Published online 17 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D801–D807***
doi: 10.1093/nar/gkv1204

InsectBase: a resource for insect genomes and transcriptomes

Chuanlin Yin[1,2,†], Gengyu Shen[3,†], Dianhao Guo[1,2], Shuping Wang[4], Xingzhou Ma[2], Huamei Xiao[2,5], Jinding Liu[6], Zan Zhang[2,7], Ying Liu[2,8], Yiqun Zhang[6], Kaixiang Yu[2], Shuiqing Huang[6] and Fei Li[1,*]

[1]Ministry of Agriculture, Key Lab of Agricultural Entomology and Institute of Insect Sciences, Zhejiang University, 866 Yuhangtang Road, Hangzhou 310058, China, [2]Department of Entomology, Nanjing Agricultural University, Nanjing 210095, China, [3]Library of Nanjing Agricultural University, Nanjing 210095, China, [4]Technical Center for Animal Plant and Food Inspection and Quarantine, Shanghai Entry-Exit Inspection and Quarantine Bureau, Shanghai 200135, China, [5]Department of City Construction, Shaoyang University, Shaoyang 422000, China, [6]College of Information Science and Technology, Nanjing Agricultural University, Nanjing, 210095 Jiangsu, China, [7]Hubei Insect Resources Utilization and Sustainable Pest Management Key Laboratory, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, Hubei, China and [8]Department of Entomology, China Agricultural University, Beijing 100193, China

Received August 14, 2015; Revised October 25, 2015; Accepted October 26, 2015

ABSTRACT

The genomes and transcriptomes of hundreds of insects have been sequenced. However, insect community lacks an integrated, up-to-date collection of insect gene data. Here, we introduce the first release of InsectBase, available online at http://www.insect-genome.com. The database encompasses 138 insect genomes, 116 insect transcriptomes, 61 insect gene sets, 36 gene families of 60 insects, 7544 miRNAs of 69 insects, 96 925 piRNAs of *Drosophila melanogaster* and *Chilo suppressalis*, 2439 lncRNA of *Nilaparvata lugens*, 22 536 pathways of 78 insects, 678 881 untranslated regions (UTR) of 84 insects and 160 905 coding sequences (CDS) of 70 insects. This release contains over 12 million sequences and provides search functionality, a BLAST server, GBrowse, insect pathway construction, a Facebook-like network for the insect community (iFacebook), and phylogenetic analysis of selected genes.

INTRODUCTION

Insects are essential to maintain agricultural ecosystems, but they are also pests that damage >30% of agricultural, forestry and livestock production and cause billions in economic losses annually. Insects are vectors of many devastating diseases leading to the loss of numerous human lives. The availability of insect genomes and transcriptomes provides valuable resources for entomological research. However, the insect community lacks an integrated, up-to-date database of gene resources.

Currently, the genomes of at least 138 insects have been sequenced and deposited in public databases such as the NCBI genome database (1), FlyBase (2), i5k Workspace@NAL (3), VectorBase (4), SilkDB (5), ButterflyBase (6), BeetleBase (7), MonarchBase (8), AphidBase (9), NasoniaBase, BeeBase and Ant Genomes Portal (10), Hessian Fly Base and Manduca Base (http://www.agripestbase.org), ChiloDB (11), DBM-DB (12), KAIKObase (13) and KONAGAbase (14). Since the cost of whole-genome sequencing has decreased dramatically, many genome-sequencing projects on insects have been initiated in recent years (Supplementary Figure S1). These projects were completed by small research groups with the technical assistance of sequencing companies. These research groups normally construct an organism-specific database for organizing and managing the genome data (8,11,12). Some organism-specific databases also contain the gene data of closely-related insect species (2,4,6,10). Though a copy of the genome Scaffolds, Contigs and Official Gene Sets (OGSs) should be submitted to the NCBI genome database because of publication requirements, it is not updated frequently and often lacks good annotation. The lack of availability of an integrated, up-to-date collection of insect genomes and transcriptomes has hampered entomological research.

Here, we introduce InsectBase (http://www.insect-genome.com/), which is intended to meet the needs of the insect community, especially for studies on molecular

*To whom correspondence should be addressed. Tel: +86-571-88982427; Fax: +86-571-88982868; Email: lifei18@zju.edu.cn
†These authors contributed equally to the paper as first authors.

