环境昆虫学报 2017，39（1）：1－18
Journal of Environmental Entomology
http：//hjkcxb. alljournals. net
doi：10. 3969/j. issn. 1674－0858. 2017. 01. 1

特邀稿件
Invited Review

尹传林，李美珍，贺康，等. 昆虫基因组及数据库研究进展［J］. 环境昆虫学报，2017，39（1）：1－18.

昆虫基因组及数据库研究进展

尹传林，李美珍，贺　康，丁思敏，郭殿豪，席　羽，李　飞*
（浙江大学昆虫科学研究所，杭州 310058）

摘要：基因组序列为昆虫分子生物学研究提供丰富的数据资源，推动系统生物学在古老的昆虫学中蓬勃发展。昆虫基因组学研究已经成为当前的研究热点，目前在 NCBI 登录注册的昆虫基因组测序计划有 494 项，其中已提交原始测序数据的昆虫有 225 种，完成基因组拼接的有 215 种，具有基因注释的有 65 种，公开发表的昆虫基因组有 43 篇。本文综述了测序技术发展的历史及其对昆虫基因组研究的推动作用、昆虫基因组的组装和注释及其存在的问题、昆虫基因组测序进展、昆虫基因组数据库的发展及基因数据挖掘利用的基本思路和对策，以及昆虫基因大数据在害虫防治和资源昆虫利用中的应用前景。

关键词：昆虫基因组；组装与注释；数据挖掘与分析；基因组数据库；害虫防治；资源昆虫利用
中图分类号： Q963；S43　　**文献标志码：** A　　**文章编号：** 1674－0858（2017）01－0001－18

The progress of insecg genomic research and the gene database

YIN Chuan-Lin，LI Mei-Zhen，HE Kang，DING Si-Min，GUO Dian-Hao，XI Yu，LI Fei*（Institute of Inesct Science，Zhejiang University，Hangzhou 310058，China）

Abstract: With huge amount of insect genome sequencing data was generated，entomology has entered a new era of systematic biology. Up to now，467 insect genome projects have been registered on NCBI，among which 225 have submitted with sequencing raw reads，215 have been assemblied，65 have been annotated and 43 have been published. Here，we reviewed the development of different sequence technologies，methods and problems of genome assembly，genome annotation and analysis，and important achievements in the field of insect genome projects. In addition，we summarized the development of insect genome databases. Insect genomics is now a hotspot of scientific study，which has wide applications in pest control and utilization of the resource insects.

Key words: Insect genome；genome database；big DATA；biological databases

昆虫是生物界种类数量最多、最古老的类群之一，距今 3.5 亿年的古生代泥盆纪就已出现，大约构成所有生物种类的 50% 左右（Robinson，*et al.*，2011），目前已经被描述鉴定的昆虫种类有一百万多种。作为重要的活化石，昆虫的进化研究可以探秘生命的起源以及地球环境的变更。昆虫与人类的活动息息相关，既有令人烦恼的农业害虫和卫生害虫，也有让人赏心悦目的观赏昆虫。农业生态系统离不开昆虫，地球上 75% 以上的开花植物都依靠昆虫来授粉（Robinson *et al.*，2011）。

基金项目：国家重点研发计划“主要入侵生物的生物学特性分析”重大课题（2016YFC1200602）
作者简介：尹传林，男，1989 年生，博士研究生，研究方向为昆虫基因组学，E－mail：yincl2013@126. com
* **通信作者** Author for correspondence，E－mail：lifei18@zju. edu. cn
收稿日期 Received：2016－12－10；**接收日期** Accepted：2016－12－20

昆虫学作为一门独立的分支进入科学领域，迄今已有 300 多年历史。

随着测序技术的快速发展，在生物大数据的潮流下，古老的昆虫学逐渐迈入基因组时代。昆虫学者利用各种组学研究手段如基因组、转录组、蛋白组、代谢组等产生了大量的生物数据，从系统生物学的角度来解决昆虫学研究中的问题，为昆虫学研究带来了新的视角，焕发了新的生机。本文围绕昆虫基因组学研究中的组装、注释、数据挖掘和基因数据库等方面进行了综述，对目前存在的问题进行了总结，对未来的发展趋势进行了展望。

1 测序技术的进步和生物信息学的发展

昆虫基因组学研究得益于测序技术的巨大进步和生物信息学的逐渐普及。测序技术根据其发展的历史可以分为三个不同的时代：以链终止法或链降解法为原理的一代测序技术（如 Sanger 测序技术）、以边合成（边链接）边测序为原理的二代测序技术（主要包括 ABI 公司的 SOLiD 技术、Illumina 公司的 Solexa 技术和 Roche 公司的 454 技术等），以及单分子测序的三代测序技术（如 PacBio 公司的 SMRT 技术和 Oxford Nanopore 公司的纳米孔单分子测序技术等）（Heather *et al.*，2016）（图 1）。

1975 年由桑格（Sanger）和考尔森（Coulson）发明的链终止法（Sanger *et al.*，1975），以及 1976 年由马克西姆（Maxam）和吉尔伯特（Gilbert）发明的链降解法（Maxam *et al.*，1977），开启了核酸测序的新纪元。利用第一代测序技术，测定了噬菌体 X174 的基因组序列，全长 5375 个碱基，这是首个生命体的基因组序列（Sanger *et al.*，1977）。2001 年，利用 Sanger 测序技术完成了人类基因组计划（Venter *et al.*，2001）。果蝇是第一个被测序的昆虫（Adams *et al.*，2000），之所以被优先选择进行基因组测序，是因为果蝇一直被视为生命科学研究中最重要的模式生物之一。但其实更重要的原因，是果蝇基因组比较小（仅 180 Mb 左右），可以用来检测全基因组鸟枪法（Whole Geome Shotgun，WGS）在人类基因组测序中的可行性。在没有其他测序技术可供选择情况下，第一代 Sanger 测序技术是唯一的技术主角，其具有明显的优势，读长最高可达 1000 bp，准确性高达 99.999%。然而，其缺点也十分明显，测序成本过高，通量低，无法实现真正的大规模应用。

