MOLECULAR ECOLOGY RESOURCES

Supplemental Information for:

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

Jung Lee, Tomoaki Nishiyama, Shuji Shigenobu, Katsushi Yamaguchi, Yutaka Suzuki, Toru Shimada, Susumu Katsuma and Takashi Kiuchi

Table of Contents:

MOLECULAR ECOLOGY
RESOURCES

Table S1 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.

Cell_name	Number of Subreads	Total base
PAC0606 11pm PAC0599*2 6pm_C01-Cell3	707,641	6,400,957,316
PAC0606 11pm PAC0599*2 6pm_B01-Cell2	778,298	6,389,349,540
PAC0599 6PM DIFF-Cell1	855,740	7,384,988,455
PAC0599 10PM DIFF-Cell1	925,576	6,587,181,920
	3,267,255	26,762,477,231

Table S2 Summary of statistics of illumina short read data.
* Illumina sequencing was conducted by pooling together libraries and running these across two HiSeq1500 High Output lanes. A total of 401,799,912 read pairs were obtained, which equates to approximately 252-fold coverage.

	Sample Index	Target fragment size	Read-pair count on lane 5	Read-pair count on lane 6	sex	Library preparation kit
paired-end	2	200—250 bp	84,216,357	105,207,338	ZZ	Illumina: Truseq DNA PCR-Free Sample Prep kit
	8	310—530 bp	3,748,902	4,562,399	ZO	Kapa hyper prep kit
Mate-pair	18	29.9–38.4 kbp	1,998,039	2,316,990	ZZ	Illumina Nextera Mate-Pair library prep kit
	14	17.0-29.9 kbp	4,517,482	5,309,208	ZZ	Illumina Nextera Mate-Pair library prep kit
	13	12.0–17.0 kbp	9,006,739	10,783,518	ZZ	Illumina Nextera Mate-Pair library prep kit
	12	9.0–12.0 kbp	16,707,198	19,641,462	ZZ	Kapa Hyper prep for the index ligation
	15	9.0–12.0 kbp	8,683,528	10,331,150	ZZ	Illumina Nextera Mate-Pair library prep kit
	6	7.0–9.0 kbp	12,575,090	14,772,609	ZZ	Illumina Nextera Mate-Pair library prep kit
	16	5.0–7.0 kbp	14,117,745	16,583,628	ZZ	Illumina Nextera Mate-Pair library prep kit
	7	4.0–5.0 kbp	16,993,251	14,772,609	ZZ	Illumina Nextera Mate-Pair library prep kit
	4	3.0–4.0 kbp	11,632,269	13,322,401	ZZ	Illumina Nextera Mate-Pair library prep kit
		grand total	401,799,912			

Table S3 Summary of statistics of Illumina RNA-seq data.
* hpo stands for "hours post oviposition."

Tissue	Sex	Read length [bp]	Read count	Total base [bp]
24 hpo embryo	male	100	44,539,911	4,453,991,100
	female	100	39,068,068	3,906,806,800
36 hpo embryo	male	100	45,561,356	4,556,135,600
	female	100	40,607,832	4,060,783,200
48 hpo embryo	male	100	43,771,786	4,377,178,600
	female	100	48,538,789	4,853,878,900
105 hpo embryo	male	101	40,468,680	4,087,336,680
	female	101	44,888,638	4,533,752,438
Midgut	Mix	76	48,152,085	3,659,558,460
Anterior Silk gland	Mix	100	42,311,688	4,231,168,800
Middle Silk gland	Mix	100	21,870,312	2,187,031,200

MOLECULAR ECOLOGY RESOURCES

Table S4 BUSCO assessment of lepidopteran genome assemblies.

Species	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Total searched BUSCOs	% complete BUSCO
Samia ricini	1615	8	16	19	1658	97.9
Bombyx mori	1611	20	9	18	1658	98.4
Danaus plexippus	1591	33	17	17	1658	97.9
Papilio xuthus	1596	22	16	24	1658	97.6
Plutella xylostella	1421	205	12	20	1658	98.1

Table S5 Sequences of genetic markers for linkage analysis.

Asterisks mean the corresponding marker was also utilised for linkage mapping.
***please refer to the attached file that corresponds to the same name.**

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 utilised this gene as an outgroup.
***please refer to the attached file that corresponds to the same name.**

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.

Accession No.	Gene model	References
LC001866	evm.model.Sr_HGAP_JL_scaf_28.54	Tsubota *et al*. (2015)
LC001867	o*	
LC001868	evm.model.Sr_HGAP_JL_scaf_28.98	
LC001869	evm.model.Sr_HGAP_JL_scaf_28.90	
LC001870	o*	
GBZD01002008.1	evm.model.Sr_HGAP_JL_scaf_28.98	Dong *et al*. (2015)
GBZD01003513.1	evm.model.Sr_HGAP_JL_scaf_36.6	
GBZD01001705.1	evm.model.Sr_HGAP_JL_scaf_28.59	
GBZD01020841.1	evm.model.Sr_HGAP_JL_scaf_18.243	

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

***please refer to the attached file that corresponds to the same name.**

Table S9 BUSCO assessment of the predicted genes of *S. ricini* and *B. mori*

Species	Complete and single-copy BUSCOs	Complete and duplicated BUSCOs	Fragmented BUSCOs	Missing BUSCOs	Total searched BUSCOs	% complete BUSCO
Samia ricini	1513	10	89	46	1658	91.9
Bombyx mori	1561	17	43	37	1658	95.2

MOLECULAR ECOLOGY
RESOURCES

Table S10 Annotation of *S. ricini* specific 205 Orthogroups (OGs) and genes included in each OGs

The first columns of rows of 46 retrotransposon related OGs were filled in yellow.

***please refer to the attached file that corresponds to the same name.**

Table S11 *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.

