GENETICS

1000 spider silkomes: Linking sequences to silk physical properties

Kazuharu Arakawa[1,2,3,4]*†, Nobuaki Kono[1,3]†, Ali D. Malay[5], Ayaka Tateishi[5,6], Nao Ifuku[5], Hiroyasu Masunaga[7], Ryota Sato[5,8], Kousuke Tsuchiya[5,6], Rintaro Ohtoshi[5,8], Daniel Pedrazzoli[8], Asaka Shinohara[8], Yusuke Ito[8], Hiroyuki Nakamura[5,8], Akio Tanikawa[9], Yuya Suzuki[10,11], Takeaki Ichikawa[12], Shohei Fujita[13], Masayuki Fujiwara[1], Masaru Tomita[1,2,3], Sean J. Blamires[14]‡, Jo-Ann Chuah[5], Hamish Craig[5,14], Choon P. Foong[5,6], Gabriele Greco[15], Juan Guan[16], Chris Holland[17], David L. Kaplan[18], Kumar Sudesh[19], Biman B. Mandal[20,21,22], Y. Norma-Rashid[23], Nur A. Oktaviani[5], Rucsanda C. Preda[18], Nicola M. Pugno[15,24], Rangam Rajkhowa[25], Xiaoqin Wang[26], Kenjiro Yazawa[5], Zhaozhu Zheng[26], Keiji Numata[5,6]*

Spider silks are among the toughest known materials and thus provide models for renewable, biodegradable, and sustainable biopolymers. However, the entirety of their diversity still remains elusive, and silks that exceed the performance limits of industrial fibers are constantly being found. We obtained transcriptome assemblies from 1098 species of spiders to comprehensively catalog silk gene sequences and measured the mechanical, thermal, structural, and hydration properties of the dragline silks of 446 species. The combination of these silk protein genotype-phenotype data revealed essential contributions of multicomponent structures with major ampullate spidroin 1 to 3 paralogs in high-performance dragline silks and numerous amino acid motifs contributing to each of the measured properties. We hope that our global sampling, comprehensive testing, integrated analysis, and open data will provide a solid starting point for future biomaterial designs.

INTRODUCTION

Modern genomics combined with advanced bioinformatics methodologies allow us to understand much more about complex living systems than was ever previously possible. In the realm of human biology, for instance, recent developments have given us the ability to pinpoint the genes influencing diseases such as cancers. One area where these novel technologies can be anticipated to exert a huge impact but have thus far remained underused is the study of structural biomaterials. Spider silk is a prime example of an extended phenotype, whose extraordinary mechanical properties are governed by the underlying composition and structure of protein building blocks called spidroins.

All spiders use silk for various critical purposes, including foraging, locomotion, nesting, mating, egg protection, and communication (*1*). Different types of threads are used for diverse purposes, each produced in specific glands in the abdomen (*2*). For example, orb-weaving spiders use up to seven different types of silks, named after the gland that produces these threads. Major ampullate silk is the toughest silk used as draglines and as frames of orb webs, minor ampullate silk is used as scaffold during orb web weaving, piriform silk adheres the frame of the orb web to wood or other substrates, and capture thread of the orb web is composed of flagelliform silk backbone and aggregate glue. Aciniform silk is used for prey wrapping and sometimes for decorations of the web, and tubiform (or cylindrical) silk is used to make an egg sac. While spiders are successful predators and are often associated with orb webs, orb-weaving spiders of superfamily Araneoidea only comprise about 25% of spider species. A more ancestral clade of spiders such as those belonging to the infraorder Mygalomorphae is comprised mostly of ground-wandering spiders that produce sheet and maze webs for prey capture. Wandering hunters and abandoned silk capture webs make up a more modern clade of spiders in the retrolateral tibial apophysis (RTA) clade; this

[1]Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan. [2]Faculty of Environment and Information Studies, Keio University, Fujisawa, Kanagawa 252-8520, Japan. [3]Graduate School of Media and Governance, Keio University, Fujisawa, Kanagawa 252-8520, Japan. [4]Exploratory Research Center on Life and Living Systems (ExCELLS), National Institutes of Natural Sciences, Okazaki, Aichi 444-8787, Japan. [5]Biomacromolecules Research Team, RIKEN Center for Sustainable Resource Science, Wako, Saitama 351-0198, Japan. [6]Department of Material Chemistry, Kyoto University, Nishikyo, Kyoto 615-8510, Japan. [7]Japan Synchrotron Radiation Research Institute, Sayo-gun, Hyogo 679-5198, Japan. [8]Spiber Inc., Tsuruoka, Yamagata 997-0052, Japan. [9]Graduate School of Agricultural and Life Sciences, University of Tokyo, Yayoi, Bunkyo, Tokyo 113-8657, Japan. [10]Graduate School of Life and Environmental Sciences, University of Tsukuba, Tennodai, Tsukuba, Ibaraki 305-8572, Japan. [11]The United Graduate School of Agricultural Sciences, Kagoshima University, Korimoto, Kagoshima 890-0065, Japan. [12]Kokugakuin Kugayama High School, Suginami, Tokyo 168-0082, Japan. [13]Graduate School of Agriculture, Saga University, Saga 840-8502, Japan. [14]Evolution and Ecology Research Centre, University of New South Wales, Sydney, NSW 2052, Australia. [15]Department of Civil, Environmental and Mechanical Engineering, University of Trento, Via Mesiano 77, I-38123 Trento, Italy. [16]Beijing Advanced Innovation Center for Biomedical Engineering, School of Materials Science and Engineering, Beihang University, Beijing 100191, China. [17]Natural Materials Group, Department of Materials Science and Engineering, The University of Sheffield, Mappin Street, Sheffield S1 3JD, UK. [18]Department of Biomedical Engineering, Tufts University, Medford, MA 02155, USA. [19]School of Biological Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia. [20]Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati (IITG), Guwahati, 781 039 Assam, India. [21]Center for Nanotechnology, IITG, Guwahati, 781 039 Assam, India. [22]School of Health Sciences and Technology, IITG, Guwahati, 781 039 Assam, India. [23]Institute of Biological Sciences, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia. [24]School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, E1 4NS London, UK. [25]Institute for Frontier Materials, Deakin University, Waurn Ponds, VIC 3216, Australia. [26]College of Textile and Clothing Engineering, Soochow University, Suzhou 215123, China.
*Corresponding author. Email: gaou@sfc.keio.ac.jp (K.A.); numata.keiji.3n@kyoto-u.ac.jp (K.N.)
†These authors contributed equally to this work.
‡Present address: NMR Facility, Mark Wainwright Analytical Centre, University of New South Wales, Sydney, NSW 2052, Australia.

group comprises as much as 50% of all spider species (*3*). Therefore, spiders have diversified, selected, and specialized various uses of silk adapting to their ecological needs. Such extraordinary plasticity and university of silk and silk proteins is an ideal target to model the link between sequence and its physical property to fully understand the underlying design principles to apply the wide range of physical properties as biomaterials.

Spider silks are renowned for their diverse and impressive mechanical properties, frequently displaying a combination of high tensile strength, extensibility, and exceptional toughness that is unmatched industrially. Hence, the processing-property space that these silks occupy makes them a unique source of inspiration for protein biopolymer materials with low embodied energy and high performance (*4*–*6*). However, this property space has yet to be fully explored, defined, and exploited. Silk fiber diversity scales rapidly, as spiders produce multiple types of silk, each of which are composed of specific proteins known as spidroins, whose mostly monophyletic origins (*7*) endow them with specific mechanical properties (*2*, *8*). One type of spider silk protein, major ampullate spidroin (MaSp; which is often included in dragline threads), has received substantial academic and industrial attention, as this silk typically shows strength and toughness comparable to those of synthetic high-performance fibers, with an approximately 1-GPa breaking strength, a 30% breaking strain, and a toughness of 130 to 200 MJ/m^3 (*9*–*11*). However, there are lesser-known taxa and species of spiders, suggesting that the limits of silk properties are yet to be defined (*12*). On the other hand, a unique property known as supercontraction, where the dragline silk shrinks in length by up to 60% when wetted, is often considered undesirable industrially, and expectations are high for protein engineering methods to reduce such property by modifying the primary sequence. Hence, a comprehensive, coordinated global effort combining taxonomy, genomics, and materiomics is required to first understand and then unlock the true potential of these materials (*13*).

The diversity of spidroin sequences has been explored for decades. Pioneering work by Gatesy *et al*. (*14*) identified and analyzed spidroin sequences from several spider lineages, including basal spider groups, thus enabling a glimpse into the complex evolution of spidroin sequences. Subsequently, there have been a large number of studies that have explored the subject of spidroin sequence diversity and evolution, including focused studies on various spidroin paralogs (*15*–*29*), and those from more phylogenetic perspective (*7*, *30*–*34*), predominantly based on the conserved terminal sequences. On the other hand, the mechanical properties of silk fibers are governed largely through the repetitive regions that dominate the silk protein sequence, and the study on the diversity of spidroin repetitive regions, particularly in the more evolutionarily divergent taxa, has been limited to date. Thus, there is still an unmet need to map out the evolutionary design space of silk sequences and mechanical performance. This is especially relevant in light of the recent major breakthroughs in the field of spider phylogenomics (*3*, *35*, *36*). Undoubtedly, part of the reason for the scarcity of data on spidroin repetitive sequences has been the serious technical challenges faced when attempting to sequence highly repetitive low-complexity sequences such as found in silk proteins (compounded by the presence of multiple paralogs in the case of spider silk proteins). Recent advances in sequencing methods (*37*), however, have made such initiatives possible, as we present in this work. To address this need, we sequenced the silk genes of more than 1000 spider species encompassing the entire order Araneae using de novo transcriptome sequencing and assembly, alongside the comprehensive measurement of the material properties of their dragline silk fibers.

RESULTS AND DISCUSSION

Expanding the repertoire of silk genes

The transcriptomes of 1774 individual spiders were sequenced, which included 1098 species belonging to 441 genera and 76 families, globally sampled from four continents. Redundant sampling was performed for certain species to observe locality or sex differences in spidroin expression (*22*, *38*) and sequence variations within species. After the curation of the assembled transcripts, a total of 11,155 putative spidroin genes were identified (Fig. 1 and data file S1). All of the data are openly accessible from the Spider Silkome Database (https://spider-silkome.org).

The present study greatly expands the number and diversity of known spidroin sequences; we report sequences from 58 spider families not previously represented in public database, including members of basal taxa (Mesothelae, Mygalomorphae, Synspermiata, and allied groups), Araneoidea (which comprises the ecribellate orb weavers), and previously poorly sampled but extremely diverse groups such as the RTA clade and other taxa. At the time of writing this manuscript, spidroin sequences in the National Center for Biotechnology Information (NCBI) Protein database come from only 52 species in 18 families, and 23% of these sequence is derived from the single genus *Trichonephila*, and majority (73%) of the registered sequences are of major/minor ampullate spidroins (MiSps). In Fig. 1, family names colored in red indicate those with species where spidroin sequence is previously unreported, and family names with orange circles in front indicate those without previously reported transcriptome data. As the number of species indicates to the right of the family names, the vast majority of species reported in this work is previously unreported for spidroins and transcriptome data.

