*D1040–D1045* *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 18 November 2021*
*https://doi.org/10.1093/nar/gkab1090*

# InsectBase 2.0: a comprehensive gene resource for insects

**Yang Mei, Dong Jing, Shenyang Tang, Xi Chen, Hao Chen, Haonan Duanmu, Yuyang Cong, Mengyao Chen, Xinhai Ye, Hang Zhou, Kang He and Fei Li***

State Key Laboratory of Rice Biology & Ministry of Agricultural and Rural Affairs Key Laboratory of Molecular Biology of Crop Pathogens and Insect Pests, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China

Received September 04, 2021; Revised October 18, 2021; Editorial Decision October 19, 2021; Accepted November 08, 2021

## ABSTRACT

**Insects are the largest group of animals on the planet and have a huge impact on human life by providing resources, transmitting diseases, and damaging agricultural crop production. Recently, a large amount of insect genome and gene data has been generated. A comprehensive database is highly desirable for managing, sharing, and mining these resources. Here, we present an updated database, InsectBase 2.0 (http://v2.insect-genome.com/), covering 815 insect genomes, 25 805 transcriptomes and >16 million genes, including 15 045 111 coding sequences, 3 436 022 3′UTRs, 4 345 664 5′UTRs, 112 162 miRNAs and 1 293 430 lncRNAs. In addition, we used an in-house standard pipeline to annotate 1 434 653 genes belonging to 164 gene families; 215 986 potential horizontally transferred genes; and 419 KEGG pathways. Web services such as BLAST, JBrowse2 and Synteny Viewer are provided for searching and visualization. InsectBase 2.0 serves as a valuable platform for entomologists and researchers in the related communities of animal evolution and invertebrate comparative genomics.**

## INTRODUCTION

Insects represent one of the largest and most diverse group of animals on earth and play important roles in ecological stability (1), agriculture (2), the economy (3) and human health (4). With rapid technological developments, a sea of insect gene data has been generated, including genomes, transcriptomes, proteomes, metabolomes and chromatin interaction information detected by the Hi-C method (5,6). Although most of these data are available in public databases such as National Center for Biotechnology Information (NCBI) (7), many are not well organized, and some are available only as raw data without annotation information. This hampers the full use of these insect gene resources. Several databases have been constructed to provide well-curated annotations and well-designed data organization in entomological field, including i5k Workspace@NAL (8), Bioinformatics Platform for Agroecosystem Arthropods (BIPAA) (https://bipaa.genouest.org/is/), VectorBase (9), FlyBase (10), LepBase (http://lepbase.org/), Hymenoptera Genome Database (11), Butterfly Genome Database (12), FireflyBase (13), SilkDB (14), KAIKObase (15), KONAGAbase (16), MonarchBase (17), LocustBase (18), BeetleBase (19), etc. Most of these databases focus on only one species or a group of closely related species, and few provide a well-designed and user-friendly platform for curating, visualizing, and sharing insect gene data. To fill this gap, we built InsectBase in 2016, which collected almost all insect genome data available at that time (20).

Due to the emergence of third-generation sequencing technology, the quantity and quality of insect gene and genome data have greatly increased in recent years. Therefore, to provide a revised and more convenient platform, we have updated InsectBase to version 2.0 with three significant improvements: (i) The quantity and quality of insect gene data are significantly increased. In total, InsectBase 2.0 contains >16 million sequences from 815 species with 207 chromosome-level genomes and 134 full-length transcriptomes. (ii) Multi-level gene and genome data are now provided, including RNA–RNA interactions, gene families, KEGG pathways and HGT genes (21). (iii) The user interface features have been enhanced to improve the web server.

## MATERIALS AND METHODS

### Data source

We collected insect gene and genome data from several databases (as described below) and developed standardized pipelines for annotation and identification of UTRs, miRNAs, lncRNAs, RNA–RNA interactions, gene families, KEGG pathways and genes likely derived from horizontal gene transfer (referred to as 'potential HGT genes').

