Genómica de insectos

Beatriz López Monroy

Universidad Autónoma de Nuevo León (UANL), Facultad de Ciencias Biológicas. Av. Universidad s/n Cd. Universitaria, San Nicolás de los Garza, N.L. 66455 México.

INTRODUCCIÓN

Los insectos juegan un papel primordial en todos los ecosistemas pues pueden fungir como polinizadores o fuente de alimento para organismos superiores, en el ciclaje de nutrientes y hasta como controladores de plagas. Desafortunadamente también se ven implicados en la transmisión de enfermedades al hombre, animales y plantas, se comen nuestros cultivos o granos almacenados, entre otros.

Lo anterior nos lleva a reflexionar en el área de oportunidad que se tiene en investigación al tratarse de insectos ya sea en la búsqueda de medidas para su control o bien el aprovechamiento que se le puede dar a estas formas de vida o los procesos que estos llevan a cabo.

La genómica se refiere al estudio del genoma completo, de todos los genes que se encuentran en un organismo (23). En una analogía, el genoma es el manual de instrucciones para la vida donde se encuentran los genes a partir de los cuales se forman las proteínas, que a su vez llevan a cabo las funciones de un organismo. Su tamaño es variable y este no es proporcional al número de genes y mucho menos se correlaciona con la posición evolutiva de los organismos. Por lo tanto, la genómica es una ciencia que se enfoca al estudio de los genomas, así como los genes que contienen, sus funciones, las interacciones entre ellos y con los factores ambientales. Dicho estudio involucra: mapas genómicos, secuencias genómicas y funciones génicas.

Con el rápido desarrollo de tecnologías de secuenciación de genes los científicos han trabajado arduamente en diferentes proyectos de secuenciación de genomas, como el Genoma 10K el cual se estableció en 2009 por un consorcio de biólogos y genómicos determinados a facilitar la secuenciación y análisis de 10,000 genomas completos de vertebrados; o el Bird 10K similar al anterior. En 2017, se propuso el Proyecto Earth BioGenome en la conferencia de BioGenomics con el propósito de lanzar el genoma de todos los organismos vivos (12).

En el caso de insectos, fue en 2011 cuando Robinson y colegas propusieron la iniciativa i5k que busca secuenciar el genoma de 5000

insectos y otros artrópodos con importancia biológica significativa o valor económico, esto antes del 2017. Desafortunadamente los objetivos de la iniciativa i5k están aun muy lejos de cumplirse debido a la dificultad de ensamblar los genomas y el limitado apoyo financiero. No obstante, este ambicioso proyecto generó una gran cantidad de datos e interés en genomas de insectos (17).

Dada la importancia de los insectos y la información contenida en sus genomas, es que a continuación se describen algunos de los principales avances dentro de la genómica de insectos.

Secuenciación, ensamblaje y anotación de genomas de insectos

Las razones por las cuales se secuencia el genoma de los insectos son casi tan diversas como el mismo grupo. Podemos mencionar que principalmente se debe a que algunos de estos tienen importancia como modelos de estudio, en ecología, evolución o bien porque son plagas o transmiten alguna enfermedad. Cualquiera que sea la razón de secuenciar el genoma de un insecto, hasta el año 2019, se han registrado 1219 proyectos de secuenciación del genoma de insectos en el Centro Nacional de Información Biotecnológica (NCBI). De este total 401 especies de insectos tienen conjuntos genómicos completos con calidad variada; 155 poseen la anotación del genoma la cual ha sido publicada; y más de 100 genomas de insectos han sido publicados en revistas científicas (12). Sobre la base de estos recursos fundamentales, entomólogos han generado abundantes datos genómicos funcionales de insectos, incluidos transcriptomas, proteomas y metabolomas.

Resulta interesante que de los 1219 genomas de insectos en el NCBI Bio Projects (Project type: primary submission) se cubren casi todos los órdenes (Figura 1). El orden de insectos con más proyectos de secuenciación es Phthiraptera, seguido de Diptera y Lepidoptera.