Within the "haplogyne" spider groups (Synspermiata and allied groups), we obtained sequences from nine previously unexplored families, including the first aciniform spidroin (AcSp), pyriform spidroin (PySp), and cribellar spidroin (CrSp) from these taxa. These proteins are consistent with more specialized silk types tuned to distinct biological functions in contrast to the undifferentiated spidroins identified from more ancestral Mesothelae and Mygalomorphae.

The most extensive sampling was conducted within Araneoidea, including family Araneidae, where we identified previously-unidentified spidroin sequences from the major subdivisions within the family (*39*), and likewise from underrepresented web-building families such as Tetragnathidae and Linyphiidae. Our results showed that the greatest diversity of spidroin types existed within the araneoid taxa, and spidroins associated with the capture spiral and aggregate glue of orb webs [flagelliform spidroin (Flag) and aggregate spidroin (AgSp)] were conserved only within the superfamily. Enrichment of the diversity of paralogs of the MaSp dragline gene was also observed in the group, and clear distinctions were possible among the different ampullate sequences (MaSp and MiSp) in terms of both terminal domain and repetitive sequences [for instance, MaSp2 is characterized by the presence of glutamine (Q)–containing dipeptide motifs in diglutamine (QQ)/proline-glutamine (PQ)/serine-glutamine (SQ)] (*40*). This is in contrast to Synspermiata and RTA clade sequences, where it is often difficult to distinguish between MaSp and MiSp types. The existence of a third type of MaSp (MaSp3), including

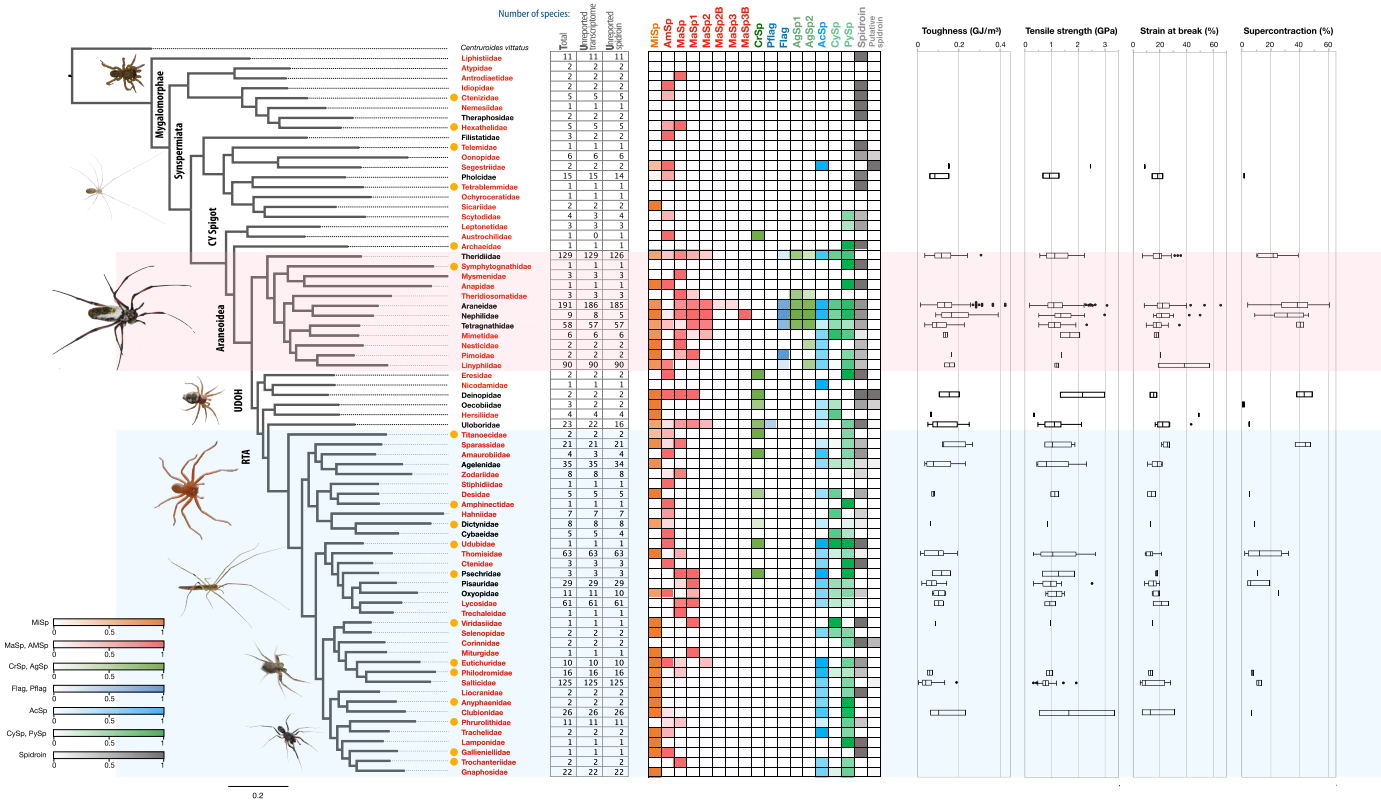

Fig. 1. Overview of the taxonomic distribution of spidroins and physical properties of dragline silks. Left: The phylogenetic tree of spider families constructed from the transcriptome data obtained from 1000 spiders in this work. The Araneoidea superfamily and the RTA clade are highlighted in red and blue, respectively. Family names in red represent those without previous report of spidroins in the NCBI Protein Database. Family names marked with orange circle represent those without previous transcriptome data. Total number of species sequenced in this work for each family, as well as species-level decomposition of unreported spidroin and transcriptome, are shown to the right of the family names. As this table shows, the vast majority of species reported in this work is previously unreported for their spidroin sequences or transcriptome. Middle: Heatmap of the conservation level of spidroin types within the spider families. For example, MaSp3 of family Araneidae has a value of around 0.5, as can be seen from the color code shown in the bottom left corner, which indicates that around 50% of the 191 species studied in this work contains MaSp3. The orb-weaving spiders in the superfamily Araneoidea (highlighted in pink) have greater diversity of spidroin types, and the RTA clade (highlighted in light blue) lost the capture web silks Flag and AgSp. MaSp sequence subtypes are not well differentiated in the RTA clade, where MiSp, ampullate spidroin (AmSp), and MaSp are more conserved than MaSp1 and MaSp2. Right: Distribution of physical properties among the spider families. Mirrored with the diversity of spidroins, orb-weaving Araneoidea spiders tend to have higher performance than other clades.

the nephilid variant MaSp3B, appears to be specific to Araneidae (see below) (*41*, *42*).

The RTA clade accounts for approximately half of all spider biodiversity, yet silk sequences from these mostly non–web-building groups have thus far received little attention. Our sampling identified a wide range of spidroin types from the RTA clade. To illustrate, we have identified the first spidroin sequences originating from jumping spiders (Salticidae), which has the highest species diversity among all spider families, with multiple representatives of MaSp, MiSp, AcSp, and cyllindrical spidroin (CySp), as well as unclassified spidroins from 63 different genera. We also extensively sampled spider groups situated between the araneoid and RTA clades [the so-called Uloboridae, Deinopidae, Oecobiidae, and Hersiliidae (UDOH) grade] and obtained the first reported spidroin sequences for Nicodamidae, Oecobiidae, and Hersilidae.

Insights from sequence analysis: Some highlights

The sequencing and annotation of the huge number and high diversity of spidroin genes from diverse spider taxa enable a deeper look into the more poorly resolved spidroin classes than previously possible. Here, we provide some examples of analyses made possible by access to such an extensive spidroin sequence database.

Cribellar spidroins: Highly conserved through evolution

From analysis of data from the most basal spider group (suborder Mesothelae: family Liphistiidae), we identified several new spidroin sequences that include the N-terminal domain region. We found that these sequences bear a close similarity with cribellar spidroins (CrSps), recently identified as a main constituent of the nonsticky capture threads of cribellate spiders (Fig. 2A). In addition, on the basis of analysis of sequences from the C-terminal side, we identified CrSp sequences from eight new families that encompass a wide phylogenetic spread (Fig. 1). Notably, we obtained the CrSp sequences from *Hickmania troglodytes* (Austrochilidae), a basal araneomorph species, in addition to representatives from Eresidae, Deinopoidae, and Oecobiidae, and a number of families from the diverse RTA clade. Analysis of the core repetitive regions of CrSp sequences showed a high degree of conservation of the amino acid composition even among widely separated groups (Fig. 2B). The most