Table 1. Summary of the data content of InsectBase

Category	Species	Sequences
Genome	138	1 090 915
Transcriptome	79	5 140 642
EST	235	4 108 911
Pathway	78	352 700
Ortholog	7	6811
Gene Family	60	39 105
miRNA	69	7544
mir-family	54	4637
piRNA	2	96 925
LncRNA	1	2439
Transposon	2	2880
UTR	84	679 881
CDS	74	160 905

biology, evolution, development, immunity, pest control and insecticide resistance. To the best of our knowledge, InsectBase collects almost all insect genomes and most insect transcriptomes from publicly available databases. Besides offering widely used Web-services such as a search tool, BLAST and GBrowse, InsectBase is also a platform for comparative genomics analysis on gene families, pathways and orthologs. Additionally, iFacebook is designed to provide a Facebook-like social network for the insect community by constructing relationships between researchers, genes and insect species.

MATERIALS AND METHODS

Data sources

InsectBase harvests insect gene data from tens of databases. We also developed software or pipelines to identify miRNA, piRNA, lncRNA, insect pathways, and orthologous groups from insect genomes and transcriptomes (Table 1).

Insect genomes sequences were downloaded from the NCBI genome database (1), Ensembl, VectorBase (4), FlyBase (2), Hessian Fly Base (www.agripestbase.org), AphidBase (9), Ant Genomes Portal (10), BeeBase (10), NasoniaBase (10), SilkDB (5), ChiloDB (11), Heliconius Genome project (http://www.butterflygenome.org/), Manduca Base (www.agripestbase.org) and DBM-DB (12). Stick insect genomes were downloaded from http://nosil-lab.group.shef.ac.uk/ (15) (Supplementary Table S1). This yielded a collection of 138 insect genomes (Table 2, Supplementary Figure S2). Among these, the gene annotation files were obtained for 61 insect genomes (Supplementary Table S2). However, 31 insect gene sets lacked gene annotation information. Therefore, these gene sets were annotated by BLASTP against the Swiss-Prot database to get annotations.

Protein information of 61 insects were obtained by InterProScan analysis (16), including Coils, Gene3D, Hamap, Pfam, PIRSF, PRINTS, ProDom, ProSitePatterns, ProSiteProfiles, SMART, SUPERFAMILY and TIGRFA.

Insect transcriptomes were assembled using Trinity with default parameters (17) and then annotated by BLASTX against the NCBI nr database (1) (Supplementary Table S3). The raw reads of 46 insect transcriptomes were downloaded from the NCBI SRA database (18). Seventy assembled insect transcriptomes were obtained from the NCBI TSA database (1) (Supplementary Figure S3).

Insect pathways information were obtained by analyzing transcriptome data of 78 insects using KAAS (19) and iPathCons (20).

Expressed sequence tags (ESTs) of 235 insects were downloaded from the NCBI EST database (1).

Insect orthologous were obtained by we analyzing the official gene sets (OGS) of seven insects with the software orthoMCL (21), including *Bombyx mori*, *Danaus plexippus*, *Linepithema humile*, *Nasonia vitripennis*, *Tribolium castaneum*, *Aedes aegypti*, and *Pediculus humanus*. This produced 973 1:1:1 ortholog groups.

Insect miRNA sequences were downloaded from the miRBase (22) and were also collected from the supplemental files of published references because many miRNAs were not submitted to miRBase. The conserved miRNAs of 54 insects were obtained by homology searching with RNA-seq data. After removing the redundancy, the miRNAs of 69 insects were stored in InsectBase (Supplementary Table S4). **For piRNA**, 987 piRNAs of *D. melanogaster* were downloaded from the NCBI GenBank database (GI: 157361675–157362817) and 13 299 *Drosophila* piRNAs were from the NCBI Gene Expression Omnibus with the accession number GSE9138 (1). The piRNAs of *Chilo suppressalis* were obtained by Piano prediction (23). **For long noncoding RNA (lncRNA)**, we developed a pipeline to find lncRNAs from 12 transcriptomes of *N. lugens*.

Transposons

1572 *Drosophila* transposons were downloaded from the Berkeley Drosophila Genome Project (http://www.fruitfly.org/p_disrupt/TE.html) (24) and 1308 silkworm transposons were obtained from the BmTEdb database (25).