***please refer to the attached file that corresponds to the same name.**

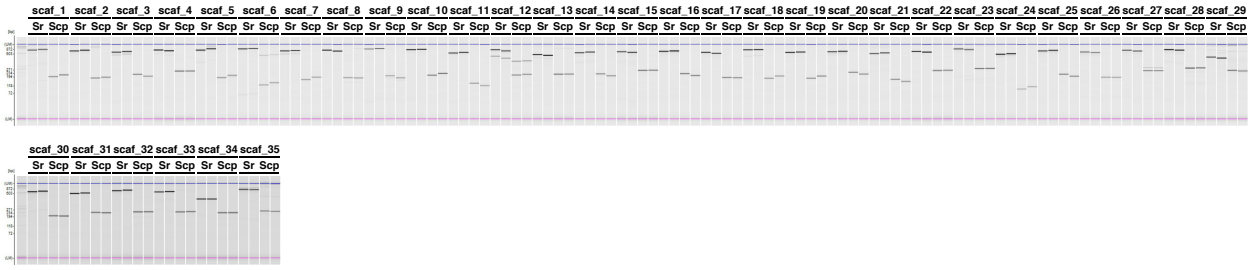
Table S12 Cysteine contents of *S. ricini* chorion proteins and *B. mori* High-cysteine (Hc) chorion proteins

***please refer to the attached file that corresponds to the same name.**

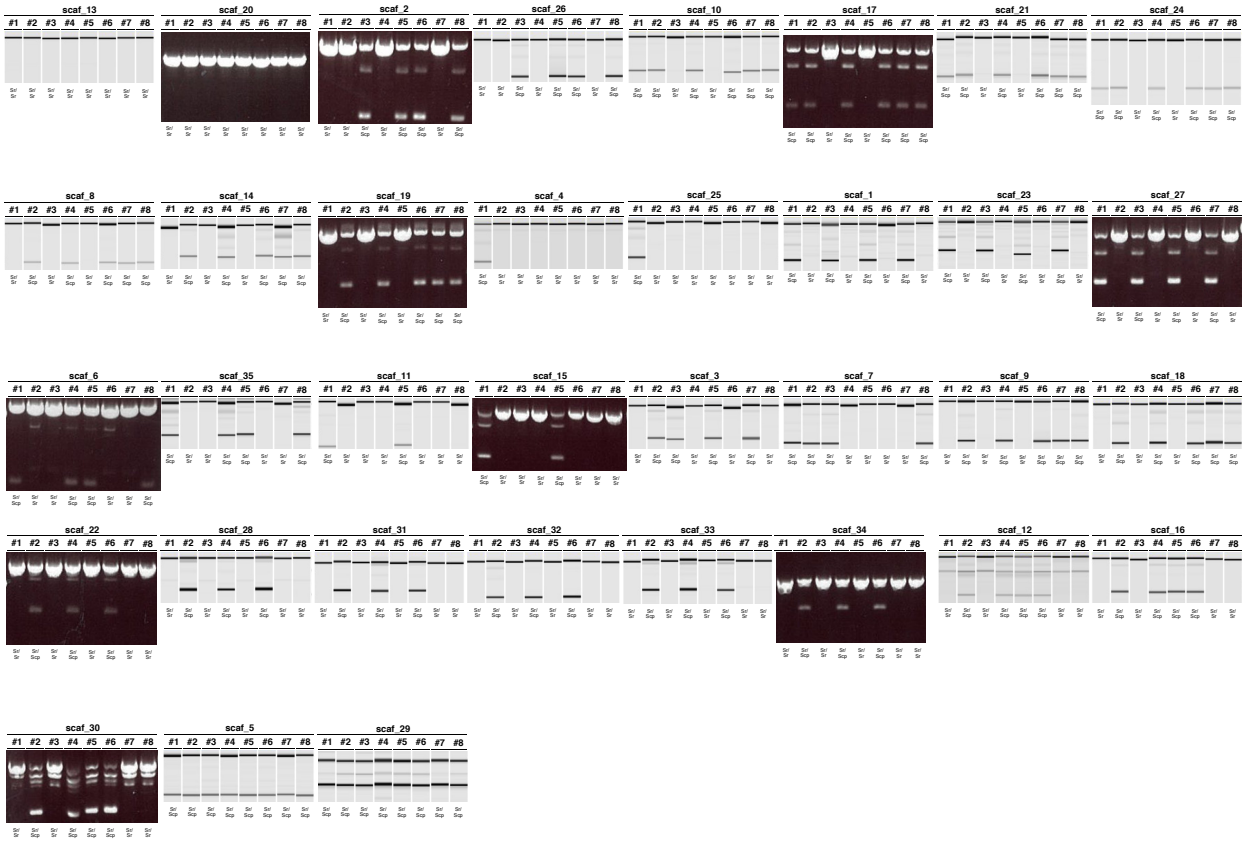
MOLECULAR ECOLOGY
RESOURCES

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

A



B



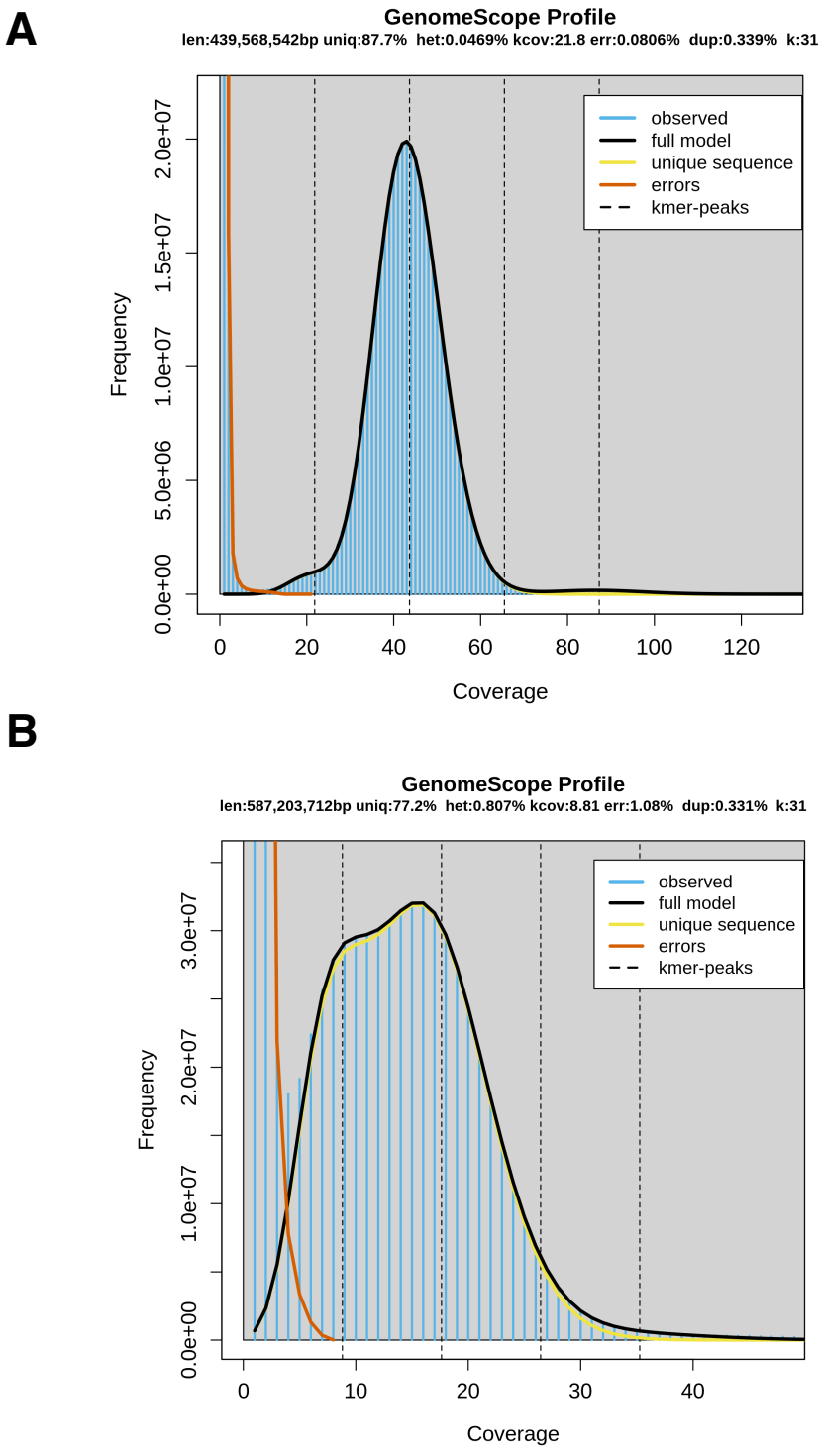
(A) Electrophoresis of genomic PCR using genetic markers to distinguish *S. ricini* and *S. c. pryeri*.