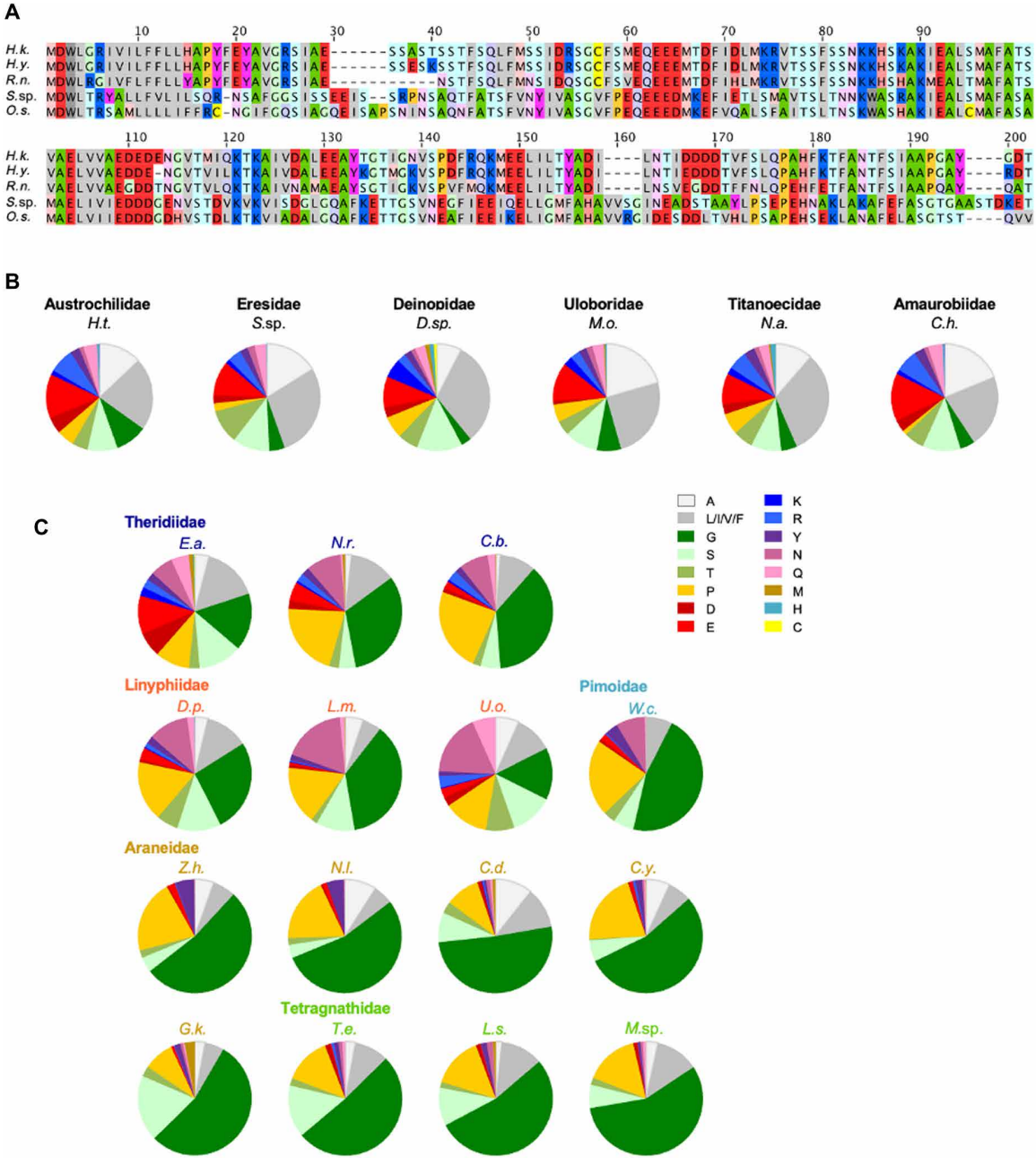

Fig. 2. Some insights from analysis of large spidroin dataset. (**A**) Spidroin N-terminal domains obtained from basal Mesothelae bear close resemblance to CrSp sequences. *H.k.*, *Heptathela kimurai* (Liphistiidae); *H.y.*, *Heptathela yanbaruensis* (Liphistiidae); *R.n.*, *Ryuthela nishihirai* (Liphistiidae); *S.*sp., *Stegodyphus* sp. (Eresidae); *O.s.*, *Octonoba sybotides* (Uloboridae). (**B** and **C**) Analysis of residue composition in spidroin repetitive regions, with residue types colored according to the legend. (B) Conservation of amino acid abundance in CrSp repetitive sequences across spider taxa. *H.t.*, *H. troglodytes*; *D.*sp., *Deinopis* sp.; *M.o.*, *Miagrammopes orientalis*; *N.a.*, *Nurscia albofasciata*; *C.h.*, *Callobius hokkaido*. (C) Conservation of amino acid abundance in Flag repetitive sequence among araneoid species. *E.a.*, *E. affinis*; *N.r.*, *Nesticodes rufipes*; *C.b.*, *Coleosoma blandum*; *D.p.*, *Doenitzius penículus*; *L.m.*, *Lepthyphantes minutus*; *U.o.*, *Ummelatia osakaensis*; *W.c.*, *Weintrauboa contortripes*; *Z.h.*, *Zygiella hiramatsui*; *N.l.*, *Nephilingis livida*; *C.d.*, *Caerostris darwini*; *C.y.*, *Cyrtarachne yunoharuensis*; *G.k.*, *Gasteracantha kuhli*; *T.e.*, *Tetragnatha extensa*; *L.s.*, *Leucauge subgemmea*; *M.*sp., *Mesida* sp.

notable feature of these repetitive sequences is the high abundance of charged residues (around 20%) and particularly of negatively charged glutamate (E) residues that occur as clusters interspersed throughout the sequence along with a relatively high proportion of hydrophobic amino acids leucine, isoleucine, valine, and phenylalanine (L, I, V, and F, respectively; collectively around 25%), a combination unique to CrSp sequences and not seen in other spidroin types.

Flag: One framework, diverse compositions

Flagelliform silk refers to the stretchable silk fibers produced by araneoid spiders (superfamily Araneoidea) and known particularly as making up the prey capture spirals of orb-weaver spiders. The sequence of the constituent Flag had previously been reported from only two families (Araneidae and Theridiidae), with the core repetitive sequences only available from Araneidae. Here, we have

considerably expanded the availability of Flag sequences by including previously unrepresented core repetitive and terminal sequences from the web-building families Theridiidae, Linyphiidae, Pimoidae, and Tetragnathidae (Fig. 1). Figure 2C shows the amino acid composition of Flag repetitive regions from a number of species from different families, wherein a diversity in the abundance of amino acid residues is clearly apparent. The most divergent repetitive sequences were found in Theridiidae, at the base of the araneoid clade, which also showed a larger number of residues represented compared to the more derived families. Some Flag repeat sequences from Theridiidae showed a marked resemblance to CrSp in terms of amino acid composition [as exemplified by *Episinus affinis* in Fig. 2C; compare with Fig. 2B]; this might reflect the close evolutionary link between Flag and CrSp, as previous studies have suggested (*21*, *26*). The Flag repetitive regions from other theridiid species tend to have a more reduced set of residues, with an abundance of proline and glycine residues. The species-rich Linyphiidae, predominantly sheet web builders, also exhibited somewhat divergent Flag sequences that feature short repeating motifs enriched in glycine (G), proline (P), asparagine (N), and serine (S). In contrast, Flag repeat sequences from the canonical orb-weaving families Araneidae and Tetragnathidae showed the most compositionally simplified Flag sequences, converging on a design that features a hyperabundance of glycine (G) residues (sometimes exceeding 50%) as well as proline (P) and/or serine (S) residues. It might be hypothesized that different araneoid spider groups have adapted the Flag repetitive sequences to fulfill different prey capture strategies; for instance, spiders that build orb webs designed to catch insects in flight (e.g., Araneidae and Tetragnathidae), where fiber extensibility is most important, correlate with the highest proportion of glycine in the repetitive regions.

Spider silkome: An integrated database of sequences and material properties

Along with the spidroin sequence data, dragline silk fibers were collected from selected spider species, which were then subjected to a comprehensive array of analyses to obtain the individual profiles across 12 index parameters, including mechanical performance (toughness, Young's modulus, tensile strength, and strain at break), morphological and structural properties [fiber diameter, birefringence, and degree of crystallinity based on wide-angle x-ray scattering (WAXS) analysis], thermal degradation profiles (onset temperature for 1, 5, and 10% weight loss), and hydration properties (fiber water content and degree of maximum supercontraction), for the reeled dragline silk of 446 spider species (Fig. 3 and fig. S1). Spiders belonging to Araneoidea show particularly diverse uses of threads (*43*), and the majority of the dragline samples included in this project was obtained from this superfamily, because the relatively large body size and copious fiber production of these species facilitate extended fiber collection.

Together, these data represent the largest collection obtained to date linking genotype to phenotype for a particular type of protein biopolymer (Spider Silkome Database; Fig. 3C), a fully searchable platform with integrated Basic Local Alignment Search Tool (BLAST) search capability. All sequence data are also available from DNA Data Bank of Japan (data files S1 and S2).

Study on the sequence to property linkage of spider silk has been a challenge, since the source of variability is threefold: interspecific, intraspecific, and intraindividual (*44*). Varying protocols for silking and mechanical property measurement also complicate meta-analysis, for which the silking strain rate and humidity is known to have significant effects (*45*). Our data are entirely obtained under a single standardized protocol and realize comprehensive comparisons. We therefore first observed the distribution of mechanical properties by families and genera. The mechanical property data obtained in this project represent an almost continuous spectrum of toughness reaching up to 0.45 GJ/m^3, a strain at break up to 60%, and a tensile strength up to 3 GPa (Fig. 3B and figs. S1 and S2); thus, this dataset seems promising for ascertaining relationships between the amino acid sequences of silk proteins and the physical properties of draglines across the spider phylogeny [see also Craig *et al.* (*46*)]. Toughness is highly correlated with the tensile strength and strain at break, as expected from its definition. Notably, the correlation between tensile strength and strain at break is low, indicating that the strength and elasticity of silk are independent factors (Fig. 3A and table S1). Birefringence reflects the degree of molecular orientation of silk protein chains and is a good predictor of tensile strength; crystallinity is a similar predictor for strain at break. Silk diameter is correlated with strain at break and supercontraction, but the latter probably represents a pseudo-correlation with Sparassidae and Araneidae silks, which tend to exhibit large diameters and high supercontraction. Overall, web-weaving spiders, or those belonging to the superfamily Araneoidea, tend to express superior mechanical, physical, structural, thermal, and water-based properties relative to basal spider groups (Fig. 3B and fig. S1). Diversity in the mechanical properties was also the largest in the family Araneidae, mirrored by the high variability in the repetitive region sequences of MaSp-type spidroins (fig. S3), whose diversity nearly covers the entire variability within the 1000 spiders encompassing 76 families.

We conducted variable selection to probe structure-function associations in dragline silks based on the mean differences in the physical properties of the silks according to taxonomic categories and spidroin types (figs. S4 to S6). Briefly, the different ampullate-like spidroin sequences found across the different spider taxa were classified according to conserved patterns within repetitive domains; this led to the categorization into 20 sequence groups, which comprised seven MiSp subtypes, seven MaSp1 subtypes, four MaSp2 subtypes, and two MaSp3 subtypes, including MaSpN. We then analyzed the contributions of the different groups to the different physical properties of the corresponding dragline fibers. For instance, the silks of spiders from the genus *Argiope* and family Araneidae showed significantly higher toughness (mean differences of +0.068 and +0.039 GJ/m^3, respectively) and expressed unique spidroins, including MaSp3 (group 19), MaSp2 (group 11), and MaSp1 (group 17), resulting in mean differences in silk toughness of +0.041, +0.031, and +0.035 GJ/m^3, respectively (Fig. 4A and fig. S7A). This suggests that the possession of MaSp3 (group 19) resulted in an increase in toughness of at least 0.041 GJ/m^3, corresponding to an increase of approximately 32% relative to the overall average of 0.127 GJ/m^3. However, this was most likely as combined effect of Araneidae-type MaSps, including MaSp2 (group 11) and MaSp1 (group 17), coinciding with the existence of MaSp3 (group 19). A similar significant superiority of Araneidae dragline fibers was observed in terms of strain at break, crystallinity, diameter, thermal degradation temperature, and supercontraction. Strain at break and supercontraction were the only properties for which the possession of the MaSp2 subtype was a greater determinant than belonging to family Araneidae, as tensile strength increased 3.7% in association with MaSp2 (group 13)