Coding sequences (CDS) and untranslated regions (UTR)

We downloaded UTR sequences of ten insects from the UTRBase (26), including Acyrthosiphon pisum, Aedes aegypti, Anopheles gambiae, Apis mellifera, Bombyx mori, Culex quinquefasciatus, D. melanogaster, Nasonia vitripennis, Pediculus humanus, Tribolium castaneum. A pipeline was developed to predict CDS and UTR from transcriptome data, producing CDS and UTR sequences of 74 insects (Supplementary Table S5).

RESULTS

Structure of InsectBase

InsectBase provides Web services such as a search tool, BLAST, visualization with GBrowse, insect pathway construction and phylogenetic analysis. The gene information for widely-studied gene families, noncoding RNA (ncRNA), transposons, UTRs and CDSs are collected and presented in the database. Genome annotation tools optimized for insects are incorporated into InsectBase. InsectBase also includes iFacebook, a Web-based construction for gene–researcher–species networking (Figure 1).

Web services

The **search** tool can be used to find interesting information from genes, transcriptomes, pathways, gene families, and or-

Table 2. The distribution of insect genome resource

Database		Species	Genome with Gene Sets	URL
NCBI(GeneBank & Refseq)		134	42	http://www.ncbi.nlm.nih.gov
Ensembl		31	31	http://metazoa.ensembl.org
Flybase		12	12	http://flybase.org/
i5kworkspace		35	35	http://i5k.nal.usda.gov/
VectorBase		42	33	https://www.VectorBase.org/
Hymenoptera Genome Database	HymenopteraMine	17	12	http://hymenopteragenome.org/hymenopteramine
	BeeBase	1	1	http://hymenopteragenome.org/beebase/
	NasoniaBase	1	1	http://hymenopteragenome.org/nasonia
	Ant Genomes Portal	8	8	http://hymenopteragenome.org/ant_genomes
Agripestbase	Hessian Fly Base	1	1	http://agripestbase.org/hessianfly/
	Manduca Base	1	1	http://agripestbase.org/manduca/
	BeetleBase	1	1	http://beetlebase.org/
DBM-DB		1	1	http://www.iae.fafu.edu.cn/DBM/
KONAGAbase			1	http://dbm.dna.affrc.go.jp/px/
MonarchBase		1	1	http://monarchbase.umassmed.edu/
APHIDBASE		1	1	http://www.aphidbase.com/
SilkDB		1	1	http://www.silkdb.org/silkdb/
KAIKObase			1	http://sgp.dna.affrc.go.jp/KAIKObase
Heliconius Genome Project		1	1	http://butterflygenome.org/
ChiloDB		1	1	http://ento.njau.edu.cn/ChiloDB
InsectBase		138	61	http://www.insect-genome.com

