The genome sequence of Samia ricini, a new model species of lepidopteran insect

Jung Lee[1], Tomoaki Nishiyama[2], Shuji Shigenobu[3], Katsushi Yamaguchi[4], Yutaka Suzuki[5], Toru Shimada[1], Susumu Katsuma[5], and Takashi Kiuchi[5]

[1]Gakushuin University
[2]Kanazawa University Advanced Science Research Center
[3]National Institute for Basic Biology
[4]Affiliation not available
[5]The University of Tokyo

April 27, 2020

Abstract

Samia ricini, a gigantic saturniid moth, has the potential to be a novel lepidopteran model species. Since S. ricini is much more tough and resistant to diseases than the current model species Bombyx mori, the former can be easily reared compared to the latter. In addition, genetic resources available for S. ricini rival or even exceed those for B. mori: at least 26 eco-races of S. ricini are reported and S. ricini can hybridise with wild Samia species, which are distributed throughout Asian countries, and produce fertile progenies. Physiological traits such as food preference, integument colour, larval spot pattern, etc. are different between S. ricini and wild Samia species so that those traits can be the target for forward genetic analysis. In order to facilitate genetic research in S. ricini, we determined the whole genome sequence of S. ricini. The assembled genome of S. ricini was 458 Mb with 155 scaffolds, and the N50 length of the assembly was approximately 21 Mb. 16,702 protein coding genes were predicted in the assembly. Although the gene repertoire of S. ricini was not so different from that of B. mori, some genes, such as chorion genes and fibroin genes, seemed to have specifically evolved in S. ricini.

Keywords: De novo genome assembly, Eri silkmoth, *Samia ricini*, Saturniidae

Introduction

Bombyx mori have been the best ‘model organism’ in Lepidoptera and allowed researchers to make remarkable discoveries. For example, Toyama (1906) confirmed Mendel’s laws of heredity is valid for *B. mori* and this was the first case which proved the validity of Mendel’s laws for an animal species. When Beadle and Tatum proposed ‘one gene–one enzyme hypothesis’ (1941), Kikkawa (1941) almost simultaneously reached the similar concept by using egg colour mutants of *B. mori* . There is no doubt that availability of hundreds of mutant strains contributed to those discoveries. Since the whole genome sequence of *B. mori* was determined (International Silkworm Genome Consortium, 2008), the usability of *B. mori* as model species has significantly increased. Whole genome sequences of other lepidopteran species, such as *Papilio polytes* , *Danaus plexippus* ,*Lymantria dispar* and so on, are now available and some considerable studies were conducted by making the best of the genome information (Gu et al., 2019; Nishikawa et al., 2015; Zhang et al., 2019). However, species mentioned above have few mutant strains and forward genetic analysis on those species have not been conducted.

However, being a model organism does not necessarily mean being an ordinary species in its own taxon. Domestication had deprived *B. mori* of many traits which majority of lepidopteran species possess, such as the larval integument colour, foraging ability, adult flight ability and so on. It is obvious that *B. mori* is not suitable for researches attempting to elucidate the genetic basis of these traits, demonstrating that researches using *B. mori* does not always lead to better understanding of lepidopteran insects.

One possible answer to the question 'Which species is better than*B. mori* in genetic research?' is *Samia ricini* (Fig. 1).*S. ricini* , also known as 'Eri silkmoth,' one of the Saturniid moth species which was originated in Assam, India, and has been artificially transferred to many Asian countries and other regions. Although this species has been domesticated for the purpose of silk production, *S. ricini* still retains the traits that are lost in*B. mori* . *S. ricini* is a multivoltine species while majority of saturniids are univoltine or bivoltine (Brahma, Swargiary, & Dutta, 2015; Sternburg, & Waldbauer, 1984), which means that research of *S. ricini* is free from seasonal limitation (Singh, Kumar, Ahmed, & Pathania, 2017). Also, *S. ricini* grows uniformly and can be reared synchronously in large scale, resulting in efficient egg production. Thus, we have already succeeded in establishing a genome-editing system in this species using Transcription activator-like effector nucleases (TALENs) and successfully obtained several gene knockout lines (Lee, Kiuchi, Kawamoto, Shimada, & Katsuma, 2018), meaning that functional analysis of genes of interest is now achievable.

Utilising *S. ricini* for genetic research has another advantage. It will enable us to access to substantially high genetic diversity. First of all, *S. ricini* reportedly consists of at least twenty-six morphologically different eco races (Singh et al., 2017). In addition, *S. ricini* is able to produce fertile hybrids with wild*Samia* species (Brahma et al., 2015; Peigler, & Naumann, 2003), such as *Samia canningi* or *Samia cynthia pryeri* . Because the populations of *S. canningi* and *S. c. pryeri* are distributed throughout south and east Asian countries, different endemic nature such as larval integument colour, larval marking patterns, cocoon colour and host plant preference can be observed among populations (Brahma et al., 2015; Peigler, & Naumann, 2003). Genetic diversity in genus *Samia* rivals or even exceeds that in *B. mori* .

In order to facilitate genetic research of *S. ricini* , we decided to determine the whole genome sequence of *S. ricini* . We employed both long-read and short-read sequencers, namely Pacbio Sequel system and illumina HiSeq1500 to construct high-quality genome assembly. After the assembly was completed, to confirm whether genetic research aiming at identifying trait-related genes is feasible or not, we attempted to identify the responsible chromosomes for several phenotypes in *S. ricini* and *S. c. pryeri* .

Materials and Methods

Insects

UT strain of *S. ricini* and Nagano strain of *S. c. pryeri*larvae were provided from the National BioResource Project (NBRP;*http://shigen.nig.ac.jp/wildmoth/*). While *S. ricini* larvae were reared on *Ricinus communis* leaves under long-day conditions (16 h light/8 h dark) at 25°C, *S. c. pryeri* larvae were reared on*Ailanthus altissima.* F_1 interspecific hybrids were obtained by crossing *S. c. pryeri* female and *S. ricini* male. F_1 individuals and BC_1individuals were reared on *Ricinus communis* leaves under long-day conditions (16 h light/8 h dark) at 25°C.

DNA sample preparation for whole genome sequencing (WGS)

Posterior silk gland was sampled from fifth-instar larvae. Genomic DNA was prepared using Genomic-tip 100/G (QIAGEN, Hilden, Germany) according to the manufacturer's protocol.

Library preparation and genome sequencing

For WGS, PacBio Sequel System (Pacific Bioscience, California, USA) and Illumina HiSeq (Illumina, California, USA) were employed. For PacBio, a 20-kb library was prepared and four SMRT cells were used for sequencing; 3,267,255 subreads were finally obtained. Illumina paired-end and mate-pair libraries were prepared using Illumina PCR-Free library prep kit, Nextera Mate Pair library prep kit, and Kapa Hyper prep kit. Paired-end libraries were constructed from DNA fragmented with Covaris S2 and separated with an agarose gel at 200–250-bp (male ZZ) and with Sage-ELF at 310-530 bp (female ZO). Mate-pair libraries were separated with CHEF-electrophoresis after tagmentation and DNA recovered from gel slices 3-kbp to approximately 40-kbp were used for the subsequent process. All the libraries had different index and mixed to be sequenced in the same lane. In total, 511,569,649 reads were obtained by Illumina v4 sequencing of 2 lanes. Table S1 and Table S2 summarize information on the results of WGS, e.g. read length, read count and total bases.

RNA sequencing (RNA-seq)

Embryo-derived libraries for RNA-seq were prepared using TruSeq RNA Library Prep Kit (Illumina) and were sequenced using the Illumina HiSeq 2500 platform with 100-bp and 101-bp paired-end reads. The library for midgut-derived RNA samples was prepared using TruSeq RNA Library Prep Kit (Illumina) and sequenced using the Genome Analyser IIx System with 76-bp paired-end reads. The library for anterior silk gland- and middle silk gland-derived RNA samples were prepared using SureSelect Strand Specific RNA Library Prep Kit (Agilent) and were sequenced using the Illumina HiSeq 2500 platform with 100-bp paired-end reads. Table S3 summerizes information on the results of RNA-seq.

Quality check and trimming

The quality of Illumina short reads was examined using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ Based on the quality check results, trimming of reads were conducted using Trimmomatic v0.36 (Bolger, Lohse, & Usadel, 2014).