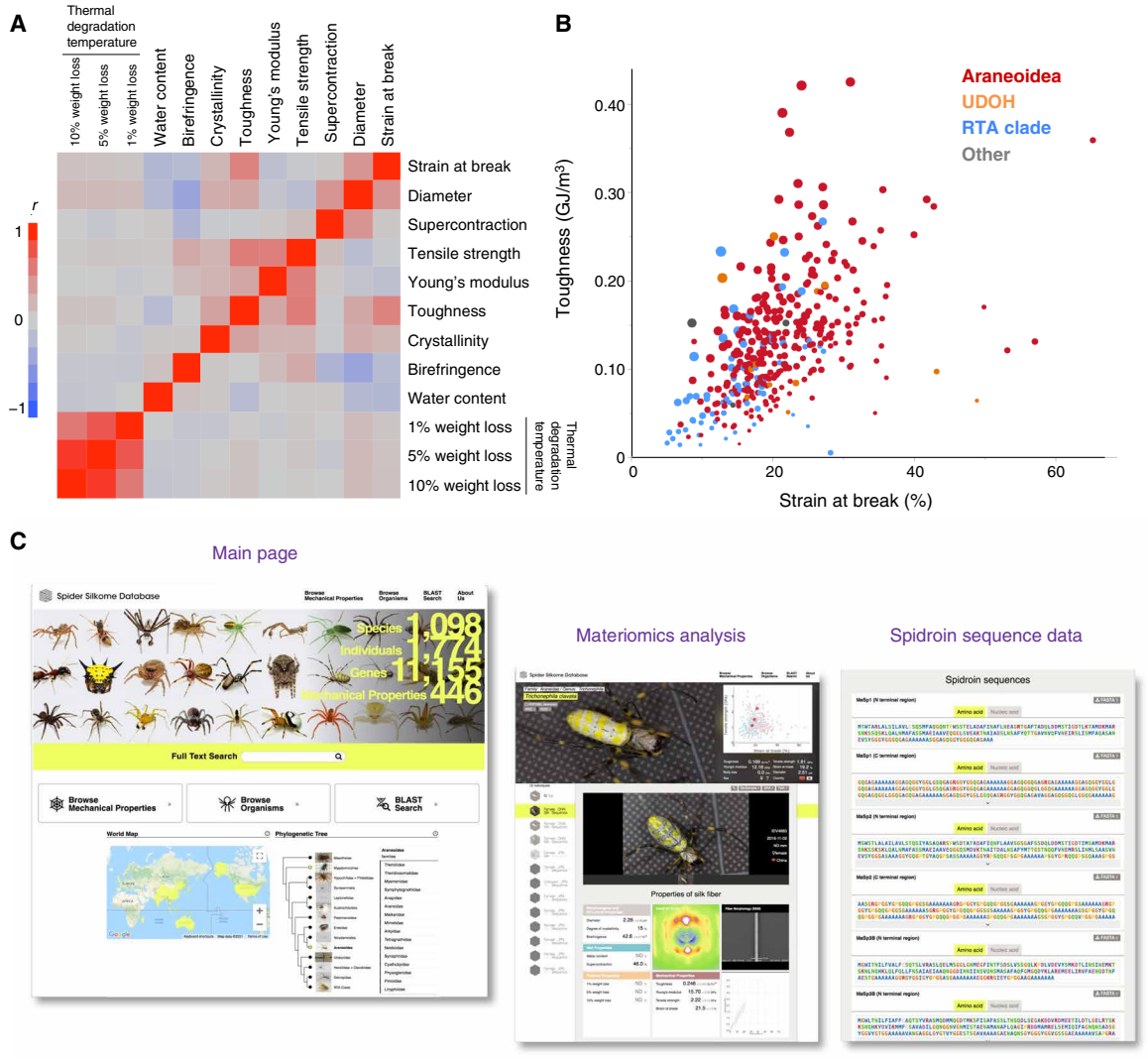

Fig. 3. Overview of the physical properties of 446 spider silk samples. (**A**) Pearson correlation heatmap of the physical properties of dragline silk fibers measured in this work. Toughness is not only correlated with tensile strength and strain at break but also correlated with Young's modulus. Supercontraction is correlated with strain at break. (**B**) Scatter plot of toughness versus strain at break (with spot size proportional to tensile strength). The collected samples represent an almost continuous spectrum of toughness from <0.01 to >0.40 GJ/m^3. Spots are colored according to broad phylogenetic grouping: Araneoidea (red) includes the orb-weaving spiders and tends to show a relatively high toughness distribution relative to wandering species (such as the RTA clade, indicated in light blue). (C) Screenshots of the Spider Silkome Database (https://spider-silkome.org), a fully searchable, public repository of all spidroin sequences and material property data generated from the 1000 spider silkome project (the main page and individual profile data for *Trichonephila clavata* are shown).

and supercontraction increased 15.7, 15.8, 14.3, and 11.0% in association with MaSp2 (group 14), MaSp2 (group 13), MaSp2 (group 11), and MaSp1 (group 17), respectively (Fig. 4B and fig. S7B). The significant contribution of MaSp2 to spider dragline supercontraction and elasticity was in line with previous suggestions regarding the different roles of MaSp1 and MaSp2 (*47*, *48*), but one spidroin subtype, MaSp2 (group 15), conversely influenced supercontraction (−8.3%; see fig. S7B). A close inspection of the repetitive motifs of MaSp2 (group 15) revealed longer polyalanine regions. Accordingly, the average β sheet region length (typically the polyalanine region but defined as stretches of multiple A, S, and V for more than five amino acid residues, as these amino acids tend to substitute for polyalanine) was negatively correlated with supercontraction (−0.508 for MaSp1 and −0.306 for MaSp2 β sheet regions). Furthermore, the correlation was higher when both the amorphous region and the polyalanine lengths were taken together in the ratio (figs. S8 and S9). The average amorphous to β sheet region length ratios for all repeats within the spidroins of interest were 0.526 for MaSp1 and 0.394 for MaSp2. Therefore, the proportion of amorphous regions within the spidroin is the key factor contributing to supercontraction. The contribution of the relaxation of orientation in the amorphous region of spidroins to supercontraction was suggested in previous works (*49*, *50*) and was confirmed by the analysis of our comprehensive dataset. Considering the effects of the amorphous and crystalline regions on the measured physical properties, as described above, the repetitive sequences, rather than the terminal domains, can be considered to play the main roles in determining these physical and mechanical properties. Shrinkage of artificial spider silk threads

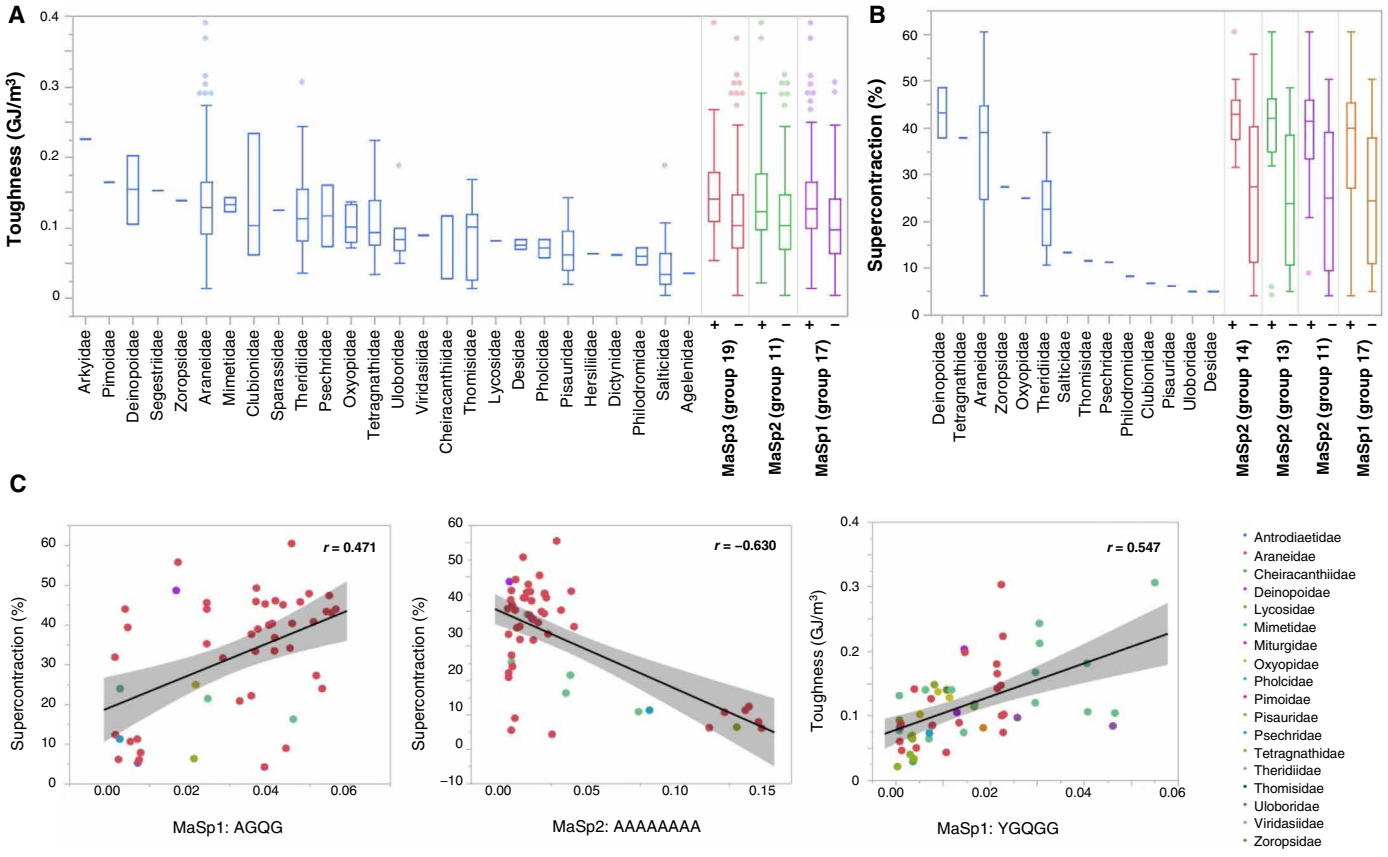

Fig. 4. Linking sequences to the physical properties of dragline silk. The different ampullate-like spidroin sequences found across the different spider taxa were classified according to conserved patterns within repetitive domains; this led to the categorization into 20 sequence groups, which comprised seven MiSp subtypes, seven MaSp1 subtypes, four MaSp2 subtypes, and two MaSp3 subtypes (figs. S3 to S5). MaSp groups most strongly contributing to the physical properties were selected through statistical screening (see Materials and Methods). (**A**) Toughness distribution among different spider families, as correlated with the presence or absence of selected MaSp subtypes: MaSp3 (group 19), MaSp2 (group 17), and MaSp1 (group 17). (**B**) Supercontraction distribution among different spider families, compared with the presence (+) or absence (−) of specific MaSp subtypes. Four MaSp2 groups (groups 14, 13, 11, and 17) showed higher average supercontraction than Araneidae. (**C**) Scatterplot of physical properties (toughness or supercontraction) as a function of the average abundance per repeat (%) of certain amino acid motifs. See data file S4 for comprehensive screening of amino acid sequence motifs contributing to the physical properties. Abundance of motifs was normalized by the number of repetitive sequences within a spidroin fragment, and this normalized abundance was correlated with the physical properties to screen for highly contributing motifs. Spot color denotes the spider family, and Pearson correlation values are shown in the top right corners. Here, AGQG motif in MaSp1 is positively correlated with supercontraction, and AAAAAAAA motif of MaSp2 is negatively correlated. Likewise, YGQGG motif in MaSp1 is positively correlated with toughness.

and textiles due to supercontraction is often considered an undesirable property for industrial use, and these findings may contribute in designing primary sequences, avoiding supercontraction while preserving toughness of the material.