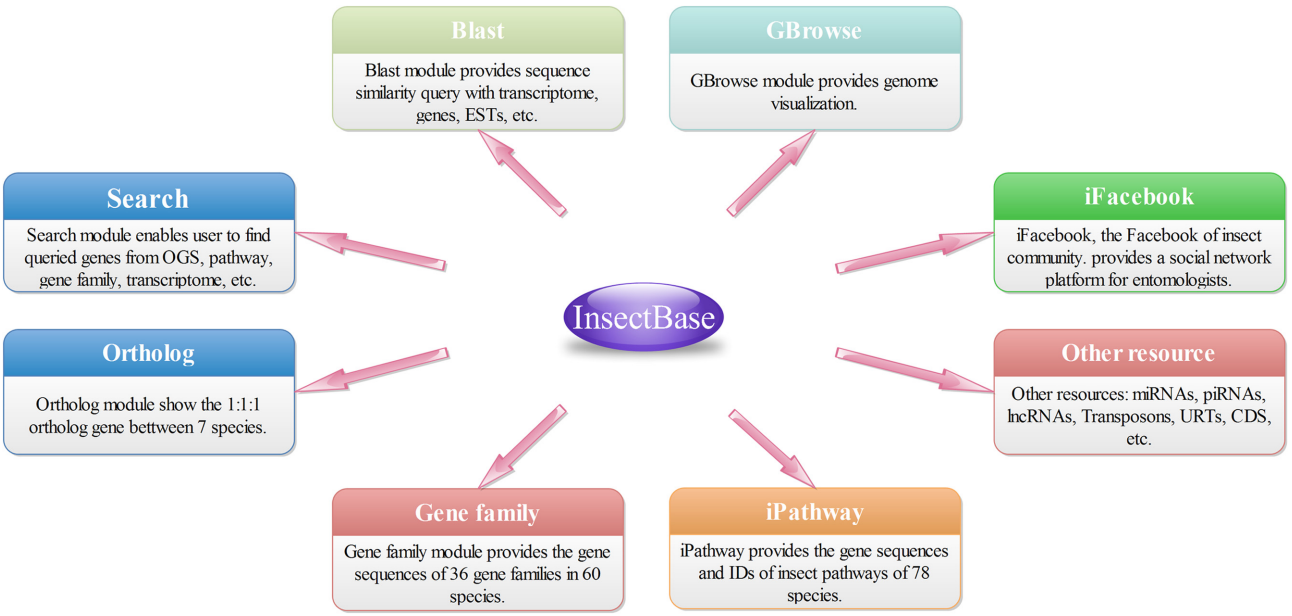

Figure 1. The structure of InsectBase. It provides Search, Blast, GBrowse, iFacebook, Ortholog, Gene family, iPathway and insect gene information.

thologs using either a gene ID or a gene name. Besides gene sequences, users can obtain related information for a gene. When a specific gene is found using a gene ID, the Swiss-Prot annotation, super family, Gene3D, Pfam, SMART, ProSite Profiles and Coils are provided. The 'Advanced' option enables users to select one or multiple species when searching.

BLAST is provided using the Web-based BLAST server 2.2.28+ (27). The data used for nucleotide BLAST (BLASTN, TBLASTN) searches contains 138 insect genomes, 61 insect OGSs and 116 insect transcriptomes. The protein data used for amino acid BLAST (BLASTP, TBLASTX, BLASTX) searches contains the 61 insect protein sequences. In the BLAST results webpage, the top five BLAST hits are presented (users can find all BLAST results by clicking 'more' under the table). InsectBase 'guesses' the gene from the BLAST results and recommends 'Your interested Gene set' by presenting Swiss-Prot annotation, KEGG, super family, Gene3D, Pfam, PRINTS and ProSitePatterns information for the gene. According to the BLAST results, InsectBase also recommends related researchers ('You might be interested in these researchers') and references ('You might be interested in these references').

GBrowse provides visualization of 58 insect genomes (28). The GBrowse tracks are customized according to the genome annotation information of various species. Three basic tracks, mRNA, CDS and exon, are provided for all insects but more tracks are provided for those insects with more annotations. For example, *A. gambiae* has 10 more tracks including tRNA, miRNA, snRNA, snoRNA, tRNA_pseudogene, rRNA, misc_RNA, RNase_p_RNA, pseudogene, and SRP_RNA. *C. suppressalis* has seven more tracks including piRNA, miRNA, repeat_sequence, ho-

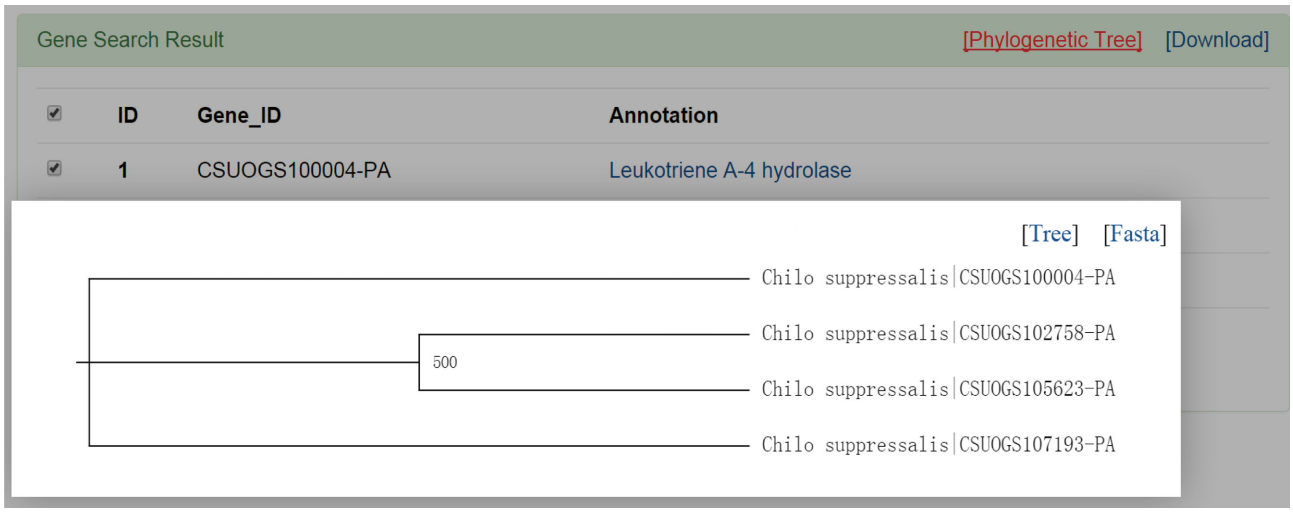

Figure 2. The server of phylogenetic analysis of selected genes were provided.