*To whom correspondence should be addressed. Tel: +86 571 88982679; Fax: +86 571 88982679; Email: lifei18@zju.edu.cn

*Genome.* We collected and downloaded 815 genomes from NCBI (7), BIPAA (https://bipaa.genouest.org/is/), GigaDB (22), i5k Workspace@NAL (8), InsectBase (20), LepBase (http://lepbase.org/), VectorBase (9), National Genomics Data Center (NGDC) (23), FireflyBase (13), DNA Data Bank of Japan (DDBJ) (24), SilkDB 3.0 (14), Assembled Searchable Giant Arthropod Read Database (ASGARD) (25), DNA Zoo (26), LocustBase (18), DRYAD (https://datadryad.org/stash) and Zenodo (https://zenodo.org/) (Supplementary Tables S1 and S2). Among these, 231 insect genomes were obtained with known annotated official gene sets. A further 482 genomes were annotated using our in-house genome annotation pipeline. First, we identified and masked the repeat sequences by RepeatModeler2 (v.2.0.1) (27) and RepeatMasker (http://www.repeatmasker.org) (v.4.0.7) with both *de novo* and homology-based methods. Next, three evidences of gene annotation were generated. BRAKER2 (v.2.1.5) (28–34) was used to generate the *de novo* gene models. HISAT2 (v.2.1.0) (35) and StringTie2 (v.2.1.5) (36) were used for transcripts assembling. And homology-based evidence was generated by GenomeThreader (v.1.7.1) (37). Finally, we integrated three types of evidences by EVidenceModeler (v.1.1.1) (38) to obtain the official gene sets (OGS).

*Transcriptome.* 25 805 transcriptomes of 439 species were downloaded from the NCBI SRA database (Supplementary Table S3) (7). The raw reads were pre-processed using fastp (v.0.21) (39) and mapped to reference genomes with HISAT2 (v.2.1.0) (35). StringTie2 (v.2.1.4) (36) was used for transcript assembly.

*ncRNA.* 1674 small RNA libraries of 60 species were download from the NCBI SRA database (Supplementary Table S4) (7). miRNAs were predicted by miRDeep2 (v.0.1.3) (40) and MapMi (v.1.5.0) (41). TargetScan 70 (42), RNAhybrid (v.2.1.2) (43) and miRanda (v.3.3a) (44) were used for miRNA target prediction. LncRNAs and partner genes were predicted with FEELnc (v.0.2) (45) using the default parameters.

*Gene family, KEGG pathway and potential HGT gene.* One hundred and sixty-four gene families were annotated by BLASTP against the Swiss-Prot protein database using DIAMOND (v.2.0.0.138) (31,46). For KEGG pathway, the reference KOs of each gene were identified by BLASTP against the KEGG database, and the KEGG pathway genes were obtained by extracting the KO information of each gene (21). Potential HGT genes were filtered by using insect genes to blast against the NCBI non-redundant protein (nr)/nucleotide (nt) database, if at least 15 of the best 20 BLAST hits are from non-insect species, we treated these genes as potential HGT genes (7). It should be noted that this pre-filtering method might have high false positive and further analysis of these genes should consider this.

*Insect virus.* Genome information of 1524 insect viruses was obtained and organized from the NCBI genome database (7).

**Table 1.** Data summary of InsectBase 1.0 and 2.0

| Feature | Units | v1.0 | v2.0 | Fold Increase |
|---|---|---|---|---|
| Genomes | Species | 138 | 815 | 5.9 |
| Transcriptomes | Runs | 116 | 25 805 | 222.4 |
| Coding sequences | Transcripts | 160 905 | 15 045 111 | 93.5 |
| UTRs | - | 678 881 | 7 781 686 | 11.4 |
| miRNAs | - | 7544 | 112 162 | 14.9 |
| lncRNAs | - | 2439 | 1 293 430 | 530.3 |
| Pathways | - | 78 | 419 | 5.4 |
| Gene families | - | 54 | 164 | 3.0 |
| HGT genes | - | - | 215 986 | New |
| Insect viruses | - | - | 1524 | New |
| miRNA–mRNA interactions | - | - | 197 533 | New |
| lncRNA–mRNA interactions | - | - | 5 147 543 | New |

### Implementation of database

InsectBase 2.0 runs on a nginx (v.1.16.1) web server (http://nginx.org/) based on the CentOS 7.4.1708 platform with a MySQL (v.5.7.17) database (https://dev.mysql.com/). Django (v.3.1.3) framework (https://www.djangoproject.com/) and Vue (v.3.0) JavaScript framework (https://v3.vuejs.org/) were used for the web construction. JBrowse2 (47), the platform for visualizing and integrating biological data, was used for genome visualization. DIAMOND (v.2.0.0.138) (31), NCBI BLAST (v.2.11.0+) and BLAT (v.36) (48) were installed for sequence alignment of genes, proteins, miRNAs and lncRNAs. SynVisio (49) was hosted for visualization of genome synteny files constructed by MCScanX (50).

