Database, 2019, 1–10
doi: 10.1093/database/baz134
Original article

Original article

SGID: a comprehensive and interactive database of the silkworm

Zhenglin Zhu[1,*], Zhufen Guan[1], Gexin Liu[1], Yawang Wang[1,2] and Ze Zhang[1,*]

[1]School of Life Sciences, Chongqing University, No.55 Daxuecheng South Rd., Shapingba, Chongqing, 401331, China and [2]Khoury College of Computer Sciences, Northeastern University, 401 Terry Ave N, Seattle, WA, 98109, USA

*Corresponding author: Tel: (86)23-6567-8492; Fax: (86)23-6567-8492; Email: zhuzl@cqu.edu.cn
Corresponding address may also be addressed to Ze Zhang. Tel: (86)23-6567-8491; Fax: (86)23-6567-8491; Email: zezhang@cqu.edu.cn

Citation details: Zhu,Z., Guan,Z., Liu,G. *et al.* SGID: a comprehensive and interactive database of the silkworm. *Database* (2018) Vol. 2019: article ID baz134; doi:10.1093/database/baz134

Received 26 September 2019; Revised 27 October 2019; Accepted 1 November 2019

Abstract

Although the domestic silkworm (*Bombyx mori*) is an important model and economic animal, there is a lack of comprehensive database for this organism. Here, we developed the silkworm genome informatics database (SGID). It aims to bring together all silkworm-related biological data and provide an interactive platform for gene inquiry and analysis. The function annotation in SGID is thorough and covers 98% of the silkworm genes. The annotation details include function description, Gene Ontology, Kyoto Encyclopedia of Genes and Genomes pathway, subcellular location, transmembrane topology, protein secondary/tertiary structure, homologous group and transcription factor. SGID provides genome-scale visualization of population genetics test results based on high-depth resequencing data of 158 silkworm samples. It also provides interactive analysis tools of transcriptomic and epigenomic data from 79 NCBI BioProjects. SGID will be extremely useful to silkworm research in the future.

Introduction

The silkworm, *Bombyx mori*, domesticated from its wild ancestry, *B. mandarina*, nearly 5000 years ago, contributes to silk industry, pest control (1, 2) and evolutionary biology. It is a promising model organism in life sciences (3). Because of its importance, the silkworm genome was sequenced and annotated in 2004 (4, 5) and supplementarily annotated in 2012 (6). According to the records in NCBI PubMed, >9000 silkworm-related works have been published. The chromosome-level assembly of the silkworm genome is accomplished in 2017 and published in 2019 (7). Meanwhile, with the development of the sequencing technology, massive silkworm transcriptomic or epigenomic data were produced (8–12). Up to now, there are already >1000 records of silkworm DNA-seq or RNA-seq data documented in NCBI SRA database. A comprehensive archiving and synthesis of these data is important to silkworm research.

Many model organisms have their own online bioinformatics analysis platform, such as TAIR for *Arabidopsis* (13), Flybase (14) for *Drosophila* and MGI (15) for mouse. Currently, SilkBase (16) is the only workable dedicated database for silkworm. It mainly focuses on the archive of sequences and is lacking in analysis tools. Ensembl Silkworm (http://metazoa.ensembl.org/Bombyx_mori) and InsectBase (17) are absence of update and without workable whole genome browser. Until now, there is still not a comprehensive online analysis platform of the silkworm.

To assist silkworm research, we developed the silkworm genome informatics database (SGID) through collecting and cataloging comprehensive genomics, transcriptomics, proteomics and epigenomics data. On the basis of previous works (7, 16, 18), we thoroughly annotated silkworm genes in the contents of function, protein structure, homolog and transcription factor (TF). We also incorporated repeat elements, population statistic tests and epigenomic analysis results into the genome browser, which will help users to get a more comprehensive picture of genome segments. We developed interactive and click-one type analysis tools in SGID, letting users to obtain one or more genes' overall information swiftly.

Materials and Methods

Data processing for basic gene annotation

We used the high-quality assembly of the silkworm genome (7) as the reference. Based on gene models (2017) in SilkBase, we re-annotated all silkworm genes by UniProt (https://www.uniprot.org). Generally speaking, we BLAST the protein sequence of each gene against the UniProt protein database and took hits with significant similarity (E-value < 0.05 and coverage > 0.7). In this way, we made connections of the silkworm genes and UniProt proteins and obtained UniProt annotations of the silkworm genes, including Pubmed ID, EMBL ID, Proteomes ID, Pfam (http://pfam.xfam.org), Interpro (http://www.ebi.ac.uk/interpro), Gene Ontologies (GOs), Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg) and so on.

To validate gene expression in protein level, we manually collected all sequenced silkworm peptides referred in published proteomics related works and did alignments of them and all silkworm genes. We filtered out results with cutoffs of E-value < 0.05 and coverage > 0.8. If the predicted protein of one gene matches two or more peptides, we consider this gene has expression evidence.

We used CD-HIT (19) to search for homologous genes, requiring identity > 50% and coverage > 70%, and identified 1064 gene clusters. We did multiple alignments of the sequences within each cluster by MUSCLE (20), and built the phylogenetic tree by FastTree 2.1 (21) with the parameter '-boot 5000' to test trees' likelihoods. We called repeat elements by RepeatMasker (http://www.repeatmasker.org) and predicted TFs by the pipeline referred in AnimalTFDB 3.0 (22).