mologs in silkDB, exon structures in SilkDB, Homologs in FlyBase, exon structures in FlyBase, and similar polypeptides in ChiloDB.

Phylogenetic analysis enables users to construct an evolutionary tree with gene sequences of interest (Figure 2). The evolutionary tree is constructed with ClustalW2 (29) using the neighbor-joining clustering method and bootstrap value of 500. The tree is displayed with Newick Utilities 1.6 (30). The phylogenetic analysis function is incorporated into the search results, BLAST results, ortholog, and gene family webpages.

Insect pathway construction is indispensable for gene function analysis. InsectBase incorporates a Web-service, iPathCons, for knowledge-based construction of pathways from the transcriptomes or OGSs of genomes (20). A voting system is used for Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthology assignment. The pathways of 52 insects are used as templates. Users can select one or multiple templates to map the queried sequences to known pathways based on their requirements. When the number of sequences is less than 10, the results are displayed directly. If there are more than 10 sequences, a URL link to the iPathCons results will be sent to the user's e-mail.

iFacebook: Facebook for the insect community

A major roadblock in the insect field is the lack of efficient communication. iFacebook is intended to construct a gene–researcher–species network for entomologists (Figure 3). In total, 94,758 references of 143 insects were downloaded from the PubMed database (1). The gene names, researchers and species were extracted from these references. A gene–researcher–species network has been constructed and deposited in InsectBase. Users can find the collaborators, studied species and publications of a researcher. For a queried gene/species, InsectBase returns a result page with researchers who studied this gene/species and their publications. In this way, the first group of gene–researcher, researcher–researcher, and species–researcher relationships is constructed. InsectBase is still at primary status and we encourage the users to submit their information to improve iFacebook.

Insect gene information

Considering the research hotspot in the insect field, data mining of important gene information from all insect genomes, genes, transcriptomes and ESTs was carried out. Reference mining of important genes was also conducted. These efforts produced a batch of insect gene collections that should be of great interest to entomologists. These genes include 22 536 pathways from 78 insects, 96 925 piRNAs from *D. melanogaster* and *C. suppressalis*, 2880 transposons from *D. melanogaster* and *B. mori*, 7544 miRNAs from 69 insects, 118 miRNA families from 54 insects, 2439 long noncoding RNAs (lncRNA) of *N. lugens*, 973 orthologs groups of seven insects, 36 protein-coding gene families from 60 species, 679 881 UTRs from 84 insects and 160,905 CDSs from 74 insects. All these gene sequences are available for download and search in InsectBase.

Tools, software and insect databases

Several tools that can be used for insect genome annotation or RNA-seq analysis are collected and provided in InsectBase. Optimized Make-based Insect Genome Annotation (OMIGA) can be used to annotate insect genomes (31). iPathCons is a tool for constructing insect pathways (20). Triplet-SVM is a support vector machine (SVM)-based de novo classifier for miRNA identification (32). Piano is a SVM-based tool for piRNA annotation (23). We developed these tools and intend to provide additional Web-based services for entomologists in the future.

In the links webpage, 180 software tools are collected and presented, including genome assemblers, gene predictors, genome browsers, multiple sequence aligners, sequence analyzers, and structure modeling modules. The URL links to 18 insect genome databases are given, which enables the user to visit other useful insect gene databases.