Heterozygosity assessment

Using Jellyfish (Marçais, & Kingsford, 2011) and GenomeScope (Vurture et al., 2017), heterozygosity in the one of the sequenced individuals was estimated. For comparison, heterozygosity of *A. yamamai* (Kim et al., 2018) was also estimated. Short read data used for Jellyfish were available under accession numbers DRR213145 (*S. ricini*) and SRR5641445 (*A. yamamai*).

Genome assembly and completeness assessment

Long reads derived from Sequel System were assembled using the HGAP4 assembler (https://www.pacb.com/support/software downloads/). To construct consensus sequences from draft contigs from HGAP4, Racon (Vaser, Sovic, Nagarajan, & Sikic, 2016) with minimap (Li, 2016) was employed. Racon treatment was repeated until the output FASTA file showed no difference from that of the previous run. In this case, four repeats were sufficient to obtain final results. Then, in order to polish the assembly, Pilon (Walker et al., 2014) was utilized with Illumina short reads. The final assembly was submitted to BUSCO v3 software (Waterhouse et al., 2018) to assess the completeness of the assembly. For comparison, the latest genome assembly of 4 lepidopteran species, including *B. mori* , *Papilio xuthus* , *Danaus plexippus* and *Plutella xylostella* , was also sumitted to BUSCO.

Linkage analysis of scaffolds

To clarify the linkages between scaffolds, we adapted the genetic approach. First, we obtained backcross generation 1 (BC_1) individuals between *S. ricini* and *S. c. pryeri* , a closely related species to *S. ricini.* Crossing scheme was (*Samia cynthia pryeri* x *S. ricini*) x *S. ricini.* Since meiotic recombination does not occur in ovaries of lepidopteran species, chromosomes in BC_1 individuals should be *S. ricini* -*S. c. pryeri* heterozygotic or *S. ricini* -*S. ricini* homozygotic.

We designed thirty-five genetic markers which can specifically detect thirty-five scaffolds longer than 1 Mbp and can molecularly distinguish*S. ricini* and *S. c. pryeri* (Fig. S1A) and then perform genomic PCR. Genomic DNA was extracted from the legs of eight BC_1 larvae using the DNeasy Blood and Tissue Kit (QIAGEN). The genomic PCR program was as follows: 40 cycles of 10 s at 98°C, 5 s at 60°C and 5 s at 68°C. KOD One^{TM} PCR Master Mix (TOYOBO) was used to perform genomic PCR. Then, the allele combinations of scaffolds in 8 BC_1 individuals were examined. According to the result, we designated the identified linkage groups according to Yoshido *et al.* (2011). Electrophoresis was conducted using 2.0% agarose gel or MultiNA microchip electrophoresis system (Shimadzu, Kyoto, Tokyo). Table S4 listed all primers used for linkage analysis.

Repeat identification and comparative analysis

To identify the repeat elements of the *S. ricini* genome, a custom repeat library was constructed using RepeatModeler v1.0.11 (http://www.repeatmasker.org/RepeatModeler/) with RECON v1.0.8 (Bao, & Eddy, 2002), RepeatScout v1.0.5 (Price, Jones, & Pevzner, 2005) and TRF v4.0.4 (Benson, 1999). To mask and annotate repetitive sequences in*S. ricini* , the constructed custom repeat library was utilised by RepeatMasker v4.0.7 (*http://www.repeatmasker.org/RMDownload.html*; Tarailo-Graovac, & Chen, 2009) with Repbase (Jurka et al., 2005). RepeatMasker was conducted with RMBlast (http://www.repeatmasker.org/RMBlast.html).

Ab initio gene prediction

BRAKER2 pipeline (Camacho et al., 2009; Hoff, Lange, Lomsadze, Borodovsky, & Stanke, 2016; Hoff, Lomsadze, Borodovsky, & Stanke, 2019; Lomsadze, Burns, & Borodovsky, 2014; Stanke, Schöffmann, Morgenstern, & Waack, 2006; Stanke, Diekhans, Baertsch, & Haussler, 2008) was employed for gene prediction. First of all, repetitive sequences in the genome identified by RepeatMasker were soft-masked. To generate extrinsic evidence for gene prediction, eleven sets of RNA-seq reads (Table S3) was mapped to the genome sequence by using HiSAT2 (Kim, Langmead, & Salzberg, 2015). The resultant BAM files generated by HiSAT2 were submitted to BRAKER2 by using '–bam' option. Parallelly, we assembled the RNA-seq reads using Trinity assembler (Haas et al., 2013). Then, the tr2aacds.pl program bundled in EvidentialGene suite (http://arthropods.eugenes.org/EvidentialGene/evigene/) was used to merge the assemblies from multiple transcriptome data sets. The merged transcriptome assemblies were aligned to the genome sequence using PASA (Haas et al., 2008; Haas et al., 2013) for identifying the exon regions. In addition to tr2aacds.pl program, StringTie (Pertea et al., 2015) was also used to merge multiple transcriptome data for exon region prediction. Furthermore, amino acid sequences of manually annotated sequences of *S. ricini* deposited in the Universal Protein Resource database (UniProt,*http://www.uniprot.org*) (Bateman, 2019) were aligned to genome sequence using exonerate v2.2.0 (Slater, & Birney, 2005) to obtain protein spliced alignment information. Finally, multiple predictions generated by BRAKER2, PASA, StringTie and exonerate were integrated by EvidenceModeler (Haas et al., 2008).

Functional annotation

Amino acid sequences of the predicted genes were aligned to Uniprot database with BLASTP program (Camacho et al., 2009). Protein classification and domain search were achieved by InterProScan program (Finn et al., 2017) with Pfam database (El-Gebali, 2019). These analyses were done in OmicsBox software through trial mode (Conesa, & Götz, 2008).

Comparative genome analysis

To identify ortholog groups among multiple species, including *B. mori* , *D. plexippus* , *P. xuthus* and *P. xylostella* , OrthoFinder (Emms, & Kelly, 2015) was used. Each gene set corresponded to the genome assembly, which was used for BUSCO analysis (Table S5). Regarding to *D. plexippus* , *P. xuthus* and *P. xylostella* , the proteome data was obtained from Lepbase (*http://lepbase.org*) (Challis, Kumar, Dasmahapatra, Jiggins, & Blaxter, 2016). The proteome data of *B. mori* was obtained from SilkBase (http://silkbase.ab.a.u-tokyo.ac.jp/cgi-bin/download.cgi).

Constructing species tree

Extracted amino acid sequences of Single Copy Ortholog, one of the outputs of OrthoFinder, were aligned by MAFFT (Katoh, Misawa, Kuma, & Miyata, 2002). To remove regions where sequence alignments were incomplete, trimAl (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009) was used. Finally, trimmed alignments were submitted to IQ-TREE ver. 1.6.11 (Nguyen, Schmidt, Haeseler, & Minh, 2015; Hoang, Chernomor, Haeseler, Minh, & Vinh, 2018) for maximum likelihood tree reconstruction, with '-sp' option (Chernomor, Haeseler, & Minh, 2016) to allow different substitution model and branch lengths for each alignment. The model of each partition was determined by ModelFinder (S. Kalyaanamoorthy, Minh, Wong, Haeseler, & Jermiin, 2017). Branch supports were evaluated based on 1,000 bootstrap replicates.

Drawing circular ideogram for *B. mori* and *S. ricini* genomes

In order to assess the similarity of *B. mori* and *S. ricini*genomes, a circular ideogram was drawn using Clico (Cheong, Tan, Yap, & Ng, 2015) with the Circos program (Krzywinski et al., 2009). Single-copy orthologs, identified by OrthoFinder in each genome, were connected. To simplify the ideogram, short scaffolds in the *B. mori* genome assembly which were not assigned to 28 chromosomes, were filtered out.

Identifying the chorion gene cluster and phylogenetic analysis of *chorion* genes

OrthoFinder found that *chorion* genes were tandemly arrayed on chromosome 1 (Chr. 1). For more detailed information, we performed BLASTP search against non-redundant protein database, in addition to the Uniprot database (Bateman, 2019) with an e-value less than 1e-5, using the predicted gene models within and around the chorion gene region. As a result, 80 *chorion* genes were found on Chr. 1 and these genes made a gene cluster. Within this cluster, five non-chorion gene models (evm.model.Sr_HGAP_JL_scaf_2.1123,1128,1135,1136 and 1137) were also identified (Table S6). Phylogenetic analysis of *chorion* genes was conducted with 80 *S. ricini chorion* genes, 121 *B. mori chorion* genes, 21 *P . xylostella* chorion genes, 29*P . xuthus* chorion genes, 24 *D . plexippus*chorion genes registered at the Uniprot and NCBI database and one non-chorion gene (evm.model.Sr_HGAP_JL_scaf_2.1135) as outgroup. Muscle was used to generate alignments of protein sequences (Edgar, 2004). Aligned sequences were subjected to phylogenetic analysis by maximum likelihood and ultrafast bootstrap methods (Minh, Nguyen, & Haeseler, 2013) with 1000 replicates using IQ-TREE ver. 1.5.5 (Nguyen, Schmidt, Haeseler, & Minh, 2015). The phylogenetic tree was constructed based on PMB+F+R5 model.

