitive polypeptide. Patterns of variation among sequence repeats at the amino acid and nucleotide levels indicate that the interaction of selection, intergenic recombination, and intragenic recombination governs the evolution of these highly unusual, modular proteins. Phylogenetic footprinting revealed putative regulatory elements in non-coding flanking sequences. Conservation of both upstream and downstream flanking sequences was especially striking between the two paralogous black widow major ampullate silk genes. Because these genes are co-expressed within the same silk gland, there may have been selection for similarity in regulatory regions. Our new data provide complete templates for synthesis of recombinant silk proteins that significantly improve the degree to which artificial silks mimic natural spider dragline fibers.**

Citation: Ayoub NA, Garb JE, Tinghitella RM, Collin MA, Hayashi CY (2007) Blueprint for a High-Performance Biomaterial: Full-Length Spider Dragline Silk Genes. PLoS ONE 2(6): e514. doi:10.1371/journal.pone.0000514

## INTRODUCTION

Spider silks have received much economic and biomedical attention because of their outstanding mechanical properties [e.g. 1–3]. For example, the dragline silk of araneoids (ecribellate orb-weaving spiders and their relatives) displays both high tensile strength and extensibility, making it tougher than nearly all other natural or synthetic materials [4–6]. Spider silks are primarily composed of proteins that are synthesized in specialized abdominal glands. An individual orb-weaving spider spins up to five different types of silk fibers, each serving critical ecological functions, including prey capture, shelter, predator avoidance, egg protection, and dispersal [7–8]. Each distinct fiber is made from one or two unique types of silk structural proteins (fibroins), almost all of which are encoded by members of a single gene family [9–12]. Thus, the spectacular diversity of spider silk proteins evolved through successive rounds of gene duplication and divergence.

Spider fibroins have very high molecular weights, estimated at 200–350 kiloDaltons [13] with transcript sizes of approximately 10,000 base pairs (bp) or larger. Such considerable size is conserved over a diverse range of spider species and fibroin types [13–16]. Partial-length complementary DNA (cDNA) sequences indicate that silk proteins are highly modular; each polypeptide is primarily composed of an uninterrupted block of repetitive sequence that is flanked on both sides by ~100 amino acids (aa) of non-repetitive amino- (N-) and carboxy- (C-) termini. The sequence attributes of the repetitive region vary according to silk protein type, with some fibroins containing short, simple repeat units, and others composed of longer, more complex repeats [10,14]. Because of the difficulty associated with cloning long stretches of repetitive DNA, only two full-length cDNA silk sequences have been characterized [16]. These cDNAs encode the silk proteins that form the egg case fibers of the orb-weaving spider, *Argiope bruennichi*. Egg case fibers, however, have substantially lower tensile strength and toughness than dragline silk [8,16]. Complete gene sequences are still unknown for any spider silk.

Because of its extremely high tensile strength and toughness, dragline (major ampullate) silk has received the most attention of the spider silks. This silk is composed of two types of fibroins, MaSp1 [17] and MaSp2 [18]. The genes encoding these proteins are co-expressed in the major ampullate silk glands, and both proteins are found throughout the fiber [15,19]. Short glycine-rich regions (GGX, where X represents a subset of aa) followed by a stretch of multiple alanines (poly-A) characterize both proteins. The ubiquitous poly-A stretches are hypothesized to form hydrophobic crystalline domains that are responsible for the high tensile strength of the fiber [20–24]. In contrast, the glycine-rich regions are hydrophilic with runs of the peptide motif GGX conforming to a $3_1$-helix [21,25]. While poly-A and GGX motifs describe almost all of MaSp1, MaSp2 also has a large proportion of GPG motifs [18]. These proline-containing repeats likely form type II beta-turns, and such kinks in part explain the reversible extensibility of dragline fiber [13,19,26–27].

Much of the applied research on spider silks has focused on mass-producing silk fibers for industrial use [e.g. 28]. However, unlike domesticated silkworm caterpillars, spiders cannot be

**Academic Editor:** Robert DeSalle, American Museum of Natural History, United States of America

**Received** January 9, 2007; **Accepted** May 12, 2007; **Published** June 13, 2007



**Funding:** The research was funded by the Army Research Office (DAAD19-02-1-0358 and W911NF-06-1-0455) and the National Science Foundation (DEB-0236020).

**Competing Interests:** The authors have declared that no competing interests exist.

* **To whom correspondence should be addressed.** E-mail: nadiaa@ucr.edu

readily farmed for silk because they are predatory and cannibalistic. Instead, researchers have created biomimetic silks through transgenic technologies using partial-length silk cDNA sequences [e.g. 29–32]. These manmade silks have, thus far, fallen short of native dragline silk in both sequence and mechanical properties [e.g. 30,33–35]. All known arthropod fibroins are extremely large, including the convergently evolved heavy-chain fibroin of lepidopterans (~370 kiloDaltons [36]). Such evolutionary convergence among distantly related silk spinning species suggests that large size is a critical molecular feature for silk fiber mechanical performance. For example, larger fibroins possess more repeat units (such as poly-A motifs that crosslink to form crystalline domains) than shorter fibroins, thereby increasing the number of interactions among monomers. However, large size is probably not the only key functional attribute of spider fibroins. The evolutionarily conserved C-termini of MaSp1 and MaSp2 aid in conversion of the liquid silk dope into a solid fiber [37], and facilitate assembly of the fiber's characteristic crystalline domains [35]. The few known N-termini are even more conserved than the C-termini [38–39], which suggests that N-termini are also necessary for proper fiber assembly and may influence the mechanical properties of spider silk. Thus, determining the entire coding sequences for MaSp1 and MaSp2 is a key step in the generation of recombinant silks that closely mimic natural spider dragline silk.

Full-length silk sequences are also crucial for understanding the molecular evolutionary dynamics of long, repetitive genes. To date, it has been assumed that the modular organization of fibroins is maintained throughout the entirety of the amino acid sequence. Yet, some molecular evidence is consistent with past recombination events between *MaSp1* and *MaSp2* [10,12,40], and thus repeat units may have transferred from one gene to the other. Moreover, there is sparse information about the exon-intron structure of spider silk genes. *Flag* (the gene encoding the capture spiral filament) from *Nephila clavipes* [41] and *MaSp2* from *Argiope trifasciata* [38] both show highly repetitive exon-intron gene organizations in which sequential introns within a gene have nearly identical nucleotide sequences. The presence of iterated introns and exons is evidence that intragenic recombination can swiftly homogenize (eliminate or spread new variants throughout) the sequence of an entire silk gene. It is unknown if this atypical gene architecture is a common feature of spider silk genes.

We have constructed a genomic library for the black widow spider, *Latrodectus hesperus* (Theridiidae), in order to identify full-length silk genes and their associated regulatory regions. Black widows, notorious for their neurotoxic venom, are members of the Araneoidea, a superfamily of orb-weaving spiders and their close relatives. Black widows are descended from orb-web weaving ancestors, but they build three-dimensional cobwebs rather than the symmetrical, wagon-wheel shaped orb-web [42]. Despite this difference in web architecture, the breaking strength and extensibility of *Latrodectus* dragline silk are equal to or higher than those of true orb-weaving spiders [43–45]. Here we report complete gene sequences of *MaSp1* and *MaSp2* as well as adjacent non-coding regions. We document the existence of higher-order repeat units that range from ~70 to over 2,000 bp, and show that the repetitive sequences of *MaSp1* are more homogenized than those of *MaSp2*. We also demonstrate marked evolutionary conservation of N-terminal and upstream non-coding regions between paralogs within a species and across orthologs from divergent species. Based on these multi-gene comparisons, we identify putative regulatory sequences that may be involved in co-expression of the two major ampullate silk genes. Collectively, our data provide the first templates for complete recombinant major ampullate fibroins and illustrate the dramatic effects of intragenic and intergenic recombination in the evolution of these extraordinarily modular genes.

## RESULTS

### Large, single exon gene organizations

We sequenced two fosmid clones each containing ~37,000 bp of the black widow genome. One clone (GenBank accession EF595246) encompassed the complete coding sequence for the dragline silk gene *MaSp1* as well as 9,928 bp upstream of its start codon and 14,728 bp downstream of its stop codon. The *MaSp1* gene is composed of a single exon with 9,390 bp encoding 3,129 aa (Figure 1). The second clone (EF595245) includes the entire coding sequence for *MaSp2* plus 17,205 bp of upstream and 8,546 bp of downstream flanking sequence. Like *MaSp1*, the *MaSp2* gene contains one enormous exon with 11,340 bp encoding 3,779 aa (Figure 2). Both *MaSp1* and *MaSp2* genes contain sequences that match partial-length cDNAs from *L. hesperus* silk gland expression libraries [40,43], indicating that these genes are transcribed. The C-terminal coding region (~300 bp) of the *MaSp1* gene is 97% identical to the corresponding 3′ partial *MaSp1* cDNA clones (AY953074, DQ409057) and the N-terminal coding region (~450 bp) is 99.8% identical to our 5′ partial cDNA clone (EF595247). Both the C-terminal coding region and the 3′ untranslated region (UTR) of the *MaSp2* gene share 99% sequence identity with 3′ partial *MaSp2* cDNA clones (AY953075, DQ409058). Similarly, the N-terminal coding regions of the *MaSp2* gene and our 5′ partial cDNA (EF595248) are 95.5% identical.

### Extreme sequence modularity

Glycine and alanine are by far the most abundant amino acids in our predicted *L. hesperus* MaSp1 and MaSp2 fibroins. These two amino acids constitute greater than 64% of both sequences, followed by glutamine in MaSp1 and proline in MaSp2 (Table 1). These values closely match published amino acid compositions of major ampullate silk from black widows [46] and other araneoid spiders [47–48], further confirming that our genes encode the two dominant protein components of major ampullate silk. Because the first two codon positions for alanine, glycine, and proline are guanine or cytosine, the base compositions of these genes are guanine/cytosine-rich (*MaSp1*–61%; *MaSp2*–59%). However, overall base compositions are not highly skewed because the third positions for these codons in the *L. hesperus MaSp1* and *MaSp2* are extremely biased towards adenine and also strongly biased, but less dramatically, towards thymine (86% of *MaSp1*, 91% of *MaSp2* glycine, alanine, and proline codons end with adenine or thymine; Table 1).