在科研需求和市场利润的双重驱动下，催生了 3 个重要的二代测序技术（SOLiD 技术、Solexa 技术和 454 技术）。在人类基因组测序计划要惊动各国领导人的时代，美国 NIH 启动了“1000 美元基因组计划”，资助 2 亿美金来推动测序技术的进步。正是这种前瞻性的资助计划，改写了生命科学研究的进程，也是当前生命科学各个研究领域的基因组计划发展如火如荼的重要基础。第二代测序技术极大地降低了测序成本，提高了测序通量和测序速度，同时保持了高准确性。在启动人类基因组计划时，预计要花费 30 亿美金、历经 15 年才能完成，而二代测序技术可在一个星期内完成，仅需 1000 美元。Solexa 技术和 454 技术是基于连合成边测序的原理，而 SOLiD 技术是基于边连接边测序和双色法的原理。如前所述，二代测序技术的优点非常明显，但其缺点是在 PCR 扩增中增加了测序的错误率，具有明显的系统偏向性，读长较短（早期仅 70 多 bp，最新技术也只有 200 多 bp）。其中，读长较短给基因组的拼接带来了困难，虽然开发了大量的生物信息学算法用于二代基因组数据的拼接，但对于高杂合物种，仍然没有满意的解决途径，而绝大部分昆虫具有高杂合性。二代测序技术目前仍是市场上的主流技术，其中 Illunima 公司的 Solexa 技术因其技术优势占据了市场的半壁江山。

技术的进步是无止境的。近年来，测序技术又有了新的突破，其中主要以 PacBio 公司的 SMRT 和 Oxford Nanopore Technologies 公司的纳米孔单分子测序技术为代表，被称为第三代测序技术。第三代测序技术的特点是单分子测序，无需进行 PCR 扩增，能有效避免因 PCR 偏好性而导致的系统误差，同时显著提高了读长，并保持了二代测序技术高通量的优点。虽然三代测序技术已经开始走向了市场，但其准确性仍然有待高。

科研人员产生数据的能力明显地增强，海量生物数据不断积累，因此对数据管理和分析提出了更高的要求，生物信息学即在此基础上诞生。当时生物学家第一次面临超出想象的基因组数据，有点无所适从，不知所措，生物信息学俨然以“救世主”身份拯救了人类基因组计划。最被广泛接受的生物信息学定义是，综合利用生物学、计

算机科学和信息科学等多学科的理论与技术，产生和创造生物数据，管理和存储生物数据，以及挖掘和分析生物数据，揭示生物数据蕴含的生物学意义。近年来，生物信息学得到了空前的充分

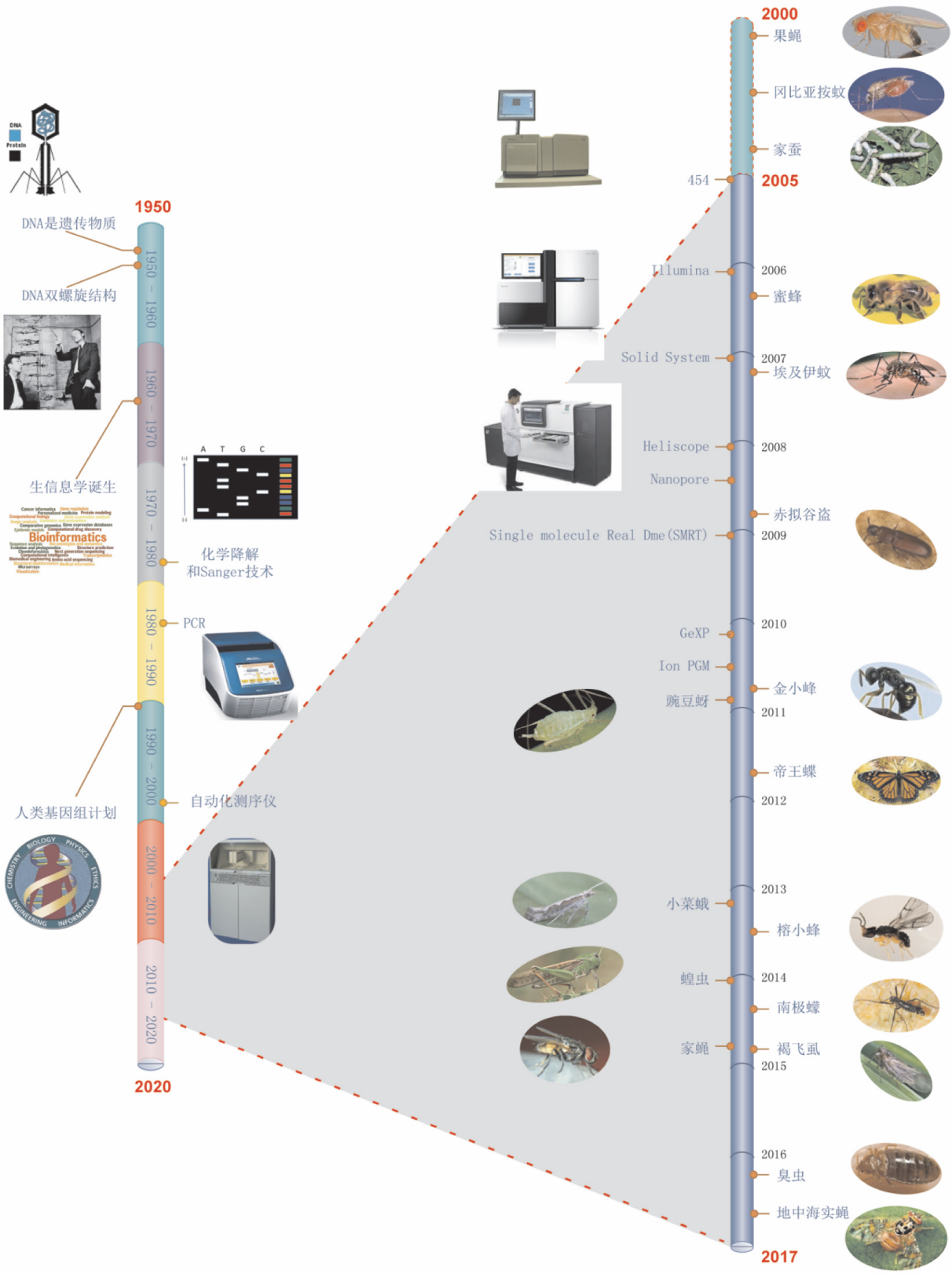

图 1　测序技术的发展和昆虫基因组测序历程

Fig. 1　The process of technology development and insect genome sequencing

发展，并被不断普及。早期的生物信息研究和算法开发主要针对普遍存在的科学问题，而现在各种衍生的生物信息学算法和软件层出不穷，针对单个学科的具体科学问题进行了优化和提高，这极大地带动了大数据时代的生物信息学研究（Ouzounis *et al.*，2003）。

