Nucleic Acids Research, 2005, Vol. 33, Database issue ***D399–D402***
doi:10.1093/nar/gki116

SilkDB: a knowledgebase for silkworm biology and genomics

Jing Wang[1], Qingyou Xia[2], Ximiao He[3,4], Mingtao Dai[3], Jue Ruan[3,4], Jie Chen[3], Guo Yu[3], Haifeng Yuan[3], Yafeng Hu[3], Ruiqiang Li[3], Tao Feng[3], Chen Ye[3], Cheng Lu[2], Jun Wang[1,3,5], Songgang Li[1], Gane Ka-Shu Wong[3,6], Huanming Yang[3,5], Jian Wang[3,5], Zhonghuai Xiang[2], Zeyang Zhou[2] and Jun Yu[3,5,*]

[1]College of Life Sciences, Peking University, Beijing 100871, China, [2]The Key Sericultural Laboratory of Agricultural Ministry, Southwest Agricultural University, Chongqing, 400716, China, [3]Beijing Genomics Institute (BGI), Chinese Academy of Sciences (CAS), Beijing Airport Industrial Zone-B6, Beijing 101300, China, [4]Graduate School of the Chinese Academy of Sciences, Yuquan Road 19A, Beijing 100039, China, [5]Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou Genomics Institute, James D. Watson Institute of Genome Sciences of Zhejiang University, Hangzhou 310007, China and [6]Department of Medicine, University of Washington Genome Center, Seattle, WA 98195, USA

Received August 15, 2004; Revised and Accepted October 20, 2004

ABSTRACT

The Silkworm Knowledgebase (SilkDB) is a web-based repository for the curation, integration and study of silkworm genetic and genomic data. With the recent accomplishment of a ~6X draft genome sequence of the domestic silkworm (*Bombyx mori*), SilkDB provides an integrated representation of the large-scale, genome-wide sequence assembly, cDNAs, clusters of expressed sequence tags (ESTs), transposable elements (TEs), mutants, single nucleotide polymorphisms (SNPs) and functional annotations of genes with assignments to InterPro domains and Gene Ontology (GO) terms. SilkDB also hosts a set of ESTs from *Bombyx mandarina*, a wild progenitor of *B.mori*, and a collection of genes from other Lepidoptera. Comparative analysis results between the domestic and wild silkworm, between *B.mori* and other Lepidoptera, and between *B.mori* and the two sequenced insects, fruitfly and mosquito, are displayed by using *B.mori* genome sequence as a reference framework. Designed as a basic platform, SilkDB strives to provide a comprehensive knowledgebase about the silkworm and present the silkworm genome and related information in systematic and graphical ways for the convenience of in-depth comparative studies. SilkDB is publicly accessible at http://silkworm.genomics.org.cn.

INTRODUCTION

The silkworm (*Bombyx mori*), domesticated over the last 5000 years from a wild progenitor *Bombyx mandarina* (1), is an important source of livelihood for subsistence farmers engaged in silk production in many countries. It is believed to be a central model for Lepidopteran genomics and genetics, and second only to fruitfly (*Drosophila melanogaster*) (2) as an insect model for genetic studies (3). As many basic physiological processes of insects are conserved through evolution, study of silkworm will help further elucidate the function of gene homologs and facilitate studies of insect domestication, morphogenesis, endocrinology, reproduction, behavior and immunity.

Bombyx mori has an estimated haploid nuclear genome size of 530 Mb (4) broken into 28 chromosomes. A 3X coverage draft sequence was reported previously (5) and many expressed sequence tag (EST) sequences have been released (6). Many other resources will be generated by the International Lepidopteran Genome Project (http://www.ab.a.u-tokyo.ac.jp/lep-genome). At the Beijing Genomics Institute (BGI), the major genome sequencing center in China, we produced a ~6X coverage draft genome sequence for the silkworm *B.mori*. The silkworm genome sequence is an important

*To whom correspondence should be addressed. Tel: +86 10 80481455; Fax: +86 10 80498676; Email: junyu@genomics.org.cn
Correspondence may also be addressed to Zeyang Zhou. Tel: +86 23 68251123; Fax: +86 23 68251128; Email: zyzhou@swau.cq.cn

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

contribution to functional genomics for the silkworm and comparative and functional genomics for Lepidopteran species, and will provide a solid foundation for integrating biological information for Lepidopteran and even insects in general. In order to facilitate the usage of most up-to-date knowledge about the silkworm genome, we developed the Silkworm Knowledgebase (SilkDB) as a highly integrated information system for silkworm data storage, retrieval, visualization and analysis. The current version of SilkDB is focused on assembling contigs and anchoring contigs onto scaffolds based on mapped genetic markers and BAC-based physical maps. In the process, we have developed software packages for sequence assembly, identification and annotation of genes and transposable elements (TEs). We utilize silkworm as a framework genome to organize information for other Lepidoptera, which represent a diverse and important group of insect pests in agriculture, so as to bridge the model insect and its family members. SilkDB, together with its database, search engine and genome-oriented MapView provides both an information resource and a comparative analysis workbench for genomic research of silkworm and other insects.