In order to check whether *S . ricini* has high-cysteine chorion gene or not, amino acid sequences of 38 high-cysteine chorion protein of *B. mori* was aligned to deduced amino acid sequences of 80 *S. ricini* chorion genes via BLASTP program.

Identifying *fibroin* and *sericin* genes in *S. ricini* genome

Fib-H (BAQ55621.1) and p25 (LC001863.1, LC001864.1 and LC001865.1) of*S. ricini* were already registered in Genbank, thus using those sequences as query, BLASTP search against 16,702 gene models of *S. ricini* was conducted with an e-value less than 1e-5 and '-seg no' option. In cases of BLASTP result being 'No hits found,' TBLASTN search against nucleotide sequences of *S. ricini* genome was conducted with the same parameter. In this report, '-evalue 1e-5' and '-seg no' options were always added when BLASTP and TBLASTN search were conducted with silk proteins (Fib-H, Fib-L, p25, sericin) as query. In order to investigate the homolog of *Fib-L* is present or not in *S. ricini* genome, *B. mori* Fib-L (NP_001037488.1) was utilised as query for BLASTP and TBLASTN search. In addition, we performed TBLASTN search against *A. yamamai* genome using *B. mori* Fib-L sequence as query.

Tsubota *et al* . (2015) and Dong *et al* . (2015) reported that 5 and 4 *sericin* genes are expressed in anterior silk gland and middle silk gland, respectively (Table S7). The deduced amino acid sequences of putative *sericin* transcripts were submitted to the gene model set of *S. ricini* through BLASTP. Regarding LC001867 and LC001870, because the corresponding gene models were not found, TBLASTN was conducted to confirm whether both transcripts were present or not.

When we tried to comprehend the repertoire of silk protein encoding genes in *D. plexippus* and *P. xylostella* , TBLASTN search against the genome assemblies was conducted with *B. mori* Fib-H (NP_001106733.1), Fib-L, p25 (NP_001139413.1) and sericin-1, 2, 3 (AB112019.1, NP_001166287.1, NP_001108116.1) sequences as queries because any transcripts or amino acid sequences were not previously reported as Fib-H, Fib-L, p25 and sericin in *P. xylostella* and*D. plexippus* . Genome assemblies which were used for TBLASTN search was the ones used in BUSCO analysis (Table S5). As the transcripts of *Fib-H* , *Fib-L* and *p25* of *P. xuthus* were already registered (see Table 3), those sequences were mapped to the *P. xuthus* genome sequence to confirm the presence. Regarding *sericin* genes in *P. xuthus* , no sequences were previously registered in Genbank, thus the same procedure as the case of*P. xylostella* and *D. plexippus* , was taken. Phylogenetic analysis of sericin was conducted with seven *S. ricini* putative*sericin* genes, three *B. mori sericin* genes and five*A. yamamai sericin* genes (LC08587, LC08588, LC08589, LC08590 and LC08591; Zurovec et al., 2016). Muscle was used to generate alignments of protein sequences (Edgar, 2004). Aligned sequences were subjected to phylogenetic analysis by maximum likelihood and bootstrap methods with 1,000 replicates using MEGAX (Kumar, Stecher, Li, Knyaz, & Tamura, 2018).

Identifying responsible chromosomes of 'Blue', 'Yellow', 'Spot' and 'Red cocoon' phenotypes in BC$_1$ individuals

BC$_1$ individuals, which has any one of four morphological traits, 'Blue,' 'Yellow,' 'Spot' and 'Red cocoon,' were picked up in order to identify responsible chromosomes for each trait. Genetic markers designed for scaffold linkage analysis were utilized (see Table S4). DNeasy Blood and Tissue kit (QIAGEN) and MightyAmp DNA Polymerase Ver.3 (TaKaRa) was used for DNA extraction and genomic PCR of BC$_1$ individuals, respectively. The genomic PCR program was as follows: 2 min at 98°C and 40 cycles of 10 s at 98°C, 15 s at 60°C and 1 min at 68°C.

Results and Discussion

Overview of *S. ricini* genome assembly

Final assembly of *S. ricini* genome was 450,479,495 bp long with 155 scaffolds. The N50 length of the assembly was approximately 21 Mb (Table 1). GC content was 34.3%. The longest scaffold length was approximately 33 Mbp. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis using BUSCO v3.0 with insecta odb9, including 1,658 BUSCOs from 42 species revealed that 97.8% of BUSCOs were completely detected in the assembled genome (1614, complete and single-copy; 8, complete and duplicated) among 1,658 tested BUSCOs (see Table S5). To the best of our knowledge, these statistic scores are the best among ever–constructed lepidopteran genome assemblies (Challis, Kumar, Dasmahapatra, Jiggins, & Blaxter, 2016; Kim et al., 2018; Triant, Cinel, & Kawahara, 2018). Low heterozygosity in *S. ricini* strain used for this project might be the key to the successful assembly: k-mer distribution analysis (k = 31) estimated that heterozygosity in one male individual of *S. ricini* was 0.0466 to 0.0472% (Table 1, see Fig. S2), considerably lower than that of the *Antheraea yamamai*genome, (0.807 to 0.808%) a related species belonging to the family Saturniidae (Fig. S2; Kim et al., 2018). This might result from the difference in voltinism between *S. ricini* and *A. yamamai* : It is quite difficult to establish an inbred line of *A. yamamai* because *A. yamamai* is a univoltine species and emerge only once per year, whereas multivoltine *S. ricini* can generate at least six generations per year. The difference of diapause strategy mentioned above indicates that at least six times longer periods are required to establish an inbred line of *A. yamamai* .

Linkage analysis of thirty-five scaffolds (> 1 Mbp) revealed that the scaffolds are grouped into fourteen linkages (Table 2, Fig. S1B), which corresponds to the previous report (Yoshido, Yasukochi, & Sahara, 2011) where BAC-FISH was conducted and concluded that*S. ricini* has thirteen autosomes and one Z chromosome (male: 2n = 28, female: 2n = 27). These thirty-five scaffolds counted up 443,618,927 bp, meaning that approximately 98.5% of the genome was assigned to the chromosomes (Table 2).

Repetitive sequences found in the assembled genome

RepeatMasker program (Tarailo-Graovac, & Chen, 2009) estimated that repeat elements occupy 43.5% (196,045,652 bp) of the assembled genome (Table 1). Except for 'unclassified' repeats, LINE is the largest superfamily of repetitive sequences in *S. ricini* (Figs. 2A, B). Interestingly, although the total length of LINE and its proportion to all repetitive sequences in the genome were similar between *S. ricini* and *B. mori* (Figs. 2A, B), the components of families of LINE were different. Table S8 shows the copy number of each LINE family in *S. ricini* and *B. mori* genomes. For example, while the CR1-Zenon family was the largest LINE family in *S. ricini* , the largest family in *B. mori* was Jockey. Given these results, although both *S. ricini* and *B. mori* have larger amounts of repetitive sequences in the genome than other lepidopteran species do (Fig. 2A), the expansion of repetitive sequences seems to have occurred in parallel and independently on their own phylogenetic branches.

Another noteworthy feature was that the *S. ricini* genome contains considerably small amounts of SINE (Fig. 2A). While the *B. mori*genome showed a large proportion of SINE (19.4% of all repetitive sequences), SINEs in *S. ricini* genome occupied only 0.0588%. This finding also supported the hypothesis of parallel and independent expansion of repetitive sequences.

Ab initio gene prediction and comparative genome analysis

BRAKER2 (Camacho et al., 2009; Hoff et al., 2016; Hoff et al., 2019; Lomsadze et al., 2014; Stanke et al., 2006; Stanke et al., 2008) software predicted 16,702 protein-coding genes in the genome assembly of*S. ricini* (Table 1). InterProScan (Finn et al., 2017) analysis shows that Reverse transcriptase domain (IPR 000477) and Integrase catalytic core domain (IPR001584) are the top 2 populous domains among*S. ricini* genes (Fig. S3), which may reflect that the large proportion of *S. ricini* genome is occupied by retrotransposable elements (SINE, LINE, LTR in Fig. 2).