The repetitive region of the *L. hesperus MaSp1* translation is dominated by amino acid sequence motifs commonly found in MaSp1 of other spider species: GGX (X = A, Q, or Y), GX (X = Q, A, or R), and poly-A (4–10 consecutive alanines, mean number = 7.7) [10,17,49]. These motifs are organized into four types of ensemble (higher order) repeat units, with each ensemble consisting of a glycine-rich region followed by a poly-A region (Figure 1). Starting at residue 542, the different ensemble types are tandemly arrayed in a consistent pattern, and this aggregate of four ensembles is iterated 20 times with near perfect fidelity. Pairwise amino acid differences between aggregates are extremely low, ranging from 0.0 to 4.3% and averaging 1.9%. This remarkable sequence homogeneity is also maintained at the nucleotide level with average uncorrected pairwise differences of only 2.5% (range = 0.3–6.3%).


**Figure 1. Complete amino acid sequence for *L. hesperus* major ampullate spidroin 1 (MaSp1).** The sequence is read from left to right and then top to bottom. The diamond marks the start position and the asterisk denotes the stop position. The protein is dominated by poly-A (red) and GGX (green) motifs. The majority of the sequence can be categorized into four types of ensemble repeat units. Repeats of each type are aligned within a box. Gaps (-) have been inserted in order to align repeat units within a type.
doi:10.1371/journal.pone.0000514.g001

The repetitive region of the *L. hesperus* MaSp2 amino acid sequence is characterized by a larger suite of motifs than MaSp1. The common MaSp2 motifs include GPX (X = G or S), QQ, GGX (X is usually A), GSG, and poly-A (3–9 consecutive alanines, mean number = 6.7). Similar to MaSp1, these motifs are organized into four types of ensemble repeat units that each contain one poly-A motif (Figure 2). However, the four types of MaSp2 ensembles are more variable than those of MaSp1, with pairwise amino acid differences between ensembles of the same type as high as 36% (Table 2). In addition, the MaSp2 ensemble types are not always strung together in the same order and do not form clearly discernible higher-level aggregates (Figure 2). Nevertheless, there is a pair of 778 amino acid long tandem repeats that differ by a scant five aa (Figure 2). The 2,334 nucleotides encoding each repeat vary at only six positions (>99.7% identity).

Using the method of Kyte and Doolittle [50], we predicted the hydrophilicity of *L. hesperus* MaSp1 and MaSp2. Both fibroins show regions of hydrophobicity (corresponding to the poly-A motifs) and hydrophilicity (corresponding to the glycine-rich regions) that alternate throughout the entire repetitive portions of the two proteins (Figure 3A). Both MaSp1 and MaSp2 repetitive regions are slightly hydrophilic when averaged across all residues (MaSp1 average = 0.13 on the Kyte-Doolittle scale; MaSp2 average = 0.14) but MaSp2 displays higher amplitude of hydrophilicity (MaSp1 max = 2.0; MaSp2 max = 2.6) (Figure 3A). The N- and C-terminal domains show a similar pattern of alternating hydrophobicity and hydrophilicity but are generally more hydrophobic than the repetitive regions (average hydrophilicity across residues: MaSp1 N-terminus = -0.29, C-terminus = −0.44; MaSp2 N-terminus = −0.34; C-terminus = −0.31; negative values indicate degree of hydrophobicity). The most hydrophobic region of both fibroins is found at the beginning of the N-terminus (Figure 3B).

## Congruence between silk N- and C-termini evolutionary relationships

We aligned the N-terminal regions of *L. hesperus* MaSp1 and MaSp2 with N-termini from other spider fibroins (Figure 4A). These proteins are constituents of three spider silk fiber types: the dragline silk composed of MaSp1 and MaSp2, the capture spiral filament of flagelliform silk protein (Flag), and the eggcase fibers produced from tubuliform (also called cylindrical) gland proteins (TuSp1, CySp1 and CySp2). For each of these N-termini, we also aligned the corresponding C-termini, if available (Figure 4B). However, only in the case of the *L. hesperus MaSp1* and *MaSp2* that we report here, and the full-length *CySp1* and *CySp2* cDNAs from *Argiope bruennichi*, is it certain that the N- and C-termini coding regions belong to the same gene. All others were partial 5′ or 3′ sequences that were assumed to represent the ends of the same gene.

We assessed evolutionary relationships among the N-termini and C-termini encoding sequences using maximum likelihood (ML), maximum parsimony (MP), and Bayesian phylogenetic

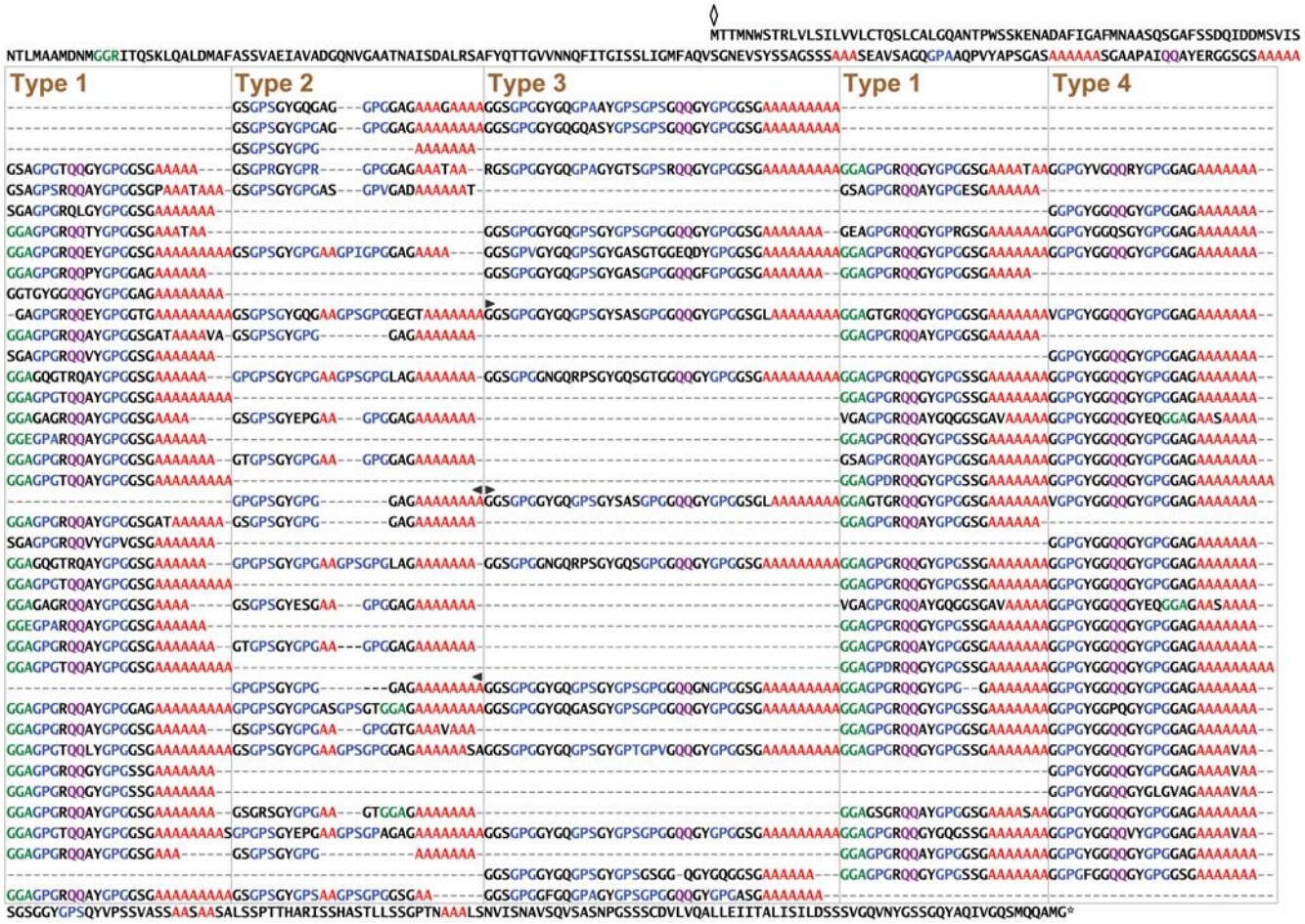


**Figure 2. Complete amino acid sequence for *L. hesperus* major ampullate spidroin 2 (MaSp2).** The sequence is read from left to right and then top to bottom. Start position, stop position, and alignment gaps are indicated as for MaSp1 (Figure 1). MaSp2 is characterized by poly-A (red), GGX (green), GPX (blue), and QQ (purple) motifs. There are four types of ensemble repeats. Repeats of each type are aligned within a box, except for Type 1, which is separated into two boxes because it is approximately twice as prevalent as any other repeat type. Right and left pointing arrows mark beginnings and ends of two near-perfect repeats of 778 aa.
doi:10.1371/journal.pone.0000514.g002

methods. All methods produced similar relationships among N-terminal sequences (Figure 4C). A clade of eggcase silks (TuSp1, CySp1 and CySp2) was always well-supported. A major ampullate silk clade (MaSp1 and MaSp2) was consistently recovered with greater than 90% bootstrap support and 100% posterior probability. Despite the distinct differences between the repetitive portions of MaSp1 and MaSp2 (Figures 1, 2, 4D), all N-termini analyses strongly supported a grouping of *Latrodectus* MaSp1 and MaSp2, rather than a multi-species MaSp1 clade that is distinct from a MaSp2 clade. Within *Latrodectus*, however, *L. hesperus* MaSp1 grouped with *L. geometricus* MaSp1.

Relationships among the corresponding C-terminal encoding sequences typically mirrored those of the N-terminal encoding sequences (Figure 4C). However, placement of the *L. hesperus* TuSp1 C-terminus was unstable. Depending on the type of analysis, it grouped with Flag, CySp1 and CySp2, or MaSp1 and MaSp2, but always with low support (less than 75% bootstrap support or 95% posterior probability). A MaSp1 and MaSp2 C-terminal clade was consistently recovered with high support values, and nested within it, a *Latrodectus* MaSp1 and MaSp2 subclade. The only difference among analyses was that *L. hesperus* MaSp1 grouped with either *L. geometricus* MaSp1 (ML and Bayesian trees) or *L. hesperus* MaSp2 (in the MP trees).