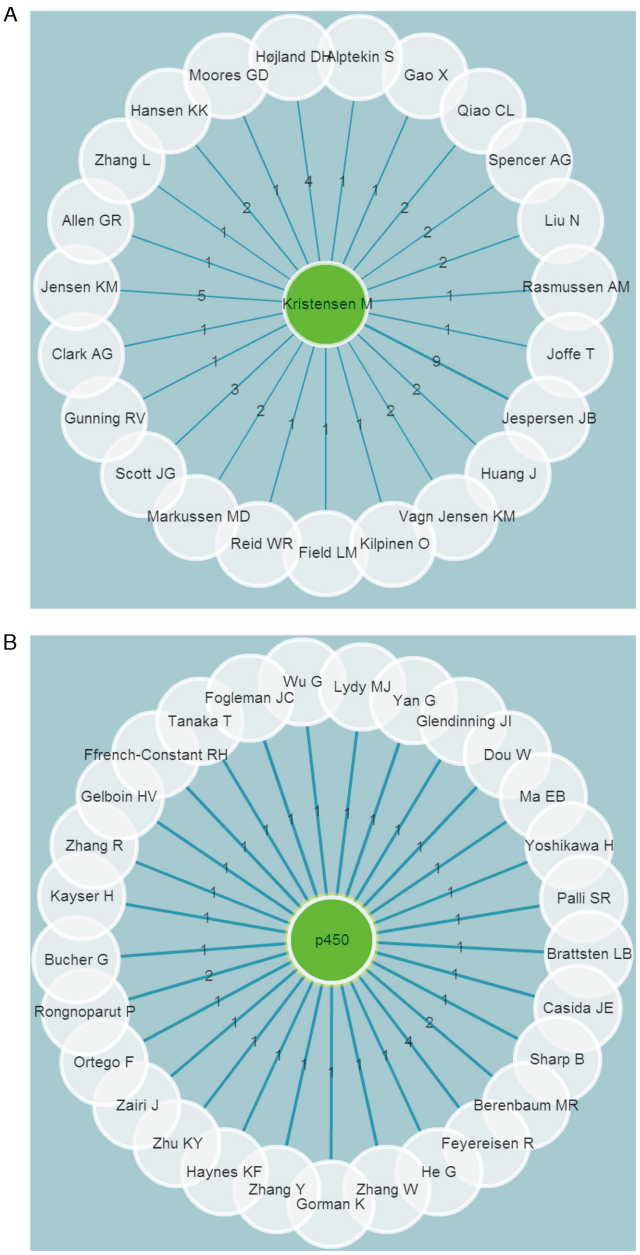

Figure 3. iFacebook provides the researcher-researcher (A) and gene-researcher (B) network.

The features of InsectBase

InsectBase is intended to provide related information for a queried gene to users, including sequences, gene features, domains, homologs, gene family, researchers, and references, etc. This 'data consumer'-oriented function enables entomologists to get an overview of a queried gene or a gene family. It should be greatly helpful for designing experiments or even finding potential collaborators.

At present, 18 insect genome databases have been constructed and reported. Most of these databases contain one or several closely related species with good annotations. InsectBase, to the best of our knowledge, has an almost complete list of insect genomes and genes. More than 100 Gb of insect gene data have been collected and deposited in InsectBase. We also provided some users-oriented software and Web servers, which could significantly facilitate the gene analysis for insect molecular biologists.

In comparison, the i5k Workspace@NAL has a wide range of arthropod genomes (3), which are mainly collected from species in the Baylor College of Medicine's i5k pilot project. At present, 42 arthropod species are included. The major goal of the i5k Workspace@NAL is to provide an efficient and well-organized platform for genome assembly, annotation, and RNA-seq mapping in a new genome-sequencing project. Data consumers can easily see the progress of the annotated genome and access the data. Therefore, the i5k Workspace@NAL is 'data producer' or 'genome-sequencer' -oriented whereas InsectBase is 'data consumer'-oriented.

Submitting data to InsectBase

We strongly encourage researchers to submit genome and transcriptome sequences. Technical assistance will be provided for uploading or handling of sequences.

Implementation

InsectBase was developed on an Apache HTTP server in a Linux (Redhat 6.5) operating system, and the database was deployed on PostgreSQL. The webpages were written using PHP, HTML language, Bootstrap, Cascading Style Sheets (CSS), the JavaScript (JS) framework and Layer JS. Perl scripts were used to make the database user-friendly with a good interaction interface. The Apache server handles queries from Web clients through PHP scripts to perform searches. Chado, a relational database schema that has been designed to handle complex representations of biological knowledge, is used to store the data (33). The generic Genome Browser (GBrowse 2.0) package, a component of the Generic Model Organism Project, was used for genome visualization (28). This tool allows researchers to obtain gene structure information. A local Basic Local Alignment Search Tool (BLAST 2.2.28+) server has been installed in the InsectBase system (27). ClustalW 2.1 has also been installed in the database for multiple alignment of nucleic acid and protein sequences (29).