Each marker is specific to 35 scaffolds (> 1 Mb).

(B) Scaffold segregation patterns in BC_1 individuals.

MOLECULAR ECOLOGY
RESOURCES

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).

MOLECULAR ECOLOGY
RESOURCES

Fig. S3. InterProScan IDs distribution

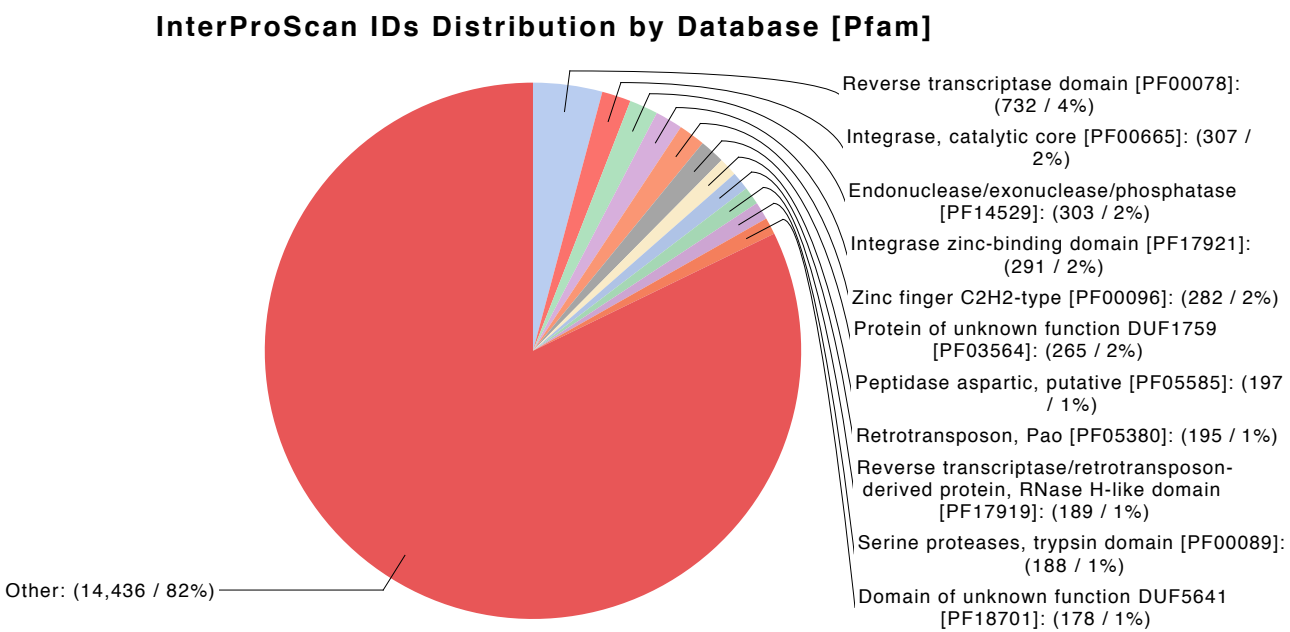


Fig. S4. Results of TBLASTN search

```
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.


Database: Sr_genome.fa
           155 sequences; 450,479,495 total letters


Query= BAQ55621.1 fibroin [Samia ricini]

Length=2880
                                                                    Score        E
Sequences producing significant alignments:                        (Bits)     Value

Sr_HGAP_JL_scaf_3                                                   2117       0.0


Query_1           3        VTAFVILCCVLQYVTARSLDDDIHSLERGYRETSKTGYDEYDVDKSGRLYERLTTRKKFE  62
Sr_HGAP_JL_scaf_3 4108320  I*SNTTFN.S- ..........................................    4108144

Query_1           63       RDSAPSRVPGGGTLVEKIVIERAPTGHETIYEEDVVIKQVPQGGAASSAASSASAGSGSG  122
Sr_HGAP_JL_scaf_3 4108143   ..........................................................  4107964

Query_1           123      APTIIVERGSGAGSGRSHGAGSAAGAAAAAAAAAAAAAGGAGRGGGGGYGRGHGTAGSAAA  182
Sr_HGAP_JL_scaf_3 4107963  ............W ..........................................    4107784

Query_1           183      AAAAAAAAAASSEGGSAGGYWQGYGSDSSAAAAAAAAAAAAGSAGGSYGSDSAEAAAAAAA  242
Sr_HGAP_JL_scaf_3 4107783  ....................................-- ..................    4107610

Query_1           243      AAAAAGTAAGGSGGGYGGDGGAAAAAAAAAAAAAASGAGSGYGGGYGHGYGSDGGAAAAAA  302
Sr_HGAP_JL_scaf_3 4107609   ..........................................................  4107430

Query_1           303      AAAAAAAGGAGRGYGAGSAAAAAAAAAAAAAGGSGGGYGGDGGAAAAAAAAAAAAASGAGS  362
Sr_HGAP_JL_scaf_3 4107429   ..........................................................  4107250

Query_1           363      GYGGGARGGYGHGYGSDGGAAAAAAAAAAAAAAGGAGGGYGGGYGGGYGGDGGAAAAAAAA  422
Sr_HGAP_JL_scaf_3 4107249   ..........................................................  4107070

Query_1           423      AAAAAGGAGSGYGGGSWHSYGSDSGAAAAAAAAAAAAAAGGSGGGYGGDGGAAAAAAAAA  482
Sr_HGAP_JL_scaf_3 4107069   ..........................................................  4106890

Query_1           483      AAAAGGSGGGYGGDGGAAAAAAAAAAAAAAGGAGDGYGAGSAAAAAAAAAAAAGGAGGGYG  542
Sr_HGAP_JL_scaf_3 4106889   ..........................................................  4106710

Query_1           543      GDGGAAAAAAAAAAAAAAGGAGSGYGGGARRGYGHGYGSDGGAAAAAAAAAAAAAGGSGGG  602
Sr_HGAP_JL_scaf_3 4106709   ..........................................................  4106530

Query_1           603      YGGDGGAAAAAAAAAAAAAGGAGSGYGGGARGGYGHGYGSDGGAAAAAAAAAAAAAGGSG  662
Sr_HGAP_JL_scaf_3 4106529   ..........................................................  4106350

Query_1           663      GGYGGDGGAAAAAAAAAAAAAGGAGSGYGGSARRGYGHGYGSDGGAAAAAAAAAAAAAGG  722
Sr_HGAP_JL_scaf_3 4106349   ..........................................................  4106170

Query_1           723      SGGGYGGDGGAAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAAGGAGDGYGAGSA  782
Sr_HGAP_JL_scaf_3 4106169   ..........................................................  4105990

Query_1           783      AAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAAAGGAGSGYGGGARRGYGHGYGSDG  842
Sr_HGAP_JL_scaf_3 4105989   ..........................................................  4105810

Query_1           843      GAAAAAAAAAAAAAGGSGGGYGGDGGAAAAAAAAAAAAAGGAGSGYGGGARRGYGHGYGS  902