Subcellular localization and structure prediction

Annotations from UniProt only cover part of the silkworm proteins. Thus, we re-do protein function and structure prediction for all genes. We predicted subcellular locations of ASFV proteins through CELLO v2.5 (23) and transmembrane helixes within proteins' sequences using TMHMM 2.0 (24). The output images were converted into PNG format for display in websites by Magick (www.imagemagick.org).

We BLAST all protein sequences against the PDB database (http://www.rcsb.org) and extracted significant alignment results as we did for UniProt. We also did protein structure prediction by InterproScan (25) and put significant results into SGID. We used SignalP (26) to predict signal peptides for silkworm proteins.

Gene Ontology

We put the GO from InterproScan, UniProt and SilkBase together as SGID GO data set. We made alignments of the silkworm genes and KEGG silkworm proteins and extracted the hits with E-value < 0.05 and identity > 0.9. In this way, we made connections between the silkworm genes and KEGG proteins. We also extracted KEGG pathway information from KEGG and made connections between pathways and silkworm genes using KEGG protein ID as a bridge. We obtain Entrez IDs of silkworm genes by KOBAS (27).

To enable a gene search using old gene models (4, 28), we made connections between old gene models and SilkBase gene models 2017. Like we did for KEGG proteins, we made alignments of predicted proteins between old gene models and gene models 2017 and selected the best hit of each alignment.

Pre-processing of transcriptomic and epigenomic data

We collected transcriptomic and epigenomic data of silkworm-related projects from NCBI. For transcriptomes, we classified them into three categories, 'DEG' (differentially expressed genes), 'Stage' and 'Tissue'. 'DEG' means the project is to identify differentially expressed genes in different experimental conditions. 'Stage' means the project is to observe gene expression at stages of different time points. Tissue means the project is to obtain expression profiling in different tissues. Following the standard RNA-seq analysis protocol (29), we mapped transcriptomic reads onto the reference genome by bowtie (30) and called Fragments per Kilobase Million (FKPM) by cuffnorm (31).

For epigenomic data, according to experimental methodologies, we classified them into ChIP-Seq, Bisulfite-Seq and miRNA. ChIP-Seq stands for combining chromatin immunoprecipitation (ChIP) assays with next-generation sequencing. Bisulfite-Seq is the use of bisulfite treatment of DNA before routine sequencing to determine the pattern of methylation (32). Small RNA means small RNA sequencing. For ChIP-Seq data, we used Bowtie to align reads onto the reference genome and inspected signatures by MAC2 (33). We used Bismark (34) to pre-process Bisulfite-Seq data. We aligned small RNA reads onto the reference genome by Bowtie. Epigenomic analysis results are converted into bigWig format by the UCSC Genome tool bedGraphToBigWig for display in terminals.

Identification of domestication genes

We used Bowtie to map the genome resequencing data of 142 domesticated and 16 wild silkworm samples, including PRJDB4743, PRJNA402108 and an unpublished resequencing data (depth = ×30) of 15 silkworm samples produced by our lab, onto the reference genome and made bam files by SAMtools (35). To illustrate the evolution pressure at a genome-wide scale, we slid along the silkworm genome with a window size of 2000 bp and a step size of 200 bp. In each sliding window, we calculated Pi, Theta (36), Tajima's D (37) and the composite likelihood ratio (CLR) (38) by ANGSD (39) and SweapFineder2 (40). We also did the four population genetics test for each gene and made coalescent simulation (41) ranking test (CSRT) (42) according to silkworm domestication mode (43). We called domestication genes in a strict method, requiring a Tajima's $D_{\text{domesticated}} < -1$, a $\text{CSRT} < 0.05$, a Tajima's $D_{\text{domesticated}} <$ Tajima's D_{wild}, a Tajima's $D_{\text{min, domesticated}} >$ top 5% point value in ascending order, a $\text{Fst}_{\text{max}} >$ the top 5% point value in descending order and a $\text{CLR}_{\text{max, domesticated}} >$ the top 5% threshold in the whole genome. 'Min' or 'max' indicates the minimal or max value in the genic region extended by 10% of gene length to accommodate the situation that the 5′ or 3′ terminals of a gene is under evolutionary forces. A subscript of 'domesticated' and 'wild' means the population genetics test is performed on domesticated or wild silkworms. We also appended identified domestication genes in (44) and (8) to our domestication genes dataset. We searched for genes possibly under balancing selection in the criteria that a Tajima's $D_{\text{domesticated}} >$ the top 5% point value in ascending ranking, a Tajima's $D_{\text{wild}} < 0.5$, a Tajima's $D_{\text{domesticated}} > 1$ and a $\text{CSRT} < 0.95$.