The circos plot which links single copy orthologs among *B. mori*and *S. ricini* shows large scale rearrangement of chromosomes, such as translocation and chromosome fusion, happened in the ancestor of*S. ricini* (Fig. 3A) (Cheong et al., 2015; Krzywinski et al., 2009). However, despite frequent chromosomal rearrangements, genomic regions of no links or extremely few links were hardly observed in the plot, suggesting that almost entire regions of *S. ricini* and*B. mori* genomes are reciprocally corresponding and there barely exist 'species-specific' regions.

The number of orthogroups (OGs) among 5 Lepidoptera species is shown in Fig. 3B, and a phylogenetic tree of 3,907 single copy orthologs explains the genetic relationships among the 5 species (Fig. 3C). Ortholog analysis using OrthoFinder (Emms, & Kelly, 2015) identified 205*S. ricini* -specific OGs (Fig. 3B). Of 205 *S. ricini* -specific OGs, forty-six OGs are related to retrotransposable elements (Fig. 2). Thus, *S. ricini* specific non-retrotransposon related OGs were 159. Of these OGs, two OGs (OG0000113 and OG0000131) are consist of 33 and 30 chorion protein genes, respectively. These*S. ricini* specific chorion genes are located in close proximity on chromosome 1 as a gene cluster, which can be the ground of the high apparent duplication rate through tandem duplication or gene conversion. In addition to the above-mentioned 63 *S. ricini* -specific*chorion* genes, 17 *chorion* genes were found in this cluster. Table S6 summarized all 80 *chorion* genes present in*S. ricini* genome. A phylogenetic analysis of these genes along with chorion genes of *B. mori, P. xylostella, P. xuthus and D. plexippus* suggests that gene duplication could have resulted in diversification of chorion proteins because chorion genes from OG0000113 and OG0000131 fell into distinct clades (Fig. 3D).

Chorion proteins comprise eggshell and protect embryos from the environment, suggesting that chorion proteins are likely to evolve to reflect adaptations to the environment (Lecanidou, Rodakis, Eickbush, & Kafatos, 1986; Papantonis, Swevers, & Iatrou, 2015; Rodakis, & Kafatos, 1982). Based on sequence homology, chorion proteins can be categorized into two groups (α and β), which include three subfamilies, respectively (Lecanidou et al, 1986; Papantonis et al., 2015). Among the three subfamilies, high-cysteine (Hc) chorion is considered to play an important role for embryonic diapause, because Hc chorion proteins increase hardness of eggshells for embryos to survive diapause in the winter (Rodakis, & Kafatos, 1982). Interestingly, according to the BLAST search and phylogenetic analysis, Hc chorion protein genes seemed to be absent in the *S. ricini* genome (Fig. 3D, Table S6 and Table S9). Given that *S. ricini* is a non-diapause species, it is highly plausible that *S. ricini* lacks Hc chorion genes.

Fibroin and sericin

Fibroin is the major component of silk protein. Although fibroin of*B. mori* consists three polypeptides, namely heavy-chain (Fib-H), light-chain (Fib-L) and fibrohexamerin (p25) (Inoue et al., 2000), it was biochemically confirmed that fibroin of *S. ricini* lacks Fib-L and p25 and it consists of Fib-H/Fib-H homodimer (Tamura, & Kubota, 1988).The complete amino acid sequence of Fib-H (SrFib-H) was already determined by Sezutsu and Yukuhiro (2014), but our gene prediction was unable to properly construct the gene model for *SrFib-H* , mainly because of its repetitive sequences. However, TBLASTN search using SrFib-H as query detected the near-complete coding sequence of*SrFib-H* (Fig. S4A), supporting the accuracy of the assembly. The genome information also revealed that *S. ricini* genome has three copies of *p25* , in addition to *Fib-H* , but lacks*Fib-L* (Table 3). In addition, we confirmed that *Fib-L* is absent in the genome of *A. yamamai* (Kim et al., 2018), another saturniid moth, through TBLASTN search (Fig. S4B). Because other lepidopteran species, including *B. mori* , *P. xylostella* ,*P. xuthus* and *Corcyra cephalonica* (Chaitanya, & Dutta-Gupta, 2010), possess *Fib-L* gene, absence of *Fib-L*in saturniid moths can be ascribed to the loss of *Fib-L* in the common ancestor of saturniid moths.

As described above, silk fibroin of *B. mori* consists H-chain, L-chain and P25. 3 fibroin polypeptides assemble with a 6:6:1 molecular ratio, which is considered to be indispensable for proper secretion of fibroin: mutations in *Fib-H* or *Fib-L* cause fibroin secretion deficiency (Inoue et al., 2000; Ma et al., 2014). *B. mori* strains with deletions in *Fib-H* or *Fib-L* cannot properly secrete fibroin protein to lumen in silk gland, and their cocoons are mainly composed of sericin. Therefore, it has been hypothesized that *B. mori* has a mechanism which recognizes three-dimensional structure of fibroin assembled by the three polypeptides with 6:6:1 molecular ratio and selectively transport the fibroin polypeptide complex to lumen in silk gland. Since saturniid species lack *Fib-L* gene, fibroin transportation and secretion system in saturniid species must be different from that in *B. mori* .

So far, biological function of p25 is still unclear. Whether knockout of*p25* affects the secretion of fibroin or not remains to be answered. Since p25 protein is undetectable in *S. ricini* silk,*p25* could take on different function other than being the part of complex structure of fibroin. The presence of multi-copies of *p25*in *S. ricini* genome are posing the possibility of functional differentiation among paralogous *p25* s (Table 3).

Sericin occupies the second largest proportion of silk protein, following fibroin. Unlike fibroin, sericin is soluble to water and consisting the most outer layer of silk. *B. mori* has three*sericin* genes, *Ser1* , *Ser2* , and *Ser3* (Tsubota et al., 2015). While Ser1 and Ser3 were the components of cocoon protein, Ser2 is not present in cocoon (Takasu, Hata, Uchino, & Zhang, 2010). Two proteins derived from alternative splicing of *Ser2* can be found in larval silk produced during the growing stages (Takasu et al., 2010). So far, nine transcripts are registered to NCBI genbank as*Sericin* -encoding genes or *Sericin* -like genes in *S. ricini* (Table S8) (Dong et al., 2015; Tsubota et al., 2015). BLAST analysis successfully confirmed that all of them are present in *S. ricini* genome and transcribed from seven locus, meaning that *S. ricini* has 7 putative *Sericin* genes. Phylogenetic analysis showed that four out of seven genes are categorized into *Ser1/3*class and the other three genes were included in *Ser2* class (Fig. S5). Despite belonging to the same family (Saturniidae), sericin gene repertoires of *A. yamamai* and *S. ricini* were quite different: Ser1/3 class genes seemed to multiplicated in *A. yamamai* . Phylogenetic analysis revealed that all sericin genes in*A. yamamai* belong to Ser1/3 class and Ser2 class genes were not identified while *S. ricini* possess three Ser2 class genes (Fig. S5). The diversity of sericin genes among these saturniids may reflect the differences of their indigenous environments. However, whether proteins encoded by seven putative *Ser* genes in *S. ricini*are present in cocoons remains to be elucidated. Proteomic analysis on*S. ricini* cocoons should be carried out to reveal the protein composition.

Identification of the responsible chromosomes of the larval and cocoon phenotypes in *S. ricini*

In order to examine the feasibility of research aiming at identifying trait-related genes in *S. ricini* , we tried to carry out forward genetic analysis using *S. ricini* and *S. c. pryeri* . Some morphological traits are different between *S. ricini* and *S. c. pryeri* (Figs. 1 and 4A), thus such traits can be good targets for forward genetic analysis.

As shown in Figs. 4B and 4C, the phenotypes originally derived from*S. c. pryeri* were isolated in backcross generation 1 (BC$_1$) individuals which were obtained by the crossing (*S. ricini* x *S. c. pryeri*) x *S. ricini*

. Here, we tried to identify the responsible chromosomes for 4 phenotypes, namely, 'Blue,' 'Yellow,' 'Spot' and 'Red cocoon.' 'Blue' and 'Yellow' refer to blue and yellow larval integument, respectively. 'Spot' refers to black spots on larval integument. 'Red cocoon' phenotype literally illustrates the colour of cocoons which some BC_1 individuals produce.