DATA CONTENT AND SOURCING

Owing to the complexity and the large-scale nature of the genomic data, the strategy of comprehensive organization and effective management are of essence for successive analyses. In SilkDB, we organize the genomic data in three different modules of scaffold/contig, gene/cDNA and TE classes, and link the data of different modules through genome-oriented MapView. In scaffold module, SilkDB contains the 428.7 Mb *B.mori* genomic sequences covering 90.9% of all known silkworm genes. The raw sequences were produced by using a whole genome shotgun (WGS) (7) technique and sequence reads were assembled by using an updated version of our RePS software (8). There are 23 156 scaffolds for the 28 chromosomes. The average contig and scaffold sizes, by using N50 statistics, are 12.5 and 26.9 kb, respectively. Genomic sequences were annotated for gene content by using BGF (BGI Gene Finder) and database searches against public resources. BGF is a self-developed *ab initio* program based on GenScan (9) and FgeneSH (10), and was successfully utilized for our rice genome annotation (11). After correction of partial and erroneous predictions, the estimated gene count is 18 510, which exceeds the 13 379 genes reported for *D.melanogaster* (12). InterPro domains (13) were annotated by InterProScan Release 7.0 and functional assignments were mapped onto Gene Ontology (GO) (14). To further the study of silkworm genome biology, we investigated the biologically important genes in comparison with spider and butterfly, such as silk gland, wing patterning, development, immunity and defense, hormones and receptors, etc., which are detailed in the gene module. Besides the 18 510 annotated genes, the gene module hosts a collection of 212 known silkworm genes (with full-length cDNA sequences) from GenBank (15), 16 425 EST clusters based on our sequencing of 80 470 ESTs from different *B.mori* tissues, 554 GenBank genes of other Lepidoptera and 521 *B.mori* homologs of other Lepidopteran genes. SNPs mined from the *B.mori* EST sequences (16) were collected and mapped onto the genome. A set of *B.mandarina* ESTs was also produced, clustered and compared with the *B.mori* dataset for domestication study. Genome expansion is believed to be due to TE insertions. To explore the increase in genome size from fruitfly (116.8 Mb) (17) to silkworm (428.7 Mb), we applied Repeat Masker (http://www.repeatmasker.org/) for identifying TEs and tagging TE classes. A total of 601 225 TEs were identified, most of which are from a single *gypsy-Ty3*-like retrotransposon (18). Classes of TEs and their detailed information are stored in the third module, the TE_class module. All the data described above are available for download through our FTP site.

ACCESS AND WEB QUERY INTERFACE

A simple way for users to access data stored in the SilkDB database is through the 'Data' module, where users can get an overview of the data content, data statistics and the correlations between each data type. The provided hyperlinks facilitate users to browse the details of each data entry directly. MapView and Search Engine are two self-developed tools built on top of the database for rapid visualization and querying of the data at many levels. As an efficient visualization tool, MapView currently displays the *B.mori* genome assembly on the scaffold scale with sequence contigs aligned to, and allows users to browse a series of tracks aligned with the genomic sequence (Figure 1). Users may center the map upon a point on the scaffold of interest and expand to obtain a more detailed view of genetic markers, predicted genes, cDNAs, EST clusters of both the domestic and wild silkworm, and classes of TEs. *B.mori* gene homologs of other Lepidoptera are also marked out with distinct color-coding. Every sequence record is linked to several display options in MapView system. A text-based tabular report for each element contained in the visualization system is displayed automatically by clicking. Cross-referenced links to related database entries, such as InterPro, GO and GenBank, are also provided if available. The SilkDB search engine is the entry point for searching the major data types stored in the SilkDB. It provides two kinds of searches for users: keyword-based subject search and BLAST-based homology search, including searches for scaffolds, contigs, genes, cDNAs, classes of TEs, etc. Users can define concrete limitations to extract records that are best suited to their research needs.

SYSTEM IMPLEMENTATION

SilkDB is implemented in the Oracle9i relational database management system. The front end consists of a set of JSP scripts running on TomCat web server. A large set of Java Servlets and Javabeans mediate the user's interaction with the database. To handle the large amount of yet complex silkworm genome data, we developed a standard set of genome-based Bio-XML format that lays the foundation for our research work and allows SilkDB to accommodate the fast-accumulating data and to integrate new data types when encountered.