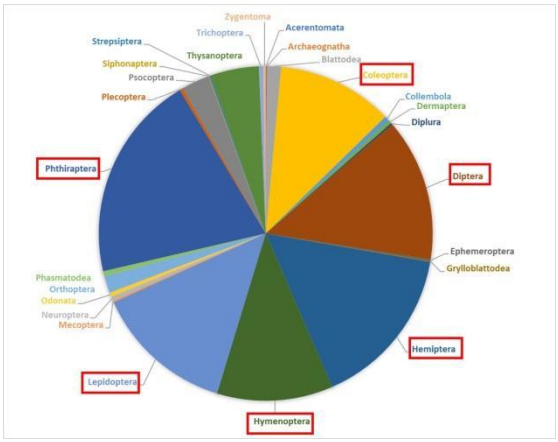

Figura 1. Proyectos de secuenciación de genomas de insectos (Elaboración propia con datos de Li et al. 2019)

De estos proyectos de secuenciación de genomas, solo algunos se han ensamblado y se encuentran en 18 órdenes, incluidos Archaeognatha, Blattodea, Coleoptera, Collembola, Diptera, Ephemeroptera, Hemiptera, Hymenoptera, Lepidoptera, Odonata, Orthoptera, Phasmatodea,

Phthiraptera, Plecoptera, Siphonaptera, Strepsiptera, Thysanoptera y Trichoptera. El número mayor de genomas ensamblados lo tiene Diptera, seguido de Lepidoptera, Hymenoptera y Hemiptera (Figura 2).

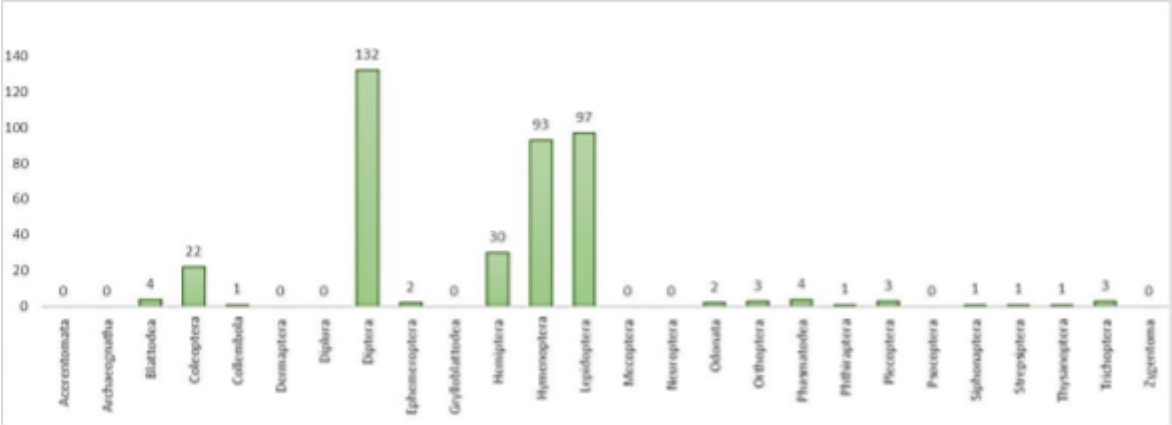

Figura 2. Genomas de insectos ensamblados (Elaboración propia con datos de Li et al. 2019)

A pesar de esta diversidad hay únicamente 10 órdenes de insectos con genomas anotados, lo que representa el 12.7% de todos los proyectos de secuenciación de genomas de insectos (Figura 3), esto es reflejo de las dificultades adicionales encontradas en el ensamblaje y anotación de genomas. A la fecha la baja calidad de ensamblaje se debe típicamente a las secuencias repetitivas y la alta heterocigocidad, siendo esta última el obstáculo para la anotación de genomas de insectos (12).

Idealmente, los datos de secuenciación para el ensamblaje del genoma se pueden obtener de individuos homocigotos endogámicos (o machos haploides en el caso de himenópteros) para evitar complejidad de datos creada por heterocigosidad. Sin embargo, dado que la obtención de tales muestras a menudo es difícil o imposible, recientemente se han realizado esfuerzos para desarrollar métodos diseñados explícitamente para manejar el ensamblaje de genomas heterocigotos. Por otro lado, una gran cantidad de secuencias repetitivas en el genoma puede causar una ambigüedad sustancial en el proceso de ensamblar contigs