Since meiotic recombination does not occur in lepidopteran females, all chromosomes of the BC_1 individuals should be *S. ricini* -*S. c. pryeri* heterozygotic or *S. ricini* -*S. ricini* homozygotic, and not chimeric. Considering that the above-mentioned four phenotypes derived from *S. c. pryeri* are dominant, responsible chromosomes should be heterozygotic in all BC_1 individuals.

Genomic PCR with chromosome-specific markers, which can molecularly distinguish *S. ricini* and *S. c. pryeri* , revealed that chromosome 8, 13, 3 and 12 were uniformly heterozygotic in all examined 'Blue,' 'Yellow,' 'Spot' and 'Red cocoon' individuals, meaning that those chromosomes were responsible for 'Blue,' 'Yellow,' 'Spot' and 'Red cocoon' traits, respectively (Fig. 5). This is the first report which demonstrated that forward genetic analysis is achievable in *S. ricini* (and *S. c. pryeri*). Although the responsible genes of the four phenotypes have not been identified yet, it will not be long before those were identified.

Larvae of many lepidopteran species are sometimes called 'green caterpillar' because of their greenish integument. Greenish colour found in lepidopteran larvae is mixture of yellow and blue pigments, namely carotenoids and bilins. Regarding 'Blue,' the substance which forms blue colour was elucidated to be biliverdin IXγ (Saitoh, 2011). In the larval integument of *S. c. pryeri* , Biliverdin IXγ binds to Biliverdin-Binding protein II (BBP-II). However, BBP-II protein amount in the larval integument of *S. c. pryeri* and *S. ricini*does not significantly differ (Saitoh, 2011), indicating that BBP-II encoding gene is not a responsible gene for 'Blue' phenotype.

In contrast to 'Blue', the exact substances which form 'Yellow' colour has yet to be identified. There are a few mutant strains of *B. mori* which show yellow integument colour phenotype, and 'lemon' and 'lemon lethal' represent one example (Meng et al., 2009). The yellow colour of 'lemon' and 'lemon lethal' has already been elucidated to be xanthopterin, not carotenoid. Furthermore, 'lemon' and 'lemon lethal' are recessive phenotypes, indicating these two mutant phenotypes of*B. mori* has nothing to do with 'Yellow' of *S. ricini*because 'Yellow' is a dominant phenotype. Identifying the responsible genes for 'Blue' and 'Yellow' phenotype will allow us to reveal the genetic basis of formation of larval 'green' colour.

Spot pattern of larvae of *S. c. pryeri* and 'Spot' individuals is uniformly formed in every segment of larval body. As shown in Fig. S6, lateral sides of a single larval body segment have three spots on each side. In addition, the dorsal side of a segment have five spots. In total, eleven spots can be found in a single segment. Although there are a number of silkworm strains which show various patterns of black spots on larval integument, none of those strains have uniform spot patterns similar to 'Spot' individuals.

'Red cocoon' phenotype exhibits brightly shining red colour. Interestingly, 'Red cocoon' phenotype has never been observed in F_1 individuals. Colour of cocoons of *S. c. pryeri*and F_1 individuals is grey, and thus, this 'Grey cocoon' is the epistatic trait to 'Red cocoon' phenotype (Fig. 4C). 'Red cocoon' is one of the economic traits of *S. ricini* . Some eco-races of*S. ricini* in India are known to spin red cocoons, similar to our 'Red cocoon' individuals. However, it is difficult to produce red silk from red cocoons because red pigments are mainly present in sericin layer and are easily swept away during the degumming step (Fig. S7). Therefore, Eri-silk made from red cocoons is not as red as they appeared to be before the degumming step. We are now attempting to identify the red pigment and reveal the genetic basis of red colour formation in cocoons, which will provide us with much information about improvement on the degumming condition.

Conclusion

We determined the whole genome sequence of *S. ricini* . The gene repertoire of *S. ricini* was not so distinct from other lepidopteran species, but some ortholog groups, such as *chorion*genes, and *sericin* genes, seemed to evolve specifically in*S. ricini* . The quality of the genome assembly met the criteria which forward genetic analysis is applicable, thus we successfully identified the responsible chromosomes for the certain traits. Now, we are anticipating that this report has paved the way for 'forward genetics of wild silkmoth.'

Acknowledgements

We thank Masahiro Kasahara for valuable advice on the assemble strategy. This study was supported by MEXT KAKENHI grant number JP22128004 and 17H06431 and JSPS KAKENHI grant numbers JSPS KAKENHI grant numbers 15H02482, 17H05047 and 18H03949.

References

Bao, Z., & Eddy, S. R. (2002). Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Research, 12(8), 1269–1276. https://doi.org/10.1101/gr.88502

Bateman, A. (2019). UniProt: A worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506–D515.*https://doi.org/10.1093/nar/gky1049*

Bateman, A. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research, 47(D1), D506–D515.*https://doi.org/10.1093/nar/gky1049*

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research, 27(2), 573–580.*https://doi.org/10.1093/nar/27.2.573*

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120.*https://doi.org/10.1093/bioinformatics/btu170*

Brahma, D., Swargiary, A., & Dutta, K. (2015). A comparative study on morphology and rearing performance of Samia ricini and Samia canningi crossbreed with reference to different food plants. Journal of Entomology and Zoology Studies, 3(5), 12–19.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. BMC Bioinformatics, 10, 1–9.*https://doi.org/10.1186/1471-2105-10-421*

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972–1973.*https://doi.org/10.1093/bioinformatics/btp3*

Chaitanya, R. K., & Dutta-Gupta, A. (2010). Light chain fibroin and P25 genes of Corcyra cephalonica: Molecular cloning, characterization, tissue-specific expression, synchronous developmental and 20-hydroxyecdysone regulation during the last instar larval development. General and Comparative Endocrinology, 167(1), 113–121.*https://doi.org/10.1016/j.ygcen.2010.02.007*

Challis, R. J., Kumar, S., Dasmahapatra, K. K., Jiggins, C. D., & Blaxter, M. (2016). Lepbase: The Lepidopteran genome database. BioRxiv, 056994. https://doi.org/10.1101/056994

Chen, Z., Nohata, J., Guo, H., Li, S., Liu, J., Guo, Y., . . . Mita, K. (2015). A comprehensive analysis of the chorion locus in silkmoth. Scientific Reports, 5, 1–11.*https://doi.org/10.1038/srep16424*

Cheong, W. H., Tan, Y. C., Yap, S. J., & Ng, K. P. (2015). ClicO FS: An interactive web-based service of Circos. Bioinformatics, 31(22), 3685–3687.*https://doi.org/10.1093/bioinformatics/btv433*

Chernomor, O., Von Haeseler, A., & Minh, B. Q. (2016). Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Systematic Biology, 65(6), 997–1008. https://doi.org/10.1093/sysbio/syw037

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. Bioinformatics, 30(1), 31–37.*https://doi.org/10.1093/bioinformatics/btt310*

Conesa, A., & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics, 2008.*https://doi.org/10.1155/2008/619832*

Dong, Y., Dai, F., Ren, Y., Liu, H., Chen, L., Yang, P., Liu, Y., Li, X., Wang, W., & Xiang, H. (2015). Comparative transcriptome analyses on silk glands of six silkmoths imply the genetic basis of silk structure and coloration. BMC Genomics, 16(1), 1–14.*https://doi.org/10.1186/s12864-015-1420-9*

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acid Research, 32(5), 1792–1797.*https://doi.org/10.1093/nar/gkh340*

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1), D427–D432.*https://doi.org/10.1093/nar/gky995*

Emms, D. M., & Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology, 16(1), 1–14.*https://doi.org/10.1186/s13059-015-0721-2*

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, Hsin Yu

Dosztanyi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, Silvio C.E., Wu, Cathy H., Xenarios, I., Yeh, L. S., Young, S. Y., & Mitchell, A. L. (2017). InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Research, 45(D1), D190–D199.*https://doi.org/10.1093/nar/gkw1107*

Gu, L., Reilly, P. F., Lewis, J. J., Reed, R. D., Andolfatto, P., & Walters, J. R. (2019). Dichotomy of Dosage Compensation along the Neo Z Chromosome of the Monarch Butterfly. Current Biology, 29(23), 4071-4077.e3. https://doi.org/10.1016/j.cub.2019.09.056

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Robin, C. B., Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology, 9(1), 1–22.*https://doi.org/10.1186/gb-2008-9-1-r7*

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., & Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8, 1494. Retrieved from*https://doi.org/10.1038/nprot.2013.084*