## Multi-species comparisons identify conserved non-coding sequences

Phylogenetic footprinting is a powerful approach for discovering putative gene regulatory regions. This method generally relies on alignments of orthologous, non-coding sequences from multiple species [51]. The presence of conserved non-coding nucleotide stretches implies that a region is under selective constraint and therefore is likely to perform an important function. A similar approach can be applied to the non-coding sequences of co-regulated genes [52]. We thus compared flanking sequences of *L. hesperus MaSp1* and *MaSp2*, paralogous genes which are simultaneously expressed [15,19]. We also analyzed these sequences with available flanking sequences of *MaSp1* and *MaSp2* from other spider species. Because the *L. hesperus MaSp2* clone contained another open reading frame (ORF) 2,611 bp upstream of the *MaSp2* start codon, we limited comparisons to ~2,500 bp of upstream sequence. Using MultiPipMaker [53], we identified regions that could be reliably aligned among *L. hesperus MaSp1* and *MaSp2, L. geometricus MaSp1* (5′: DQ059133S1, 3′: DQ059133S2), *Argiope trifasciata MaSp2* (DQ059136), and *Nephila inaurata madagascariensis MaSp2* (DQ059135; only 700 bp upstream sequence available). Downstream genomic sequences were not available for

**Table 1.** Amino acid content and codon usage for the most common amino acids of black widow MaSp1 and MaSp2.

| Amino Acid | Codon | MaSp1 | | MaSp2 | |
|---|---|---|---|---|---|
| | | % aa | % codon | % aa | % codon |
| Glycine | GGA | 42.3 | 54 | 33.5 | 65 |
| | GGT | | 38 | | 30 |
| | GGC | | 7 | | 4 |
| | GGG | | 1 | | 1 |
| Alanine | GCA | 32.7 | 59 | 31.1 | 66 |
| | GCT | | 18 | | 18 |
| | GCC | | 17 | | 7 |
| | GCG | | 6 | | 9 |
| Glutamine | CAA | 11.3 | 98 | 6.9 | 97 |
| | CAG | | 2 | | 3 |
| Proline | CCA | 0.4 | 69 | 8.6 | 64 |
| | CCT | | 23 | | 33 |
| | CCC | | 8 | | 1 |
| | CCG | | 0 | | 2 |

doi:10.1371/journal.pone.0000514.t001

*Argiope* and *Nephila MaSp2*. MultiPipMaker generates local alignments using the BLASTZ algorithm and only produces an alignment if identity among sequences exceeds a threshold, below which alignments are considered random [54−55]. Margulies et al. [56] argued that pairwise alignments are unreliable for detecting regulatory elements. Thus, we focused on conserved regions found in at least three sequences. When attempting to align only upstream non-coding sequence, MultiPipMaker produced alignments among *Latrodectus* sequences but not between *Latrodectus* and *Argiope* or *Nephila*. When the coding sequences were included as an anchor, a span of ~90 bp directly upstream of the start codon could be aligned among all 5 genes. This region included the conserved motif CACG and the TATA box, which were also identified by Motriuk-Smith et al. [38]. While the TATA box is thought to guide RNA polymerase II to the transcription initiation site in many eukaryotic genes, the motif CACG represents a potentially novel regulatory element for spider silk genes. Approximately 150 bp of sequence upstream from the start codon could be aligned among the three *Latrodectus* genes and ~300 bp upstream sequence between *L. hesperus MaSp1* and *MaSp2*. Additionally, ~180 bp of sequence downstream of the stop codon could be aligned among all three *Latrodectus* genes.

We further investigated the regions of similarity identified among the *Latrodectus* non-coding sequences by creating global alignments of the ~300 bp region upstream of the start codon and of the ~180 bp segment downstream of the stop codon. In addition to the CACG motif and TATA box found among all sequences examined, the three *Latrodectus* upstream sequences share a 15 bp motif found ~110 bp upstream of the start codon that has only 2 variable positions. When scanned against the TRANSFAC database [57], this conserved region perfectly matches a 6 bp binding site for the Achaete-Scute family of transcription factors.

We also compared nucleotide substitution rates for various regions of the *Latrodectus* sequences (Figure 5). To detect selection on protein coding sequences, we calculated the ratio of the number of nonsynonymous substitutions per nonsynonymous site ($K_n$) to the number of synonymous substitutions per synonymous site ($K_s$) [58]. As expected for evolutionarily conserved proteins, we found $K_n/K_s$ was very low, ranging from 0.05 to 0.20 for *Latrodectus MaSp1* and *MaSp2* terminal coding regions, suggesting strong purifying selection (Figure 5). We applied an analogous approach (as in Wong&Nielsen [59]) to estimate selective pressures in non-coding sequences by calculating the ratio of the number of substitutions per site (K) to $K_s$ for the adjacent coding sequence. We found $K_{(150\text{ bp upstream})}/K_{s(\text{N-terminus})}$ ranged from 0.26 to 0.63, which is higher than for coding sequence but still considerably less than 1. In contrast, $K_{(300-150\text{ bp upstream})}/K_{s(\text{N-terminus})}$ ranged from 0.82 to 1.45 (Figure 5), suggesting that the 150 bp directly upstream of coding sequence are under selective constraints while regions farther upstream are not. We also found $K_{(3'\text{ UTR})}/K_{s(\text{C-terminus})} = 0.27$ for *L. hesperus MaSp1* and *MaSp2*, consistent with strong purifying selection on the 3′ UTR.

**Table 2.** Prevalence (#) and average pairwise amino acid differences between MaSp2 ensemble repeats of the same type.

| Ensemble Type* | # | Average % aa difference (min-max) |
|---|---|---|
| 1 | 62 | 11.8 (0.0–36.0) |
| 2 | 24 | 11.7 (0.0–28.0) |
| 3 | 16 | 11.4 (0.0–22.0) |
| 4 | 30 | 5.6 (0.0–20.8) |

*Ensemble repeat types shown in Figure 2.

doi:10.1371/journal.pone.0000514.t002

## Global comparisons of genomic clones

We compared the entire clones containing *MaSp1* (34,046 bp) and *MaSp2* (37,092 bp) using MultiPipMaker and the global alignment program AVID [60]. We also compared the flanking sequences of the genes using BLASTN [61] to search for repetitive elements in the *L. hesperus* genome. As expected, the N- and C-terminal coding regions are significantly conserved between the two genes (Figure 6). Within the genes themselves, there are also multiple regions of significant similarity at the DNA level. These regions correspond to the poly-A, GG, GGXG and GQ motifs found in both proteins. Additionally, there were numerous significant matches between regions of non-silk-protein-coding sequence. Each of these regions, when translated, was similar to transposable elements in the NCBI nr protein database (based on BLASTX [61] scores: $E<e^{-10}$). Most notably, there is a significantly conserved region spanning ~700 bp that is found ~10,000 bp downstream of the *MaSp1* and *MaSp2* ORFs (Figure 6). The translated sequence of this region from the *MaSp2* clone significantly matched TCb1-transposase. The translated sequence from the *MaSp1* clone significantly matched gag-pol polyprotein, which contains a retrotransposon. Although both clones contain ORFs in this region, they do not encode full-length proteins. Thus, these genomic regions appear to be inactive transposable elements.

# DISCUSSION

## Gene structure

Black widow dragline silk is an exceptionally tough biomaterial, even compared to the high-performance draglines spun by other spiders [43,45]. Here we report the complete gene sequences for the MaSp1 and MaSp2 fibroins that form this silk. We found that both genes lack introns and thus *MaSp1* and *MaSp2* each possess only one enormous exon containing either 9,390 bp (*MaSp1*) or 11,340 bp (*MaSp2*) of coding sequence. No other full-length spider

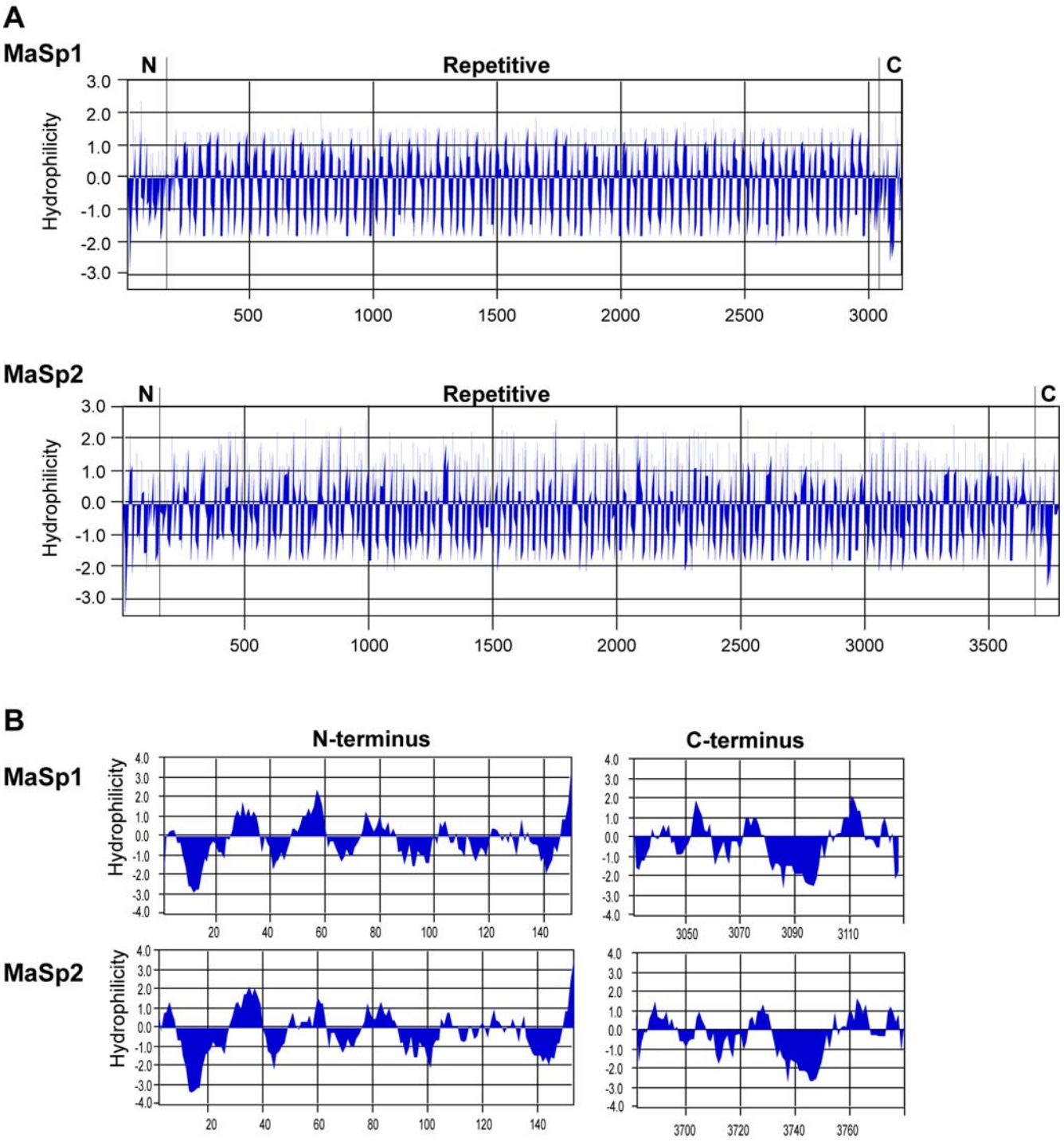


**Figure 3. Kyte and Doolittle [50] hydrophilicity plots for *L. hesperus* MaSp1 and MaSp2.** Scan window size = 7. Negative values indicate hydrophobicity. (A) Complete proteins. (B) Non-repetitive terminal regions.
doi:10.1371/journal.pone.0000514.g003

silk genes have been characterized, but the few known partial-length gene sequences fit into two categories of exon-intron structure. First, based on *L. geometricus MaSp1* and *Nephila MaSp2* fragments [38], and the full-length genes described here, some silk genes are composed of single exons. Second, *Nephila Flag* and *Argiope MaSp2* have introns that are peculiar because successive introns within the same gene are nearly identical in sequence [38,41]. Thus, all known spider silk genes have unusual architectures.