To further extract the sequence features contributing to the physical properties of spider silk, we screened the amino acid motifs correlated with the measured properties (data file S4), and the main findings are summarized in Table 1. Confirming the above analysis of categorical variable selection according to gene class and taxonomy, the degree of supercontraction was strongly negatively correlated with the frequency of the appearance of polyalanine sequences and was correlated with short (one– to four–amino acid) motifs corresponding to amorphous regions such as G, GG, and AGQG (Fig. 4C and data file S4). Likewise, strain at break was negatively correlated with polyalanine prefixed with Ser in MaSp2 and positively correlated with MaSp1/2 amorphous regions including Pro, which presumably adds to the elasticity of this region (*51*). Concerning tensile strength, the inclusion of Ala in the amorphous region of MaSp1 and Pro in that of MaSp2 had a negative effect, while the inclusion of Ser in the amorphous region of MaSp1 had a positive influence. The GYGQGG motif in MaSp1 was most strongly correlated with both tensile strength ($r = 0.377$) and strain at break ($r = 0.416$) and was consequently also correlated with toughness [YGQGG was ranked 1 ($r = 0.547$), and GYGQGG was ranked 2 ($r = 0.531$)] (Fig. 4C). The Tyr residues in the amorphous regions of MaSp1 may play a critical role in intermolecular chain packing in the spider dragline, similar to the intermolecular interactions suggested from the structural analysis of silkworm silk (*52*) . The inclusion of Pro in the MaSp2 amorphous region, along with the SY and SV motifs in MaSp1, was negatively correlated with toughness. The presence of GGS after the polyalanine region in MaSp1 was positively correlated with toughness. Confirmation of the contribution of these motifs to the physical properties using recombinant properties would be a future direction to fully understand the primary sequence designs, leading to the extraordinary mechanical properties of spider silk.

Table 1. Feature extraction summary. Amino acid sequence features of the underlying MaSp repetitive domains that have positive and negative effects on the different physical properties of spider dragline silks are presented. Poly-Ala, polyalanine.

	Positive effect	Negative effect
Toughness	MaSp1-GYGQGG	P, SQGP in MaSp2
	MaSp1–poly-Ala ending with GGS	SY, SV in MaSp1
	MaSp1-GGGQ	
Tensile strength	MaSp1-GYGQGG	MaSp2-PQ
	MaSp1-SS before poly-Ala	Lacking S in GQG motif in MaSp1
	MaSp1-QGGS	A before GQG motif in MaSp1
Strain at break	MaSp1-GYGQGG	ASA before poly-Ala
	QGP, PGA in MaSp1	
Young's modulus	PA in MaSp2	Q in MaSp2
	GL in MaSp1 and MaSp2	MaSp1-GGQ
	MaSp1-GQ	
Crystallinity	PA, N, A, GA in MaSp2	GT in MaSp1
	MaSp1-GQ	MaSp1-GGQ
Birefringence	SS, N, GQQ in MaSp2	MaSp1-GQGGAGAA
	TGG in MaSp1	
Diameter	MaSp1-GAAAAAAG	MaSp2-PSGPGS
	MaSp1-AAGGAGQG	MaSp2-SQG
	MaSp2-PQG	MaSp2-AAGGY MaSp1-QS
N% water loss	MaSp2-PGGYGP	MaSp1-SQGAG
	MaSp2 poly-Ala	V in MaSp2
		GT in MaSp1
Water content	MaSp1-GSG	MaSp2-QQGPG
	MaSp2-GAS	MaSp1-PGAA
		A in MaSp1 and MaSp2
Supercontraction	MaSp2 presence	Poly-Ala in MaSp1 and MaSp2
	MaSp1-AGQG	
	MaSp1-GLG	

Together, our findings provide a thorough mechanistic evaluation of the pathways of spidroin evolution. First, the physical properties of spider dragline silk have significantly diversified and specialized with the deployment of orb webs related to Araneoidea species (*43*), and this is mirrored by the diversification of MaSp paralogs, as previously suggested through meta-analysis of silk mechanics and sequence motifs (*46*) . We propose that MaSp1 is specialized to increase fiber strength, while MaSp2 is specialized to increase fiber elasticity, and the combination of these paralogs results in the high toughness of dragline silk. Furthermore, species requiring extraordinary fiber toughness have evolved to produce a third paralog, MaSp3, whose presence was clearly shown to be one of the strongest determinants of high toughness in our analysis. The full complexity of the proteome composition of dragline silk is beginning to be elucidated. However, MaSp3 was shown to be the major component of Nephilinae and *Araneus* dragline silks, and the complexity of these silks extends beyond the composition of spidroins (*42*), involving other essential components referred to as spider silk-constituting elements (SpiCE), which has been shown to double the tensile strength of an artificial spider silk–based film in vitro (*41*) . Elasticity and supercontraction are related properties of dragline silk that are likely linked to the sequence features of MaSp2, in which the ratio of amorphous to β sheet regions plays critical roles. Similarly, the compositions of several amino acid motifs in the amorphous regions of MaSp1 were shown to be highly correlated with the toughness of dragline silk; these sequence-level design elements derived from the comprehensive analysis of 1000 spiders provide a foundation for the design and production of artificial spider silks. Many of these designs may also be applicable to other protein-based and polymeric materials.

In this study, we have provided a comprehensive dataset encompassing the genotypes and phenotypes (including the mechanotypes) of spider silks and identified the design elements responsible for the extraordinary mechanical and physical performances of these silks. Silk proteins have convergently evolved in various lineages (*53*), but the sequence motifs (*54*), amino acid composition (*55*), and the trade-off between tensile strength and elasticity as a function of ratio between amorphous and crystalline regions (*56*) have been shown to have a certain degree of shared characteristics, something supported by our spider silk data. Therefore, these data will serve as a framework for the future analysis of silk proteins and other structural proteins as biomaterials. Similar data-driven approaches encompassing protein materials excelling in properties other than toughness, such as elastomers and adhesive proteins, could also accelerate our understanding on the genetic design principles of the biomaterials. Methods including computational modeling and simulation that allow the prediction of the outcomes of molecular interactions between the multiple components of these biomaterials, such as multiple MaSp-type spidroins and SpiCE proteins, would be an important future direction. We focused on the silk mechanics in this work, but the 1000 spider transcriptome data should also facilitate arachnid and arthropod phylogenomics.

MATERIALS AND METHODS

Spider sampling

Field work took place from 2014 to 2019 in Japan, Malaysia, United States, China, India, United Kingdom, Australia, Madagascar, and Italy (data file S1). In each field work session, the collected spiders were stored in a New PP Sample Tube (Maruemu Corporation, Osaka, Japan) and transported live back to the laboratory. Spider specimens were identified by the method described in the "Species identification" section. Immediately after arrival at the laboratory, photographs of the collected spiders were taken on 1-cm by 1-cm grids to measure their total body length, silk was sampled by the method described in the "Silk sampling" section, and specimens were preserved or RNA was extracted for transcriptome sequencing. In field work conducted in countries other than Japan, preserved specimens or RNA samples were transported back to the laboratory. The total body length of spiders was measured from the photographs by using ImageJ.

Silk sampling

Spider silks were forcibly silked from captured spiders immobilized using two pieces of sponge and locked with rubber bands. After immobilization, the silks spun from spinnerets were obtained with tweezers and attached to the end of the bobbin. Dedicated reeling devices were used for silking with a constant reeling speed (1.28 m/min). The duration of forcible silking was 1 hour at most. After silking, the bobbin with the reeled silk was placed in a plastic bag and stored in a cardboard preservation box at room temperature.

Species identification

All spiders were morphologically identified by A. Tanikawa and subsequently confirmed by Cytochrome c oxidase subunit I (COI) sequencing from the transcriptome assembly based on BLAST searches in the Barcode of Life database with a 90% identity threshold. If the COI sequence could not be recovered from the transcriptome assembly, additional Sanger sequencing was conducted by amplifying the cDNA with the primer sets COI1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′)–COI2198 (5′-TAAACTTCAGGGTGACCAAAAAATCA-3′) and COI1718 (5′-GGAGGATTTGGAAATTGATTAGTTCC-3′)–COI2776 (5′-GGATAATCAGAATATCGTCGAGG-3′) (*57*, *58*). Spiders that were difficult to identify on the basis of morphology or COI database searches were clustered into groups of operational taxonomic units (OTUs). The OTU clusters were defined by a 98% identity threshold in BLAST searches.

Transcriptome sequencing and assembly

Sample preservation, RNA extraction, sequencing, and assembly were conducted on the basis of methods previously described (*59*) with some modifications. Briefly, a single specimen of each of the spiders brought to the lab alive was flash frozen with liquid nitrogen and stored at −80°C until use. Samples of spiders that were difficult to transport alive were stored in RNAlater, in which they were initially held at 4°C for 24 hours and then stored below −20°C. RNA was extracted after homogenization on a multibead shocker with metal cones (Yasui Kikai) using TRIzol (Thermo Fisher Scientific), followed by purification with an RNeasy Plus Mini kit. Small specimens (body size, <5 mm) were extracted using a Direct-zol RNA Microprep kit (Zymo Research). RNA quality was checked using RNA ScreenTape on a TapeStation 2100 (Agilent) according to an RNA integrity number (RIN) of >6 and was quantified using Qubit v.3 (Life Technologies) and NanoDrop 2000 (Life Technologies) systems. The Illumina library was prepared using the NEBNext Ultra II RNA Library Prep Kit for Illumina (New England Biolabs); however, for samples with available amounts below the required input amount (<20 ng of total RNA), preparation was performed using the SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Clontech), followed by fragmentation and cDNA library preparation with the KAPA HyperPlus Kit (KAPA Biosystems). The sequence library was then sequenced as paired-end reads on the NextSeq 500 platform (Illumina) via 300 cycles in high-output mode. The sequences were subjected to base calling and demultiplexing, and adaptor sequences were removed with bcls2fastq v.2 software (Illumina). Transcriptome assembly was performed using Bridger software with the default parameters using Illumina reads (*60*). To eliminate possible cross-contamination, transcripts with mapped read count per million values of less than 1 and "comp" numbers greater than 30,000 in the Bridger assembly were removed. See data file S1 for the list of Sequence Read Archive (SRA) and Transcriptome Shotgun Assembly (TSA) accession numbers.