## UPDATES IN INSECTBASE 2.0

### More insect gene data with high assembly quality and standard annotations

Recent advances in third-generation sequencing techniques and chromosome conformation capture (3C) methods have provided a valuable platform for generation of high-quality genomes and full-length transcriptomes (5,6). In InsectBase 2.0, we collected 815 genomes from 457 genera and 25 805 well-assembled transcriptomes from 439 species. Among these, 207 genomes were assembled at the chromosome level and 134 full-length transcriptomes from 31 species were generated by nanopore sequencing.

Using an in-house pipeline, we annotated 482 insect genomes, yielding standard official gene sets for these species. In total, we generated 15 045 111 coding sequences of 713 insects, 112 162 miRNAs from 807 insects, 1 293 430 lncRNAs representing 376 insects, 419 KEGG pathways, 7 781 686 UTRs in 374 insects and 164 gene families in 713 insects. Overall, this represents a substantial increase in insect gene and genome data from InsectBase 1.0 (Table 1).

### ncRNAs, HGT genes and insect viruses

ncRNAs participate in many important biological processes by interacting with RNAs either directly or indirectly through protein intermediates (51). Here, we predicted 197 533 miRNA–mRNA interactions and identified

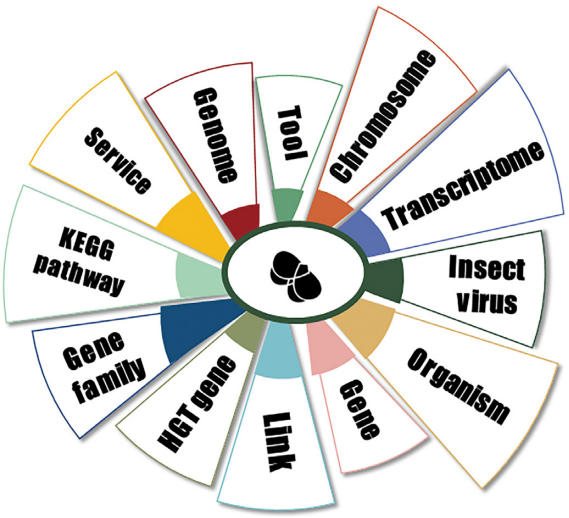


**Figure 1.** Main modules of InsectBase 2.0. It provides information about an organism, genome, transcriptome, chromosome, gene information about protein coding genes, miRNA and lncRNA, gene family, HGT genes, insect pathways, insect viruses, online tools, links and additional services.

1 293 737 lncRNA partner genes. HGT is a key evolutionary force which has constantly reshaped genomes throughout evolution (52). We identified 215 986 potential HGT genes from five kingdoms (Bacteria, Fungi, Metazoa [excluding insecta], Viridiplantae and Virus; Table 1). We also collected 1524 insect viruses which are important pathogens of many arthropod species and are potential microbial control agents. These data will benefit researches in the fields of gene networks, evolution and comparative analysis.

### Enhanced user interface features

InsectBase 2.0 contains 12 modules, namely 'organism', 'chromosome', 'genome', 'transcriptome', 'gene', 'gene family', 'HGT gene', 'KEGG pathway', 'insect virus', 'tools', 'links' and 'service' (for searching, browsing, and downloading) (Figure 1).

The 'organism' module shows a species tree modified from the NCBI Taxonomy common tree (https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) (7). Each order, family, genus, and species are introduced with pictures from public sources such as Wikipedia (https://www.wikipedia.org/) and iNaturalist (https://www.inaturalist.org/). Users can click on the species name to access the species page, which shows information about multiple aspects of the selected species. This includes a basic introduction, genome statistics, gene information, and related publications in PubMed (https://pubmed.ncbi.nlm.nih.gov/) (Figure 2A).