依据研究方向，生物信息学可分为 3 个主要部分：（1）研发有效利用和管理数据的新工具，构建新平台，例如构建各种各样的生物信息学数据库；（2）新算法的开发，例如各类基因组测序数据的拼接和比对算法等；（3）生物数据的挖掘与分析，从海量生物数据中挖掘和发现规律，帮助生物学家从“大海捞针”变为“池塘捞鱼”，为揭示生物表型的分子机制提供有益的参考。前两个研究方向偏“信息”，而第三个研究方向偏“生物”，这与计算机科学的“偏硬”和“偏软”两个方向有异曲同工之处。生物学家更加熟悉和倚重”生物数据挖掘与分析”这一方向。但必须强调的是，数据平台和算法开发是生物信息学重要的基础，没有准确的数据，没有合适的算法，生物学意义的挖掘就无从谈起，甚至会被引至错误的方向。

2 昆虫基因组拼接、注释与数据分析

2.1 昆虫基因组组装

基因组鸟枪法是将 DNA 随机打断成较短的序列，构建测序载体进行测序，获得了大量的小片段序列。因此，基因组组装是基因组测序中最为关键的一步。尤其困难的是，基因组组装算法需要根据测序平台、文库构建策略和测序读长等进行优化（Richards *et al.*，2015）。由于测序策略的设计缺陷或优化不足，往往导致昆虫基因组拼接失败，这样的例子并不鲜见。

根据是否有参考序列，可把基因组拼接分为从头拼接（De novo assembly）和比较拼接（comparative assembly）两大类（Wajid *et al.*，2012）。从头组拼接指完全依赖 reads 间的重叠信息拼接出基因组序列，而比较拼接综合了 reads 间的重叠信息和 reads 在参考序列中的位置信息，相比而言，从头拼接更难更复杂。按照算法的原理，从头拼接大致可以分以下几类：第一类是 overlap/layout/Consensus（OLC）法，这类组装算法有 CABOG、Newbler、Shorty、Edena、Celera 等，其适应于读长较长的测序数据，如 Sanger 法测序和第三代测序技术，果蝇基因组的组装采用的就是 Celera 软件；第二类是 De Bruijn Graph（DBG）法，一种基于图论的算法，软件有 SOAPdenovo、Euler、Velvet 等，这类算法需要不断调整 k-mer 的值来达到一个最佳的组装效果；第三类是 Greey graph alogorithms 法，这类算法有 SSAKE、SHARCGS、VCAKE 等（Wajid *et al.*，2012）。

已发表的昆虫基因组组装算法主要使用了 CABOG（Miller *et al.*，2008）、SOAPdenove（Luo *et al.*，2012）、ALLPATH-LG（Butler *et al.*，2008）、ABySS（Simpson *et al.*，2009）等方法。SOAPdenove 是华大基因开发的基因短序列拼接，运行速度快，依赖于搜索 k-mer 来寻求最优解。ALLPATH-LG 近年来使用率越来越高，特别适合于读长 100 – 200 bp、覆盖倍数 200X 左右的测序策略。和 SOAPdenove 比，不需要设定 K-mer 值。但是由于其依赖穷举法，因此对硬件要求很高，运行时间非常长。

生物信息学发展至今，不断诞生了新的软件。然而，基因组组装一直都面临着巨大的挑战，无法取得理想的效果。分析认为，影响昆虫基因组拼接质量的主要原因有，一是重复序列，基因组中含有大量的重复序列，对拼接造成非常大的干扰，而昆虫基因组有可能产生了大量新的重复序列，产生了明显的影响；二是物种杂合度，当来自父本或母本染色体 DNA 之间的差异大时，后代可能具有更大的环境适应性优势，但给拼接造成了困难。昆虫基因组拼接困难的解决，一方面依赖于测序技术的继续进步，另一方面也依赖于算法的不断优化和提高。

2.2 昆虫基因组质量评估

目前，主要从完整性、正确性、拼接长度等几个方面进行基因组组装结果的评价（Wajid *et al.*，2012）。

（1）组装序列的完整性

组装序列的完整性指组装得到的基因组大小与实际基因组大小之间的差异，通常采用两者的比值来衡量。检测基因组大小的常用方法有流式细胞仪技术和 K-mer 分析法。

（2）拼接正确性

拼接正确性反应了组装结果和真实基因组的一致性。通常采用已知大片段序列来检测组装结

果的正确性。如果没有大片段序列，可把 paired-end 或者 mate-pair 序列比对到组装结果上，检查序列在组装上的位置以及两者间的距离，以此评估拼接正确性。

(3) N50

N50 是衡量基因组拼接质量的重要标准，其计算方法是，把所有序列按照从长到短进行排序，并对序列长度进行累加，当累加值达到基因组序列总数的一半时所对应的序列长度即为 N50。通过计算组装基因组的 contigs 和 scaffolds 的 N50，可以非常直观的评价拼接质量。

(4) CEGMA 评估

CEGMA (Parra *et al.*, 2007) 是目前使用最广泛的评估基因组甚至是转录组拼接质量的方法，其首先确定了真核生物中极其保守的 248 个核心基因 (CEG)，然后在基因组 Scaffold 序列中搜寻这些 CEG 基因，计算具有全长序列的 CEG 百分比、仅有部分片段的 CEG 百分比和完全缺失的 CEG 百分比，以此来判断基因组的拼接质量。

(5) BUSCO 评估

BUSCO (Simao *et al.*, 2015) 是在 CEGMA 上进行更新的新算法。BUSCO 的其本原理与 CEGMA 类似并进行了优化，其按照不同的大类群选取不同的直系同源基因集，在节肢动物中挑选了 2647 个直系同源基因，通过检索缺失率来反映基因组质量。

2.3 昆虫基因组的注释

基因组注释是指对基因组特征进行描述，包括结构注释和功能注释。结构注释主要包括预测基因组重复序列、非编码 RNA 和蛋白编码基因; 功能注释是根据基因序列信息预测基因的功能。

(1) 重复序列注释

重复序列识别方法分为序列比对和从头预测两大类。序列比对法是根据相似性程度在基因组中识别同源的重复序列。该方法预测的结果往往比较可靠，但不全面。目前广泛使用的比对预测软件有 Repeatmasker (Tarailo-Graovac *et al.*, 2009)。从头预测方法利用重复序列的结构特征在基因组中进行预测，这种方法对结构特征明确的重复序列具有非常好的预测效果，比如 MITEs、LTR 等，常见的从头预测方法有 Recon (Bao *et al.*, 2002)，Piler (Edgar *et al.*, 2005)，Repeatscout (Price *et al.*, 2005)，LTR-finder (Xu *et al.*, 2007) 等。一般而言，采用同源比对和从头预测两者相结合的方法进行重复序列识别，比较可靠全面 (刘金定，2014)。