Direct RNA sequencing

The direct RNA sequencing library was constructed with the SQK-RNA001 kit (Oxford Nanopore Technologies). More than 500 ng of mRNA was prepared from the extracted total RNA using the NucleoTrap mRNA Mini Kit (Clontech), and library generation was completed following the manufacturer's protocol. Appropriate numbers of individual samples were prepared according to the amount of total RNA from each species (data file S2). Direct RNA sequencing was performed using a MinION device, and one v9.4 SpotON MinION flow cell (FLO-MIN106, Oxford Nanopore Technologies) per species was used. The produced reads were corrected by using proovread (v2.13.4) (*61*) .

Spidroin curation and nomenclature

Spidroin gene curation was performed using a previously reported spidroin motif collection algorithm (*37*, *42*) . The BLAST search detected contigs, including the spidroin gene N/C termini (nonrepetitive region). The obtained spidroin terminus contigs were used as seeds to screen the short reads harboring exact matches of extremely large *k*-nucleotide oligomers (approximately 100) up to the 5′-end of the seed. The selected short reads were aligned on the 3′-side seed of the matching *k*-nucleotide oligomer to build a position weight matrix (PWM). Using very stringent thresholds, the seed sequence was extended on the basis of the PWM until there was a split in the graph (i.e., the neighboring repeats were not resolvable). By iterating this overlap-based extension process, we obtained the full subsets of the repeat units. The collected repeat units were mapped onto error-corrected long reads obtained by direct RNA sequencing. Last, the spidroin gene length or architecture data were manually curated on the basis of the mapped long reads. The obtained spidroins were categorized into the following groups based on sequence homology with known spidroin data: AcSp, AgSp1/AgSp2, CrSp, CySp, Flag, MaSp1 to MaSp3, MiSp, Pflag, PySp, ampullate spidroin (MaSp or MiSp), spidroin, and putative spidroin (no homology but a spidroin-like structure).

Spidroin grouping

Curated MaSp/MiSp/spidroins were categorized into groups based on repetitive motifs. The repetitive regions and terminal sequences of the curated spidroins were separated computationally using the frequency of 5-nucleotide oligomer amino acids. The tree containing all spidroins was created on the basis of the N-terminal region sequences. The phylogenetic trees were created with FastTree (v2.1.10, default option) (*62*) after alignment with MAFFT (v. 7.273, maxiterate option 1000) (*63*) and trimming with trimAl (v. 1.2rev59, gt option 0.2) (*64*). Because the definition of MaSp and MiSp proteins based on sequence information was ambiguous, very MiSp-like MaSp (or vice versa) proteins were scattered. Therefore, we redefined the groups by clade to discuss them separately. A clade consisting of only spidroins of the same type was defined as a group.

Measurements of morphological and structural properties

The surface morphology and cross sections of the dragline silk fibers were assessed via scanning electron microscopy (SEM) (JCM-6000, JEOL Ltd., Tokyo, Japan) according to a previous report (*65*) . The samples were mounted on an aluminum stub with conductive tape and sputter-coated with gold for 1 min with a Smart Coater (JEOL, Tokyo, Japan) before SEM visualization at 5 kV.

Birefringence measurements

The retardation provided by the silk fiber was measured with a WPA-100 birefringence measurement system (Photonic Lattice Inc., Miyagi, Japan) and was analyzed with WPA-VIEW (version 1.05) software in accordance with a previously reported method (*65*). The birefringence of the dragline silk fiber was calculated from the retardation value and silk fiber diameter, which was determined via SEM.

Measurements silk of mechanics

Tensility tests of the single dragline silk fibers were conducted with a mechanical testing apparatus (EZ-LX/TRAPEZIUM X, Shimadzu, Kyoto, Japan) at 25°C and a relative humidity of approximately 50% according to a previous report (*45*) . The initial length of the single dragline silk fiber was set to 5 mm. The extension speed was applied at 10 mm/min, and the force during testing was measured with a 1-N load cell. The tensile strength, Young's modulus, elongation at break, and toughness were obtained from the resultant stress-strain curves. To assess the tensile strength, the cross-sectional areas of the fiber samples were calculated on the basis of the diameters determined by SEM observations.

Thermal property measurements

Simultaneous thermogravimetric analysis (TGA) and differential scanning calorimetry (DSC) were conducted in triplicate using spider silk samples with a total mass of 0.5 to 1.0 mg according to a previous report (*65*) . Samples were encapsulated in aluminum pans and heated under a nitrogen atmosphere at a rate of 20°C/min from 30° to 500°C using a TGA/DSC 2 instrument (Mettler Toledo, Greifensee, Switzerland). The device was calibrated with an empty cell baseline and with indium for heat flow and temperature. The degradation temperatures that yielded 1, 5, and 10% weight losses in the silk samples were defined as degradation temperatures of 1, 5, and 10% (T_{d1}, T_{d5}, and T_{d10}). The water content was calculated from the percent weight loss associated with the evaporation of bound water from the TGA data based on a previous silkworm silk study (*65*).

Synchrotron WAXS measurements

Spider silk fibers were aligned in bundles and subjected to synchrotron WAXS at 12.4 keV at the BL45XU beam line at SPring-8 (Harima, Japan), as described in previous literature (*45*, *66*). The data collection parameters included a wavelength of 1.00 Å, a beam size of 250 μm by 150 μm ($H \times V$), and an exposure time of 10 s at 25°C and 40% relative humidity. Diffraction patterns were recorded using Pilatus 2 M (Dectris Ltd., Switzerland) with a sample-to-detector distance of 179.6 mm. The module gaps of the detector according to offset measurement were complemented. The two-dimensional (2D) diffraction patterns were converted into 1D profiles using FIT2D (*67*), with corrections made for background scattering and detector geometry. The degree of crystallinity of the silks was calculated from the 1D profile. Each dataset was separated into crystalline and amorphous scattering components by curve fitting using Gaussian functions. The ratio of the total area of the separated crystalline scattering components to that of the crystalline and amorphous scattering components was used to determine the degree of crystallinity.

Maximum supercontraction

The supercontraction of spider silks was evaluated according to a previous method (*68*). Individual dragline fibers were prepared by cutting fragments of 5 to 10 cm (L_0), to which a small piece of vinyl tape was affixed on either end. The fibers were immersed in Milli-Q water for 1 min to allow supercontraction and then allowed to air dry overnight in an unrestrained state. The final length of the fiber (L_f) was measured, and the maximum supercontraction (%) was calculated as $(L_0 - L_f)/L_0 \times 100$. At least six replicates were performed for each sample; all measurements were performed with a caliper.

Silkome database (https://spider-silkome.org)

Top page

On the top page of Spider Silkome Database, there are full-text search menu buttons linked to the "Browse Mechanical Properties," "Browse Organisms," and "BLAST Search" pages. Under the menu buttons, there is a world map and phylogenetic tree. The world map shows the regions where we performed field work, indicated in yellow. Clicking the indicated areas results in a pop-up display of the numbers of collected spiders and a link to a list of these individuals. Users can search individual spiders with the area where they were collected. The phylogenetic tree shows the names of clades, infraorders, superfamilies, and families of spiders. Clicking the branches of the phylogenetic tree shows a list of families in the clicked branch next to the tree. These family names are linked to organism pages so users can view the spiders in the phylogenetic tree.

Browse mechanical properties

On the Browse Mechanical Properties page, there is a table of the properties of the silk samples. The properties include mechanical properties, thermal properties, morphological properties, and structural properties as well as wet properties (water content and supercontraction). By clicking the checkboxes for each property, the visibility of property columns in the table can be toggled. The interactive search box at the top of the page can be used to narrow down the results by scientific name interactively. The "Scatter graph" button next to the property check boxes opens scatter graphs of the properties of silks in a new window. Each data point in the scatter graph is linked to an individual page. Users can change the type of properties for the x and y axes to easily view the relationships between properties. The "Download CSV" button can be used for exporting data on the properties of silk samples. By selecting check boxes on the left of each row, users can select data to export. There are sliders on the top of the property columns to narrow the results by the value of the property. To narrow the results according to a lower threshold, the user first right clicks "<" and the number and then moves the slider. To narrow the results according to an upper threshold, the user first left clicks the "<" symbol and the number before moving the slider.

BLAST searches

On the BLAST Search page, users can search spidroins with protein sequences or nucleotide sequences. The "DB: Protein" and "DB: Nucleotide" tabs are used to select the database of the BLAST search, and the "Query type" select button is used for selecting the query type of the search. The program for searching is automatically selected by the combination of the tab (database) and select button (query). When the tab is DB: Protein and the select button is "Protein," then blastp is used for searching. When the tab is DB: Protein and the select button is "Nucleotide," then tblastn is used for searching. When the tab is DB: Nucleotide and the select button is Protein, then blastx is used for searching. When the tab is DB: Nucleotide and the select button is Nucleotide, then blastn is

used for searching. By clicking the "Download FASTA" button on the results page, users can export result sequences to FASTA format files.

Spider entity page

The entity page represents the species of the spider. The scatter plot on the right side of the top area is a plot of the tensile strength and strain at break data of all spiders. Pink circles indicate data from the same family as the entity page species. Large red circles indicate the data of the species of the entity page. The property table under the scatter plot shows the median values of each property.

Below the top area, photographs, silk sample properties, and spidroin sequences of each individual are provided. The links to the top right of the individual are external links to the NCBI BioSamples, SRA, and TSA databases. In the properties of silk fibers section, tables of each type of property, WAXS 2D profiles, SEM images, and stress-strain curves of the silk samples are provided. In the spidroin sequences section, the amino acid sequences of spidroins are provided. By clicking the "Amino acid" and "Nucleic acid" tabs, users can toggle the sequence panels. Users can download FASTA sequences by clicking the "FASTA" button on the upper right of the sequence.

Statistical analyses

For categorical variable selection, mechanical properties were tested to evaluate differences between their mean values for those belonging to the category and those not in the category, with the unpaired Student's *t* test with a *P* value threshold of <0.01 using the G-language Genome Analysis Environment v.1.9.1 (*69*–*71*) . The categories used in this analysis were the family and genus of the spiders as well as the spidroin type. For example, the toughness value distribution of family Araneidae was compared with those of all other families. A minimum of five samples was required to belong to a category.