The advent of high-quality genome has greatly advanced the study of entomology. To aid in investigate of chromosome evolution, the 'chromosome' module displays 207 genomes with information at the chromosome level (Figure 2B). Chromosomes in 155 genomes are displayed for browsing and downloading.

Transcriptomes are an essential data resource for understanding biological processes under different conditions. The 'transcriptome' module contains 25 805 assembled transcriptomes with sample information, including species, gender, tissue, stage, and condition to help researchers conduct genetic investigations with different conditions or treatments.

The 'gene information' module allows the user to conduct an advanced search for protein coding genes, miRNAs, and lncRNAs by species, gene name, and gene description. Beyond the basic information of selected gene, gene structure, gene sequence and gene interactions such as mRNA–miRNA and mRNA–lncRNA interactions are displayed. By clicking on the interacting genes, users can access the related gene page (Figure 2C).

Gene families often exhibit apparent expansion or contraction in terms of gene numbers or structures. Gene family analysis is not only essential for uncovering gene functions, but also frequently used in revealing the evolutional mechanism of gene gain and loss. Hence, InsectBase 2.0 analysed 164 gene families by annotating them with an in-house pipeline. The 'gene family' module allows users to easily search and download gene families of interest in a given species. In addition to conventional tools such as DIAMOND (31), BLAT (48) and BLAST, we constructed a comprehensive genome browser with all annotated genomes by JBrowse2 (47) (Figure 2D). Moreover, InsectBase 2.0 provides a genome synteny visualization tool. Genome synteny between 155 chromosome-level genomes is visualized for chromosome evolution analysis (Figure 2E).

## DISCUSSION AND FUTURE DEVELOPMENT

At present, insect genome and gene data are stored in multiple databases once they are generated (53). InsectBase 2.0 uses standard pipelines to predict protein coding genes, miRNAs, lncRNAs and UTRs, promoting standardisation of comparative genomics. In addition, gene families, KEGG pathways and genes potentially involved in many crucial biological processes (such as pesticide detoxification metabolism and host-seeking) are annotated. In summary, InsectBase 2.0 is a substantially improved database for insect gene resources and serves as a valuable resource to meet the needs of entomologists and the related research communities of animal evolution and invertebrate comparative genomics.

We will continue to add newly-available data and new features. For example, the three-dimensional (3D) organization of genomes plays an essential role in gene regulation. With the development of the 3C technique, such as Hi-C, ChIA-PET, Capture-C and Capture Hi-C, chromosome interaction information has provided an unprecedented opportunity to study spatial organization in a genome-wide fashion (54). We plan to analyse these data and add associated features in the next update. The recently-developed AlphaFold2 (55) predicts protein structure with high accuracy, which would be greatly valuable in investigating protein-protein binding, enzyme active sites, and the functional implications of genetic mutations. We thus plan to integrate this tool in the next update.