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. Molecular Biology and Evolution, 35(2), 518–522. https://doi.org/10.5281/zenodo.854445

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. Bioinformatics, 32(5), 767–769.*https://doi.org/10.1093/bioinformatics/btv661*

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation with BRAKER (pp. 65–95).*https://doi.org/10.1007/978-1-4939-9173-0_5*

Inoue, S., Tanaka, K., Arisaka, F., Kimura, S., Ohtomo, K., & Mizuno, S. (2000). Silk fibroin of Bombyx mori is secreted, assembling a high molecular mass elementary unit consisting of H-chain, L-chain, and P25, with a 6:6:1 molar ratio. Journal of Biological Chemistry, 275(51), 40517–40528.*https://doi.org/10.1074/jbc.M006897200*

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research, 110(1–4), 462–467.*https://doi.org/10.1159/000084979*

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods, 14(6), 587–589. https://doi.org/10.1038/nmeth.428

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14), 3059–3066.*https://doi.org/10.1093/nar/gkf436*

Kikkawa, H. (1941). Mechanism of Pigment Formation in Bombyx and Drosophila. Genetics, 26(6), 587–607. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17247024

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature Methods, 12, 357. Retrieved from*https://doi.org/10.1038/nmeth.3317*

Kim, S.R., Kwak, W., Kim, H., Caetano-Anolles, K., Kim, K. Y., Kim, S. B., Choi, K. H., Kim, S. W., Hwang, J. S., Kim, M., Kim, I., Goo, T. W., & Park, S. W. (2018). Genome sequence of the Japanese oak silk moth, Antheraea yamamai: the first draft genome in the family Saturniidae. GigaScience, 7(1), 1–11.*https://doi.org/10.1093/gigascience/gix113*

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., & Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. Genome Research, 19(9), 1639–1645.*https://doi.org/10.1101/gr.092759.109*

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution, 35(6), 1547–1549.*https://doi.org/10.1093/molbev/msy*

Lecanidou, R., Rodakis, G. C., Eickbush, T. H., & Kafatos, F. C. (1986). Evolution of the silk moth chorion gene superfamily: gene families CA and CB. Proceedings of the National Academy of Sciences of the United States of America, 83(17), 6514–6518.*https://doi.org/10.1073/pnas.83.17.6514*

Lee, J., Kiuchi, T., Kawamoto, M., Shimada, T., & Katsuma, S. (2018). Accumulation of uric acid in the epidermis forms the white integument of Samia ricini larvae. PLOS ONE, 13(10), e0205758.*https://doi.org/10.1371/journal.pone.*

Li, H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics, 32(14), 2103–2110.*https://doi.org/10.1093/bioinformatics/btw152*

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Research, 42(15), 1–8.*https://doi.org/10.1093/nar/gku557*

Ma, S., Shi, R., Wang, X., Liu, Y., Chang, J., Gao, J., Lu, W., Zhang, J., Zhao, P., & Xia, Q. (2014). Genome editing of BmFib-H gene provides an empty Bombyx mori silk gland for a highly efficient bioreactor. Scientific Reports, 4, 1–7.*https://doi.org/10.1038/srep06867*

Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics, 27(6), 764–770.*https://doi.org/10.1093/bioinformatics/btr011*

Meng, Y., Katsuma, S., Daimon, T., Banno, Y., Uchino, K., Sezutsu, H., Tamura, T., Mita, K., & Shimada, T. (2009). The silkworm mutant lemon (lemon lethal) is a potential insect model for human sepiapterin reductase deficiency. Journal of Biological Chemistry, 284(17), 11698–11705.*https://doi.org/10.1074/jbc.M900485200*

Minh, B. Q., Nguyen, M. A. T., & Von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. Molecular Biology and Evolution, 30(5), 1188–1195. https://doi.org/10.1093/molbev/mst024

Nguyen, L.T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution, 32(1), 268–274.*https://doi.org/10.1093/molbev/msu300*

Nishikawa, H., Iijima, T., Kajitani, R., Yamaguchi, J., Ando, T., Suzuki, Y., ... Fujiwara, H. (2015). A genetic mechanism for female-limited Batesian mimicry in Papilio butterfly. Nature Genetics, 47(4), 405–409. https://doi.org/10.1038/ng.3241

Papantonis, A., Swevers, L., & Iatrou, K. (2015). Chorion Genes: A Landscape of Their Evolution, Structure, and Regulation. Annual Review of Entomology, 60(1), 177–194.*https://doi.org/10.1146/annurev-ento-010814-020810*

Peigler, R. S., & Naumann, S. (2003). A revision of the silkmoth genus Samia. University of the Incarnate Word.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. -C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology, 33, 290. Retrieved from*https://doi.org/10.1038/nbt.3122*

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. Bioinformatics (Oxford, England), 21 Suppl 1(suppl_1), i351-8.*https://doi.org/10.1093/bioinformatics/bti1018*

Rodakis, G. C., & Kafatos, F. C. (1982). Origin of evolutionary novelty in proteins: how a high-cysteine chorion protein has evolved. Proceedings of the National Academy of Sciences of the United States of America, 79(11), 3551–3555.*https://doi.org/10.1073/pnas.79.11.3551*

Saitoh, H. (2011). Yamamayuga-ka youtyu no aoirosikisoketugoutanpakusitu –sinjusanyoutyutaieki to hihusosiki ni okeru Biliverdin ketugoutanpakusitu no seisitu. Kankyo, (23), 26–34. Retrieved from*http://hdl.handle.net/10212/203*

Sezutsu, H., & Yukuhiro, K. (2014). The complete nucleotide sequence of the eri-silkworm (Samia cynthia ricini) fibroin gene. Journal of Insect Biotechnology and Sericology, 83(3), 59–70.

Singh, B. K., Kumar, R., Ahmed, S. A., & Pathania, P. C. (2017). Diversity and their clarification of species Genus Samia (Lepidoptera : Saturniidae) in India and their prospects for utilization (LEPIDOPTERA : SATURNIIDAE) IN INDIA AND THEIR PROSPECTS. Journal of Insect Science, 30(1), 43–52.

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 6(5), 31.*https://doi.org/10.1186/1471-2105-6-31*

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics, 7(1), 62.*https://doi.org/10.1186/1471-2105-7-62*

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24(5), 637–644.*https://doi.org/10.1093/bioinformatics/btn013*

Sternburg, J., & Waldbauer, G. (1984). Diapause and Emergence Patterns in Univoltine and Bivol Tine Populations of Promethea (Lepidoptera: Saturniidae). The Great Lakes Entomologist, 17(3), 8.

Takasu, Y., Hata, T., Uchino, K., & Zhang, Q. (2010). Identification of Ser2 proteins as major sericin components in the non-cocoon silk of Bombyx mori. Insect Biochemistry and Molecular Biology, 40(4), 339–344.*https://doi.org/10.1016/j.ibmb.2010.02.010*

Tamura, T., & Kubota, T. (1988). A determination of molecular weight of fibroin polypeptides in the saturnid silkworms, Antheraea yamamai, Antheraea pernyi and Philosamia cynthia ricini by SDS PAGE. In International Society for Wild Silkmoths.

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics, (SUPPL. 25), 1–14.*https://doi.org/10.1002/0471250953.bi0410s25*

Toyama, K. (1906). Studies on the hybridology of insects. I. On some silkworm crosses, with special reference to Mendel's law of heredity. The Bulletin of the College of Agriculture, Tokyo Imperial University., 7, 259–393.*https://doi.org/10.1192/bjp.111.479.1009-a*

Triant, D. A., Cinel, S. D., & Kawahara, A. Y. (2018). Lepidoptera genomes: current knowledge, gaps and future directions. Current Opinion in Insect Science, 25, 99–105.*https://doi.org/10.1016/j.cois.2017.12.004*

Tsubota, T., Yamamoto, K., Mita, K., & Sezutsu, H. (2015). Gene expression analysis in the larval silk gland of the eri silkworm Samia ricini. Insect Science, 1–14.*https://doi.org/10.1111/1744-7917.12251*

Vaser, R., Sovic, I., Nagarajan, N., & Sikic, M. (2016). Fast and accurate de novo genome assembly from long uncorrected reads. BioRxiv, 068122.*https://doi.org/10.1101/068122*

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics, 33(14), 2202–2204.*https://doi.org/10.1093/bioinformatics/btx153*

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE, 9(11).*https://doi.org/10.1371/journal.pone.0112963*