(2) 非编码 RNA 的识别

非编码 RNA 指不生成蛋白产物、以 RNA 形式发挥功能的 RNA 基因，如 tRNA、rRNA、piRNA、miRNA、snoRNA、rasiRNA 等。非编码 RNA 没有蛋白质编码基因的典型特征，因此一般对其二级结构序列和特征进行预测，常用的软件有 miRdeep (Friedlander *et al.*, 2008)、RNAstructure (Bellaousov *et al.*, 2013)、TripletSVM (Xue *et al.*, 2005) 等，常用的非编码 RNA 数据库有 RNAdb (Pang *et al.*, 2007)、NONCODE (Zhao *et al.*, 2016)、Rfam、miRBase (Kozomara *et al.*, 2014) 和 snoRNABase 等 (陈勇等，2014)。

(3) 编码基因组注释

蛋白编码基因的识别是基因组注释中最为重要的部分。常见的编码基因预测方法有基于基因模型的从头预测方法、基于比对的蛋白同源预测方法以及基于转录组比对的表达证据方法等。这 3 类方法各有优点和缺点: 从头预测方法理论上可以覆盖全面基因集，但假阳性高; 同源比对方法预测结果准确，但局限于物种间保守基因; 转录组比对方法直接来自表达证据，但受限于转录组的数据质量和数量。研究人员通过整合多种预测结果来提高编码基因注释的准确性，比如 Glean (Elsik *et al.*, 2007)、Evigan (Liu *et al.*, 2008)、PASA (Xu *et al.*, 2006)、MAKER (Cantarel *et al.*, 2008)、jigsaw (Allen *et al.*, 2006) 等。虽然多证据整合方法可以提高编码基因注释可靠性，但是仍然也存在一些问题需要解决，比如新测序物种缺少必要数量的可靠基因用于从头预测软件训练，难以获得足够的表达证据等。真核生物广泛存在可变剪接和多个转录起始位点，导致编码基因预测更加复杂。

(4) 功能注释

基因组功能注释是依据“序列决定结构，结构决定功能”的基本原理，利用序列相似性来推断基因的功能。基因功能预测是利用序列同源比对软件如 Blast 等搜索序列相似的已知基因，再利用已知基因的功能进行注释。常用于基因功能注释的基因集有 NCBI 的非冗余蛋白序列数据库 (Non-redundant protein sequences，NR)、参考蛋白数据库 (refseq protein)、SWISS-PROT 数据库等，这些数据库中蛋白序列一般都带有注释信息。

2.4 比较昆虫基因组分析

比较基因组学是对近缘物种和同一物种的不同个体的基因组序列，从基因结构、共线性及基因家族等方面进行分析，揭示不同物种之间的基因家族扩增与丢失、基因的起源及进化等，协助阐明重要性状的分子机制。比较基因组可分为种间比较基因组和种内比较基因组，种间比较基因组是近缘物种之间的基因组比较，重点研究基因家族和基因进化；种内比较基因组比较的是同一个物种之间不同个体的遗传差异性，通过将重测序序列与参考基因组序列进行比较后，进行关联性分析，挖掘可能与重要性状关联的单核苷酸多态性和结构差异，为分子机制研究奠定基础（陈勇等，2014）。

2.5 直系同源和共线性分析

直系同源基因具有相似的生物学功能，确定直系同源基因是功能基因鉴定、比较基因组、功能基因分类、信号通路预测等的基础。预测直系同源基因的方法大致可分为 3 类：一是比较序列相似性来识别直系同源基因；二是通过构建系统发育树来识别直系同源关系；三是混合利用序列相似性和系统发育树的方法。

基因共线性（synteny）是指基因在染色体上排列顺序的一致性。在进化过程中，由于转座、插入、染色体重排、区段加倍和缺失等原因，会发现基因序列的重排，进化距离越远的物种，基因共线性越差。通过比较物种间同源基因的相对位置，可以确定不同物种间基因组的共线性，揭示所比较物种间基因结构以及基因顺序的异同。

2.6 基因家族的扩张和收缩

基因家族是来源于同一个祖先，由一个基因通过基因重复而产生两个或更多的拷贝而构成的一组基因，它们在结构和功能上具有明显的相似性，编码相似的蛋白质产物，同一家族基因可以紧密排列在一起，形成一个基因簇（gene cluster）。但多数时候，它们分散在同一染色体的不同位置，或者分布于不同染色体上，各自具有不同的表达调控模式。在长期进化过程中，基因家族会有扩张和收缩，这通常与物种的性状密切相关。

3 昆虫基因组测序计划及其进展

3.1 i5k 计划

i5k 计划由 Gene Robinson 等人（2011）在 Science 上发文提出，倡议在 2020 年前后完成 5000 种节肢动物基因组的测序和分析工作，建议选定的物种应该广泛分布于各种生态系统，对世界范围的农业、食品安全、药物研究、能源再生、模式生物研究等有着非常重要的影响，能够作为昆虫分类各分支上的代表物种，有助于全面理解节肢动物的进化历程和系统发育关系。我国昆虫学者积极响应 i5k 全球性计划，以我国昆虫学者为主导，先后完成了家蚕、小菜蛾、蝗虫、褐飞虱、榕小蜂、二化螟等昆虫的基因组测序。迄今已经召开了两届国际昆虫基因组学学术会议，分别为 2013 年 12 月 15 日在中国科学院动物研究所举办了“首届中国昆虫基因组学及国际 i5k 计划研讨会”，及于 2015 年 9 月 18 日在重庆召开了“第二届国际昆虫基因大会”，从基因组测序、功能基因组学、比较和进化基因组学、生物信息学技术等多个方面讨论了昆虫基因组学的发展及发展趋势，探讨了基因组学在害虫防治、资源昆虫利用、药物靶点开发及进化生物学等方面的应用前景。