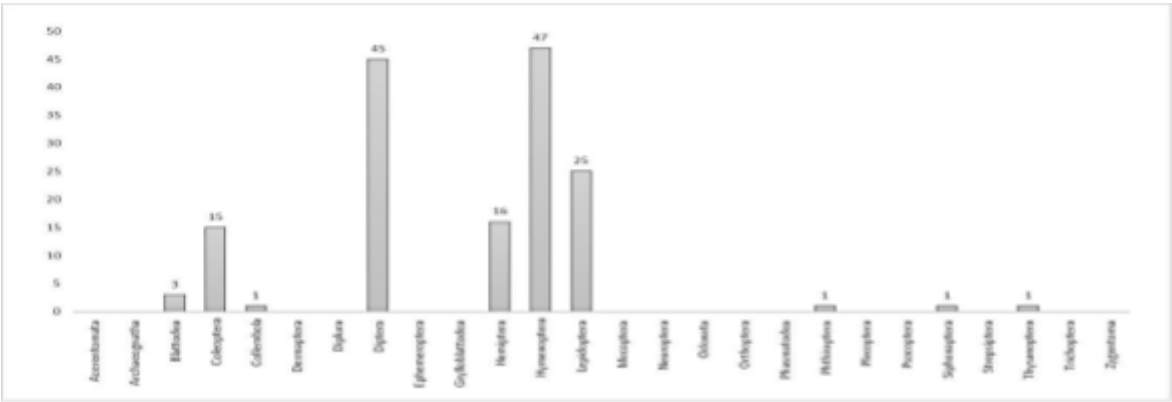

Figura 3. Genomas de insectos anotados (Elaboración propia con datos de Li et al. 2019)

y andamios. La anotación del genoma es indispensable para caracterizar los elementos funcionales en el genoma. Se puede clasificar en dos pasos: anotación estructural y anotación funcional. La anotación estructural viene primero, identificando qué regiones del conjunto corresponden a características específicas, como los genes (incluidos los límites intrón-exón) y los elementos transponibles (TE). Una vez que se delinean las características estructurales, la anotación funcional tiene como objetivo inferir la función e identidad de los genes y otros elementos, en función de las similitudes de secuencia.

Bases de datos de genomas de insectos

Con el avance y desarrollo de tecnologías de secuenciación, los datos genómicos de insectos se están acumulando rápidamente, y la forma de gestionar, almacenar, mostrar y compartir estos datos se convierte en un problema muy urgente (2,20). Las bases de datos genómicas y los recursos relacionados juegan un papel importante en la gestión, el intercambio y la extracción de datos biológicos.

De acuerdo con los datos y recursos que contienen las bases de datos, estas se pueden clasificar en tres categorías. La primera categoría la conforman grandes bases de datos de almacenamiento integradas, que contienen tipos de datos muy diversos, a menudo con variaciones sustanciales en la calidad de los datos. Dichas bases de datos tienen altos costos de mantenimiento y desempeñan principalmente el papel de almacenamiento de datos. Este tipo de base de datos está representada por el NCBI (18), el Instituto Europeo de Bioinformática (4) y la Base de datos japonesa de ácidos nucleicos (11), que se reconocen como las tres principales bases de datos de genes. La segunda categoría la conforman bases de datos del genoma con un enfoque particular en un grupo de especies. Este tipo de base de datos es comúnmente mantenida por un grupo de investigación. La cantidad de datos se reduce significativamente en comparación con las tres bases de datos principales, pero la calidad de los datos es alta. La tercera categoría tiene como objetivo presentar los datos genéticos de una sola especie o un solo género. Normalmente, estas bases de datos son mantenidas por el grupo que secuencia el genoma, el cual es fácil y las actualizaciones son rápidas.

En la actualidad, la mayoría de los datos genéticos de insectos se almacenan en bases de datos completas. La base de datos NCBI RefSeq es la mayor fuente de anotaciones del genoma de los insectos, con un navegador genómico, BLAST, base de datos de genes y otros recursos (14).

Muchos investigadores de insectos han establecido bases de datos para un grupo de especies, como VectorBase (8), FlyBase (1),

Butter flyBase (15), AphidBase (7), Hessian Fly Base (16), Ant Genomes Base (6), Hymentoptera Genome Database (13), BeeBase y LepBase (3).

Por otro lado, las bases de datos de genomas de insectos individuales incluyen MonarchBase (27), ChiloDB (25), WaspAtlas (5), el Proyecto Genoma Heliconius (9), NasoniaBase (6), DBM-DB (21), KONAGAbase (10), KAIKObase (19), SilkDB (22), BeetleBase (22) y ManducaBase (16). Estas bases de datos proporcionan los datos genéticos de una especie o un género, satisfaciendo los requisitos de diferentes usuarios.