Waterhouse, R. M., Seppey, M., Simao, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution, 35(3), 543–548.*https://doi.org/10.1093/molbev/msx319*

Yoshido, A., Yasukochi, Y., & Sahara, K. (2011). Samia cynthia versus Bombyx mori: Comparative gene mapping between a species with a low-number karyotype and the model species of Lepidoptera. Insect Biochemistry and Molecular Biology, 41(6), 370–377.*https://doi.org/10.1016/j.ibmb.2011.02.005*

Zhang, J., Cong, Q., Rex, E. A., Hallwachs, W., Janzen, D. H., Grishin, N. V., & Gammon, D. B. (2019). Gypsy moth genome provides insights into flight capability and virus–host interactions. Proceedings of the National Academy of Sciences of the United States of America, 116(5), 1669–1678. https://doi.org/10.1073/pnas.1818283116

Zurovec, M., Yonemura, N., Kludkiewicz, B., Sehnal, F., Kodrik, D., Vieira, L. C., Kucerova, L., Strnad, H., Konik, P., & Sehadova, H. (2016). Sericin Composition in the Silk of Antheraea yamamai. Biomacromolecules, 17(5), 1776–1787.*https://doi.org/10.1021/acs.biomac.6b00189*

Data Accessibility

NGS data obtained in this study are available under the accession numbers DRA009717 and DRA009718 (DDBJ), respectively. Genome assembly is available at https://datadryad.org/stash/share/5hfxrylqOiDgzbXoM_8esv8rQGyC

Author Contributions

J.L. designed the experiments and analyzed the data and wrote the manuscript. T.N. prepared Illumina libraries for genome sequencing. K.Y. and S.S. performed the sequence runs. Y.S. prepared the library for RNA sequencing and performed the sequence runs. T.S., T.K., S.K. and J.L. discussed the results.

Table and Figures

Table 1 Features of *S. ricini* genome.

Table 2 The result of linkage analysis.

Table 3 The presence of *Fib-H*, *Fib-L*,*fibrohexamerin*, *p25* and *sericin* genes in 5 lepidopteran genomes.

The numbers in the column of *S. ricini* and *B. mori* stand for copy number of each genes. The accession numbers in the column of*P. xuthus* were derived from the transcripts of the corresponding genes, registered at Genbank. Circles in the columns of *P. xylostella* and *D. plexippus* indicate that genome assembly of each species has at least one genomic region showing high similarity to the *B. mori* silk proteins with an e-value less than 1e-5. Question marks means BLAST search failed to identify any region with high similarity.

Fig 1. Graphical view of *S. ricini.*

(A) Fifth-instar larva of *S. ricini* .

(B) Adult male moth of *S. ricini* .

Fig 2. Amount and proportion of repeat sequences in the*S. ricini* genome.

Amount (bp; A) and proportion (%; B) of repetitive sequences in five lepidopteran species, including *S. ricini* , *B. mori* ,*D. plexippus* , *P. xuthus* and *P. xylostella* .

Fig 3. Comparison between *S. ricini* and *B. mori* genome

(A) Left side of ideogram represents chromosomes of *B. mori* and right side represents scaffolds of *S. ricini* . 'bm_1' to 'bm_28' corresponds to the chromosomes of *B. mori* . As for *S. ricini* , 35 scaffolds, 'Sr_HGAP_ JL_scaf_1 (sr_1)' to 'Sr_HGAP_ JL_scaf_35 (sr_35)' are shown. Outer ring (black) indicates putative chromosomes of *S. ricini* . The chromosome numbers of *S. ricini* are given according to Yoshido *et al.* (2011).

(B) Venn diagram of protein orthogroups in five lepidopteran species. Number in each section indicates the number of orthogroups.

(C) Phylogenetic tree of five lepidopteran species based on alignments of 3,907 single copy orthologs. *Plutella xylostella* was used as the outgroup for rooting the tree. Bootstrap values are shown on the branches.

(D) Phylogenetic tree of *S. ricini* chorion proteins (SrCho),*B. mori* chorion proteins (BmCho), P. xylostella chorion proteins (PxyCho), *P. xuthus* chorion proteins (PxuCho) and *D. plexippus* chorion proteins (DpCho). Branch colors are: Red– BmCho; Blue–SrCho; Purple–PxyCho; Green–PxuCho; Orange–DpCho.

Fig 4. Graphical view of *S. c. pryeri* and hybrid progenies.

(A) Fifth-instar larva of *S. c. pryeri* .

(B) 5th-instar larvae of *S. ricini* and three form of BC_1 obtained by the crossing (*S. ricini* x*S. c. pryeri*) x *S. ricini* .

(C) Cocoon of *S. ricini* , *S. c. pryeri* , F_1individuals and 'Red cocoon' individuals in BC_1.

Fig 5. Linkage mapping of 'Blue,' 'Yellow,' 'Spot' and 'Red cocoon.'

Segregation patterns of PCR-based markers in the BC_1progenies which showed 'Blue (A),' 'Yellow (B),' 'Spot (C),' and 'Red cocoon (D)' phenotypes.

Supporting Information

Table S1 Summary of statistics of illumina short read data.

*Illumina sequencing was conducted twice re-using the same libraries because amounts of obtained reads via the first sequencing were not sufficient for assembly. Rounds 1 and 2 mean the first and second sequencing, respectively.

Table S2 Summary of statistics of Pacbio long read data.

4 SMRT cells were used for obtaining long reads. The number of subreads and total bases per SMRT cell are shown.

Table S3 Summary of statistics of Illumina RNA-seq data.

Table S4 Sequences of genetic markers for linkage analysis.

Table S5 BUSCO assessment of lepidopteran genome assemblies.

Table S6 Annotations of genes in chorion gene cluster.

The best-hit results of BLASTP search to non-redundant protein database were shown. The top hit of evm.model.Sr_HGAP_JL_scaf_2.1091 was not annotated as 'chorion,' but some superior hits were annotated as 'chorion,' so we decided for this gene to be in 'chorion.' Because evm.model.Sr_HGAP_JL_scaf_2.1135 showed no similarity to any registered sequences, we utilized this gene as an outgroup.

Table S7 Putative *sericin* genes of *S. ricini*registered in NCBI genbank.

* BLASTP search could not identify the corresponding gene models, but TBLASTN search was able to find the identical genomic regions with an e-value less than 1e-5. Because LC001867 and LC001870 were elucidated to be mapped to the same locus, we concluded that LC001867 and LC001870 were splicing variants of a single gene.

Table S8 Proportion and amount of LINE elements in the genomes of *S. ricini* and *B. mori*.

Table S9 *S. ricini* chorion showing the highest similarity to High-cysteine (Hc) chorion of *B.mori*

The best-hit results of BLASTP search with Hc chorion of *B.mori*as query to *S. ricini* chorion proteins were shown.

Fig. S1. Genetic markers for linkage analysis and the result of linkage analysis

(A) Electrophoresis of genomic PCR using genetic markers to distinguish*S. ricini* and *S. c. pryeri* . Each marker is specific to 35 scaffolds (> 1 Mbp).

(B) Scaffold segregation patterns in BC_1 individuals.

Fig. S2. k-mer distribution analysis of the *S. ricini* and*A. yamamai* genomes.

GenomeScope k-mer profile plots of the *S. ricini* (A) and *A. yamamai* (B) genomes showing the fit of the GenomeScope model (black) to the observed k-mer (k=31) frequencies (blue).

Fig. S3. InterProScan IDs distribution

Fig. S4. Results of TBLASTN search

(A) Result of TBLASTN search against the genome assembly of *S. ricini* using *S. ricini* Fib-H amino acid sequence as query.

(B) Result of TBLASTN search against the genome assembly of *A. yamamai* using *B. mori* Fib-L amino acid sequence as query.

Fig. S5. Phylogenetic tree of sericins of *S. ricini* and*B. mori*.

Fig. S6. Larval black spots on the integument of *S. c. pryeri*.

Dorsal (A, C) and lateral (B) views of a *S. c. pryeri* larva. The red arrowheads indicate the black spots while the blue arrowhead indicates a spiracle.

Fig. S7. Pigment extracts of 'Red cocoon.'

In order to extract pigments, red cocoons were boiled for 30 min in extraction solution. Extraction solution was a mixture of 100 mg of sodium carbonate, 250 mg of Triton-X and 50 mL of water.