3.2 已经完成的昆虫基因组测序

截至 2016 年 11 月 1 日，从美国国立生物技术信息中心（National Center for Biotechnology Information，NCBI）BioProject 数据库统计，共有 494 种昆虫的基因组测序项目在开展，覆盖了几乎所有目的昆虫。在这些的基因组测序项目中，有 215 个基因组完成组装并且数据已经提交到 NCBI 数据库，占总提交昆虫基因组测序项目的 43.5%。这些物种共涵盖了 15 目的昆虫（图 2A），包括捻翅目 Strepsiptera、蜻蜓目 Odonata、蜚蠊目 Blattodea、直翅目 Orthoptera、毛翅目 Trichoptera、虱目 Phthiraptera、缨翅目 Thysanoptera、襀翅目 Plecoptera、等翅目 Isoptera、内华达古白蚁 *Zootermopsis nevadensis*，蜉蝣目 Ephemeroptera、鞘翅目 Coleoptera、半翅目 Hemiptera、鳞翅目 Lepidoptera、膜翅目 Hymenoptera 和双翅目 Diptera（表 1）。从目的分布来看，47.17% 的物种为双翅目昆虫（达 100 种），膜翅目占 21.86%，鳞翅目占 11.63%，半翅目占 9.30%，鞘翅目占 4.18%，其他目仅有 1 – 2 种昆虫。在双翅目昆虫中，主要为模式昆虫黑腹果蝇及其近缘种，医学昆虫蚊子等；在膜翅目昆虫中，主要为蚂蚁、蜂等；鳞翅目昆虫主要为重要农业害虫和蝶类。其中，果蝇、蚊子、蚂蚁等三类昆虫占 70% 以上，表明目前昆虫基因组测序仍主要为模式生物和医学昆虫等。

图 2B 显示了 215 种昆虫基因组完成测序或提交序列的时间。统计结果表明，2002 – 2010 年期

间的昆虫基因组测序进展缓慢。2010 年后，在二代测序技术带动下，昆虫基因组测序的物种数大幅增长，这些“旧时王谢堂前燕”，已经“飞入了寻常百姓家”，不再是“高门槛”的项目，越来越多的实验室独立开展了昆虫基因组测序分析（张传溪，2015）。

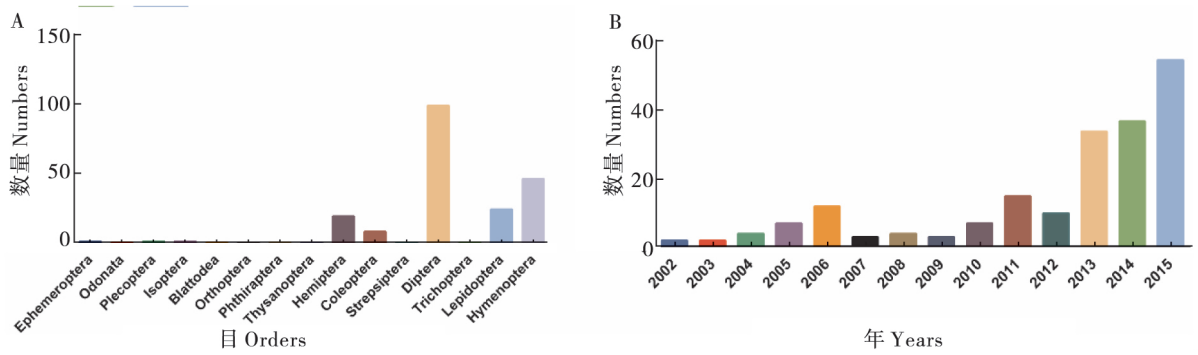

图 2　已发布昆虫基因组统计

Fig. 2　The statistics of insect genomes have been released

从昆虫基因组数据分析来看，由于早期基因组测序是一项艰难的任务，需要庞大的人力和财力投入，基因组工作多限于数据的获得和初步分析，为分子生物学研究提供序列数据。在早期测序物种较少的情况下，比较基因组学难以展开，基因组学数据的威力一时难以完全发挥。近年来，测序物种越来越多，比较基因组分析得以深入开展，从而发现了传统思路无法发现的规律，基因组数据得到了更加充分的挖掘，为解决重要的生物学问题提供了有力的支撑。例如，对褐飞虱基因组的研究揭示了胰岛受体基因在褐飞虱翅型分化中的调控作用。

值得注意的是，在 NCBI 数据库注册的昆虫数要远多于提交序列的昆虫数量，而基因组数据公开发表的数量则更少。其中最为主要的原因之一，是许多昆虫的基因组拼接质量较差，还不适宜于发表。绝大多数昆虫具有非常高的杂合度，导致无法组装出高质量的基因组，影响了基因注释和后续的基因家族分析等。

3.3　重要昆虫的基因组测序及分析

如前所述，目前公开发表的昆虫基因组文章 43 篇涉及物种 46 个，昆虫基因组测序及数据分析的思路大同小异，涉及基因组拼接、注释、基因家族分析等，但针对不同昆虫的特异性表型，不同物种的分析结果各有千秋。在此，选择了一些重要的昆虫并对其基因组测序结果进行简要介绍。

3.3.1　家蚕基因组

家蚕 *Bomyx mori* 基因组于 2004 年完成，是继果蝇、冈比亚按蚊之后的第 3 个昆虫基因组，具有历史性意义。对家蚕 Dazao 品系进行了全基因组鸟枪法测序，基因组大小为 428.7 Mb，拼接后基因组的 contig N50 为 12.9 kb，scaffold N50 为 26.9 kb，共注释了 18510 个基因。基因组分析结果发现，家蚕基因组中含有大量的转座子插入，导致家蚕的某些基因比果蝇中的同源基因更大。在家蚕丝腺中发现了 87 个神经肽激素、激素受体、激素调节相关基因。在家蚕中还发现了 69 个与免疫相关的基因，包括 *moricin*、*cecropins*、*lysozymes*、*hemolin*、*lectins*、*prophenoloxidases* 等。2008 年，国际家蚕基因组联盟对家蚕基因组进行了更新，提高了测序覆盖度，基因组 contig N50 提高为 15.5 kb，scaffold N50 提高到 3.7 Mb，87% 的 scaffold 被定位于 28 条染色体上，预测发现了 14623 个基因。对新版本的基因组进行分析，发现基因组中含大量转座子，包括 LINEs 和 SINEs 两种主要类型，分别占全基因组的 14.5% 和 13.3%。3223 个家蚕特有基因在其他昆虫和脊椎动物中没有发现同源基因。研究还发现，转运 Gly、Ala 和 Ser 的 tRNA 基因数目明显多于其他氨基酸 tRNA，这与蚕丝蛋白中各类氨基酸含量相一致；基因 *Ser*1、*Ser*2、*Ser*3 分别编码蚕丝的不同位置和不同结构的丝胶成分；家蚕在进化过程中通过水平基因转移从细菌中获得呋喃果糖苷酶基因，得以降解桑叶中的 D-AB1、DNJ 等对其他昆虫有毒的生物碱类物质，这是家蚕能够专一取食桑叶的重要原因（Xia *et al.*，2004）。