FUTURE PERSPECTIVES

InsectBase is intended to provide a platform for researchers interested in analyzing insect gene data. We anticipate development in two directions: experiment-oriented and integration. First, we wish to integrate all related information for each gene. When a queried gene is searched or BLASTed, InsectBase will be designed to provide information to answer questions such as: what is this gene? Which gene family does this gene belong to? Which pathway does this gene participate in? How many homologs of this gene are there? What is the evolutionary tree of this gene and its homologs? Who has studied this gene and in what species? What are the expression patterns of this gene or its orthologs? What is the RNAi phenotype of this gene or its orthologs? We will present the results in a webpage for users. Second, we will

assign all insect genes a unique InsectBase ID, which will facilitate the use and management of rapidly accumulated insect genes. Third, we will harvest all newly published insect genomes and transcriptomes and keep the database up-to-date. Fourthly, InsectBase will pay special attentions to the pathways and ncRNA in insects. Pathway construction and ncRNA annotation will be improved.

AVAILABILITY

All data in InsectBase are available for download. InsectBase can be accessed at http://www.insect-genome.com.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Chun Liang in Miami University, Professor Fei Ma in Nanjing Normal University, Professor Junping Zhang in Fudan University and Dr Tao He in Beijing Institute of Biotechnology for their critical suggestions. Qihuang Ran and Xiaojuan Zhang provided great helps in collecting the insect photo. We appreciate the works of all the insect genome and transcriptome producers, who kindly share their data with the insect community, and the authors of all insect-related databases, who improve and organize the data.

FUNDING

National Basic Research Program of China [2013CB127600, 2012CB114102]; The Science and Technology Research Project of the Ministry of Education [V201308]; National High Technology Research and Development Program ('863'Program) of China [2012AA101505]; National Science Foundation of China [31171843, 31301691, 31260431]. Funding for open access charge: National Basic Research Program of China [2013CB127600].
Conflict of interest statement. None declared.