In eukaryotes, proteins encoded by single exons are rare and strongly biased towards sizes much smaller (<1,000 aa, [62–63]) than the spider silk proteins (>3,000 aa). Intronless genes may reflect one type of gene duplication process that led to the diversification of the spider silk gene family; retroposition of mRNA transcripts (inherently intronless) into the genome can give rise to functional gene duplicates [64]. Alternatively, intronless genes may be selectively favored. Intron length is negatively

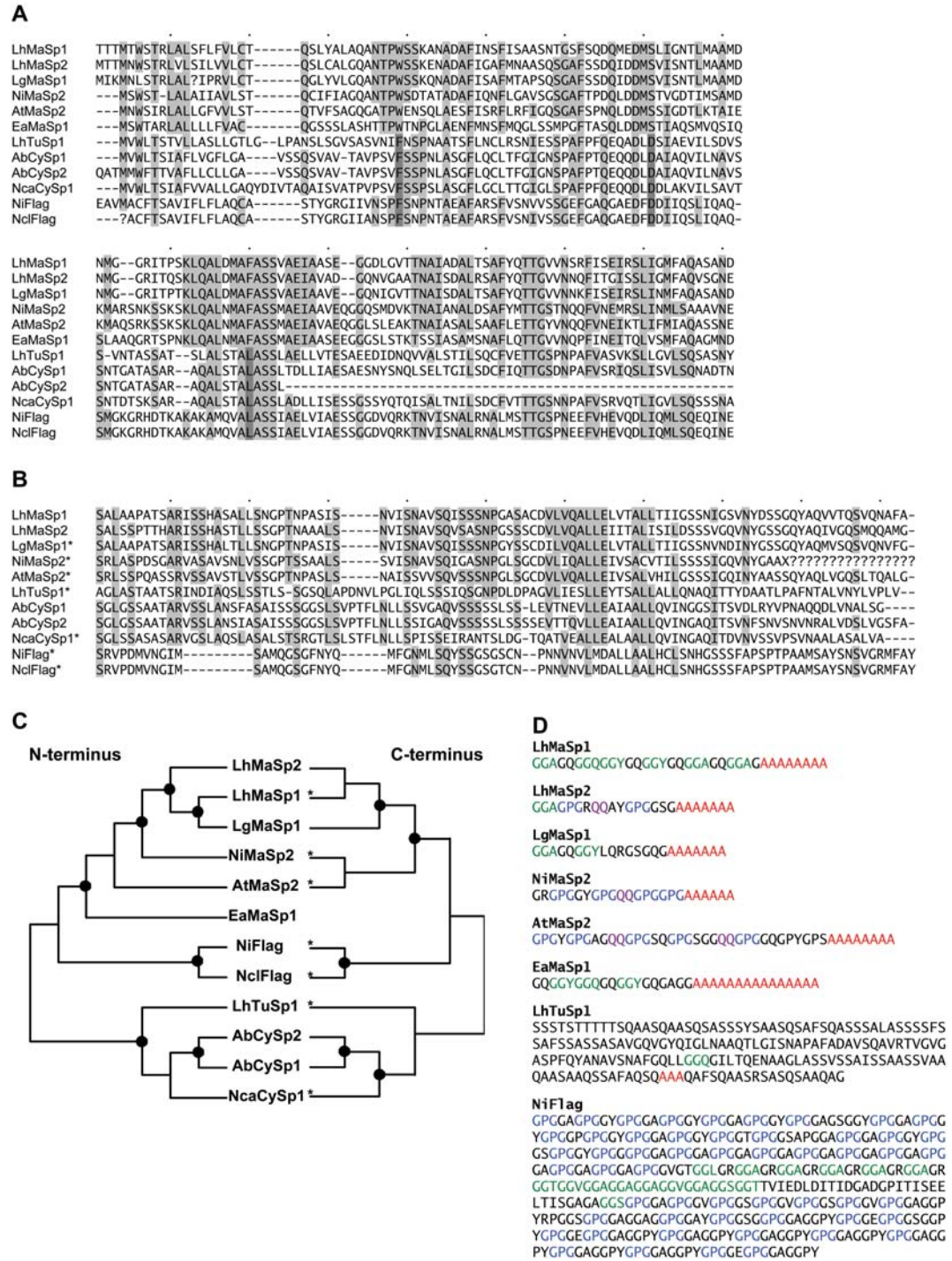


**Figure 4. Comparison of N-termini, C-termini and repeat units of spider silk proteins.** (A) Alignment of published N-terminal amino acid sequences. Amino acids shared by ≥50% of proteins are highlighted in grey. Gaps are represented by dashes and missing characters by question marks. (B) Alignment of corresponding C-terminal amino acid sequences. Taxa with an asterisk result from partial sequencing and are presumed to belong to the same locus as the N-terminal sequences. (C) MP trees of N and C-terminal encoding sequences treating gaps as a fifth state and employing midpoint rooting. Left tree length = 1449 (N-terminus); Right tree length = 838 (C-terminus). Dots represent nodes with >75% bootstrap support in all MP and ML analyses and >95% Bayesian posterior probability. (D) Exemplar repeat units for each of the major ampullate fibroins and representative TuSp1 and Flag repeats. Amino acid motifs are colored as in Figure 2. Abbreviations: LhMaSp2, *Latrodectus hesperus MaSp2* (EF595245); LhMaSp1, *L. hesperus MaSp1* (EF595246); LgMaSp1, *L. geometricus MaSp1* (5′ sequence: DQ059133S1, 3′ sequence: DQ059133S2); NiMaSp2, *Nephila inaurata madagascariensis MaSp2* (5′ sequence: DQ059135, 3′ sequence: AF350278); AtMaSp2, *Argiope trifasciata MaSp2* (5′ sequence: DQ059136, 3′ sequence: AF350266); EaMaSp1; *Euprosthenops australis MaSp1* (AM259067); LhTuSp1, *L. hesperus TuSp1* (5′ sequence: DQ379383, 3′ sequence: AY953070); AbCySp1, *A. bruennichi CySp1* AB242144; AbCySp2, *A. bruennichi CySp2* (AB242145); NcaCySp1, *N. clavata CySp1* (5′ sequence: AB218974, 3′ sequence: AB218973); NiFlag, *N. i. madagascariensis Flag* (5′ sequence: AF218623S1, 3′ sequence: AF218623S2); NclFlag, *N. clavipes Flag* (5′ sequence: AF027972, 3′ sequence: AF027973).
doi:10.1371/journal.pone.0000514.g004

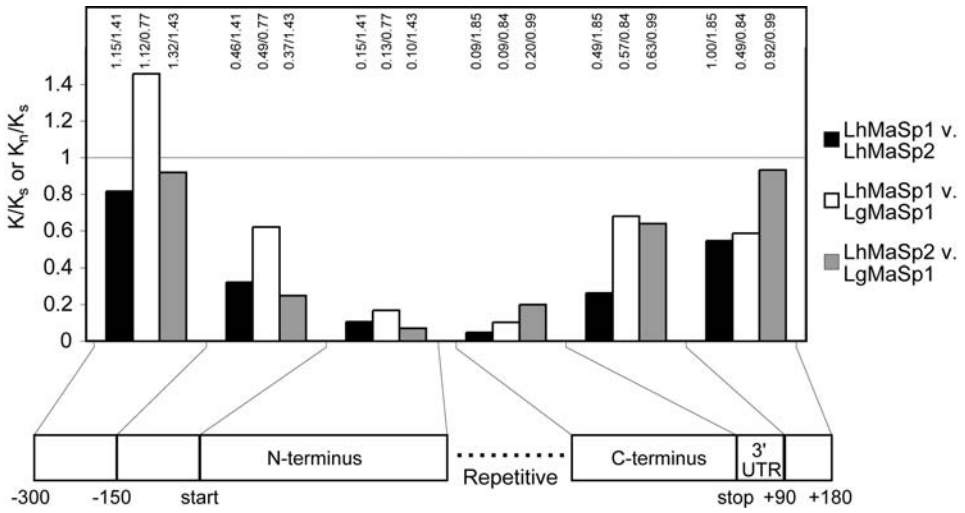


**Figure 5. $K/K_s$ or $K_n/K_s$ for flanking and terminal regions of *Latrodectus* major ampullate silk genes. $K_{s(N\text{-terminus})}$ is the denominator for upstream ratios; $K_{s(C\text{-terminus})}$ is the denominator for downstream ratios.** Actual K values shown above bars. Gene abbreviations are the same as for Figure 4.
doi:10.1371/journal.pone.0000514.g005

correlated with expression level [65–67] and major ampullate silk genes must be highly expressed throughout the lifetime of a spider. However, once an intron invades a silk gene, the intron can be rapidly propagated throughout the gene due to unequal crossing over, which appears to be common in silk genes (see Figure 2, [40–41,49]).

MaSp1 and MaSp2 are almost entirely composed of a small suite of amino acid sequence motifs, such as GGX and poly-A, which are repeated many times throughout both fibroins (Figures 1, 2). In each fibroin, the frequency and arrangement of these motifs occur in four distinct types of repeat units, termed ensemble repeats. Although the ensemble repeats of both MaSp1 and MaSp2 are similar in length (~30 aa) and composition (glycine-rich regions interspersed with alanine-rich regions), no repeat units from one protein are found in the other (Figures 1, 2). These results confirm that distinct genes encode each silk protein [e.g. 9,17–18], rather than posttranscriptional processing of a single gene leading to silk protein diversity as previously suggested by Craig et al. [68].