A continuación, se presentan las descripciones de tres bases de datos de genoma de insectos: i5k workspace @NAL (16), InsectBase (26) y VectorBase (8).

*i5k Workspace @ NAL (*https://i5k.nal.usda.gov/)

La base de datos i5k Workspace @ NAL es una base de datos genómica de artrópodos, dirigida por el Departamento de Agricultura de los Estados Unidos. Esta plataforma surge como resultado de la gran cantidad de información que se generó y se sigue generando del proyecto i5k. Poelchau y colaboradores construyeron esta base de datos, proporcionando procesos y plataformas estandarizados para el ensamblaje, anotación y mantenimiento del genoma. Se incluyen un total de 64 genomas de insectos. La base de datos proporciona la función de navegación, descarga, envío de datos, alineación de secuencias, visualización del genoma y plataforma de anotación manual genómica en línea, y dos herramientas en línea, HMMER y CLUSTAL (16). La ventaja más notable de i5k Workspace @ NAL es que esta plataforma proporciona tanto anotaciones comunitarias como curación manual de genomas de insectos y produce conjuntos de genes oficiales estándar para insectos.

*InsectBase (*http://www.insect-genome.com/)

InsectBase es una base de datos integral de genomas y transcriptomas de insectos (26). Incluye genomas de insectos de 155 especies correspondientes a 16 órdenes, de las cuales 61 genomas tienen información de anotación También contiene más de 4 millones de tecnologías ecológicamente racionales de 237 especies y 7544 miRNA de 69 especies, proporciona una variedad de funciones, que incluyen consulta, alineación, visualización del genoma, construcción de rutas y anotaciones, análisis evolutivo y construcción de árboles filogenéticos. Todos los datos genéticos se pueden descargar y un dato adicional es que a través de la minería de datos, InsectBase proporciona una función iFacebook para identificar la red de relación de investigadores, genes y especies.

*VectorBase (*https://www.vectorbase.org/)

VectorBase es una base de datos del Centro de Recursos Bioinformáticos (BRC), perteneciente al Instituto Nacional de Alergias y Enfermedades Infecciosas (NIAID) del Departamento de Salud de los Estados Unidos. Esta base de datos se enfoca en información genómica, fenotípica y poblacional de vectores invertebrados de organismos patógenos para los humanos. Se pueden encontrar genomas, transcriptomas, proteomas, secuencias mitocondriales, información poblacional y notas de los cambios realizados a la bioinformación. La información de cada tipo está organizada, además presenta filtros y herramientas que facilitan la búsqueda. Dentro de las herramientas que contiene se encuentran Apollo, BLASY, ClustalW, Genome Browser entre otras.

Desafíos y perspectivas de la genómica de insectos

Hoy en día, el mayor desafío de la secuenciación del genoma del insecto es la dificultad de obtener un ensamblaje del genoma de alta calidad a partir de las lecturas sin procesar producidas por las técnicas de secuenciación de segunda generación. Illumina HiSeq es la plataforma de secuenciación de segunda generación más utilizada actualmente, y es rápida, barata y altamente precisa. Sin embargo, solo puede producir lecturas cortas (<250 pb). Esto supone un gran desafío para el ensamblaje del genoma, porque la mayoría de los insectos, especialmente los lepidópteros, tienen una alta heterocigosidad. El problema de la heterocigosidad a menudo se complica por el tamaño de los insectos lo que hace necesario formar pools de individuos para obtener suficiente ADN para el análisis. Además, es bastante difícil producir cepas endogámicas para la mayoría de los insectos (12).

El uso de las técnicas de secuenciación de tercera generación, que pueden producir lecturas largas (> 10 kb), es un remedio prometedor para estos desafíos. PacBio es la primera técnica de secuenciación de tercera generación ampliamente utilizada. Las lecturas de PacBio son incluso más largas que muchos contigs en varios conjuntos de genomas de insectos. Los genomas de varios insectos han sido reportados usando PacBio o la combinación de Illumina y PacBio. Una de las desventajas de las tecnologías de lectura larga es la cantidad de ADN de alto peso molecular requerido para los métodos. Además, estos métodos tienen altas tasas de error en la secuencia, aunque la corrección de errores puede depender del uso de sub-lecturas crudas superpuestas para mejorar la precisión base. En general, PacBio requiere más profundidad de cobertura para mejores correcciones.