Sr_HGAP_JL_scaf_3 4105809   ..........................................................  4105630

Query_1           903      DGGAAAAAAAAAAAAAAGSAESSYGGGSWYGYGSDSSAAAAAAAAAAAGGAGGGYGGDGGSA  962
Sr_HGAP_JL_scaf_3 4105629   ..........................................................  4105450

Query_1           963      AAAAAAAAAAAGGSGGGYGGDGGAAAAAAAAAAAAAAGGAGGGYGGGYGGDGGAAAAAAAA  1022
Sr_HGAP_JL_scaf_3 4105449   ..........................................................  4105270

Query_1           1023     AAAAAGGSRSGYGGGSWHGYGSDSGAAAAAAAAAAAAAAGGAGDGYGPGSAAAAAAAAAAA  1082
Sr_HGAP_JL_scaf_3 4105269   ..........................................................  4105090

Query_1           1083     AGGAGGGYGGDGGAAAAAAAAAAAAAAGGAGSGYGGGARRGYGHGYGSDGGAAAAAAAAAA  1142
Sr_HGAP_JL_scaf_3 4105089   ..........................................................  4104910

Query_1           1143     AAAGGSGGGYGGDGGAAAAAAAAAAAAAGGAGSGYGGGSWHGYGSDSSAAAAAAAAAAAAA  1202
Sr_HGAP_JL_scaf_3 4104909  ..............................RKV-TEVV..MAM.QTVV             4104766
Sr_HGAP_JL_scaf_3 4104814                                  E ........................    4104728

Query_1           1203     GGAGGGYGAGSAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAAGGAGGGYGGGYG  1262
Sr_HGAP_JL_scaf_3 4104727   ..........................................................  4104548

Query_1           1263     GDGGAAAAAAAAAAAAAAGGSRSGYGGGSWHGYGSDSGAAAAAAAAAAAAAGGAGDGYGAG  1322
Sr_HGAP_JL_scaf_3 4104547   ..........................................................  4104368

Query_1           1323     SAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAAGGAGSGYGGGARRGYGHGYGSD  1382
Sr_HGAP_JL_scaf_3 4104367   ..........................................................  4104188

Query_1           1383     GGAAAAAAAAAAAAAGGSGGGYGGDGGAAAAAAAAAAAAGGAGSGYGGGSWHGYGSDSSA  1442
Sr_HGAP_JL_scaf_3 4104187   ..........................................................  4104008

Query_1           1443     AAAAAAAAAAAAGGAGGGYGAGSAAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAA  1502
Sr_HGAP_JL_scaf_3 4104007   ..........................................................  4103828

Query_1           1503     GGAGSGYGGGYGHGYGSDGGAAAAAAAAAAAAAAGGAGGGYGAGSAAAAAAAAAAAAGGAG  1562
Sr_HGAP_JL_scaf_3 4103827   ..........................................................  4103648

Query_1           1563     GGYGGDGGAAAAAAAAAAAAAAGGAGSGYGGGYGHGYGSDGGAAAAAAAAAAAAAGGAGGG  1622
Sr_HGAP_JL_scaf_3 4103647   ..........................................................  4103468

Query_1           1623     YGGDGGAAAAAAAAAAAAAAGGAGSGYGGGARGGYGHGYGSDGGAAAAAAAAAAAAAGGSG  1682
Sr_HGAP_JL_scaf_3 4103467   ..........................................................  4103288

Query_1           1683     GGYGGDGGAAAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAAGGAGSGYGGGYGHG  1742
Sr_HGAP_JL_scaf_3 4103287   ..........................................................  4103108

Query_1           1743     YGSDGGAAAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAAAGGAGSGYGGGARGGYG  1802
Sr_HGAP_JL_scaf_3 4103107   ..........................................................  4102928

Query_1           1803     HGYGSDGGAAAAAAAAAAAAAAGGAGSGYGGGSWNSYGSDSGAAAAAAAAAAAAAGGAGDG  1862
Sr_HGAP_JL_scaf_3 4102927   ..........................................................  4102748

Query_1           1863     YGAGSAAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAAAGGAGSGYGGGARGGYGHG  1922
Sr_HGAP_JL_scaf_3 4102747   ..........................................................  4102568