Genome browser and analysis tools

The genome browser of SGID is developed based on an open source population genetics visualization and analysis package SWAV (swav.popgenetics.org). We used MSAViewer (45) to show multiple alignments of homologous proteins, and phylotree (46) to display phylogenetic trees of gene clusters. The fuzzy text search in the home page is compatible with gene ID, gene name and gene function annotations. The alignment search tools in SGID are Perl codes to parse BLAT (47) or NCBI BLAST results (48). The interface to exhibit gene expression is built upon D3 and JQuery. The overall web structure is Mysql + PHP + CodeIgniter (www.codeigniter.com) + JQuery (jquery.com).

Result and Discussions

The biological data in SGID

Out of the 16 880 gene models predicted in the high-quality assembly of the silkworm genome (7), 13 551 are of function annotations in SilkBase, leaving 3329 with unknown function. To make annotations of genes more comprehensive, SGID incorporated protein information from UniProt and successfully annotated the functions of 15 594 genes, within which 2962 are of function descriptions for the first time. For a lot of genes, SGID gives not only simple descriptions, but also information on function details, chemical properties, related publications, protein structure, topologies, pathways and GOs. In addition to the available GO annotations of 9147 genes in SilkBase, SGID newly labeled GO IDs for 5521 genes. Besides, SGID made KEGG annotations for 16 028 genes and Entrez IDs for 16 320 genes. These are important for research, especially for gene set function enrichment analysis (Table 1).

Using peptide sequences from published experiments, we validated 2999 protein coding genes. They are of proteomics evidence. To depict one gene's function in a cell, SGID provides information on gene's subcellular localization and topology prediction. More than half (9592, 56.8%) of the silkworm genes are located in the nuclear (Figure 1), and 2878 genes (17.0%) have transmembrane regions. Furthermore, 1960 silkworm genes are predicted to have signal peptides. Encouragingly, 9844 silkworm proteins are of PDB matches with *E*-value < 0.05, which infers that more than half (58.3%) silkworm expressed proteins have structural information. External links to UniProt Proteomes, PRIDE (https://www.ebi.ac.uk/pride), Pfam, Interpro, SUPFAM (http://supfam.org), Gene 3D (http://gene3d.biochem.ucl.ac.uk), Protein Model Potal (49) and PANTHER (http://www.pantherdb.org) are also provided and they are helpful to understand the protein structure and related functions of one gene.

Table 1. A summary of the data in SGID

Item	Description	Cov.	Previous
Genome	High-quality assembly of the silkworm genome in chromosome level (7)		The same as Kawamoto et al. and Mita et al. (7, 16)
Gene models	16 880 in total, with 12 752 correlated with old gene models (4, 28)		The same as Kawamoto et al. and Mita et al. (7, 16)
Gene function annotation	15 594 genes are of function annotations	92.4%	80.3% in Kawamoto et al. and Mita et al. (7, 16)
	4937 gene feature annotations	22.0%	NA
	201 309 annotations from InterproScan	93.9%	NA
	8730 distinctive GO lists	86.9%	54.2% in Kawamoto et al. and Mita et al. (7, 16)
	16 028 correlated KEGG Gene IDs	96.4%	NA
	138 KEGG pathways	16.5%	NA
	16 320 correlated Entrez IDs	96.7%	NA
Biophysics and chemistry	2487 EC numbers	13.8%	NA
	329 biophysicochemical properties	0.6%	NA
	2445 catalytic activity annotations	12.0%	NA
	1743 cofactor information annotations	7.5%	NA
Topology	20 378 subcellular localization annotations	99.9%	NA
	2878 genes with transmembrane regions	17.0%	NA
	1960 genes with signal peptides	11.6%	NA
Proteomics and protein structure	12 394 real peptides from experiments validated 2999 protein coding genes	17.8%	NA
	9844 genes significantly correlated PDB protein structures	58.3%	NA
	1 730 892 correlated EMBL IDs	92.3%	NA
	17 762 correlated Gene3D IDs	57.2%	NA
	112 275 correlated Interpro IDs	86.9%	NA
	6257 CDD annotations	29.0%	NA
TFs	704 items	4.2%	NA
Repeat elements	571 401 segments, with 28 519 DNA transposons, 190 316 LINE, 13763 LTR and 179 435 SINE		In accordance with Osanai-Futahashi et al. (53)
Transcriptomics	306 samples from 41 projects		NA
Epigenomics	187 samples from 38 projects		NA
Populations genetics	Sliding widow analysis results based on 158 silkworm genomes		NA

'Cov.' denotes the coverage of genes with corresponding annotations in total genes. 'Previous' denotes the comparison of SGID and previous work. 'NA' denotes that related data is not available in previously built silkworm databases, such as SilkBase, Ensembl Silkworm and SilkDB.