表 1 已发表的昆虫基因组

Table 1 The published insect genomes

物种 Species	基因组大小（Mb） Genome size	测序平台 Sequencing platform	染色体 Chromosome scaffolds	N50（Kb）	基因数 Gene numbe	来源文献 References
果蝇 *Drosophila melanogaster*	~180	Sanger	15	–	13601	*Science*，2000，287（5461）：2185－2195
冈比亚按蚊 *Anopheles gambiae*	265	Sanger	13042	3800	13683	*Science*，2002，298（5591）：129－149
家蚕 *Bombyx mori*	467	Sanger	1801	3998	14623	*Science*，2004，306（5703）：1937－1940
意大利蜜蜂 *Apis mellifera*	227	Sanger	5644	984	15314	*Nature*，2006，443（7114）：931
埃及伊蚊 *Aedes aegypti*	1434	Sanger	4757	1547	15696	*Science*，2007，316（5832）：1718－1923
赤拟谷盗 *Tribolium castaneum*	222	Applied Biosystems	2322	975	16526	*Nature*，2008，452（7190）：949－955
致倦库蚊 *Culex quinquefasciatus*	562	454 FLX	3172	478	19032	*Science*，2010，330（6000）：86－88
佛罗里达弓背蚁 *Camponotus floridanus*	228	454 FLX	24026	441	15069	*Science*，2010，329（5995）：1068－1071
跳蚁 *Harpegnathos saltator*	288	454 FLX	21347	598	18563	
豌豆芽 *Acyrthosiphon pisum*	525	Sanger	23925	518	36194	*PLoS Biol.*，2010，8（2）：e1000313
体虱 *Pediculus humanus*	108	454 FLX	1882	497	11664	*PNAS*，2010，107（27）：12168－12173
金小蜂 *Nasonia vitripennis*	287	454 FLX	6181	708	18940	*Science*，2010，327（5963）：343－348
切叶蚁 *Acromyrmex echinatior*	289	454 FLX	4339	1110	17277	*Genome Res.*，2011，21（8）：1339－1348
入侵阿根廷蚁 *Linepithema humile*	213	454 FLX Titanium	3030	1428	16115	*PNAS*，2011，108（14）：5673－5678
红收获蚁 *Pogonomyrmex barbatus*	228	454 XLR	4646	816	17737	*PNAS*，2011，108（14）：5667－5672
大头切叶蚁 *Atta cephalotes*	309	Roche 454	2835	5154	18092	*PLoS Genetics*，2011，7（2）：e1002007
火蚁 *Solenopsis invicta*	343	Roche 454	10543	720	16521	*PNAS*，2011，108（14）：5679－5684
帝王蝶 *Danaus plexippus*	238	454 FLX/titanium platform	5397	715	15129	*Cell*，2011，147（5）：1171－1185
诗神袖蝶 *Heliconius melpomene*	266	454 and Illumina	4309	194	15201	*Nature*，2012

续上表

物种 Species	基因组大小 (Mb) Genome size	测序平台 Sequencing platform	染色体 Chromosome scaffolds	N50 (Kb)	基因数 Gene number	来源文献 References
松甲虫 *Dendroctonus ponderosae*	246	Illumina Hiseq	8188	628	13456	*Genome Biol.*, 2013, 14 (3): R27
隧蜂 *Lasioglossum albipes*	350	Illumina	4317	616	13448	*Genome Biol.*, 2013, 14 (12): R142
小菜蛾 *Plutella xylostella*	383	Illumina Hiseq2000	1819	737	18072	*Nature Genetics*, 2013, 45 (2): 220 – 225
榕小蜂 *Ceratosolen solmsi*	268	Illumina Hiseq2000	2457	9558	13200	*Genome Biol.*, 2013, 14 (12): R141
南极蠓 *Antarctic midge*	99	Illumina	3589	98	13517	*Nat. Commun.*, 2014, 54611
无性生殖行军蚁 *Cerapachys biroi*	206	Illumina Hiseq2000	4579	1350	26315	*Curr Biol.*, 2014, 24 (4): 451 – 458
家蝇 *Musca domestica*	728	Illumina	20487	226	20165	*Genome Biol.*, 2014, 15 (10): 466
竹节虫 *Stick insect*	1027	Illumina	14211	312	23083	*Science*, 2014, 344 (6185): 738 – 742
湿木白蚁 *Zootermopsis nevadensis*	472	Illumina Hiseq2000	31622	751	14610	*Nat. Commun.*, 2014, 53636
蝗虫 *Locusta migratoria*	6300	Illumina Hiseq2000	–	320	17307	*Nat. Commun.*, 2014, 52957
褐飞虱 *Nilaparvata lugens*	1324	Illumina Hiseq2000	45279	360	36723	*Genome Biol.*, 2014, 15 (12): 521
草地贪夜蛾 *Spodoptera frugiperda*	358	Illumina	37243	53.7	11595	*Genomics*, 2014, 104 (2): 134 – 143
麦双尾蚜 *Diuraphis noxia*	421	Illumina Hiseq2000	5641	397	19097	*BMC Genomics*, 2015, 16 (1): 429
咖啡果小蠹 *Hypothenemus hampei*	163	Illumina Hiseq2000	86848	44.7	19222	*Rep.*, 2015, 512525
铜绿蝇 *Lucilia cuprina*	458	Illumina – ALLPATHS – LG	4625	744	14554	*Nat. Commun.*, 2015, 67344
冬尺蠖蛾 *Operophtera brumata*	638	Illumina Miseq	25801	65.6	16912	*Genome Biol. Evol.*, 2015, 7 (8): 2321 – 2332
温带臭虫 *Cimex lectularius*	650	Illumina – ALLPATHS – LG	1402	7172	14220	*Nat. Commun.*, 2016, 710165
地中海实蝇 *Ceratitis capitata*	479	Illumina	1806	4060	14547	*Genome Biol.*, 2016, 17 (1): 192