Query_1           1923     YGSDGGAAAAAAAAAAAAAAGGSGGGYGGDGGAAAAAAAAAAAAAAGGAGSGYGGGSWHSYG  1982
Sr_HGAP_JL_scaf_3 4102567   ..........................................................  4102388

Query_1           1983     SDSGAAAAAAAAAAAAAAGGAGGGYGAGSAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAA  2042
Sr_HGAP_JL_scaf_3 4102387   ..........................................................  4102205
                                      \
                                      |
                                      A

Query_1           2043     AAAAAGGAGSGYGGGARGGYGHGYGSDGGAAAAAAAAAAAAAGGAGSGYGGGSWHSYGSD  2102
Sr_HGAP_JL_scaf_3 4102204   ..........................................................  4102025

Query_1           2103     SGAAAAAAAAAAAAAGGAEGGYGAGSAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAA  2162
Sr_HGAP_JL_scaf_3 4102024   ..........................................................  4101845

Query_1           2163     AAAGSGGGYGGDGGAAAAAAAAAAAAAAGGAGSGYGGGARGGYGHGYGSDGGAAAAAAAAA  2222
Sr_HGAP_JL_scaf_3 4101844   ..........................................................  4101665

Query_1           2223     AAAAGGSEGGYGGDGGAAAAAAAAAAAAAGGAGSGYGGGSWHSYGSDSGAAAAAAAAAAA  2282
Sr_HGAP_JL_scaf_3 4101664  .......G ..........................................          4101485

Query_1           2283     AAGGAGGGYGAGSAAAAAAAAAAAAGGAGGGYGGDGGAAAAAAAAAAAAAGGAGSGYGGG  2342
Sr_HGAP_JL_scaf_3 4101484   ..........................................................  4101305

Query_1           2343     ARGGYGHGYGSDGGAAAAAAAAAAAAAAGGAGSGYGGGSWHSYGSDSGAAAAAAAAAAAAS  2402
Sr_HGAP_JL_scaf_3 4101304   ..........................................................  4101125

Query_1           2403     GGAGSGYGGGSWHSYGSDSGAAAAAAAAAAGAAGGAGGGYGAGSAAAAAAAAAAAAGGAG  2462
Sr_HGAP_JL_scaf_3 4101124   ..........................................................  4100945

Query_1           2463     GGYGRDGGAAAAAAAAAAAAAAGGAGSGYGGGARGGYGHGYGSDGGAAAAAAAAAAAAAGG  2522
Sr_HGAP_JL_scaf_3 4100944   ..........................................................  4100765

Query_1           2523     AGGGYGGGYGGGYGGDGGAAAAAAAAAAAAAAGGAGSGYGGGSWHSYGSDSGAAAAAAAAA  2582
Sr_HGAP_JL_scaf_3 4100764   ..........................................................  4100585

Query_1           2583     AAAAGGAGSGYGGGSRGGYGHGYGSDGGAAAAAAAAAAAAAGGAGGGYGGGYGGDGGAAA  2642
Sr_HGAP_JL_scaf_3 4100584   ..........................................................  4100405

Query_1           2643     AAAAAAAAAAGGSGSGYGGGSWHGYGSDSDAAAAAAAAAASAAAGGAGSGYGGGYWQGYGS  2702
Sr_HGAP_JL_scaf_3 4100404   ..........................................................  4100225

Query_1           2703     NSGAAAAAAAAAAAGGAGSGYGGGARGGYGHGYGSDSAAAAAAAAAAAAAAGGAGGGAGRG  2762
Sr_HGAP_JL_scaf_3 4100224   ..........................................................  4100045

Query_1           2763     AIGGYGGGYGSDNAAAAAAAAAAAAAGGAGGDYGRGYGARSAAAAAAAAAASSGARGAVRV  2822
Sr_HGAP_JL_scaf_3 4100044   ..........................................................  4099865

Query_1           2823     HETGDGFLLRGDYGSDSSAAAAAAAAAAAAAASSASSANGYVSICCKPCRLGTSKIAVH  2880
Sr_HGAP_JL_scaf_3 4099864   ..........................................................    4099691


Lambda      K        H
   0.305    0.123    0.356

Gapped
Lambda      K        H
   0.267   0.0410    0.140

Effective search space used: 411830168828


  Database: Sr_genome.fa
    Posted date:  Sep 5, 2019  2:17 PM
  Number of letters in database: 450,479,495
  Number of sequences in database:  155


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 13
Window for multiple hits: 40
```