For motif extraction, repetitive regions of spidroin sequences were first extracted as the longest segments, spanning an amino acid motif composed of S, A, or V with a length greater than four. Subsequently, these regions were split into repeat units segmented by SAV motifs with a length greater than five. This SAV region was defined as the crystalline region, and the remaining amino acids within the repeat were considered the amorphous region. Amino acid motifs with lengths of one to eight were counted in the repetitive region and divided by the number of repeats. This occurrence value was averaged for all MaSp1, MaSp2, and MaSp3 paralogs, and the correlations of these values with the mechanical properties were calculated using Pearson correlation in the G-language Genome Analysis Environment v.1.9.1 (*69*–*71*) . Motifs appearing less than 100 times in total among all spidroin sequences and motifs appearing in less than 30 samples were discarded. Graphs were visualized using JMP v.15 software.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abo6043

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. F. Vollrath, P. Selden, The role of behavior in the evolution of spiders, silks, and webs. *Annu. Rev. Ecol. Evol. Syst.* **38**, 819–846 (2007).
2. F. Vollrath, D. Porter, Spider silk as archetypal protein elastomer. *Soft Matter* **2**, 377–385 (2006).
3. R. Fernandez, R. J. Kallal, D. Dimitrov, J. A. Ballesteros, M. A. Arnedo, G. Giribet, G. Hormiga, Phylogenomics, diversification dynamics, and comparative transcriptomics across the spider tree of life. *Curr. Biol.* **28**, 2190–2193 (2018).
4. N. C. Abascal, L. Regan, The past, present and future of protein-based materials. *Open Biol.* **8**, 180113 (2018).
5. J. A. Kluge, O. Rabotyagova, G. G. Leisk, D. L. Kaplan, Spider silks and their applications. *Trends Biotechnol.* **26**, 244–251 (2008).
6. K. Numata, *Biopolymer Science for Proteins and Peptides* (Elsevier, 2021).
7. J. E. Garb, N. A. Ayoub, C. Y. Hayashi, Untangling spider silk evolution with spidroin terminal domains. *BMC Evol. Biol.* **10**, 243 (2010).
8. M. Humenik, T. Scheibel, A. Smith, Spider silk: Understanding the structure-function relationship of a natural fiber. *Prog. Mol. Biol. Transl. Sci.* **103**, 131–185 (2011).
9. R. Madurga, G. R. Plaza, T. A. Blackledge, G. V. Guinea, M. Elices, J. Pérez-Rigueiro, Material properties of evolutionary diverse spider silks described by variation in a single structural parameter. *Sci. Rep.* **6**, 18991 (2016).
10. F. G. Omenetto, D. L. Kaplan, New opportunities for an ancient material. *Science* **329**, 528–531 (2010).
11. D. Porter, J. Guan, F. Vollrath, Spider silk: Super material or thin fibre? *Adv. Mater.* **25**, 1275–1279 (2013).
12. I. Agnarsson, M. Kuntner, T. A. Blackledge, Bioprospecting finds the toughest biological material: Extraordinary silk from a giant riverine orb spider. *PLOS ONE* **5**, e11234 (2010).
13. A. Rising, H. Nimmervoll, S. Grip, A. Fernandez-Arias, E. Storckenfeldt, D. P. Knight, F. Vollrath, W. Engström, Spider silk proteins-mechanical property and gene sequence. *Zoolog. Sci.* **22**, 273–281 (2005).
14. J. Gatesy, C. Hayashi, D. Motriuk, J. Woods, R. Lewis, Extreme diversity, conservation, and convergence of spider silk fibroin sequences. *Science* **291**, 2603–2605 (2001).
15. N. A. Ayoub, K. Friend, T. Clarke, R. Baker, S. M. Correa-Garhwal, A. Crean, E. Dendev, D. Foster, L. Hoff, S. D. Kelly, W. Patterson, C. Y. Hayashi, B. D. Opell, Protein composition and associated material properties of cobweb spiders' gumfoot glue droplets. *Integr. Comp. Biol.* **61**, 1459–1480 (2021).
16. N. A. Ayoub, J. E. Garb, A. Kuelbs, C. Y. Hayashi, Ancient properties of spider silks revealed by the complete gene sequence of the prey-wrapping silk protein (AcSp1). *Mol. Biol. Evol.* **30**, 589–601 (2013).
17. N. A. Ayoub, J. E. Garb, R. M. Tinghitella, M. A. Collin, C. Y. Hayashi, Blueprint for a high-performance biomaterial: Full-length spider dragline silk genes. *PLOS ONE* **2**, e514 (2007).
18. P. L. Babb, N. F. Lahens, S. M. Correa-Garhwal, D. N. Nicholson, E. J. Kim, J. B. Hogenesch, M. Kuntner, L. Higgins, C. Y. Hayashi, I. Agnarsson, B. F. Voight, The Nephila clavipes genome highlights the diversity of spider silk genes and their complex expression. *Nat. Genet.* **49**, 895–903 (2017).
19. D. Bittencourt, K. Dittmar, R. V. Lewis, E. L. Rech, A MaSp2-like gene found in the Amazon mygalomorph spider *Avicularia juruensis*. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **155**, 419–426 (2010).
20. S. M. Correa-Garhwal, P. L. Babb, B. F. Voight, C. Y. Hayashi, Golden orb-weaving spider (*Trichonephila clavipes*) silk genes with sex-biased expression and atypical architectures. *G3 (Bethesda)* **11**, jkaa039 (2021).
21. S. M. Correa-Garhwal, R. C. Chaw, T. H. Clarke 3rd, L. G. Alaniz, F. S. Chan, R. E. Alfaro, C. Y. Hayashi, Silk genes and silk gene expression in the spider *Tengella perfuga* (Zoropsidae), including a potential cribellar spidroin (CrSp). *PLOS ONE* **13**, e0203563 (2018).
22. S. M. Correa-Garhwal, R. C. Chaw, T. H. Clarke III, N. A. Ayoub, C. Y. Hayashi, Silk gene expression of theridiid spiders: Implications for male-specific silk use. *Zoology* **122**, 107–114 (2017).
23. S. M. Correa-Garhwal, R. C. Chaw, T. Dugger, T. H. Clarke III, K. H. Chea, D. Kisailus, C. Y. Hayashi, Semi-aquatic spider silks: Transcripts, proteins, and silk fibres of the fishing spider, *Dolomedes triton* (Pisauridae). *Insect Mol. Biol.* **28**, 35–51 (2019).
24. S. M. Correa-Garhwal, T. H. Clarke 3rd, M. Janssen, L. Crevecoeur, B. N. McQuillan, A. H. Simpson, C. J. Vink, C. Y. Hayashi, Spidroins and silk fibers of aquatic spiders. *Sci. Rep.* **9**, 13656 (2019).
25. S. M. Correa-Garhwal, J. E. Garb, Diverse formulas for spider dragline fibers demonstrated by molecular and mechanical characterization of spitting spider silk. *Biomacromolecules* **15**, 4598–4605 (2014).
26. N. Kono, H. Nakamura, M. Mori, M. Tomita, K. Arakawa, Spidroin profiling of cribellate spiders provides insight into the evolution of spider prey capture strategies. *Sci. Rep.* **10**, 15721 (2020).
27. K. W. Sanggaard, J. S. Bechsgaard, X. Fang, J. Duan, T. F. Dyrlund, V. Gupta, X. Jiang, L. Cheng, D. Fan, Y. Feng, L. Han, Z. Huang, Z. Wu, L. Liao, V. Settepani, I. B. Thogersen, B. Vanthournout, T. Wang, Y. Zhu, P. Funch, J. J. Enghild, L. Schauser, S. U. Andersen, P. Villesen, M. H. Schierup, T. Bilde, J. Wang, Spider genomes provide insight into composition and evolution of venom and silk. *Nat. Commun.* **5**, 3765 (2014).