As a domesticated insect, the silkworm is important in evolution research. Domestication genes are the genes with functions to underlie one or more domestication-related traits. They usually underwent positive selection and are of complete or near-complete fixation of causative mutations in all domestic lineages (50, 51). Totally, we identified 569 domestication gene candidates (for details, see the Materials and Methods section). Users can view and inspect theses domestication genes by a SGID tool named 'Population Genetics'. Population genetics test results are also displayed in the genome bowser, where users can do sliding window analysis of interested genomic or genic

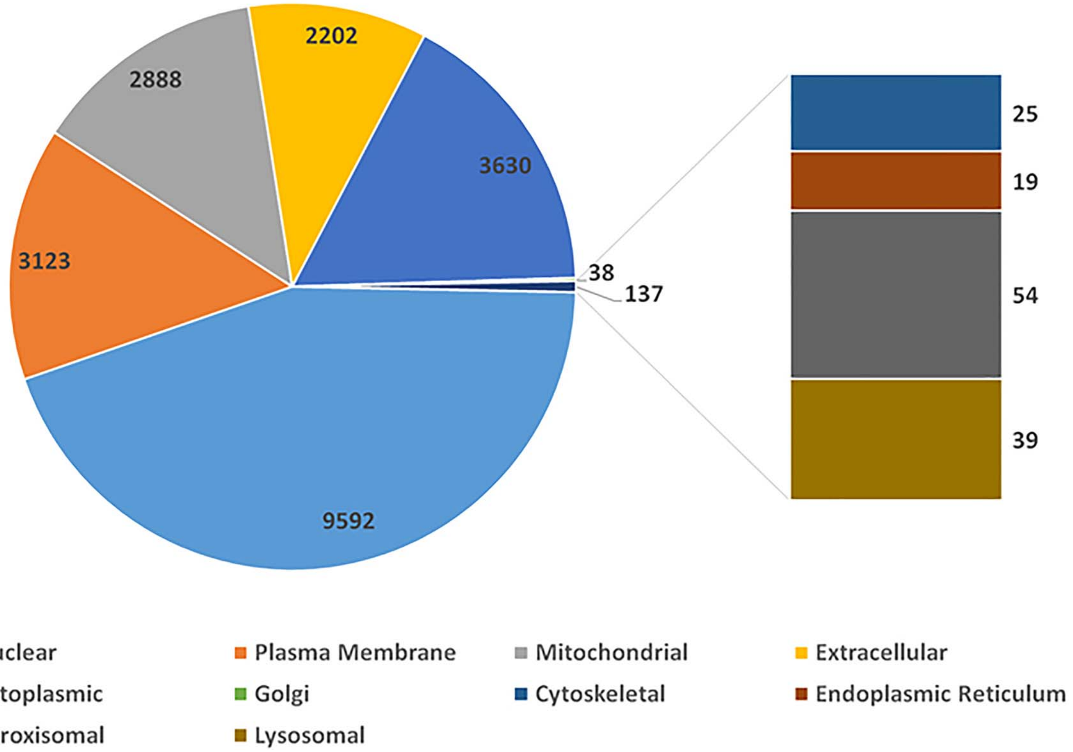

Figure 1. The distribution of silkworm genes in different cellular subunits.

segments. We also identified 81 genes possibly under balancing selection, which are evolutionary forces that cause elevated levels of nucleotide polymorphism exceeding neutral levels, maintaining multiple alleles at higher-than-expected frequencies (52).

SGID includes transcriptomic data of 41 projects and epigenomic data of 38 projects, respectively. For transcriptomes, 28 are 'DEG', 9 are 'Stage' and 4 are 'Tissue' as we described in Materials and Methods. SGID includes 704 TFs belonging to 68 TF families. It also has 571 401 repeat segments covering 27.5% of the silkworm genome, which is generally in accordance with previous records (53). There are more retrotransposons (93%) than DNA transposons (7%). For retrotransposons, most are long interspersed nuclear elements (LINE) (46%) and short interspersed nuclear elements (SINE) (44%).

The genome browser in SGID

In SGID's genome browser page (Figure 2), users can view the silkworm genes, repeat elements and population genetics test tracks subsequently. An input box and a list of buttons above the browser allow users to move, zoom in and zoom out, setting focus bar, generating figures or downloading the data of one track. A click onto a gene figure will take users to the gene detail page. Clicking on one point of some track will raise a dialog displaying the value at the point. Except for a genome browser, SGID also provides a browser to view epigenomic data. In the browser, users could view gene regulation signals at some specific genome position.

Retrieve genes' information from SGID

As a one-click type platform, SGID offers to search genes by a gene ID of SilkBase gene models 2017 or old gene models (4, 28), a gene name, a gene function or even a brief description. In the page displaying search results, there is a list of gene information buttons within each result list. With the buttons, users can jump to view gene details, a gene in genome browser, GO and pathway, gene expression, regulation elements, gene structure and population genetics analysis results in one page or in a new window (Supplementary Figure S1A). In the detail page of each gene, aside from basic annotations (such as gene name, description, subcellular location and sequences, Supplementary Figure S1B), six information groups are listed subsequently, including 'Summary', 'Ontologies', 'Topology', 'Population Genetics', 'Multiple Alignment' and 'Gene Tree'. 'Summary' mainly includes information resulted from protein sequence analysis (Supplementary Figure S1C). 'Ontologies' display a gene's annotation on GO, KEGG Function, KEGG Pathway and PANTHER (Supplementary Figure S1C). In the part of 'Topology', transmembrane regions are listed and marked