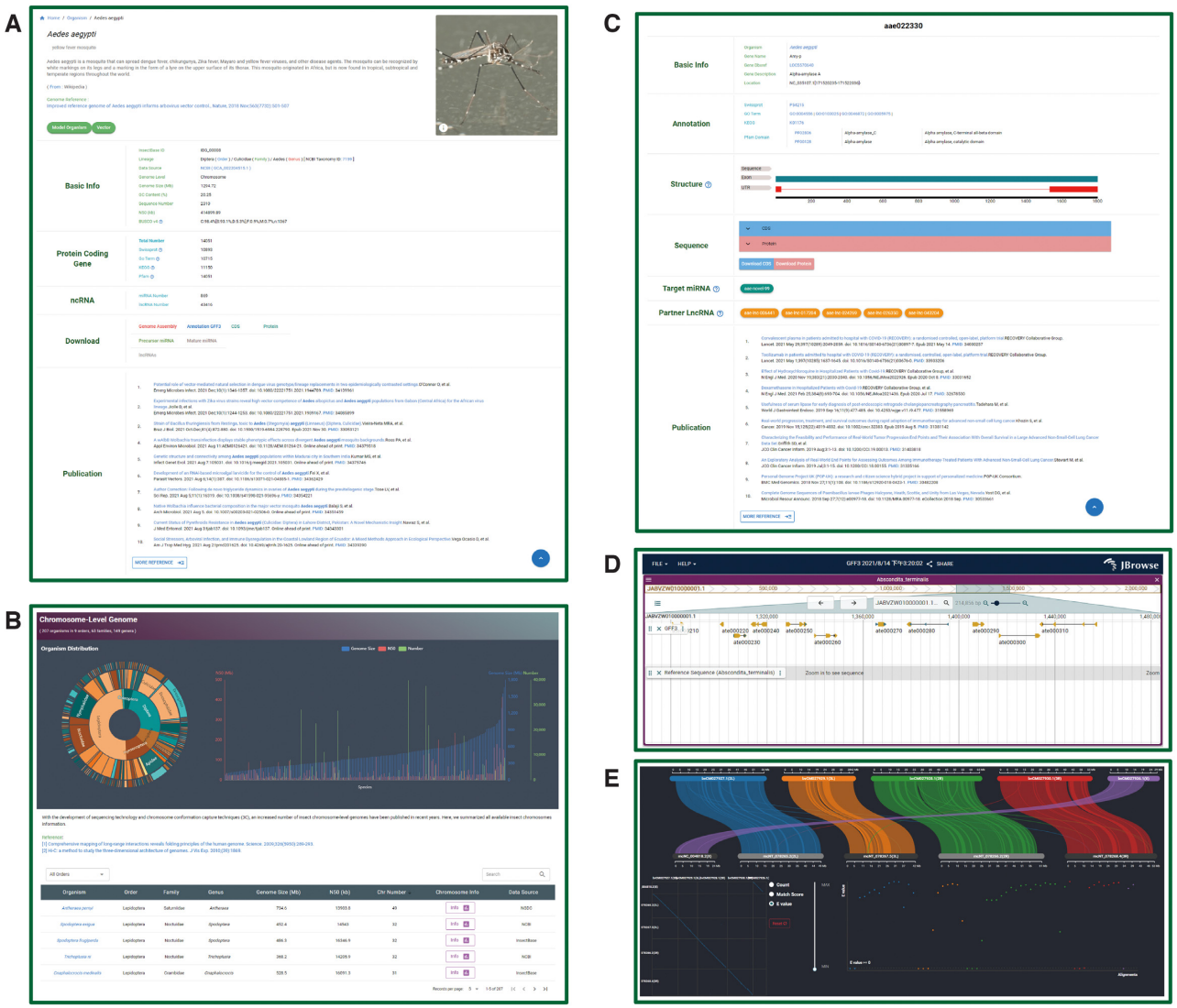


**Figure 2.** Enhanced user interface features of InsectBase 2.0. (**A**) Species: basic information, file downloading and publications related to each species. (**B**) Chromosome: information for each chromosome. (**C**) Protein coding gene: detailed information about each protein coding gene. (**D**) JBrowse2: genome browser of each annotated genome. (**E**) Genome synteny: visualization of synteny of 155 chromosome-level genomes.

## DATA AVAILABILITY

All data in InsectBase 2.0 are available for downloading. The database can be accessed at http://v2.insect-genome.com/. The genome annotation pipeline is available at https://github.com/meiyang12/Genome-annotation-pipeline.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We appreciate all researchers sharing public resources in the insect community, and constructing and organizing the databases. We kindly thank Wikipedia, iNaturalist (https://www.inaturalist.org/), the British Natural History Museum (https://www.nhm.ac.uk/), BOLDSYSTEMS (http://www.boldsystems.org/) and EHIME-Fly (https://kyotofly.kit.jp/cgi-bin/ehime/index.cgi) for publicly-available information and images.

## FUNDING

National High Technology Research and Development Program of China [2019YFD1002100]; National Science Foundation of China [31972354]; National Science & Technology Fundamental Resources Investigation Program of China [2019FY100400]; Zhejiang National Science Foundation of China [LZ18C060001]; Fundamental Research Funds for the Central Universities [2020QNA6024]. Funding for open access charge: National High Technology Research and Development Program of China [2019YFD1002100]; National Science Foundation of China [31972354]; National Science & Technology Fundamental Resources Investigation Program of China [2019FY100400]; Zhejiang National Science Foundation of