REFERENCES

1. NCBI Resource Coordinators. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.
2. dos Santos,G., Schroeder,A.J., Goodman,J.L., Strelets,V.B., Crosby,M.A., Thurmond,J., Emmert,D.B., Gelbart,W.M. and FlyBase,C. (2015) FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.*, **43**, D690–D697.
3. Poelchau,M., Childers,C., Moore,G., Tsavatapalli,V., Evans,J., Lee,C.Y., Lin,H., Lin,J.W. and Hackett,K. (2015) The i5k Workspace@NAL–enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Res.*, **43**, D714–D719.
4. Giraldo-Calderon,G.I., Emrich,S.J., MacCallum,R.M., Maslen,G., Dialynas,E., Topalis,P., Ho,N., Gesing,S., VectorBase,C., Madey,G. *et al.* (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res.*, **43**, D707–D713.
5. Duan,J., Li,R., Cheng,D., Fan,W., Zha,X., Cheng,T., Wu,Y., Wang,J., Mita,K., Xiang,Z. *et al.* (2010) SilkDB v2.0: a platform for silkworm (Bombyx mori) genome biology. *Nucleic Acids Res.*, **38**, D453–D456.
6. Papanicolaou,A., Gebauer-Jung,S., Blaxter,M.L., Owen McMillan,W. and Jiggins,C.D. (2008) ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Res.*, **36**, D582–D587.
7. Kim,H.S., Murphy,T., Xia,J., Caragea,D., Park,Y., Beeman,R.W., Lorenzen,M.D., Butcher,S., Manak,J.R. and Brown,S.J. (2010) BeetleBase in 2010: revisions to provide comprehensive genomic information for Tribolium castaneum. *Nucleic Acids Res.*, **38**, D437–D442.
8. Zhan,S. and Reppert,S.M. (2013) MonarchBase: the monarch butterfly genome database. *Nucleic Acids Res.*, **41**, D758–D763.
9. Legeai,F., Shigenobu,S., Gauthier,J.P., Colbourne,J., Rispe,C., Collin,O., Richards,S., Wilson,A.C., Murphy,T. and Tagu,D. (2010) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. *Insect Mol. Biol.*, **19 Suppl 2**, 5–12.
10. Munoz-Torres,M.C., Reese,J.T., Childers,C.P., Bennett,A.K., Sundaram,J.P., Childs,K.L., Anzola,J.M., Milshina,N. and Elsik,C.G. (2011) Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. *Nucleic Acids Res.*, **39**, D658–D662.
11. Yin,C., Liu,Y., Liu,J., Xiao,H., Huang,S., Lin,Y., Han,Z. and Li,F. (2014) ChiloDB: a genomic and transcriptome database for an important rice insect pest Chilo suppressalis. *Database (Oxford)*, 10.1093/database/bau065.
12. Tang,W., Yu,L., He,W., Yang,G., Ke,F., Baxter,S.W., You,S., Douglas,C.J. and You,M. (2014) DBM-DB: the diamondback moth genome database. *Database (Oxford)*, bat087.
13. Shimomura,M., Minami,H., Suetsugu,Y., Ohyanagi,H., Satoh,C., Antonio,B., Nagamura,Y., Kadono-Okuda,K., Kajiwara,H., Sezutsu,H. *et al.* (2009) KAIKObase: an integrated silkworm genome database and data mining tool. *BMC Genomics*, **10**, 486.
14. Jouraku,A., Yamamoto,K., Kuwazaki,S., Urio,M., Suetsugu,Y., Narukawa,J., Miyamoto,K., Kurita,K., Kanamori,H., Katayose,Y. *et al.* (2013) KONAGAbase: a genomic and transcriptomic database for the diamondback moth, Plutella xylostella. *BMC Genomics*, **14**, 464.
15. Soria-Carrasco,V., Gompert,Z., Comeault,A.A., Farkas,T.E., Parchman,T.L., Johnston,J.S., Buerkle,C.A., Feder,J.L., Bast,J., Schwander,T. *et al.* (2014) Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, **344**, 738–742.
16. Jones,P., Binns,D., Chang,H.Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G. *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
17. Haas,B.J., Papanicolaou,A., Yassour,M., Grabherr,M., Blood,P.D., Bowden,J., Couger,M.B., Eccles,D., Li,B., Lieber,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8**, 1494–1512.
18. Lipman,D., Flicek,P., Salzberg,S., Gerstein,M. and Knight,R. (2011) Closure of the NCBI SRA and implications for the long-term future of genomics data storage. *Genome Biol.*, **12**, 402.
19. Moriya,Y., Itoh,M., Okuda,S., Yoshizawa,A.C. and Kanehisa,M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–W185.
20. Zhang,Z., Yin,C., Liu,Y., Jie,W., Lei,W. and Li,F. (2014) iPathCons and iPathDB: an improved insect pathway construction tool and the database. *Database (Oxford)*, doi: 10.1093/database/bau105.
21. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–D368.
22. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
23. Wang,K., Liang,C., Liu,J., Xiao,H., Huang,S., Xu,J. and Li,F. (2014) Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics*, **15**, 419.
24. Kaminker,J.S., Bergman,C.M., Kronmiller,B., Carlson,J., Svirskas,R., Patel,S., Frise,E., Wheeler,D.A., Lewis,S.E., Rubin,G.M. *et al.* (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol.*, **3**, RESEARCH0084.
25. Xu,H.E., Zhang,H.H., Xia,T., Han,M.J., Shen,Y.H. and Zhang,Z. (2013) BmTEdb: a collective database of transposable elements in the silkworm genome. *Database (Oxford)*, bat055.

26. Grillo,G., Turi,A., Licciulli,F., Mignone,F., Liuni,S., Banfi,S., Gennarino,V.A., Horner,D.S., Pavesi,G., Picardi,E. *et al.* (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **38**, D75–D80.
27. Mount,D.W. (2007) Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc.*, pdb top17.
28. Stein,L.D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief Bioinform.*, **14**, 162–171.
29. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
30. Junier,T. and Zdobnov,E.M. (2010) The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics*, **26**, 1669–1670.
31. Liu,J., Xiao,H., Huang,S. and Li,F. (2014) OMIGA: Optimized Maker-Based Insect Genome Annotation. *Mol. Genet. Genomics*, **289**, 567–573.
32. Xue,C., Li,F., He,T., Liu,G.P., Li,Y. and Zhang,X. (2005) Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, **6**, 310.
33. Guignon,V., Droc,G., Alaux,M., Baurens,F.C., Garsmeur,O., Poiron,C., Carver,T., Rouard,M. and Bocs,S. (2012) Chado controller: advanced annotation management with a community annotation system. *Bioinformatics*, **28**, 1054–1056.