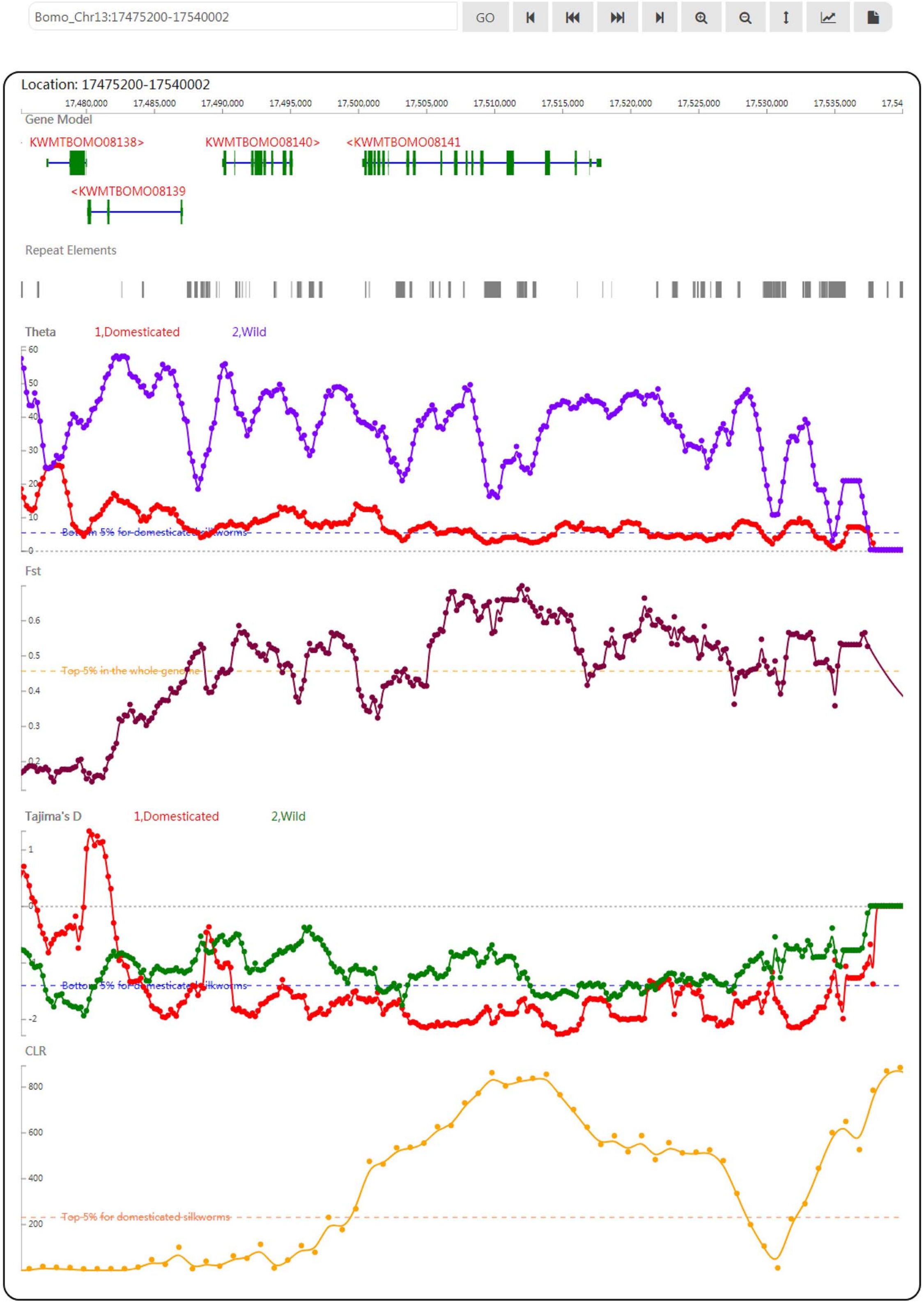

Figure 2. A snapshot of the genome browser of SGID with KWMTBOMO08141 in the center. Below are tracks of population genetics test results.