Both *L. hesperus* MaSp1 and MaSp2 have glycine and alanine-rich motifs that occur in ensemble repeats, but the fibroins differ in their higher-level repeat organization (repetitiveness) and similarity of repeat copies (homogenization). In MaSp1, the four types of ensemble repeats are strung together to form an ~120 amino acid long, higher-level repeat unit. This large aggregate is tandemly arrayed twenty times and the iterations share high identity at both the amino acid and nucleotide level (98.1% and 97.5% mean pairwise identity, respectively). In contrast, MaSp2 does not have clearly discernible higher-level repeats and has more sequence and length variation among its ensemble repeats than in MaSp1 (Figure 2, Table 2). MaSp2, however, has a tandem repetition of 778 aa that is >99.7% identical over the 2,334 encoding nucleotides (Figure 2). The modular architectures of MaSp1 and MaSp2 likely reflect concerted evolution within a single gene, as has been implicated in maintaining similarity among Flag (~440 aa) ensemble repeats and the long repeats of TuSp1 (~200 aa) and AcSp1 (aciniform silk; 200 aa [14]).

Modular architecture is also hypothesized to facilitate replication slippage in silk genes that have tandem arrays of codons for simple amino acid sequence motifs (e.g., poly-A, GGX, GA). Replication slippage would result in length variation among the ensemble repeats within a gene, as has been observed in *MaSp1, MaSp2*, and *Flag* [26,41,49]. Because previously described *MaSp1* and *MaSp2* gene or cDNA fragments are substantially incomplete (typically<2000 bp) and represent the least homogenized parts of the genes (5′ or 3′ ends), it is unknown if these genes are composed of higher level aggregates of ensemble repeats. Thus, it remains to be seen whether the extreme repetitiveness and homogenization of *L. hesperus MaSp1* compared to *L. hesperus MaSp2* is a general feature of spider dragline fibroins, or whether this pattern is peculiar to black widows.

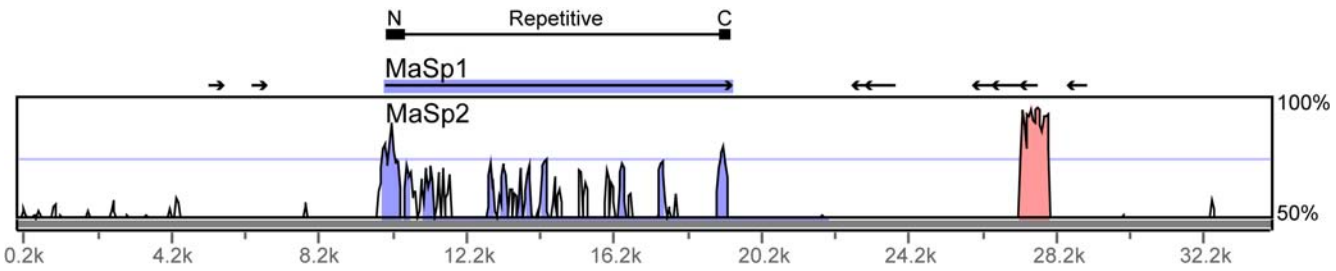


**Figure 6. Global AVID alignment of *L. hesperus* genomic clones containing *MaSp1* and *MaSp2* visualized with VISTA.** The *MaSp1*-containing clone was used as the reference sequence. Peak height corresponds to level of identity between the two clones. Regions exceeding 70% conservation over a window length of 100 bp are colored (blue for exons, red for non-coding sequence). The red peak corresponds to a putative transposable element found in both clones. Arrows mark open reading frames (ORFs) on the clone containing *MaSp1*.
doi:10.1371/journal.pone.0000514.g006

## Relationship to other silk proteins

Attempts to reconstruct evolutionary relationships among members of the spider silk gene family have relied exclusively on the non-repetitive C-terminus [11–12,14,26,40,69], but the N-terminus has great promise for phylogenetic reconstruction [38–39]. In our analyses, there was extensive congruence between trees based on N- and C-termini of silk gene family members (Figure 4C). A curious relationship found in both the N- and C-terminal phylogenetic trees is the grouping of *Latrodectus* major ampullate silk genes rather than a clade of *MaSp1* from all species separate from a *MaSp2* clade (Figure 4C). A similar sister relationship between *MaSp1* and *MaSp2* C-termini has been found for other species [10,12,40]. Given the striking conservation of repetitive amino acid motifs for each fibroin across divergent species, it seems unlikely that this pattern could result from independent duplication and convergence events. To explain the similarities in the repetitive regions by these means requires the convergence of thousands of nucleotides within a silk gene to encode either entirely MaSp1 or MaSp2 motifs, and for such convergences to have occurred multiple times in different spider lineages. Instead, recombination, selection, or the interaction of these two processes more likely explains the similarity of MaSp1 and MaSp2 N- and C-termini within species. Intergenic pairing during meiosis could be facilitated by the stretches of DNA coding for similar amino acid motifs, such as poly-A and GGX, in both *MaSp1* and *MaSp2*. For example, pairwise comparisons of the *L. hesperus MaSp1* and *MaSp2* genes show that they contain multiple regions of significant similarity spanning at least 100 bp (Figure 6). If recombination occurs between these two genes, it is less frequent than speciation events; *MaSp1* of *L. geometricus* and *L. hesperus* were clustered in the N-terminal trees and the C-terminal ML tree. Thus far, no single gene has been described that contains repeat units typical of both genes, which would provide the most convincing evidence for intergenic recombination. We did not find any clones in the *L. hesperus* genomic library that were positive for both *MaSp1* and *MaSp2*, nor did Sponner et al. [15] find double positive clones in a *Nephila clavipes* genomic library. However, there could be strong selection against proteins with a mixture of repeat units, while terminal recombinants may be tolerated. Convergent evolution could alternatively explain the grouping of *MaSp1* and *MaSp2* paralogs by their N- and C-termini. Selection could drive convergence of terminal amino acid sequences within species if similarity in these regions is necessary for accurate assembly of the two fibroins into a single fiber. Both proteins are exposed to the same environments, such as salt and pH gradients in the silk gland and duct [70], which could also favor evolutionary convergence of terminal domains.

## Non-coding sequence

Non-coding sequences upstream of major ampullate silk genes from different genera were too divergent to reliably align or identify regulatory elements other than the conserved motif CACG and the TATA-box identified by Motriuk-Smith et al. [38]. Although phylogenetic footprinting is a powerful tool for identifying novel regulatory elements, the appropriate level of divergence among species is critical [51]. The genera examined here, *Latrodectus*, *Nephila*, and *Argiope*, belong to three different families that shared a common ancestor ~135–160 million years ago (MYA) [71–72]. In contrast, some of the most successful examples of phylogenetic footprinting involve more recent divergences (e.g. *Drosophila* spp. [73], *Saccharomyces* spp. [74–75], grasses [76], and primates [77]). Comparisons of human and rodent genomes, thought to have split ~100 MYA [78], yield many novel regulatory elements, while extending divergence to mammals and birds (~310 MYA [79–80]) causes a precipitous drop off in the ability to detect motifs [81–82]. In plants, the limit of motif detection is gene specific but appears to be reached when comparing poplars and *Arabidopsis* [83], which diverged ~110 MYA [84–85]. Thus, given the divergence times of the spider taxa examined here, the fact that the promoter regions of their major ampullate silk genes retain any significant sequence similarity is notable.

In *Latrodectus*, ~300 bp of upstream sequence could be reliably aligned. However, the ~150 bp directly upstream of the start codon are more conserved than the adjacent, upstream non-coding sequence or synonymous sites in coding regions of the genes (Figure 5). We found a conserved motif in this region that matches the binding site for the Achaete-Scute family of transcription factors, which regulate neurogenesis and sensory mother cell development in *Drosophila* [86–87]. A homolog of this transcription factor family, called SGSF, shows a silk gland-restricted pattern of expression in *L. hesperus*, specifically to the tubuliform and major ampullate silk glands [88]. These are the only glands that appear to express *MaSp1* and *MaSp2* in appreciable quantities [40]. Experimental manipulation is needed to elucidate if SGSF or a related protein is, in fact, involved with regulating major ampullate silk gene expression in black widows and other spider species.

The conserved, upstream non-coding regions and the 3′ UTRs of *L. hesperus MaSp1* and *L. hesperus MaSp2* show evidence for stronger selective constraints than do *L. hesperus MaSp1* and *L. geometricus MaSp1* (lower $K/K_s$, Figure 5). Although regulatory element evolution in the 3′ UTR has received less attention than in promoter regions, many genes display significantly conserved sequence motifs in the 3′ UTR [e.g. 89–91]. Additionally, experimental evidence has shown that elements in the 3′ UTR bind factors involved in posttranscriptional regulation [e.g. 92–93]. A striking example of 3′ regulation is in *Drosophila's Enhancer of split* Complex, which belongs to the same class of genes (beta helix-loop-helix) as *achaete* and *scute* [89,94–96]. Taken together, our findings suggest selection on non-coding sequences for co-regulation of the paralogous dragline silk genes, *MaSp1* and *MaSp2*.

## Recipe for a high-performance biomaterial

The production of synthetic spider dragline silk is a major goal of biomimetics research [1,29]. Though promising advances have been made with a variety of transgene constructs and host organisms, an exact mimic of a native dragline silk fiber has yet to be produced [e.g. 30,33–35]. While artificial spinning is certainly an important consideration, a significant challenge to the efforts to create synthetic silk proteins has been the incomplete knowledge of spider silk genes. Thus far, all transgene constructs for recombinant silk proteins have relied on partial cDNA sequences from two orb-weaving species, *Nephila clavipes* and *Araneus diadematus* [e.g. 29–30,32–35,37,97]. These truncated cDNAs encode only a fraction (typically 20% or less) of the repetitive region and the C-terminal domain. Experiments on recombinant silks made with and without the C-terminal region showed that the C-terminus was required for fibroins to form aggregates. Protein aggregation is an essential step in the precipitation of liquid spinning dope into a solid silk fiber [37,98]. The C-terminus is not only necessary for aggregation of recombinant fibroins, but also for the formation of the characteristic crystalline structures that impart strength to dragline silk fibers [35]. As has been proposed for the C-terminus [37], the evolutionary conservation of the N-terminus suggests that this region is also functionally significant. For example, N-termini may

play a central role in the proper transport of fibroins from secretory cells to silk gland lumen, aid in fiber formation, and contribute to the structural properties of silk fibers. In both *L. hesperus* MaSp1 and MaSp2, the N-terminal domain contains the most hydrophobic region of the entire fibroin (Figure 3). The next most hydrophobic region is the C-terminus. Sponner et al. [37] hypothesized that the hydrophobicity of the C-terminus was a key characteristic for its role in fibroin aggregation. The hydrophobic N-terminal region could thus similarly enhance silk fiber formation and mechanical properties. Another evolutionarily conserved aspect of spider fibroins is their extremely large size, which is also a feature of independently evolved insect fibroins. Thus, large size has been repeatedly selected for in the evolution of fibroin genes. Therefore, a complete silk gene, with full representation of the N- and C-terminal regions, the intervening repetitive sequence, and the transitions among these domains, should dramatically improve recombinant silk performance.