B

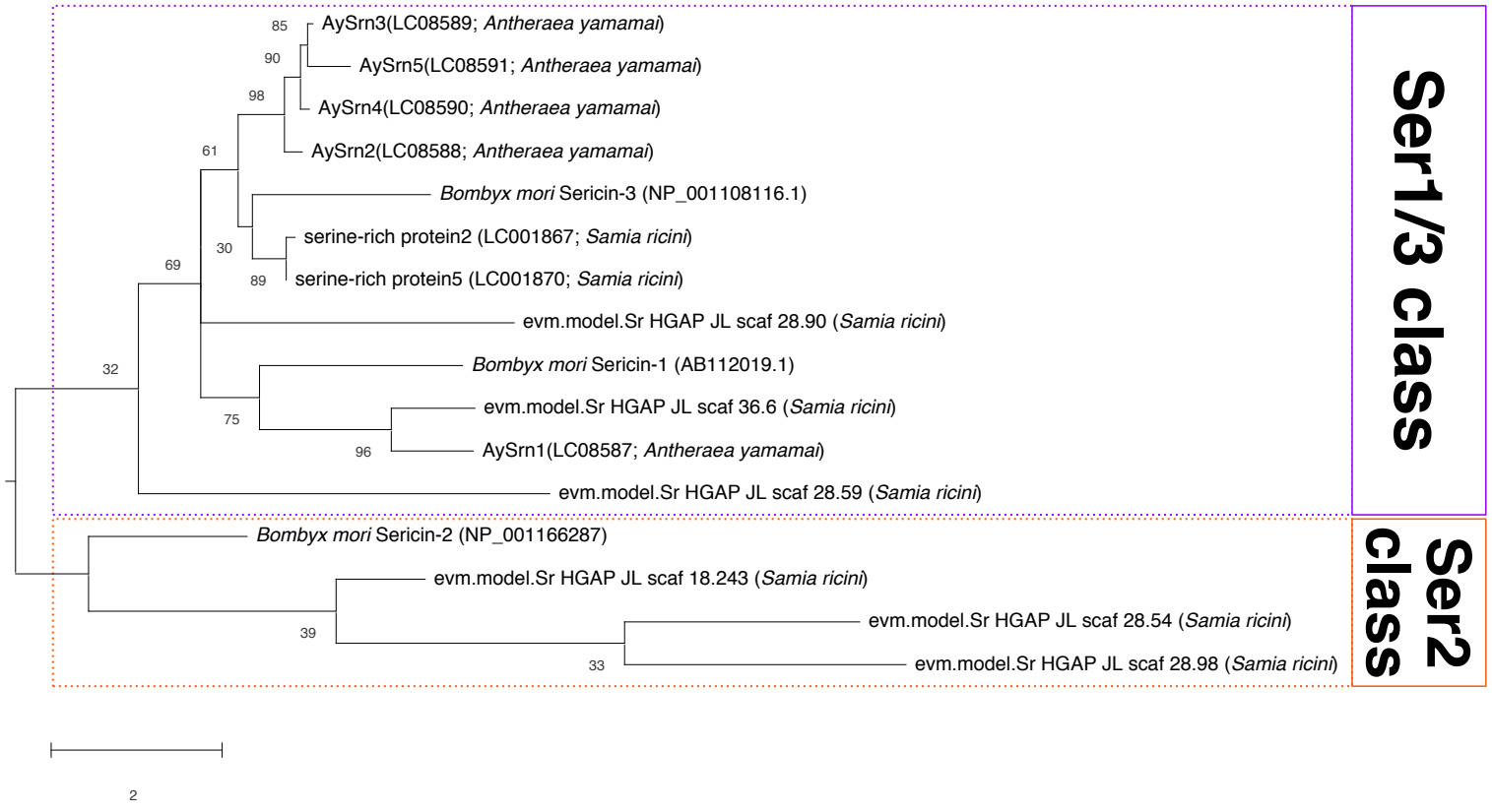
```
TBLASTN 2.9.0+


Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.


Database: AY.fasta
           7,723 sequences; 661,382,373 total letters


Query= sp|P21828|FIBL_BOMMO Fibroin light chain OS=Bombyx mori OX=7091
GN=FIBL PE=1 SV=1

Length=262


***** No hits found *****


Lambda      K        H        a         alpha
   0.317    0.132    0.381    0.792     4.96
```

(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.

MOLECULAR ECOLOGY
RESOURCES

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

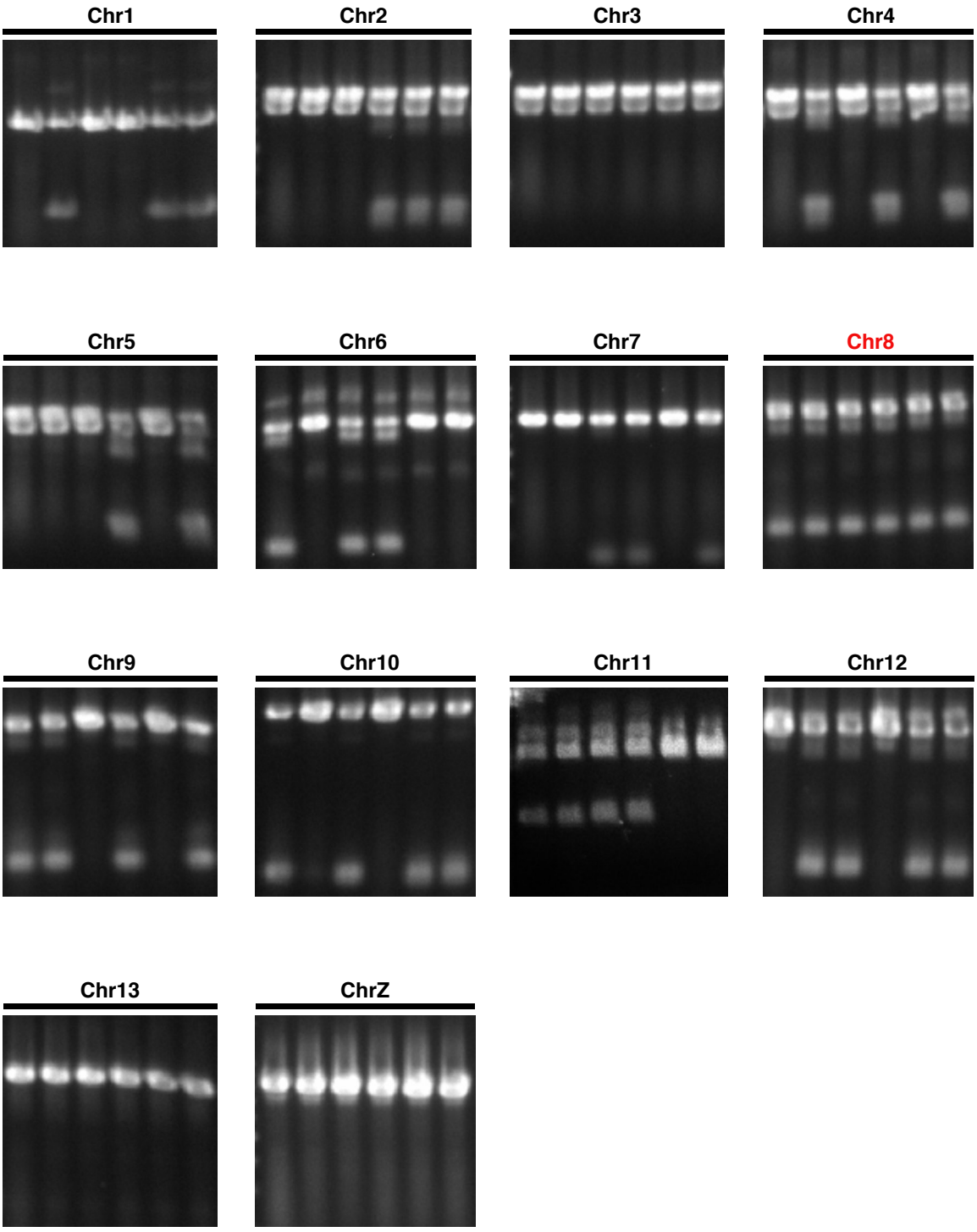


Sericin genes from *S. ricini*, *B. mori*, and *A. yamamai* were analysed. The maximum likelihood tree under Whelan And Goldman + Freq. model (Whelman and Goldman, 2001) was inferred with MEGA X. Bootstrap value are shown on each branch. Branch lengths are proportional to the number of substitutions per site. All sites containing gaps were used for the analysis. Nearest-Neighbor-Interchange (NNI) was used for heuristic tree searching. The sequences are either identified by the gene model id (*S. ricini*) or accession no. The root was manually placed between Ser-2 class and Ser-1/3 class.