3.3.2 蜜蜂基因组

蜜蜂 *Apis mellifera* 基因组由 The Honeybee Genome Sequencing Consortium 团队于 2006 年完成。文章先后注释了六版基因组数据，将 contig N50 从 19 kb 提高到 41 kb，scaffold N50 从 223 kb 提高到 362 kb。基因组大小 236 Mb。基因组注释获得了 10157 个基因，比果蝇和库蚊少 25% 左右。蜜蜂基因组为 AT-rich，高达到 67%，而黑腹果蝇 *Drosophila melanogaster* 仅为 58%，库蚊仅为 56%。在蜜蜂基因组 AT 丰富区中，基因分布反而较多，这与脊椎动物明显不同。蜜蜂基因组中的转座子明显比其他昆虫更少。蜜蜂和果蝇只有 10% 同源基因，远少于人和鸡之间有 85% 同源基因的比例，表明昆虫的进化速度很快。蜜蜂有 163 个气味受体基因，远多于果蝇（62）和库蚊（79），显示蜜蜂化学感受能力增强，用来探测外激素、辨别同伴和花香等。与此相反，蜜蜂的味觉基因只有10 个，少于其他昆虫的 50 – 76 个。与预期相反，蜜蜂免疫和抗病基因明显变少，只有 71 个与免疫和抗病相关的基因，远少于库蚊的 209 和果蝇的 196 个，分析认为这与蜜蜂的清洁行为、蜂王浆和蜂胶的抗细菌特性，以及蜂群像城堡一样的结构等有关。研究还发现，与果蝇不同，蜜蜂有完整的 DNA 甲基化酶系，包括 *Dnmt*1、*Dnmt*2 和 *Dnmt*3，DNA 甲基化在蜜蜂不同蜂型的分化中具有重要的功能（Consortium，2006）。

3.3.3 体虱基因组

体虱 *Pediculus humanus* 基因组于 2010 年完成，其基因组大小仅为 108 Mb，拼接获得的基因组 scaffold N50 为 488 kb。预测发现了 10773 个蛋白编码基因和 57 个 microRNAs。与其他昆虫基因组相比，体虱具有更少的与环境感知和响应相关的基因，包括那些嗅觉和味觉感受器以及解毒酶编码的基因等。同时，还对体虱 Riesia 菌的基因组进行了测序。Riesia 菌是体虱消化道中的一种关键细菌，它分泌营养物质作为人血的补充物质，Riesia 细菌缺乏抵抗抗生素的基因。比较基因组学分析显示，人类体虱是从头虱进化而来的，基因组分析有助于利用体虱的独特基因属性如其有限的嗅觉能力等，开发出体虱控制的新方法（Kirkness *et al.*，2010）。

3.3.4 豌豆蚜基因组

豌豆蚜 *Acyrthosiphon pisum* 由国际蚜虫基因组联盟于 2010 年完成。作者利用单个雌虫的个体后代进行测序，流式细胞仪估测基因组大小为 517 Mb，测序组装获得的基因组为 464 Mb，基因组 contig N50 为 10.8 kb，scaffold N50 为 88.5 kb，注释获得了 34604 个基因，远多于其他昆虫的 15000 – 20000 个，其中 2459 个基因家族中发现大量的基因复制，等义距离评估表明在该物种形成初期已经存在了基因复制现象，涉及功能包括染色质修饰、miRNA 合成和糖转运等。豌豆蚜基因组丢失了 IMD（免疫缺陷）免疫通路、硒蛋白利用、嘌呤补救途径及鸟氨酸循环等通路的基因。通过与蚜虫初级内共生菌 *Buchnera aphidicola* 基因组比较分析，发现两者具有代谢系统的互补性。豌豆蚜基因组中具有明显的基因横向转移现象，部分基因与细菌基因具有共同起源，其线粒体基因亦有部分在基因组中重复。基因组中发现了 12 个新的 *dynamin* 基因，可能与病毒运输、转胞等过程相关。豌豆蚜基因组中胚胎发育相关基因存在特异性的缺失，可能与其发育多型性有关。基因组中锌指结构蛋白的扩增，以及保幼激素合成酶、降解酶的 hexamerin 的缺失可能与豌豆蚜发育可塑性有关。

3.3.5 丽蝇蛹集金小蜂基因组

丽蝇蛹集金小蜂 *Nasonia vitripennis* 是双翅目蝇类的重要寄生蜂，其基因组测序完成于 2010 年。作者采用了 Sanger 测序法获得 26605 条 contigs（N50 = 18.5 kb），6181 条 Scaffolds（N50 = 709 kb），基因组大小约 295 Mb。同时对另两种近缘寄生蜂 *N. giraulti* and *N. longicornis* 采用了 Sanger 测序技术和 Illumina 测序平台进行测序，得用 *N. vitripennis* 基因组做为参考，分别有 62% and 62.6% 的 reads 比对到 *N. vitripennis* 基因组上，有 84.7% 和 86.3% 的蛋白编码区域。在 *N. vitripennis* 基因组中，注释到 17279 个基因，并预测了 52 个 miRNA 基因。研究发现，金小蜂具有完整的 DNA 甲基化“工具包”，即含有三种 DNA 甲基化基因，并且 *Dnmt1* 具有 3 个拷贝。*N. vitripennis* 基因组的 Toll 通路中发现大量的基因复制。在 *N. vitripennis* 基因组中，性别决定相关基因如 *yellow/major*、*royal*、*jelly* 基因等，表现出大量的复制；*N. vitripennis* 基因组具有与细菌 *Wolbachia* 基因相似的保守域，表明细菌基因被整合宿主基因组中，发生了基因转移现象；丽蝇蛹集金小蜂的毒液蛋白基因受到很高的进化压力。作者分析还发现，3 种金小蜂线粒体基因在不同的世代受到