28. J. Starrett, J. E. Garb, A. Kuelbs, U. O. Azubuike, C. Y. Hayashi, Early events in the evolution of spider silk genes. *PLOS ONE* **7**, e38084 (2012).
29. M. Tian, C. Liu, R. Lewis, Analysis of major ampullate silk cDNAs from two non-orb-weaving spiders. *Biomacromolecules* **5**, 657–660 (2004).
30. M. A. Collin, T. H. Clarke Iii, N. A. Ayoub, C. Y. Hayashi, Genomic perspectives of spider silk genes through target capture sequencing: Conservation of stabilization mechanisms and homology-based structural models of spidroin terminal regions. *Int. J. Biol. Macromol.* **113**, 829–840 (2018).
31. J. E. Garb, T. DiMauro, V. Vo, C. Y. Hayashi, Silk genes support the single origin of orb webs. *Science* **312**, 1762–1762 (2006).
32. J. E. Garb, C. Y. Hayashi, Modular evolution of egg case silk genes across orb-weaving spider superfamilies. *Proc. Natl. Acad. Sci.* **102**, 11379–11384 (2005).
33. M. Sarr, K. Kitoka, K.-A. Walsh-White, M. Kaldmäe, R. Metlāns, K. Tārs, A. Mantese, D. Shah, M. Landreh, A. Rising, J. Johansson, K. Jaudzems, N. Kronqvist, The dimerization mechanism of the N-terminal domain of spider silk proteins is conserved despite extensive sequence divergence. *J. Biol. Chem.* **298**, 101913 (2022).
34. M. Strickland, V. Tudorica, M. Rezac, N. R. Thomas, S. L. Goodacre, Conservation of a pH-sensitive structure in the C-terminal region of spider silk extends across the entire silk gene family. *Heredity (Edinb)* **120**, 574–580 (2018).
35. J. E. Bond, N. L. Garrison, C. A. Hamilton, R. L. Godwin, M. Hedin, I. Agnarsson, Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. *Curr. Biol.* **24**, 1765–1771 (2014).
36. W. C. Wheeler, J. A. Coddington, L. M. Crowley, D. Dimitrov, P. A. Goloboff, C. E. Griswold, G. Hormiga, L. Prendini, M. J. Ramírez, P. Sierwald, L. Almeida-Silva, F. Alvarez-Padilla, M. A. Arnedo, L. R. Benavides Silva, S. P. Benjamin, J. E. Bond, C. J. Grismado, E. Hasan, M. Hedin, M. A. Izquierdo, F. M. Labarque, J. Ledford, L. Lopardo, W. P. Maddison, J. A. Miller, L. N. Piacentini, N. I. Platnick, D. Polotow, D. Silva-Dávila, N. Scharff, T. Szűts, D. Ubick, C. J. Vink, H. M. Wood, J. Zhang, The spider tree of life: Phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. *Cladistics* **33**, 574–616 (2017).
37. N. Kono, K. Arakawa, Nanopore sequencing: Review of potential applications in functional genomics. *Dev. Growth Differ.* **61**, 316–326 (2019).
38. C. Viera, L. F. Garcia, M. Lacava, J. Fang, X. Wang, M. M. Kasumovic, S. J. Blamires, Silk physico-chemical variability and mechanical robustness facilitates intercontinental invasibility of a spider. *Sci. Rep.* **9**, 13273 (2019).
39. N. Scharff, J. A. Coddington, T. A. Blackledge, I. Agnarsson, V. W. Framenau, T. Szűts, C. Y. Hayashi, D. Dimitrov, Phylogeny of the orb-weaving spider family Araneidae (Araneae: Araneoidea). *Cladistics* **36**, 1–21 (2020).
40. A. D. Malay, K. Arakawa, K. Numata, Analysis of repetitive amino acid motifs reveals the essential features of spider dragline silk proteins. *PLOS ONE* **12**, e0183397 (2017).
41. N. Kono, H. Nakamura, M. Mori, Y. Yoshida, R. Ohtoshi, A. D. Malay, D. A. Pedrazzoli Moran, M. Tomita, K. Numata, K. Arakawa, Multicomponent nature underlies the extraordinary mechanical properties of spider dragline silk. *Proc. Natl. Acad. Sci.* **118**, e2107065118 (2021).
42. N. Kono, H. Nakamura, R. Ohtoshi, D. A. P. Moran, A. Shinohara, Y. Yoshida, M. Fujiwara, M. Mori, M. Tomita, K. Arakawa, Orb-weaving spider *Araneus ventricosus* genome elucidates the spidroin gene catalogue. *Sci. Rep.* **9**, 8380 (2019).
43. G. Hormiga, C. E. Griswold, Systematics, phylogeny, and evolution of orb-weaving spiders. *Annu. Rev. Entomol.* **59**, 487–512 (2014).
44. B. Madsen, Z. Z. Shao, F. Vollrath, Variability in the mechanical properties of spider silks on three levels: Interspecific, intraspecific and intraindividual. *Int. J. Biol. Macromol.* **24**, 301–306 (1999).
45. K. Yazawa, A. D. Malay, H. Masunaga, Y. Norma-Rashid, K. Numata, Simultaneous effect of strain rate and humidity on the structure and mechanical behavior of spider silk. *Commun. Mater.* **1**, 10 (2020).
46. H. C. Craig, D. Piorkowski, S. Nakagawa, M. M. Kasumovic, S. J. Blamires, Meta-analysis reveals materiomic relationships in major ampullate silk across the spider phylogeny. *J. R. Soc. Interface* **17**, 20200471 (2020).
47. A. E. Brooks, S. R. Nelson, J. A. Jones, C. Koenig, M. Hinman, S. Stricker, R. V. Lewis, Distinct contributions of model MaSp1 and MaSp2 like peptides to the mechanical properties of synthetic major ampullate silk fibers as revealed in silico. *Nanotechnol. Sci. Appl.* **1**, 9 (2008).
48. C. L. Tucker, J. A. Jones, H. N. Bringhurst, C. G. Copeland, J. B. Addison, W. S. Weber, Q. Mou, J. L. Yarger, R. V. Lewis, Mechanical and physical properties of recombinant spider silk films using organic and aqueous solvents. *Biomacromolecules* **15**, 3158–3170 (2014).
49. N. Cohen, M. Levin, C. D. Eisenbach, On the origin of supercontraction in spider silk. *Biomacromolecules* **22**, 993–1000 (2021).
50. J. Johansson, A. Rising, Doing what spiders cannot—A road map to supreme artificial silk fibers. *ACS Nano* **15**, 1952–1959 (2021).
51. K. N. Savage, J. M. Gosline, The effect of proline on the network structure of major ampullate silks as inferred from their mechanical and optical properties. *J. Exp. Biol.* **211**, 1937–1947 (2008).
52. T. Asakura, K. Suita, T. Kameda, S. Afonin, A. S. Ulrich, Structural role of tyrosine in Bombyx mori silk fibroin, studied by solid-state NMR and molecular mechanics on a model peptide prepared as silk I and II. *Magn. Reson. Chem.* **42**, 258–266 (2004).
53. C. L. Craig, Evolution of arthropod silks. *Annu. Rev. Entomol.* **42**, 231–267 (1997).
54. K. Kakui, J. F. Fleming, M. Mori, Y. Fujiwara, K. Arakawa, Comprehensive transcriptome sequencing of tanaidacea with proteomic evidences for their silk. *Genome Biol. Evol.* **13**, evab281 (2021).
55. K. Arakawa, M. Mori, N. Kono, T. Suzuki, T. Gotoh, S. Shimano, Proteomic evidence for the silk fibroin genes of spider mites (order Trombidiformes: family Tetranychidae). *J. Proteomics* **239**, 104195 (2021).
56. N. Kono, H. Nakamura, A. Tateishi, K. Numata, K. Arakawa, The balance of crystalline and amorphous regions in the fibroin structure underpins the tensile strength of bagworm silk. *Zoological Lett* **7**, 11 (2021).
57. O. Folmer, M. Black, W. Hoeh, R. Lutz, R. Vrijenhoek, DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* **3**, 294–299 (1994).
58. M. C. Hedin, W. P. Maddison, A combined molecular approach to phylogeny of the jumping spider subfamily Dendryphantinae (Araneae: Salticidae). *Mol. Phylogenet. Evol.* **18**, 386–403 (2001).
59. N. Kono, H. Nakamura, Y. Ito, M. Tomita, K. Arakawa, Evaluation of the impact of RNA preservation methods of spiders for de novo transcriptome assembly. *Mol. Ecol. Resour.* **16**, 662–672 (2016).
60. Z. Chang, G. Li, J. Liu, Y. Zhang, C. Ashby, D. Liu, C. L. Cramer, X. Huang, Bridger: A new framework for *de novo* transcriptome assembly using RNA-seq data. *Genome Biol.* **16**, 30 (2015).
61. T. Hackl, R. Hedrich, J. Schultz, F. Förster, proovread: Large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004–3011 (2014).
62. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2–Approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490 (2010).
63. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
64. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
65. A. D. Malay, R. Sato, K. Yazawa, H. Watanabe, N. Ifuku, H. Masunaga, T. Hikima, J. Guan, B. B. Mandal, S. Damrongsakkul, K. Numata, Relationships between physical properties and sequence in silkworm silks. *Sci. Rep.* **6**, 27573 (2016).
66. K. Numata, R. Sato, K. Yazawa, T. Hikima, H. Masunaga, Crystal structure and physical properties of *Antheraea yamamai* silk fibers: Long poly (alanine) sequences are partially in the crystalline region. *Polymer* **77**, 87–94 (2015).
67. A. P. Hammersley, S. O. Svensson, M. Hanfland, A. N. Fitch, D. Hausermann, Two-dimensional detector software: From real detector to idealised image or two-theta scan. *Int. J. High Pressure Res.* **14**, 235–248 (1996).
68. M. Elices, J. Pérez-Rigueiro, G. Plaza, G. V. Guinea, Recovery in spider silk fibers. *J. Appl. Polym. Sci.* **92**, 3537–3541 (2004).
69. K. Arakawa, K. Mori, K. Ikeda, T. Matsuzaki, Y. Kobayashi, M. Tomita, G-language Genome Analysis Environment: A workbench for nucleotide sequence data mining. *Bioinformatics* **19**, 305–306 (2003).
70. K. Arakawa, H. Suzuki, M. Tomita, Computational genome analysis using the G-language system. *Genes, Genomes Genomics* **2**, 1–13 (2008).
71. K. Arakawa, M. Tomita, G-language System as a platform for large-scale analysis of high-throughput omics data. *J. Pestic. Sci.* **31**, 282–288 (2006).

Acknowledgments: We acknowledge the MICET (especially, T. Vololontiana), the Ministry of Environment and Sustainable Development (Ministère de l'Environnement de l'Ecologie et des Forêts at that time), the MZBA, and the University of Antananarivo for spider sampling in Madagascar. We thank Y. Takai, N. Ishii, and Y. Onozawa for technical support in sequencing; H. Ozaki and M. Sato for meaningful discussion; and H. Kano, R. Sato, and H. Nishijima for the development of reeling machines. **Funding:** This work was supported by grants from the ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan) to K.A., H.N., and K.N.; by research funds from the Yamagata Prefectural Government and Tsuruoka City, Japan, to K.A., N.K., and M.T.; and by JST ERATO grant number JPMJER1602, Grant-in-Aid for Transformative Research Areas (B), and Material DX to K.N.
Author contributions: Conceptualization: K.A. and K.N. Data curation: K.A., N.K., A.D.M., H.N., M.T., and K.N. Formal analysis: K.A., N.K., A.D.M., A.Tat., N.I., H.M., R.S., H.N., K.Y., N.A.O., and K.N. Resources: K.A., A.D.M., R.S., K.T., R.O., D.P., A.S., Y.I., H.N., A.Tan., Y.S., T.I., S.F., M.F., S.J.B., J.-A.C., H.C., C.P.F., G.G., J.G., C.H., D.L.K., K.S., B.B.M., Y.N.-R., R.C.P., N.M.P., R.R., X.W., K.Y., Z.Z.,

and K.N. Writing (original draft): K.A. and K.N. Writing (review and editing): All authors. **Competing interests:** R.S., R.O., D.P., A.S., Y.I., and H.N. are employees of Spiber Inc. The authors declare that they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper, the Spider Silkome Database (https://spider-silkome.org), and/or the Supplementary Materials. The spider biological materials in Malaysia, United States, India, China, United Kingdom, Australia, and Italy can be provided by K.S., D.L.K., B.B.M., J.G., C.H., S.J.B., and N.M.P., respectively, pending scientific review and a completed material transfer agreement. Requests for the spider biological materials should be submitted to K.N.

Submitted 14 February 2022
Accepted 19 August 2022
Published 12 October 2022
10.1126/sciadv.abo6043

ScienceAdvances

1000 spider silkomes: Linking sequences to silk physical properties

Kazuharu ArakawaNobuaki KonoAli D. MalayAyaka TateishiNao IfukuHiroyasu MasunagaRyota SatoKousuke TsuchiyaRintaro OhtoshiDaniel PedrazzoliAsaka ShinoharaYusuke ItoHiroyuki NakamuraAkio TanikawaYuya SuzukiTakeaki IchikawaShohei FujitaMasayuki FujiwaraMasaru TomitaSean J. BlamiresJo-Ann ChuahHamish CraigChoon P. FoongGabriele GrecoJuan GuanChris HollandDavid L. KaplanKumar SudeshBiman B. MandalY. Norma-RashidNur A. OktavianiRucsanda C. PredaNicola M. PugnoRangam RajkhowaXiaoqin WangKenjiro YazawaZhaozhu ZhengKeiji Numata

Sci. Adv., 8 (41), eabo6043. • DOI: 10.1126/sciadv.abo6043

Science Advances (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.