China [LZ18C060001]; The Fundamental Research Funds for the Central Universities [2020QNA6024].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Losey,J.E. and Vaughan,M. (2006) The economic value of ecological services provided by insects. *Bioscience*, **56**, 311–323.
2. Meier,R. and Lim,G.S. (2009) Conflict, convergent evolution, and the relative importance of immature and adult characters in endopterygote phylogenetics. *Annu. Rev. Entomol.*, **54**, 85–104.
3. Robinson,G.E., Hackett,K.J., Purcell-Miramontes,M., Brown,S.J., Evans,J.D., Goldsmith,M.R., Lawson,D., Okamuro,J., Robertson,H.M. and Schneider,D.J. (2011) Creating a buzz about insect genomes. *Science*, **331**, 1386.
4. Lounibos,L.P. (2002) Invasions by insect vectors of human disease. *Annu. Rev. Entomol.*, **47**, 233–266.
5. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
6. van Berkum,N.L., Lieberman-Aiden,E., Williams,L., Imakaev,M., Gnirke,A., Mirny,L.A., Dekker,J. and Lander,E.S. (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.*, **39**, 1869.
7. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W. *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic. Acids. Res.*, **49**, D10–D17.
8. 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.
9. Giraldo-Calderón,G.I., Emrich,S.J., MacCallum,R.M., Maslen,G., Dialynas,E., Topalis,P., Ho,N., Gesing,S., VectorBase,Consortium, 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.
10. Larkin,A., Marygold,S.J., Antonazzo,G., Attrill,H., Dos-Santos,G., Garapati,P.V., Goodman,J.L., Gramates,L.S., Millburn,G., Strelets,V.B. *et al.* (2021) FlyBase: updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Res.*, **49**, D899–D907.
11. Elsik,C.G., Tayal,A., Diesh,C.M., Unni,D.R., Emery,M.L., Nguyen,H.N. and Hagen,D.E. (2016) Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. *Nucleic Acids Res.*, **44**, D793–D800.
12. Davey,J.W., Chouteau,M., Barker,S.L., Maroja,L., Baxter,S.W., Simpson,F., Merrill,R.M., Joron,M., Mallet,J., Dasmahapatra,K.K. *et al.* (2016) Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. *G3 (Bethesda)*, **6**, 695–708.
13. Fallon,T.R., Lower,S.E., Chang,C.H., Bessho-Uehara,M., Martin,G.J., Bewick,A.J., Behringer,M., Debat,H.J., Wong,I., Day,J.C. *et al.* (2018) Firefly genomes illuminate parallel origins of bioluminescence in beetles. *eLife*, **7**, e36495.
14. Lu,F., Wei,Z., Luo,Y., Guo,H., Zhang,G., Xia,Q. and Wang,Y. (2019) SilkDB 3.0: visualizing and exploring multiple levels of data for silkworm. *Nucleic Acids Res.*, **48**, D749–D755.
15. Yang,C., Yokoi,K., Yamamoto,K. and Jouraku,A. (2021) An update of KAIKObase, the silkworm genome database. *Database (Oxford)*, **2021**, baaa099.
16. 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.
17. Zhan,S. and Reppert,S.M. (2013) MonarchBase: the monarch butterfly genome database. *Nucleic Acids Res.*, **41**, D758–D763.
18. Wang,X., Fang,X., Yang,P., Jiang,X., Jiang,F., Zhao,D., Li,B., Cui,F., Wei,J., Ma,C. *et al.* (2014) The locust genome provides insight into swarm formation and long-distance flight. *Nat. Commun.*, **5**, 2957.
19. 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.
20. Yin,C., Shen,G., Guo,D., Wang,S., Ma,X., Xiao,H., Liu,J., Zhang,Z., Liu,Y., Zhang,Y. *et al.* (2016) InsectBase: a resource for insect genomes and transcriptomes. *Nucleic Acids Res.*, **44**, D801–D807.
21. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
22. Xiao,S.Z., Armit,C., Edmunds,S., Goodman,L., Li,P., Tuli,M.A. and Hunter,C.I. (2019) Increased interactivity and improvements to the GigaScience database, GigaDB. *Database (Oxford)*, **2019**, baz016.
23. National Genomics Data Center Members and Partners. (2020) Database resources of the National Genomics Data Center in 2020. Nucleic Acids Res., **48**, D24–D33.
24. Fukuda,A., Kodama,Y., Mashima,J., Fujisawa,T. and Ogasawara,O. (2021) DDBJ update: streamlining submission and access of human data. *Nucleic Acids Res.*, **49**, D71–D75.
25. Zeng,V. and Extavour,C.G. (2012) ASGARD: an open-access database of annotated transcriptomes for emerging model arthropod species. *Database (Oxford)*, **2012**, bas048.
26. Dudchenko,O., Batra,S.S., Omer,A.D., Nyquist,S.K., Hoeger,M., Durand,N.C., Shamim,M.S., Machol,I., Lander,E.S., Aiden,A.P. *et al.* (2017) De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, **356**, 92–95.
27. Flynn,J.M., Hubley,R., Goubert,C., Rosen,J., Clark,A.G., Feschotte,C. and Smit,A.F. (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 9451–9457.
28. Brůna,T., Hoff,K.J., Lomsadze,A., Stanke,M. and Borodovsky,M. (2021) BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom. Bioinform.*, **3**, lqaa108.
29. Hoff,K.J., Lomsadze,A., Borodovsky,M. and Stanke,M. (2019) Whole-genome annotation with BRAKER. *Methods Mol. Biol.*, **1962**, 65–95.
30. Brůna,T., Lomsadze,A. and Borodovsky,M. (2020) GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom. Bioinform.*, **2**, lqaa026.
31. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
32. Gotoh,O. (2008) A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res.*, **36**, 2630–2638.
33. Iwata,H. and Gotoh,O. (2012) Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Res.*, **40**, e161.
34. Stanke,M., Diekhans,M., Baertsch,R. and Haussler,D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.
35. Zhang,Y., Park,C., Bennett,C., Thornton,M. and Kim,D. (2021) Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Res.*, **31**, 1290–1295.
36. Kovaka,S., Zimin,A.V., Pertea,G.M., Razaghi,R., Salzberg,S.L. and Pertea,M. (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.*, **20**, 278.
37. Gremme,G., Brendel,V., Sparks,M.E. and Kurtz,S. (2005) Engineering a software tool for gene structure prediction in higher organisms. *Inf. Softw. Technol.*, **47**, 965–978.
38. Haas,B.J., Salzberg,S.L., Zhu,W., Pertea,M., Allen,J.E., Orvis,J., White,O., Buell,C.R. and Wortman,J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.*, **9**, R7.
39. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
40. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
41. Guerra-Assunção,J.A. and Enright,A.J. (2010) MapMi: automated mapping of microRNA loci. *BMC Bioinformatics*, **11**, 133.