FUTURE DEVELOPMENTS

We are aiming at building a genomic information resource and comparative analysis workbench for silkworm with an

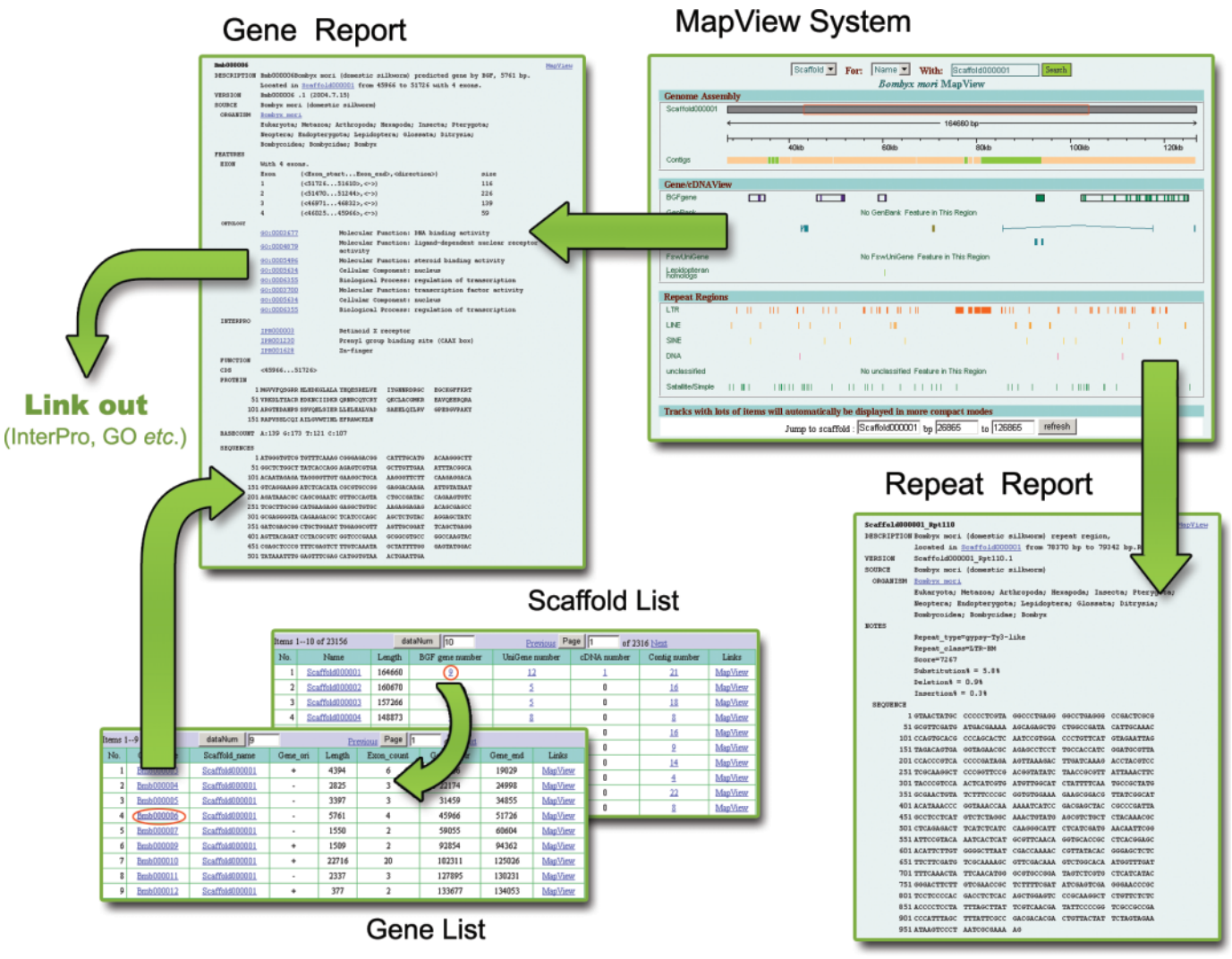

Figure 1. A screenshot of SilkDB MapView system, which displays the *B.mori* genome assembly on the scaffold scale, and allows users to browse a series of tracks of genes, cDNAs, EST clusters, homologs of other Lepidopteran genes and classes of TEs. Gene details in Gene Report can also be accessed through Gene List from the 'Data' module and the same access method can be used for cDNA/EST/ Lepidopteran homolog/TE details.

intention to expand to other Lepidoptera and model insects. Continued efforts will be made for the improvement of data quality, including anchoring scaffolds onto *B.mori* chromosomes, improving functional annotations based on phenotypically identified mutants and gene expression at the transcription and translation levels, updating EST clusters as more data are generated and annotating the clusters with respect to potential encoded protein products. Besides the timely updated silkworm genome information, SilkDB has been constantly incorporating more data, as they become available, from other Lepidoptera genomes, and different types of biological data, such as phenotype and gene expression data. Refinement of the system and addition of new applications are continuous efforts for the SilkDB project. We will introduce into SilkDB a versioning system and references around different versions. In the near future, it will be possible for users to retrieve the data of different versions, trace up and locate changes of a given entity between different versions. A key enhancement to comparative analysis will be the development of a comparative map viewer, allowing users to evaluate the alignment of conserved regions with alternative views of genome evolution. Based on the comparative map viewer, further comparative studies on genomic sequences between Lepidoptera, *D.melanogaster*, *Anopheles gambiae* (19), *Caenorhabditis elegans* (20), and other invertebrates will be conducted for the study of Lepidopteran-specific genes, many of which are potential candidates for targets of Lepidopteran-selective insecticides, and will help further our understanding of the molecular mechanism of genetic diversity among insects.