MOLECULAR ECOLOGY
RESOURCES

Fig S6. Linkage mapping of 'Blue,' 'Yellow,' 'Spot,' and 'Red cocoon.'

A

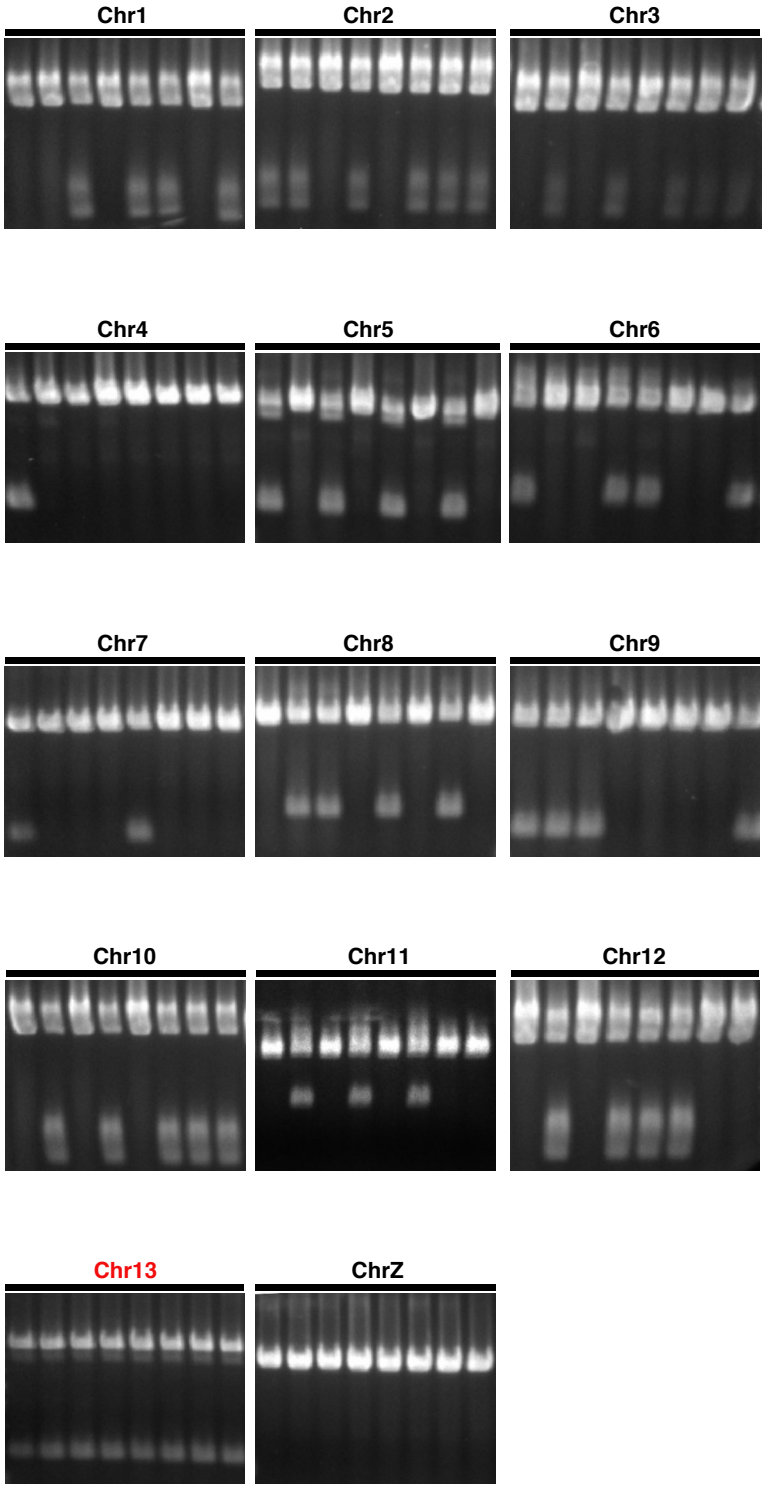


(A) Segregation patterns of PCR-based markers in the BC_1 progenies showing 'Blue' phenotype.

MOLECULAR ECOLOGY
RESOURCES

Fig S6. Linkage mapping of 'Blue,' 'Yellow,' 'Spot,' and 'Red cocoon.'