The complete gene sequences described here highlight the extraordinary molecular characteristics of spider silks. Black widow major ampullate silk genes are highly modular, exhibiting a hierarchical organization of iterated short motifs and ensemble repeats (groups of motifs). By characterizing full-length *MaSp1* and *MaSp2* genes, we were able to detect even higher-level repeats (aggregates of ensemble repeats) and uncover a striking difference in the degree of repeat homogenization between *MaSp1* and *MaSp2*. The extreme modularity of *MaSp1* (Figure 1) may reflect selection on the MaSp1 fibroin for perfect repeats, perhaps important for rapid and consistent spinning of high quality silk fibers. Sequence homogenization, however, is also due to molecular mechanisms, such as unequal crossing over (e.g., two large tandem repeats in Figure 2), and the interaction between selection and concerted evolution is a subject for further investigation. We have additionally identified putative regulatory elements that may enhance expression of transgenic silks. Thus, the clones sequenced here provide the complete genetic blueprints for the primary protein components of the major ampullate silk fiber. These designs hold critical information for the mass production of artificial fibers that accurately mimic the spectacular high-performance properties of native spider silk.

## METHODS

### Genomic Library Construction and Screening

We targeted black widow silk genes because in addition to the exemplary properties of their silk, *Latrodectus hesperus* has one of the smallest known genome sizes for a spider (C-value of 1.29 picograms [99]), meaning that fewer genomic clones must be screened to find a gene of interest. Individuals were collected from a single locality in Riverside, California (USA), live frozen in liquid nitrogen, and stored at $-80^{\circ}$C. High-molecular-weight DNA was isolated from the cephalothoraxes of eight individuals using a modified method of Sambrook and Russell [100]. Following isolation, DNA was mechanically sheared through a pipette tip and subsequently treated with End-Repair Enzyme Mix (Epicentre) to produce blunt 5′ phosphorylated ends. Fragments ranging from 38–50 kilobases were gel excised, purified, and ligated into pCC1FOS$^{TM}$ vector (Epicentre). Resulting fosmids were packaged using MaxPlax$^{TM}$ Lambda Packaging Extracts and transfected into Epi300-T1R *E. coli* cells following protocols for the CopyControl$^{TM}$ Fosmid Library Production kit (Epicentre). Approximately 100,000 recombinant *E. coli* colonies were picked and arrayed into 276 culture plates each containing 384 wells using a QPIX robotic picker (Genetix). Each culture plate was replicated and original stock plates containing 7.5% glycerol were stored at $-80^{\circ}$C.

To efficiently screen the genomic library, fosmid DNA was extracted from cell cultures combined from a single 384-well plate, and such extractions were done for every plate in the library. Polymerase chain reaction (PCR) experiments targeting genes of interest were used to identify which plate contained one or more positive clones. Once the plate was identified, that plate was replicated twice, and cell cultures from the rows were combined to form 16 templates, while cell cultures from the columns were combined to form 24 templates. Templates were then PCR screened to identify individual clones containing the gene of interest. Primers targeting *MaSp1* and *MaSp2* were designed from *L. hesperus* cDNA clones [40] (*MaSp1*–N-terminal clone, EF595247; *MaSp2*–C-terminal clone, AY953075). The primers, LhMaSp1NF254, 5′-TGGCTTTCGCATCATCTGTAGC-3′ and LhMaSp1NR607, 5′-CTCCTTGACCATAACTAACTGG-CTG-3′ amplified a 350 bp portion of the *MaSp1* 5′ region. Primers LhMaSp2_1086F, 5′-CATCAGCAGCAGGACCAAG-TG-3′, and LhMaSp2_1337R, 5′-GCGTTGTCGGTGAAGA-TAAAGC-3′, amplified a 250 bp portion of the *MaSp2* 3′ region.

### Sequencing

Seven *MaSp1*-positive clones and three *MaSp2*-positive clones were found after screening half of the library. One positive clone for each gene was shotgun sequenced and assembled by Qiagen (Hilden, Germany) to 6× coverage for the *MaSp2*-positive clone and 8× coverage for the *MaSp1*-positive clone. This resulted in three contiguous sequences (contigs) for the *MaSp2*-positive clone with two gaps within the coding sequence and one directly after the stop codon. The 707 bp gap between the stop codon and the downstream contig was closed by sequencing directly off the fosmid clone using primers designed from the C-terminal coding region of *MaSp2* and for the beginning of the downstream contig (all primer sequences used in this study are available upon request). Primer walking to close the two gaps within the *MaSp2* coding sequence was not possible due to its repetitive nature. Instead the clone was digested with NotI and BamHI (New England Biolabs) and a 9 kb restriction fragment containing almost the entire repetitive portion of *MaSp2* was subcloned into pZErO$^{TM}$-2 plasmids (Invitrogen) and electroporated into Epi-400 *E. coli* (Epicentre). The subclone was partially digested with PstI (New England Biolabs) and 2000-3000 bp fragments were gel excised and ligated into PstI digested and dephosphorylated pZErO$^{TM}$-2. Ligation products were electroporated into TOP10 *E. coli* (Invitrogen). A library of 96 PstI partial-digest clones were arrayed and sequenced in one direction. Sequences were assembled independently and using the fosmid contigs as a backbone in SEQUENCHER v4.5 (Gene Codes Corp.), requiring 100% identity for high-quality bases. Ten clones spanned the first gap (111 bp) and 18 clones spanned the second gap (632 bp) with no less than 5× sequence coverage of any base along the length of the NotI-BamHI subclone. No disagreement between the sequences of the subclone and the fosmid contigs was found.

Shotgun sequencing of a *MaSp1*-postive clone resulted in a single contig containing the entire coding sequence of *MaSp1* and the vector. However, this contig was ~7000 bp smaller than expected based on restriction digests. This missing sequence was determined by PCR amplifying with AccuPrime$^{TM}$ Taq DNA Polymerase High Fidelity (Invitrogen) and primers designed from both ends of the contig. The 7890 bp PCR product was sequenced with at least 2× coverage by primer walking. Additionally, the fosmid was directly sequenced at intervals along the gap to ensure that no mutations had been introduced by the PCR amplifcation. Experimental restriction digests of the *MaSp1*-positive and *MaSp2*-positive clones matched predicted restriction sites in the final

sequences, verifying that assembly had not erroneously excluded repetitive sequence.

## Sequence analysis

Nucleotide sequences were conceptually translated using the standard genetic code. Base composition, amino acid content, codon usage, and Kyte and Doolittle [50] hydrophilicity predictions were calculated in MacVector$^{TM}$ (Oxford Molecular Group). Amino acid sequences were considered to start at the first methionine in frame. The first M on the MaSp1 sequence corresponded to the conserved start position identified by Rising et al. [39] (see also Figure 4A). The MaSp2 sequence also displayed an M at this position, but the first in frame M codon was 9 bp upstream (Figure 4A). Pairwise K, $K_s$, and $K_n$ were calculated using DnaSp v4.0 [101] excluding gaps and missing data.

Predicted amino acid sequences of all currently published N-termini were aligned (Figure 4A), making corrections to the nucleotide sequences of *L. geometricus MaSp1, A. bruennichi CySp2*, and *N. clavipes Flag* according to the modifications described in Rising et al. [39]. Alignments of N- and C-terminal amino acid sequences were made separately using default parameters in ClustalW (MacVector$^{TM}$). The C-terminal alignment was modified slightly such that the first position of the C-terminal Flag sequences aligned with the first position of the other sequences (Figure 4B). Amino acid alignments were used to guide nucleotide alignments, which formed the basis for phylogenetic analyses. Heuristic ML and MP searches were performed in PAUP* [102] using TBR (tree bisection reconnection) branch swapping and 10,000 (MP) or 100 (ML) random stepwise addition replicates. Support for clades was evaluated with 1000 (MP) or 100 (ML) bootstrap pseudoreplicates (of all characters), and 100 (MP) or 1 (ML) random stepwise addition replicates per pseudoreplicate. ML analyses treated gaps as missing data. MP analyses were performed treating gaps as missing data and as a 5$^{th}$ state. Optimal model parameters for ML analyses were calculated with MODELTEST [103]. The N-termini fit the HKY+G [104] model of evolution (transitions/transversions = 1.24; gamma = 0.9058). The C-termini fit the TrN+G [105] model of evolution (A<>G = 2.34; C<>T = 1.27; transversions = 1; gamma = 1.34). To further evaluate tree structure and clade support in a model-based framework, Bayesian analyses were carried out using MRBAYES v.3.1.2 [106]. The same model of evolution determined by MODELTEST was used but parameter values were evaluated during the Bayesian analysis. Default priors and Metropolis-coupled, Markov-chain, Monte Carlo (MCMC) sampling procedures were executed for two independent runs, sampled every 100$^{th}$ generation, carried out simultaneously. Convergence was assessed every 1000$^{th}$ generation and the posterior distribution was considered adequately sampled when the standard deviation of split frequencies of these two runs dropped below 0.01 (<1 million generations). A second analysis was run for 10 million generations (sampling every 500) to ensure that a longer sampling time did not change the results. For each run, the first 50% of sampled trees were discarded as burnin prior to calculating the majority rule consensus tree.

Comparisons of genes with MultiPipMaker were done using the "high sensitivity low time limit" option. Each major ampullate silk gene with upstream sequence was sequentially input as the reference to obtain maximal pairwise alignments. AVID alignments were made using default parameters and viewed on the VISTA browser <www-gsd.lbl.gov/vista/> [107–108]. Global alignments of conserved non-coding sequence identified by MultiPipMaker were made using default parameters in ClustalW and modified manually. Approximately 300 bp of upstream sequence were scanned against insect transcription factor binding sites in the TRANSFAC 6.0 database using the program PATCH$^{TM}$ v1.0 [57] with a minimum match of 6 and a maximum mismatch of 2.

Open reading frames on the black widow genomic clones were identified using the ORFFinder program on the NCBI website <http://www.ncbi.nlm.nih.gov/gorf/gorf.html>, with a minimum cutoff of 300 nucleotides.

## ACKNOWLEDGMENTS

We thank Richard Baker, Clay Clark, Dayan Colon-Sanchez, Teresa DiMauro, Tim Kingan, and Lobna Shahatto for help constructing, arraying, and screening the genomic library. We also thank Richard Baker, Laura Baldo, and John Gatesy for comments on this manuscript.