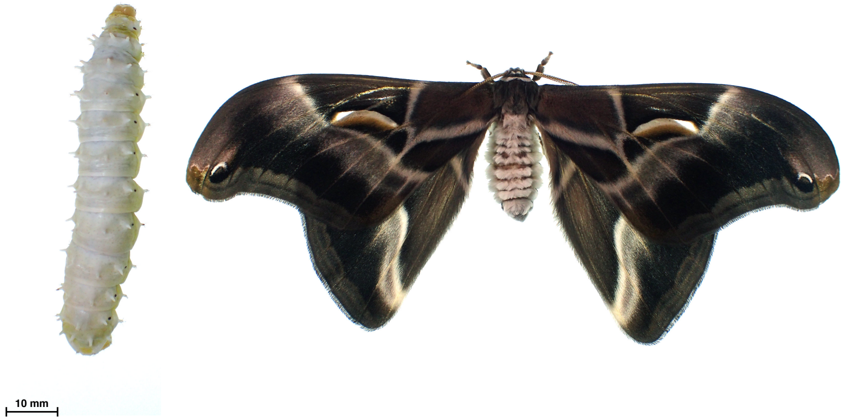

Fig. 1

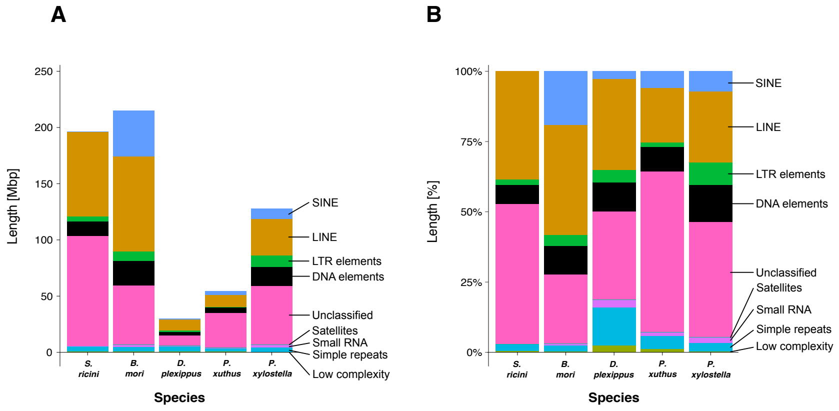

Fig. 2

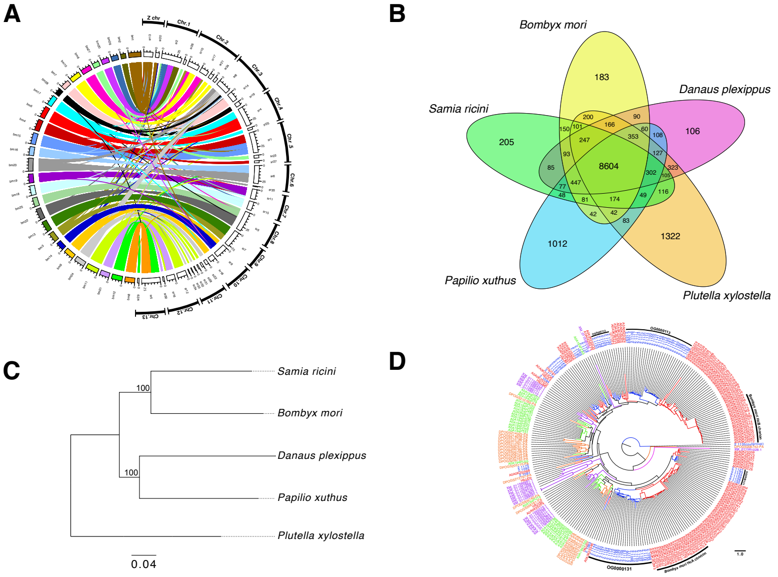
A
B
Bombyx mori
Samia ricini
Danaus plexippus
Papilio xuthus
Plutella xylostella
183
205
106
1012
1322
8604
C
100
100
Samia ricini
Bombyx mori
Danaus plexippus
Papilio xuthus
Plutella xylostella
0.04
D

Fig. 3

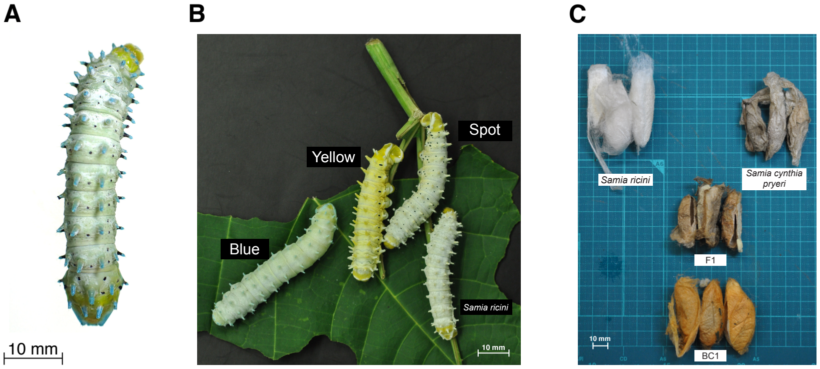
A
10 mm
B
Yellow
Spot
Blue
Samia ricini
10 mm
C
Samia ricini
Samia cynthia pryeri
F1
BC1
10 mm

Fig. 4

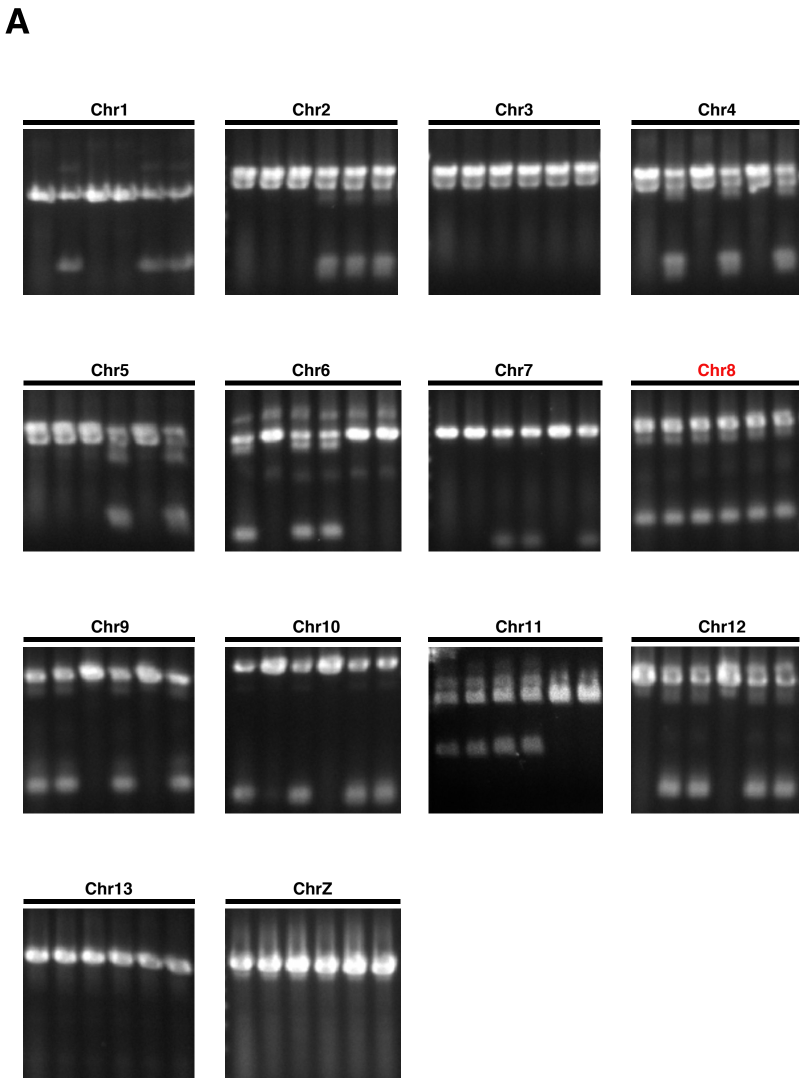

Fig. 5

Hosted file

Table1.xlsx available at https://authorea.com/users/315693/articles/446018-the-genome-sequence-of-samia-ricini-a-new-model-species-of-lepidopteran-insect

Hosted file

Table2.xlsx available at https://authorea.com/users/315693/articles/446018-the-genome-sequence-of-samia-ricini-a-new-model-species-of-lepidopteran-insect

Hosted file

Table3.xlsx available at https://authorea.com/users/315693/articles/446018-the-genome-sequence-of-samia-ricini-a-new-model-species-of-lepidopteran-insect