B

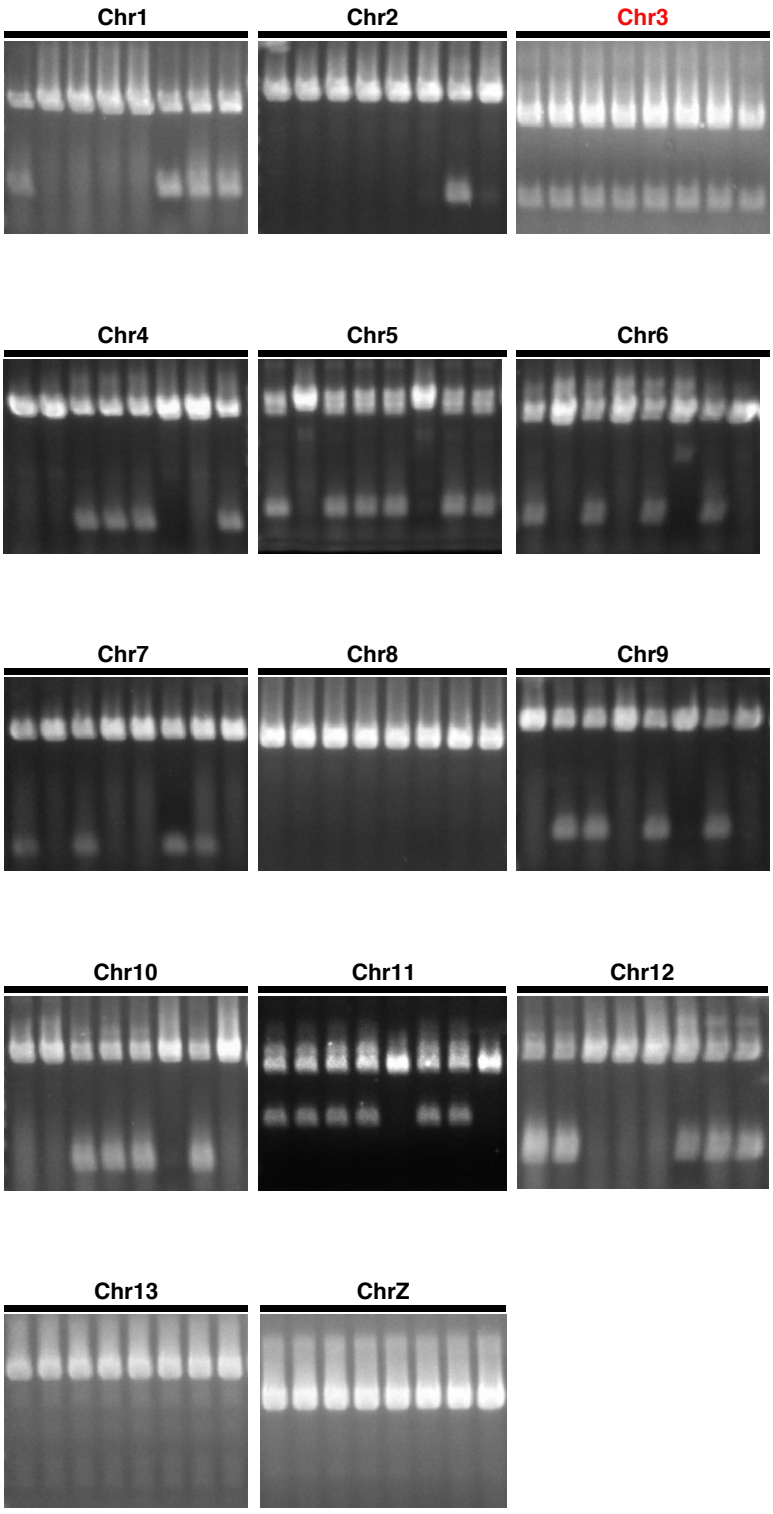


(B) Segregation patterns of PCR-based markers in the BC_1 progenies showing 'Yellow' phenotype.

MOLECULAR ECOLOGY
RESOURCES

Fig S6. Linkage mapping of 'Blue,' 'Yellow,' 'Spot,' and 'Red cocoon.'

C

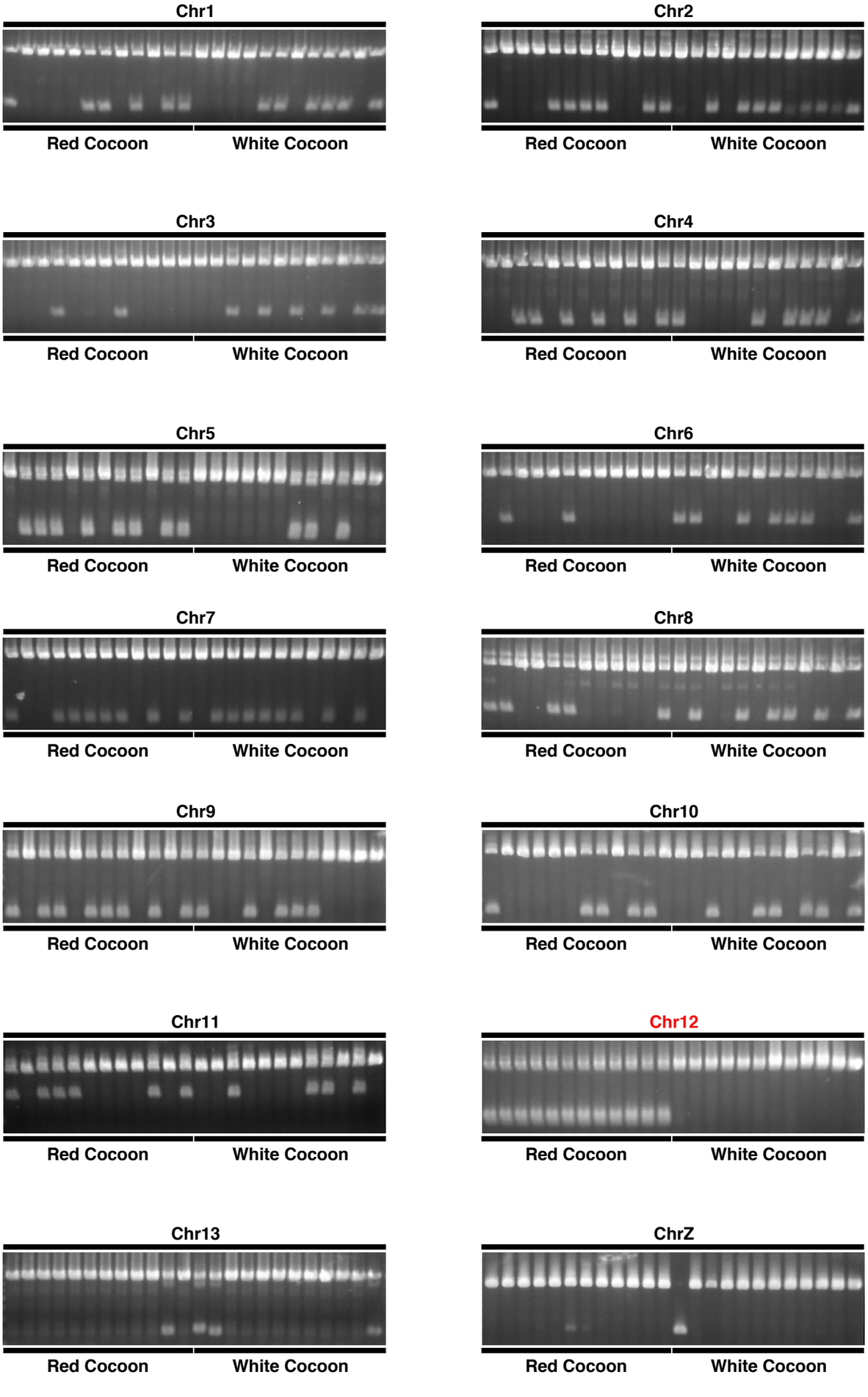


(C) Segregation patterns of PCR-based markers in the BC_1 progenies showing 'Spot' phenotype.

MOLECULAR ECOLOGY
RESOURCES

Fig S6. Linkage mapping of 'Blue,' 'Yellow,' 'Spot,' and 'Red cocoon.'

D



(D) Segregation patterns of PCR-based markers in the BC_1 progenies showing 'Red cocoon' phenotype.