Los avances regulares en las tecnologías de secuenciación han estimulado la rápida acumulación de genomas de insectos,

preparando el escenario para una nueva era de la ciencia de los insectos. Las nuevas herramientas para la manipulación genética y la edición del genoma serán un jugador importante en esta etapa. Con el apoyo de la información genómica, tales herramientas sin duda acelerarán los conocimientos sobre la base genética de los fenotipos.

Más allá de diversificar la investigación básica sobre los insectos, los datos del genoma de los insectos son recursos útiles para desarrollar políticas alternativas y ecológicas de control de plagas. La genómica funcional y poblacional aplicada a las especies de plagas está dando nuevos conceptos para la implementación del manejo integrado de plagas (MIP), lo que lleva a la incipiente subdisciplina de las MIP-omicas (12). Tales análisis permiten a los entomólogos descubrir la diversidad molecular en las poblaciones de insectos que subyacen

Además, la re-secuenciación del genoma seguida de estudios de asociación de todo el genoma se ha convertido en una estrategia eficaz para descubrir los mecanismos de muchos rasgos importantes, como la resistencia a los insecticidas, el polimorfismo geográfico y la adaptación al huésped (24).

Conclusión

Los avances en las tecnologías de secuenciación han llevado a generar diversos proyectos a fin de obtener la secuencia de genomas de diferentes organismos, en donde los insectos no han sido la excepción. La gran cantidad de datos generados por la secuenciación del genoma de los insectos está llevando a la entomología a una nueva era, promoviendo avances científicos en diversas áreas. La rápida acumulación de grandes cantidades de datos del genoma no solo brinda muchas oportunidades para importantes descubrimientos científicos, sino que también plantea muchos desafíos.

Literatura citada

1. Ashburner, M. and Drysdale, R. (1994) FlyBase – the Drosophila genetic database. Development, 120, 2077–2079.
2. Baxevanis, A.D. and Bateman, A. (2015) The importance of biological databases in biological discovery. Current Protocols in Bioinformatics, 50: 1.1.1–1.1.8.
3. Challis, RJ, Kumar, S, Dasmahapatra, KKK, Jiggins, CD and Blaxter, M (2016) Lepbase: the lepidopteran genome database. https://doi.org/10.1101/056994
4. Cook Charles E, Lopez Rodrigo, Stroe Oana, Cochrane Guy, Brooksbank Cath, Birney Ewan and Apweiler Rolf, (2019). The European Bioinformatics Institute in 2018: tools, infrastructure and training, Nucleic Acids Research. 47 (D1): D15–D22.
5. Davies, N.J. and Tauber, E. (2015) WaspAtlas: a Nasonia vitripennis gene

database and analysis platform. Database: The Journal of Biological Databases and Curation, 2015 pii: bav103. https://doi.org/10.1093/database/bav103.

6. Elsik, C. G., Tayal, A., Diesh, C. M., Unni, D. R., Emery, M. L., Nguyen, H. N., & Hagen, D. E. (2016). Hymenoptera Genome Database: integrating genome annotations in HymenopteraMine. Nucleic acids research, 44(D1): D793–D800.
7. Gauthier, J.P., Legeai, F., Zasadzinski, A., Rispe, C. and Tagu, D. (2007) AphidBase: a database for aphid genomic resources. Bioinformatics, 23: 783–784.
8. Giraldo-Calderón, G. I., Emrich, S. J., MacCallum, R. M., Maslen, G., Dialynas, E., Topalis, P., Ho, N., Gesing, S., VectorBase Consortium, Madey, G., Collins, F. H., & Lawson, D. (2015). VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic acids research, 43(Database issue): D707–D713.
9. Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature, 487(7405): 94–98.
10. Jouraku, A., Yamamoto, K., Kuwazaki, S., Urio, M., Suetsugu, Y., Narukawa, J., Miyamoto, K., Kurita, K., Kanamori, H., Katayose, Y., Matsumoto, T., & Noda, H. (2013). KONAGAbase: a genomic and transcriptomic database for the diamondback moth, Plutella xylostella. BMC genomics, 14: 464.
11. Kodama, Y., Mashima, J., Kosuge, T., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y., & Takagi, T. (2018). DNA Data Bank of Japan: 30th anniversary. Nucleic acids research, 46(D1): D30–D35.
12. Li F., Zhao X., Li M., Huang C., Zhou Y., Li Z. and Walters R. (2019) Insect genomes: progress and challenges. Insect Molecular Biology. The Royal Entomological Society, 1-20.
13. Munoz-Torres, M. C., Reese, J. T., Childers, C. P., Bennett, A. K., Sundaram, J. P., Childs, K. L., Anzola, J. M., Milshina, N., & Elsik, C. G. (2011). Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. Nucleic acids research, 39(Database issue): D658–D662.
14. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research, 44(D1): D733–D745.