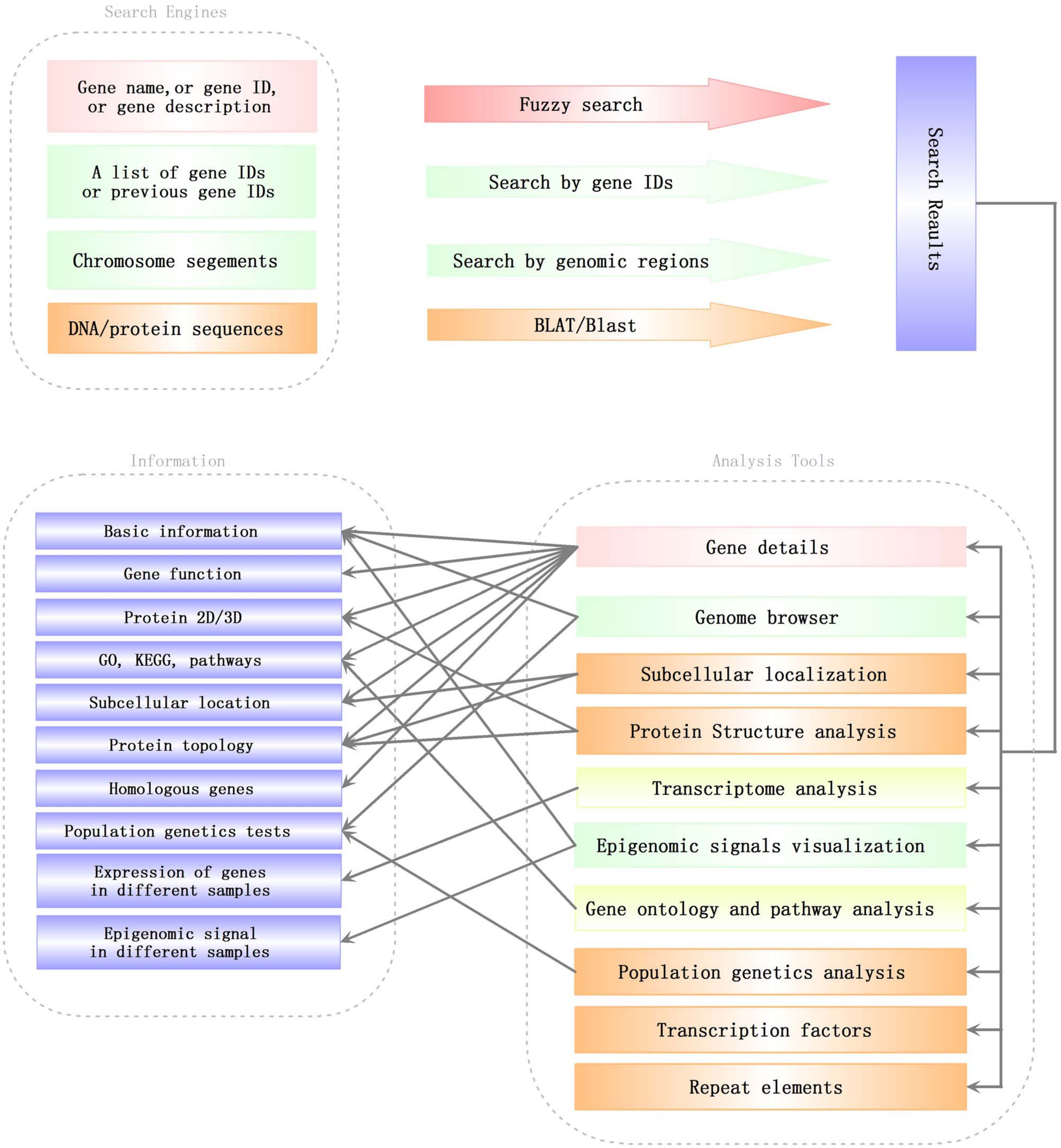

Figure 3. Search engines and analysis tools in SGID. Lines with arrows pointed out a general analysis flow in SGID.

in a diagram (Supplementary Figure S1D). If one gene's protein product is of signal peptide, the region of the signal peptide will also be listed and marked. 'Population Genetics' listed five population genetic test results (Pi, Theta, Tajima's D, CLR and CSRT) and will give an interpretation about evolutionary forces. 'Multiple Alignment' and 'Gene Tree' displayed the multiple alignments of homologous genes at protein level and the phylogenetic tree produced based on the alignment.

To facilitate users to analyze a list of genes, SGID also offers to generate a list of gene information buttons through inputting a list of gene IDs, separated by semicolons or line feeds. With the buttons, users can jump to some information view page directly like they do in search result page as referred above. Analogously, users can input a list of chromosome positions and obtain a list of genomic information links, with which users can view the genome browser or the epigenomics browser swiftly. Users can also choose to browse in gene lists to access information or to retrieve the whole data at the download page.

SGID analysis tools

To help users to visit data more quickly, we developed a list of analysis tools in SGID. As shown in the home page, ‘Gene Ontology’ is a tool to retrieve GO, KEGG or Entrez numbers using a list of gene IDs. ‘Transcriptome’ is a tool to view the expression of several genes in different experiment conditions, tissue or development stages. The results will be displayed in a heatmap figure. Stopping the mouse cursor at one cell of the heatmap will display the FKPM value of one gene at an experiment condition. The project’s name is listed at the top right and users can click it to view the project’s description. ‘Protein Structure’, ‘TF’, ‘Population Genetics’, ‘Repeat Elements’ and ‘Subcellular localization’ are interactive search tools, with which users can obtain a group of genes or items with some similar biological properties. ‘Cluster’ listed the 1064 gene clusters we identified. A summary of SGID search engines and analysis tools is shown in Figure 3.