## Author Contributions

Conceived and designed the experiments: RT NA JG MC CH. Performed the experiments: RT NA JG MC CH. Analyzed the data: NA CH. Contributed reagents/materials/analysis tools: CH. Wrote the paper: NA JG CH.

## REFERENCES

1. Vollrath F, Knight DP (2001) Liquid crystalline spinning of spider silk. Nature 410: 541–548, DOI: 10.1038/35069000.
2. Gosline J, Lillie M, Carrington E, Guerrette P, Ortlepp C, et al. (2002) Elastic proteins: biological roles and mechanical properties. Phil Trans R Soc Lond B Biol Sci 357: 121–132, DOI: 10.1098/rstb.2001.1022.
3. Foo CWP, Kaplan DL (2002) Genetic engineering of fibrous proteins: spider dragline silk and collagen. Adv Drug Deliv Rev 54: 1131–1143, DOI: 10.1016/S0169-409X(02)00061-3.
4. Gosline JM, DeMont ME, Denny MW (1986) The structure and properties of spider silk. Endeavour 10: 37–43, DOI: 10.1016/0160-9327(86)90049-9.
5. Hinman M, Dong Z, Xu M, Lewis RV (1992) Spider silk: a mystery starting to unravel. In: Case ST, ed. Biopolymers Springer: Berlin, pp 227–254.
6. Gosline JM, Guerette PA, Ortlepp CS, Savage KN (1999) The mechanical design of spider silks: from fibroin sequence to mechanical function. J Exp Biol 202: 3295–3303.
7. Foelix R (1996) Biology of Spiders. Oxford University Press: New York, pp 336.
8. Blackledge TA, Hayashi CY (2006) Silken toolkits: biomechanics of silk fibers spun by the orb web spider *Argiope argentata* (Fabricius 1775). J Exp Biol 209: 2452–2461, DOI: 10.1242/jeb.02275.
9. Guerette PA, Ginzinger DG, Weber BHF, Gosline JM (1996) Silk properties determined by gland-specific expression of a spider fibroin gene family. Science 272: 112–115, DOI: 10.1126/science.272.5258.112.
10. Gatesy J, Hayashi C, Motriuk D, Woods J, Lewis R (2001) Extreme diversity, conservation, and convergence of spider silk fibroin sequences. Science 291: 2603–2605, DOI: 10.1126/science.1057561.
11. Tian M, Lewis RV (2005) Molecular characterization and evolutionary study of spider tubuliform (eggcase) silk protein. Biochemistry 44: 8006–8012, DOI: 10.1021/bi050366u.
12. Garb JE, DiMauro T, Vo V, Hayashi CY (2006) Silk genes support the single origin of orb webs. Science 312: 1762, DOI: 10.1126/science.1127946.
13. Hayashi CY, Shipley NH, Lewis RV (1999) Hypotheses that correlate the sequence, structure, and mechanical properties of spider silk proteins. Int J Biol Macromol 24: 271–275, DOI: 10.1016/S0141-8130(98)00089-0.
14. Hayashi CY, Blackledge TA, Lewis RV (2004) Molecular and mechanical characterization of aciniform silk: uniformity of iterated sequence modules in a novel member of the spider silk fibroin gene family. Mol Biol Evol 21: 1950–1959, DOI: 10.1093/molbev/msh204.
15. Sponner A, Schlott B, Vollrath F, Unger E, Grosse F, et al. (2005) Characterization of the protein components of *Nephila clavipes* dragline silk. Biochemistry 44: 4727–4736, DOI: 10.1021/bi047671k.
16. Zhao A, Zhao T, Nakagaki K, Zhang Y-S, SiMa Y, et al. (2006) Novel molecular and mechanical properties of egg case silk from wasp spider, *Argiope bruennichi*. Biochemistry-US 45: 3348–3356, DOI: 10.1021/bi052414g.
17. Xu M, Lewis RV (1990) Structure of a protein superfiber: spider dragline silk. Proc Natl Acad Sci USA 87: 7120–7124, DOI: 10.1073/pnas.87.18.7120.

18. Hinman MB, Lewis RV (1992) Isolation of a clone encoding a second dragline silk fibroin. *Nephila clavipes* dragline silk is a two-protein fiber. J Biol Chem 267: 19320–19324.
19. Sponner A, Unger E, Grosse F, Weisshart K (2005) Differential polymerization of the two main protein components of dragline silk during fibre spinning. Nat Mater 4: 772–775, DOI: 10.1038/nmat1493.
20. Simmons A, Ray E, Jelinski LW (1994) Solid-state $^{13}C$ NMR of *Nephila clavipes* dragline silk establishes structure and identity of crystalline regions. Macromolecules 27: 5235–5237, DOI: 10.1021/ma00096a060.
21. Kümmerlen J, van Beek JD, Vollrath F, Meier BH (1996) Local structure in spider dragline silk investigated by two dimensional spin-diffusion NMR. Macromolecules 29: 2920–2928, DOI: 10.1021/ma951098i.
22. Simmons AH, Michal CA, Jelinski LW (1996) Molecular orientation and two-component nature of the crystalline fraction of spider dragline silk. Science 271: 84–87, DOI: 10.1126/science.271.5245.84.
23. Bram A, Bränden CI, Craig C, Snigireva I, Riekel C (1997) X-ray diffraction from single fibers of spider silk. J Appl Cryst 30: 390–392, DOI: 10.1107/S0021889896012344.
24. Parkhe AD, Seeley SK, Gardner K, Thompson L, Lewis RV (1997) Structural studies of spider silk proteins in the fiber. J Mol Recogn 10: 1–6, DOI: 10.1002/(SICI)1099-1352(199701/02)10:1<1::AID-JMR338>3.0.CO;2-7.
25. Dong Z, Lewis RV, Middaugh CR (1991) Molecular mechanisms of spider silk elasticity. Arch Biochem Biophys 284: 53–57.
26. Hayashi CY, Lewis RV (1998) Evidence from flagelliform silk cDNA for the structural basis of elasticity and modular nature of spider silks. J Mol Biol 275: 773–784, DOI: 10.1006/jmbi.1997.1478.
27. van Beek JD, Hess S, Vollrath F, Meier BH (2002) The molecular structure of spider dragline silk: folding and orientation of the protein backbone Proc Natl Acad Sci USA 99: 10266–10271, DOI: 10.1073/pnas.152162299.
28. O'Brien JP, Fahnestock SR, Termonia Y, Gardner KH (1998) Nylons from nature: synthetic analogs to spider silk. Adv Mater 10: 1185–1195, DOI: 10.1002/(SICI)1521-4095(199810)10:15<1185::AID-ADMA1185>3.0.CO;2-T.
29. Hinman MB, Jones JA, Lewis RV (2000) Synthetic spider silk: a modular fiber. Trends Biotechnol 18: 374–379, DOI: 10.1016/S0167-7799(00)01481-5.
30. Lazaris A, Arcidiacono S, Huang Y, Zhou J-F, Duguay F, et al. (2002) Spider silk fibers spun from soluble recombinant silk produced in mammalian cells. Science 295: 472–476, DOI: 10.1126/science.1065780.
31. Scheller J, Henggeler D, Viviani A, Conrad U (2004) Purification of spider silk-elastin from transgenic plants and application for human chondrocyte proliferation. Transgenic Res 13: 51–57, DOI: 10.1023/B:TRAG.0000017175.78809.7a.
32. Bini E, Foo CWP, Huang J, Karageorgiou V, Kitchel B, et al. (2006) RGD-functionalized bioengineered spider dragline silk biomaterial Biomacromolecules 7: 3139–3145, DOI: 10.1021/bm0607877.
33. Scheller J, Gührs K-H, Grosse F, Conrad U (2001) Production of spider silk proteins in tobacco and potato. Nat Biotechnol 19: 573–577, DOI:10.1038/89335.
34. Foo CWP, Bini E, Huang J, Lee SY, Kaplan DL (2006) Solution behavior of synthetic silk peptides and modified recombinant silk proteins. Appl Phys A 82: 193–203, DOI: 10.1007/s00339-005-3425-8.
35. Ittah S, Cohen S, Garty S, Cohn D, Gat U (2006) An essential role for the C-terminal domain of a dragline spider silk protein in directing fiber formation. Biomacromolecules 7: 1790–1795, DOI: 10.1021/bm060120k.
36. Sprague KU (1975) The *Bombyx mori* silk proteins: characterization of large polypeptides. Biochemistry 14: 925–931, DOI: 10.1021/bi00676a008.
37. Sponner A, Vater W, Rommerskirch W, Vollrath F, Unger E, et al. (2005) The conserved C-termini contribute to the properties of spider silk fibroins. Biochem Biophys Res Commun 338: 897–902, DOI: 10.1016/j.bbrc.2005.10.048.
38. Motriuk-Smith D, Smith A, Hayashi CY, Lewis RV (2005) Analysis of the conserved N-terminal domains in major ampullate spider silk proteins. Biomacromolecules 6: 3152–3159, DOI: 10.1021/bm050472b.
39. Rising A, Hjälm G, Engström W, Johansson J (2006) N-terminal nonrepetitive domain common to dragline, flagelliform, and cylindriform spider silk proteins. Biomacromolecules 7: 3120–3124, DOI: 10.1021/bm060693x.
40. Garb JE, Hayashi CY (2005) Modular evolution of egg case silk genes across orb-weaving spider superfamilies. Proc Natl Acad Sci USA 102: 11379–11384, DOI: 10.1073/pnas.0502473102.
41. Hayashi CY, Lewis RV (2000) Molecular architecture and evolution of a modular spider silk protein gene. Science 287: 1477–1479, DOI: 10.1126/science.287.5457.1477.
42. Griswold CE, Coddington JA, Hormiga G, Scharff N (1998) Phylogeny of the orb-web building spiders (Araneae, Orbiculariae: Deinopoidea, Araneoidea). Zool J Linn Soc–Lond 123: 1–99, DOI: 10.1006/zjls.1997.0125.
43. Lawrence BA, Vierra CA, Moore AF (2004) Molecular and mechanical properties of major ampullate silk of the black widow spider, *Latrodectus hesperus*. Biomacromolecules 5: 689–695, DOI: 10.1021/bm0342640.
44. Blackledge TA, Summers AP, Hayashi CY (2005) Gumfooted lines in black widow cobwebs and the mechanical properties of spider capture silk Zoology 108: 41–46, DOI: 10.1016/j.zool.2004.11.001.
45. Swanson BO, Blackledge TA, Beltrán J, Hayashi CY (2006) Variation in the material properties of spider dragline silk across species Appl Phys A 82: 213–218, DOI: 10.1007/s00339-005-3427-6.
46. Casem ML, Turner D, Houchin K (1999) Protein and amino acid composition of silks from the cob weaver, *Latrodectus hesperus* (black widow). Int J Biol Macromol 24: 103–108, DOI: 10.1016/S0141-8130(98)00078-6.
47. Anderson SO (1970) Amino acid composition of spider silks. Comp Biochem Physiol 35: 705–711.
48. Lombardi SJ, Kaplan DL (1990) The amino acid composition of major ampullate gland silk (dragline) of *Nephila clavipes* (Araneae, Tetragnathidae). J Arachnol 18: 297–306.
49. Beckwitt R, Arcidiacono S, Stote R (1998) Evolution of repetitive proteins: spider silks from *Nephila clavipes* (Tetragnathidae) and *Araneus bicentenarius* (Araneidae). Insect Biochem Mol Biol 28: 121–130, DOI: 10.1016/S0965-1748(97)00083-0.
50. Kyte J, Doolittle R (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157: 105–132.
51. Hardison RC (2003) Comparative genomics. PLoS Biology 1: 156–160, DOI: 10.1371/journal.pbio.0000058.
52. Rombauts S, Florquin K, Lescot M, Marchal K, Rouzé P, et al. (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol 132: 1162–1176, DOI: 10.1104/pp.102.017715.
53. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, et al. (2000) PipMaker–A web server for aligning two genomic DNA sequences. Genome Res 10: 577–586.
54. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, et al. (2003) Human-mouse alignments with BLASTZ. Genome Res 13: 103–107, DOI: 10.1101/gr.809403.
55. Pollard DA, Moses AM, Iyer VN, Eisen MB (2006) Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. BMC Bioinformatics 7: 376, DOI: 10.1186/1471-2105-7-376.
56. Margulies EH, Blanchette M, NISC Comparative Sequencing Program, Haussler D, Green ED (2003) Identification and characterization of multiple-species conserved sequences. Genome Res 13: 2507–2518, DOI: 10.1101/gr.1602203.
57. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108–D110, <http://www.gene-regulation.com/pub/databases.html> DOI: 10.1093/nar/gkj143.
58. Graur D, Li WH (2000) Fundamentals of Molecular Evolution, second edition. Sunderland, MA: Sinauer. pp 481.
59. Wong WSW, Nielsen R (2004) Detecting selection in noncoding regions of nucleotide sequences. Genetics 167: 949–958, DOI: 10.1534/genetics.102.010959.
60. Bray N, Dubchak I, Pachter L (2003) AVID: A Global Alignment Program. Genome Res 13: 97–102, DOI: 10.1101/gr.789803.
61. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
62. Sakharkar MK, Kangueane P, Petrov DA, Kolaskar AS, Subbiah S (2002) SEGE: A database on 'intron less/single exonic' genes from eukaryotes. Bioinformatics 18: 1266–1267.
63. Sakharkar MK, Kangueane P (2004) Genome SEGE: A database for 'intronless' genes in eukaryotic genomes. BMC Bioinformatics 5: 67, DOI: 10.1186/1471-2105-5-67.
64. Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol 18: 292–298, DOI: 10.1016/S0169-5347(03)00033-8.
65. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA (2002) Selection for short introns in highly expressed genes. Nat Genet 31: 415–418, DOI: 10.1038/ng940.
66. Urrutia AO, Hurst LD (2003) The signature of selection mediated by expression on human genes. Genome Res 13: 2260–2264, DOI: 10.1101/gr.641103.
67. Marais G, Nouvellet P, Keightley PD, Charlesworth B (2005) Intron size and exon evolution in *Drosophila*. Genetics 170: 481–485, DOI: 10.1534/genetics.104.037333.
68. Craig CL, Riekel C, Herberstein ME, Weber RS, Kaplan D, et al. (2000) Evidence for diet effects on the composition of silk proteins produced by spiders. Mol Biol Evol 17: 1904–1913.
69. Beckwitt R, Arcidiacono S (1994) Sequence conservation in the C-terminal region of spider silk proteins (spidroin) from *Nephila clavipes* (Tetragnathidae) and *Araneus bicentenarius* (Araneidae). J Biol Chem 269: 6661–6663.
70. Dicko C, Vollrath F, Kenney JM (2004) Spider silk protein refolding is controlled by changing pH. Biomacromolecules 5: 704–710, DOI: 10.1021/bm034307c.
71. Selden PA (1990) Lower Cretaceous spiders from the Sierra-de-Montsech, north-east Spain. Palaeontology 33: 257–285.
72. Ayoub NA, Garb JE, Hedin M, Hayashi CY (2007) Utility of the nuclear protein-coding gene, elongation factor-1 gamma (*EF-1γ*), for spider systematics, emphasizing family level relationships of tarantulas and their kin (Araneae: Mygalomorphae). Mol Phylogenet Evol 42: 394–409, DOI: 10.1016/j.ympev.2006.07.018.
73. Sinha S, Siggia ED (2005) Sequence turnover and tandem repeats in *cis*-regulatory modules in *Drosophila*. Mol Biol Evol 22: 874–885, DOI: 10.1093/molbev.msi090.

74. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, et al. (2003) Finding functional elements in *Saccharomyces* genomes by phylogenetic footprinting. Science 301: 71–76, DOI: 10.1126/science.1084337.
75. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241–254, DOI: 10.1038/nature01644.
76. Inada DC, Bashir A, Lee C, Thomas BC, Ko C, et al. (2003) Conserved noncoding sequences in the grasses. Genome Res 13: 2030–2041, DOI: 10.1101/gr.1280703.
77. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, et al. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299: 1391–1394, DOI: 10.1126/science.1081331.
78. Arnason U, Gullberg A, Burguete AS, Janice A (2000) Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. Hereditas 133: 217–228, DOI: 10.1111/j.1601-5223.2000.00217.x.
79. Benton MJ (1990) Phylogeny of the major tetrapod groups: morphological data and divergence dates. J Mol Evol 30: 409–424, DOI: 10.1007/BF02101113.
80. Lee MSY (1999) Molecular clock calibration and metazoan divergence dates. J Mol Evol 49: 385–391, DOI: 10.1007/PL00006562.
81. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424: 788–793, DOI: 10.1038/nature01858.
82. Prakash A, Tompa M (2005) Discovery of regulatory elements in vertebrates through comparative genomics. Nat Biotechnol 23: 1249–1256, DOI: 10.1038/nbt1140.
83. De Bodt S, Theissen G, Van de Peer Y (2006) Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. Mol Biol Evol 23: 1293–1303, DOI: 10.1093/molbev/msk016.
84. Chaw S-M, Chang C-C, Chen H-L, Li W-H (2001) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol 58: 424–441, DOI: 10.1007/S00239-003-2564-9.
85. Wikström N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: calibrating the family tree. Proc R Soc Lond B Biol Sci 268: 2211–2220, DOI: 10.1098/rspb.2001.1782.
86. Cabrera CV, Alonso MC (1991) Transcriptional activation by heterodimers of the *achaete-scute* and *daughterless* gene products of *Drosophila*. EMBO J 10: 2965–2973.
87. Ruiz-Gómez M, Ghysen A (1993) The expression and role of a proneural gene, *achaete*, in the development of the larval nervous system of *Drosophila*. EMBO J 12: 1121–1130.
88. Kohler K, Thayer W, Le T, Sembhi A, Vasanthavada K, et al. (2005) Characterization of a novel class II bHLH transcription factor from the black widow spider, *Latrodectus hesperus*, with silk-gland restricted patterns of expression. DNA Cell Biol. 24: 371–380.
89. Leviten MW, Lai EC, Posakony JW (1997) The *Drosophila* gene *Bearded* encodes a novel small protein and shares 3′ UTR sequence motifs with multiple *Enhancer of split* Complex genes. Development 124: 4039–4051.
90. Coy JF, Sedlacek Z, Bächner D, Delius H, Poustka A (1999) A complex pattern of evolutionary conservation and alternative polyadenylation within the long 3′-untranslated region of the methyl-CpG-binding protein 2 gene (*MeCP2*) suggests a regulatory role in gene expression. Hum Mol Genet 8: 1253–1262.
91. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature 434: 338–345, DOI: 10.1038/nature03441.
92. Goodwin EB, Hofstra K, Hurney CA, Mango S, Kimble J (1997) A genetic pathway for regulation of *tra-2* translation. Development 124: 749–758.
93. Wickens M, Bernstein DS, Kimble J, Parker R (2002) A PUF family portrait: 3′ UTR regulation as a way of life. Trends Genet 18: 150–157, DOI: 10.1016/S0168-9525(01)02616-6.
94. Delidakis C, Artavanis-Tsakonas S (1992) The Enhancer of split [*E(spl)*] locus of *Drosophila* encodes seven independent helix-loop-helix proteins. Proc Natl Acad USA 89: 8731–8735.
95. Lai EC, Posakony JW (1997) The Bearded box, a novel 3′ UTR sequence motif, mediates negative post-transcriptional regulation of *Bearded* and *Enhancer of split* Complex gene expression. Development 124: 4847–4856.
96. Lai EC, Burks C, Posakony JW (1998) The K box, a conserved 3′ UTR sequence motif, negatively regulates accumulation of Enhancer of split Complex transcripts. Development 125: 4077–4088.
97. Arcidiacono S, Mello C, Kaplan D, Cheley S, Bayley H (1998) Purification and characterization of recombinant spider silk expressed in *Escherichia coli*. Appl Microbiol Biotechnol 49: 31–38, DOI: 10.1007/S002530051133.
98. Huemmerich D, Helsen CW, Quedzuweit S, Oschmann J, Rudolph R, et al. (2004) Primary structure elements of spider dragline silks and their contribution to protein solubility. Biochemistry 43: 13604–13612, DOI: 10.1021/bi048983q.
99. Gregory TR, Shorthouse DP (2003) Genome sizes of spiders. J Hered 94: 285–290, DOI: 10.1093/jhered/esg070.
100. Sambrook J, Russell DW (2001) Molecular Cloning: A Laboratory Manual, third edition. Cold Spring Harbor Laboratory. 999 p.
101. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496–2497, DOI: 10.1093/bioinformatics/btg359.
102. Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates: Sunderland, Massachusetts.
103. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817–818.
104. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174, DOI: 10.1007/BF02101694.
105. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.
106. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models DOI: 10.1093/bioinformatics/btg180. Bioinformatics 19: 1572–1574.
107. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, et al. (2000) VISTA: Visualizing global DNA sequence alignments of arbitrary length. Bioinformatics 16: 1046–1047, <http://genome.lbl.gov/vista/index.shtml>.
108. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32: W273–W279.