ACKNOWLEDGEMENTS

This project was funded by Chinese Academy of Sciences (KSCX2-SW-223), National Development and Reform Commission, Ministry of Science and Technology (2002AA104250; 2002AA234011; 2001AA231061; 2001AA231011; 2001AA231101; 2004AA231050; 30200163; 90208019), China National Grid

(2002AA104250) and 863 High Technology Foundation of China (2004AA2Z1020).

REFERENCES

1. Zhou,Y. (1958) *General Entomology, 2nd edn*. High Education Publication House, Beijing, China.
2. Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D., Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F. *et al*. (2000) Finishing a whole-genome shotgun: release 3 of the *Drosophila* euchromatic genome sequence. *Science*, **287**, 2185–2195.
3. Goldsmith,M.R. (1995) *Molecular Model Systems in the Lepidoptera*. Cambridge University Press, Cambridge, pp. 21–76.
4. Gage,L.P. (1974) The *Bombyx mori* genome: analysis by DNA reassociation kinetics. *Chromosoma*, **45**, 27–42.
5. Mita,K., Kasahara,M., Sasaki,S., Nagayasu,Y., Yamada,T., Kanamori,H., Namiki,N., Kitagawa,M., Yamashita,H., Yasukochi,Y. *et al*. (2004) The genome sequence of silkworm, *Bombyx mori*. *DNA Res*., **11**, 27–35.
6. Mita,K., Morimyo,M., Okano,K., Koike,Y., Nohata,J., Kawasaki,H., Kadono-Okuda,K., Yamamoto,K., Suzuki,M.G., Shimada,T. *et al*. (2003) The construction of an EST database for *Bombyx mori* and its application. *Proc. Natl Acad. Sci. USA*, **100**, 14121–14126.
7. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, **296**, 79–92.
8. Wang,J., Wong,G.K., Ni,P., Han,Y., Huang,X., Zhang,J., Ye,C., Zhang,Y., Hu,J., Zhang,K. *et al*. (2002) RePS: a sequence assembler that masks exact repeats identified from the shotgun data. *Genome Res*., **12**, 824–831.
9. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*., **268**, 78–94.
10. Salamov,A.A. and Solovyev,V.V. (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res*., **10**, 516–522.
11. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y., Ren,X. *et al*. (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res*., **32**, D377–D382.
12. Misra,S., Crosby,M.A., Mungall,C.J., Matthews,B.B., Campbell,K.S., Hradecky,P., Huang,Y., Kaminker,J.S., Millburn,G.H., Prochnik,S.E. *et al*. (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol*., **3**, RESEARCH0083.1–RESEARCH0083.22.
13. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
14. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
15. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res*, **32**, D23–D26.
16. Cheng,T.C., Xia,Q.Y., Qian,J.F., Liu,C., Lin,Y., Zha,X.F. and Xiang,Z.H. (2004) Mining single nucleotide polymorphisms from EST data of silkworm, *Bombyx mori*, inbred strain *Dazao*. *Insect Biochem. Mol. Biol*., **34**, 523–530.
17. Celniker,S.E., Wheeler,D.A., Kronmiller,B., Carlson,J.W., Halpern,A., Patel,S., Adams,M., Champe,M., Dugan,S.P., Frise,E. *et al*. (2002) Finishing a whole-genome shotgun: release 3 of the *Drosophila* euchromatic genome sequence. *Genome Biol*., **3**, RESEARCH0079. 1–RESEARCH0079.14.
18. Abe,H., Ohbayashi,F., Shimada,T., Sugasaki,T., Kawai,S., Mita,K. and Oshiki,T. (2000) Molecular structure of a novel gypsy-Ty3-like retrotransposon (Kabuki) and nested retrotransposable elements on the W chromosome of the silkworm *Bombyx mori*. *Mol. Gen. Genet*., **263**, 916–924.
19. Holt,R.A., Subramanian,G.M., Halpern,A., Sutton,G.G., Charlab,R., Nusskern,D.R., Wincker,P., Clark,A.G., Ribeiro,J.M., Wides,R. *et al*. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, **298**, 129–149.
20. The *C.elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.