42. Lewis,B.P., Shih,I.H., Jones-Rhoades,M.W., Bartel,D.P. and Burge,C.B. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
43. Krüger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.
44. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1.
45. Wucher,V., Legeai,F., Hédan,B., Rizk,G., Lagoutte,L., Leeb,T., Jagannathan,V., Cadieu,E., David,A., Lohi,H. *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.*, **45**, e57.
46. UniProt Consortium. (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
47. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
48. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
49. Bandi,V. and Gutwin,C. (2020) Interactive exploration of genomic conservation. *Proc. Graph. Interface*, 74–83.
50. Wang,Y., Tang,H., Debarry,J.D., Tan,X., Li,J., Wang,X., Lee,T., Jin,H., Marler,B., Guo,H. *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.*, **40**, e49.
51. Engreitz,J.M., Sirokman,K., McDonel,P., Shishkin,A., Surka,C., Russell,P., Grossman,S.R., Chow,A.Y., Guttman,M. and Lander,E.S. (2014) RNA-RNA Interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell*, **159**, 188–199.
52. Husnik,F. and McCutcheon,J.P. (2018) Functional horizontal gene transfer from bacteria to eukaryotes. *Nat. Rev. Microbiol.*, **16**, 67–79.
53. Li,F., Zhao,X., Li,M., He,K., Huang,C., Zhou,Y., Li,Z. and Walters,J.R. (2019) Insect genomes: progress and challenges. *Insect Mol. Biol.*, **28**, 739–758.
54. Wang,Y., Song,F., Zhang,B., Zhang,L., Xu,J., Kuang,D., Li,D., Choudhary,M.N.K., Li,Y., Hu,M. *et al.* (2018) The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biol.*, **19**, 151.
55. Senior,A.W., Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T., Qin,C., Žídek,A., Nelson,A.W.R., Bridgland,A. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, **577**, 706–710.