KWMTBOMO08141, a case study

KWMTBOMO08141, BGIBMGA001085 in the previous annotation (4, 5), is a domestication gene referred in (8), Bmor_03834. Through searching in the home page, we found this gene is of other 17 functional-related members as transient receptor (Supplementary Figure S1A). Using the buttons listed below, genes in the search result page, we can obtain genes’ information one by one. In the detail page, we found this gene is of a full name ‘transient receptor potential-gamma protein’ (54–56) and an alternative name ‘Transient receptor potential cation channel gamma’. The gene is annotated to be located in plasma membrane (Supplementary Figure S1B) and function in interacting preferentially with trpl and to a lower extent with trp (Supplementary Figure S1C). Encouragingly, the protein product of this gene is validated by experiment peptides (Supplementary Figure S1B) and of significant similarity to a real protein structure 5Z96 recorded in PDB (Supplementary Figure S1C). In GO, we obtained the GO and KEGG IDs of this gene and found KWMTBOMO08141 plays roles in the pathway of phototransduction in cell membrane (Supplementary Figure S1C). In topology, this gene has six transmembrane regions (Supplementary Figure S1D), which is in accordance to its subcellular localization prediction. As a domestication gene, KWMTBOMO08141 has low Tajima’s D (−1.949945) and high CLR (629.851816). In the genome browser, we observed a CLR peak at the gene’s region (Figure 2) and a higher CLR peak at the right (Supplementary Figure S1E), indicating there may be genetic hitchhiking effects in this case. In transcription analysis, we found the expression of this gene is higher in brain than other tissues (Supplementary Figure S1F) and affected by ectopic expression of ecdysone oxidase (Supplementary Figure S1G). Through scanning this gene in the SGID epigenomics browser, we observed that epigenomic signals within the genic region disappear in some cell lines (Supplementary Figure S1H).

Conclusion

SGID is informative and user friendly. Under the idea of ‘Click-one’, SGID integrated different biological data and made them connective. SGID allows to search genes in fuzzy mode and to do analysis of more than one gene simultaneously. SGID pre-analyzed available transcriptomic data and developed a search tool to view the expression of genes in different conditions. Tools similar to SGID made the initial bioinformatics analysis of silkworm projects more efficient. With the advancement in sequencing and experiments of the silkworm, more and more data will be incorporated into SGID, making the platform to be more and more powerful.

Supplementary data

Supplementary data are available at *Database* Online.

Data availability

All SGID data are publicly and freely accessible at http://sgid.popgenetics.net. Feedback on any aspect of the SGID database and discussions of the silkworm gene annotations are welcome by email to zhuzl@cqu.edu.cn.

Author contributions

Z.L.Z. developed the web interface of the database. Z.Z.L., Z.G., G.L. and Y.W. collected and compiled the data and performed the analysis. Z.L.Z. and Z.Z. wrote the manuscript, conceived the idea and coordinated the project.

Acknowledgements

We thank Dr. Yong Zhang for insightful suggestions and Dr. Anyuan Guo for help on transcription factor prediction. We also thank Mr. Yun Wang for useful discussions.

Funding

The National Natural Science Foundation of China (31772524 to Z.Z.; 31200941 to Z.L.Z.); the Fundamental Research Funds for the Central Universities (106112016CDJXY290002).

Conflict of interest. None declared.

References

1. Gu,Z., Li,F., Hu,J. *et al.* (2017) Sublethal dose of phoxim and *Bombyx mori* nucleopolyhedrovirus interact to elevate silkworm mortality. *Pest Manag. Sci.*, **73**, 554–561.
2. Li,F., Ni,M., Zhang,H. *et al.* (2015) Expression profile analysis of silkworm P450 family genes after phoxim induction. *Pestic Biochem. Physiol.*, **122**, 103–109.
3. Meng,X., Zhu,F. and Chen,K. (2017) Silkworm: a promising model organism in life science. *J. Insect Sci.*, **17**.
4. Xia,Q., Zhou,Z., Lu,C. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science*, **306**, 1937–1940.
5. Mita,K., Kasahara,M., Sasaki,S. *et al.* (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res.*, **11**, 27–35.
6. Shao,W., Zhao,Q.Y., Wang,X.Y. *et al.* (2012) Alternative splicing and trans-splicing events revealed by analysis of the *Bombyx mori* transcriptome. *RNA*, **18**, 1395–1407.
7. Kawamoto,M., Jouraku,A., Toyoda,A. *et al.* (2019) High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.*, **107**, 53–62.
8. Xiang,H., Liu,X., Li,M. *et al.* (2018) The evolutionary road from wild moth to domestic silkworm. *Nat. Ecol. Evol.*, **2**, 1268–1279.
9. Li,B., Wang,X., Li,Z. *et al.* (2019) Transcriptome-wide analysis of N6-methyladenosine uncovers its regulatory role in gene expression in the lepidopteran *Bombyx mori*. *Insect Mol. Biol.*, **28**, 703–715.
10. Li,G., Zhou,K., Zhao,G. *et al.* (2019) Transcriptome-wide analysis of the difference of alternative splicing in susceptible and resistant silkworm strains after BmNPV infection. *3 Biotech.*, **9**, 152.
11. Gu,J., Li,Q., Chen,B. *et al.* (2019) Species identification of *Bombyx mori* and Antheraea pernyi silk via immunology and proteomics. *Sci. Rep.*, **9**, 9381.
12. Wu,P., Shang,Q., Huang,H. *et al.* (2019) Quantitative proteomics analysis provides insight into the biological role of Hsp90 in BmNPV infection in *Bombyx mori*. *J. Proteomics*, **203**, 103379.
13. Poole,R.L. (2007) The TAIR database. *Methods Mol. Biol.*, **406**, 179–212.
14. Thurmond,J., Goodman,J.L., Strelets,V.B. *et al.* (2018) FlyBase 2.0: the next generation. *Nucleic Acids Res.*, **47**, D759–D765.
15. Bult,C.J., Blake,J.A., Smith,C.L. *et al.* (2019) Mouse genome database (MGD) 2019. *Nucleic Acids Res.*, **47**, D801–D806.
16. Mita,K., Morimyo,M., Okano,K. *et al.* (2003) The construction of an EST database for *Bombyx mori* and its application. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 14121–14126.
17. Yin,C., Shen,G., Guo,D. *et al.* (2016) InsectBase: a resource for insect genomes and transcriptomes. *Nucleic Acids Res.*, **44**, D801–D807.
18. Duan,J., Li,R., Cheng,D. *et al.* (2010) SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Res.*, **38**, D453–D456.
19. Li,W. and Godzik,A. (2006) CD-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
20. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
21. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
22. Hu,H., Miao,Y.R., Jia,L.H. *et al.* (2015) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.*, **47**, D33–D38.
23. Yu,C.S., Chen,Y.C., Lu,C.H. and Hwang,J.K. (2006) Prediction of protein subcellular localization. *Proteins*, **64**, 643–651.
24. Krogh,A., Larsson,B., von,G. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
25. Jones,P., Binns,D., Chang,H.Y. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
26. Almagro Armenteros,J.J., Tsirigos,K.D., Sonderby,C.K. *et al.* (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.*, **37**, 420–423.
27. Wu,J., Mao,X., Cai,T. *et al.* (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, **34**, W720–W724.
28. Wang,J., Xia,Q., He,X. *et al.* (2005) SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Res.*, **33**, D399–D402.
29. Ghosh,S. and Chan,C.K. (2016) Analysis of RNA-Seq data using TopHat and cufflinks. *Methods Mol. Biol.*, **1374**, 339–361.
30. Langdon,W.B. (2015) Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Min.*, **8**.
31. Wang,Q., Zhang,Y., Guo,W. *et al.* (2017) Transcription analysis of cochlear development in minipigs. *Acta Otolaryngol.*, **137**, 1166–1173.
32. Chatterjee,A., Stockwell,P.A., Rodger,E.J. *et al.* (2012) Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Res.*, **40**, e79.
33. Liu,T. (2014) Use model-based analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. *Methods Mol. Biol.*, **1150**, 81–95.
34. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
35. Li,H., Handsaker,B., Wysoker,A. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
36. Watterson,G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**, 256–276.
37. Tajima,F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
38. Nielsen,R., Williamson,S., Kim,Y. *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.*, **15**, 1566–1575.