15. Papanicolaou, A., Gebauer-Jung, S., Blaxter, M.L., Owen McMillan, W. and Jiggins, C.D. (2008) ButterflyBase: a platform for lepidopteran genomics. Nucleic Acids Research, 36: D582–D587.
16. Poelchau, M., Childers, C., Moore, G., Tsavatapalli, V., Evans, J., Lee, C.Y. et al. (2015) The i5k Workspace@NAL – enabling genomic data access, visualization and curation of arthropod genomes. Nucleic Acids Research, 43: D714–D719.
17. 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., & Schneider, D. J. (2011). Creating a buzz about insect genomes. Science, 331:6023.
18. Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., Holmes, J. B., Kim, S., Kimchi, A., Kitts, P. A., Lathrop, S., Lu, Z., Madden, T. L., Marchler-Bauer, A., Phan, L., Schneider, V. A., … Ostell, J. (2019). Database resources of the National Center for Biotechnology Information. Nucleic acids research, 47(D1): D23–D28.
19. Shimomura, M., Minami, H., Suetsugu, Y., Ohyanagi, H., Satoh, C., Antonio, B., Nagamura, Y., Kadono-Okuda, K., Kajiwara, H., Sezutsu, H., Nagaraju, J., Goldsmith, M. R., Xia, Q., Yamamoto, K., & Mita, K. (2009). KAIKObase: an integrated silkworm genome database and data mining tool. BMC genomics, 10: 486.
20. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big data: Astronomical or genomical? PLoS biology, 13(7).
21. Tang, W., Yu, L., He, W., Yang, G., Ke, F., Baxter, S. W., You, S., Douglas, C. J., & You, M. (2014). DBM-DB: the diamondback moth genome database. Database: the journal of biological databases and curation, 2014, bat087. https://doi.org/10.1093/database/bat087
22. Wang, J., Xia, Q., He, X., Dai, M., Ruan, J., Chen, J., Yu, G., Yuan, H., Hu, Y., Li, R., Feng, T., Ye, C., Lu, C., Wang, J., Li, S., Wong, G. K., Yang, H., Wang, J., Xiang, Z., Zhou, Z., … Yu, J. (2005). SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic acid Wang, L.J., Wang, S.Z., Li, Y.H., Paradesi, M.S. and Brown, S.J. (2007) BeetleBase: the model organism database for Tribolium castaneum. Nucleic Acids Research, 35, D476–D479.s research, 33(Database issue): D399–D402.
23. WHO (2002). Genomics and World Health: Report of the Advisory Committee on Health research, Geneva.
24. Xiao, H.M., Li, M.Z. and Li, F. (2017) Progress in research on long, non-coding,

insect RNA. Chinese Journal of Applied Entomology, 54(1): 1–12.

25. Yin, C., Liu, Y., Liu, J., Xiao, H., Huang, S., Lin, Y., Han, Z., & Li, F. (2014). ChiloDB: a genomic and transcriptome database for an important rice insect pest Chilo suppressalis. Database: the journal of biological databases and curation, 2014, bau065. https://doi.org/10.1093/database/bau065
26. Yin, C., Shen, G., Guo, D., Wang, S., Ma, X., Xiao, H., Liu, J., Zhang, Z., Liu, Y., Zhang, Y., Yu, K., Huang, S., & Li, F. (2016). InsectBase: a resource for insect genomes and transcriptomes. Nucleic acids research, 44(D1): D801–D807.
27. Zhan, S. and Reppert, S.M. (2013) MonarchBase: the monarch butterfly genome database. Nucleic Acids Research, 41: D758–D763.