39. Korneliussen,T.S., Albrechtsen,A. and Nielsen,R. (2014) ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics*, **15**, 356.
40. DeGiorgio,M., Huber,C.D., Hubisz,M.J. *et al.* (2016) SweepFinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*, **32**, 1895–1897.
41. Pavlidis,P., Laurent,S. and Stephan,W. (2010) msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Mol. Ecol. Resour.*, **10**, 723–727.
42. Zhu,Q., Zheng,X., Luo,J. *et al.* (2007) Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol.*, **24**, 875–888.
43. Yang,S.Y., Han,M.J., Kang,L.F. *et al.* (2014) Demographic history and gene flow during silkworm domestication. *BMC Evol. Biol.*, **14**, 185.
44. Xia,Q., Guo,Y., Zhang,Z. *et al.* (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science*, **326**, 433–436.
45. Yachdav,G., Wilzbach,S., Rauscher,B. *et al.* (2016) MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics*, **32**, 3501–3503.
46. Shank,S.D., Weaver,S. and Kosakovsky Pond,S.L. (2018) phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics*, **19**, 276.
47. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
48. Johnson,M., Zaretskaya,I., Raytselis,Y. *et al.* (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.
49. Haas,J., Roth,S., Arnold,K. *et al.* (2013, 2013) The Protein Model Portal–a comprehensive resource for protein structure and model information. *Database (Oxford)*, bat031.
50. Sedivy,E.J., Wu,F. and Hanzawa,Y. (2017) Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytol.*, **214**, 539–553.
51. Meyer,R.S. and Purugganan,M.D. (2013) Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.*, **14**, 840–852.
52. Wang,B. and Mitchell-Olds,T. (2017) Balancing selection and trans-specific polymorphisms. *Genome Biol.*, **18**, 231.
53. Osanai-Futahashi,M., Suetsugu,Y., Mita,K. *et al.* (2008) Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.*, **38**, 1046–1057.
54. Woodard,G.E., Sage,S.O. and Rosado,J.A. (2007) Transient receptor potential channels and intracellular signaling. *Int. Rev. Cytol.*, **256**, 35–67.
55. Sanyal,S., Matthews,J., Bouton,D. *et al.* (2004) Deoxyribonucleic acid response element-dependent regulation of transcription by orphan nuclear receptor estrogen receptor-related receptor gamma. *Mol. Endocrinol.*, **18**, 312–325.
56. Selbie,L.A., King,N.V., Dickenson,J.M. *et al.* (1997) Role of G-protein beta gamma subunits in the augmentation of P2Y2 (P2U) receptor-stimulated responses by neuropeptide Y Y1 Gi/o-coupled receptors. *Biochem. J.*, **328**, 153–158.