了比较显著的进化压力（Werren *et al.*，2010）。

3.3.6　帝王蝶基因组

帝王蝶 *Danaus plexippus* 基因组于 2011 年完成，是目前唯一一篇发表于 Cell 杂志的昆虫基因组。帝王蝶具有迁徙和不迁徙两种类型，最早起源于美国南部和墨西哥北部的是迁徙型，大约两万年前数量增长开始迁移，向南进入南美，直到近期北美类群又分为跨太平洋和跨大西洋两个方向分布于全球各地。作者利用二代测序平台通过全基因组鸟枪法测序得到了 14.7 Gb 的 Illumina reads，经拼接得到了 273 Mb 的帝王蝶基因组，注释发现了 16866 个蛋白编码基因。对 12 种昆虫和 2 种哺乳动物基因组进行了同源分析，结果表明鳞翅目是目前为止进化最快的昆虫；帝王蝶和家蚕在直系同源数量、微共线性、蛋白家族大小等方面具有明显的相似性。通过对帝王蝶基因组的分析，更深入地破解了其迁飞的分子机制。在帝王蝶基因组中发现了可能与处理光信号和太阳罗盘结构有关的多种蛋白和神经递质，并注释了39 个与定位功能相关的基因，其中 2 个功能未知的基因可能是帝王蝶特有的。位于帝王蝶触角的生物钟在迁徙活动中具有重要作用，分析发现帝王蝶除了具有大量和果蝇相同的生物钟关键基因外，还具有 CRY2 基因，而果蝇只含有 CRY1 基因。保幼激素的生物合成在帝王蝶雌雄中具有两态性，表现为在雌性上调、雄性下调。研究还发现，miR-1、miR-7、miR-14 在内的 27 种 miRNA 在迁徙和非迁徙蝴蝶中的表达量有差异，可能对迁飞起调节作用。独特的 P 型钠钾泵构成了帝王蝶防御机制的分子基础，而 Ors、Grs、IRs 等化学感受器在迁飞过程中也有潜在的作用。

此后，该团队采集了不同地区的 101 个帝王蝶基因组进行了重测序分析。在与迁徙相关的 5 Mb 序列中，有大约 21 kb 的异常序列，这段序列包含 3 个基因，其中 Collagen IV α-1 在迁徙和非迁徙群体之间具有明显的不同，从而影响了 2 种类型蝴蝶的体型、飞行肌以及飞行特点的不同。相比之下，迁徙蝴蝶飞行代谢率低，飞行效率高；高代谢率更有利于非迁徙蝴蝶的生存。帝王蝶特有的警戒色被发现与肌球蛋白基因 DPOGS206617 有密切关系，表明翅色并非由色素分子的产生决定而是由色素的运输来决定（Zhan *et al.*，2011）。

3.3.7　小菜蛾基因组

小菜蛾 *Plutella xylostella* 是世界性的重要害虫，食性广，危害严重，容易对农药形成抗性，基因组大小仅为 343 Mb，但其杂合度高，导致测序困难，其基因组于 2013 年完成测序，是第一个成功测序的高杂合度昆虫基因组。作者利用 Illumina Genome Analyzer IIx 和 HiSeq2000 平台，采用 Fosmid-to-Fosmid 结合 WGS 的测序策略，最终获得了 1819 条 scaffold 序列，N50 为 737 kb。基因组注释获得了 18071 个基因和 781 ncRNA。比较基因组学分析发现，小菜蛾基因组中有 1412 个特有基因，参与感知和解毒代谢的基因家族发生了明显的扩张。基因组数据分析发现了在幼虫阶段偏好表达的 354 个基因，部分基因参与硫酸盐代谢及硫酸酯酶修饰因子基因。其中，硫代葡萄糖苷硫酸酯酶（GSS）通过催化硫代葡萄糖苷防御化合物转化为脱硫葡萄糖苷酸酯，使得小菜蛾能够在广泛的十字花科植物上进食，从而防止毒性水解产物的形成。分析认为，小菜蛾硫代葡萄糖苷硫酸酯酶（GSS）基因和硫酸酯酶修饰因子基因 1（*SUMF*1）在幼虫时期的协同表达是决定小菜蛾能够取食十字花科蔬菜的关键。除细胞色素（P450）、谷胱甘肽转移酶（GST）和羧基酯酶（COE）这三大代谢水解酶家族外，ABC 转运蛋白家族也出现了明显的扩张，进一步解释了小菜蛾容易产生抗性的基因组学特性（You *et al.*，2013）。

3.3.8　榕小蜂基因组

榕小蜂 *Ceratosolen solmsi* 在长期进化过程中，与榕属植物形成了一种密切的共生关系，是榕属植物重要的传粉媒介，以回报榕属植物为其提供栖身场所和营养来源。榕小蜂基因组于 2013 年完成测序和发表，其基因组大小 278 Mb，scaffold 数量 7397。值得一提的是由于其基因组中富含 AT（69.6%），重复序列只有 9.85%，因此组装完成后 scaffold N50 值竟达到 9.558 Mb，是目前测序昆虫中最高的。通过从头预测、同源搜索、转录组覆盖等方法，共注释获得蛋白质编码基因 11412 个。

通过比较基因组分析，发现榕小蜂的基因组进化相比于其他昆虫更快。由于榕小蜂基本上大部分时间都栖息在榕树，其基因组中 ORs、GRs、IR、OBPs、CSPs 等化学感受基因家族出现明显的收缩。由于榕树已为榕小蜂提供了安全的场所和营养来源，因此其 P450s、GSTs、CCEs 等解毒代谢基因家族基因也明显减少，以及在 Toll、imd、JAK/STAT、JNK 等免疫通路中很多基因退化。为

了了解榕小蜂雌雄异型的分子机制，通过转录组测序技术研究了其雌雄个体中基因的表达情况，发现了很多与基因在雌雄个体中出现差异表达，推测与其这种两性差异有关。榕小蜂在长期与肠道共生菌协同进化过程中，通过基因组数据证实其可以从细菌和病毒中获得一些基因片段或完整基因，总共在榕小蜂基因组鉴定出 12 个水平转移基因（Xiao *et al.*，2013）。

3.3.9 蝗虫基因组

蝗虫 *Locusta migratoria* 是世界范围的具有严重危害性的昆虫，其周期性的大爆发，具有长距离迁飞和两型变化的习性。蝗虫基因组达 6.52 Gb，是迄今为止最大的昆虫基因组，因此完成测序极其困难，来自中国科学院动物所康乐院士所带领的团队于 2014 年首次解开了蝗虫的遗传密码，破解了这一难题。蝗虫基因组 scaffold N50 为323 kb，通过从头预测、同源预测以及表达证据共获得 17307 个蛋白质编码基因。基因组分析发现，蝗虫的基因组之所以如此之大，主要体现在重复序列增多，占基因组 60% 以上，蝗虫基因内含子的长度是其他昆虫的 10 倍左右，这也是造成其基因组变大的一个重要因素。

通过比较基因组学研究，发现了大量与变态发育相关的调控基因，蝗虫进化获得了 55 个新的基因家族，共有 25 个基因家族显著扩增，参与解毒代谢、化学感受、营养代谢等。蝗虫具有 *Dnmt*1 两个以及 *Dnmt*2 和 *Dnmt*3 完整的 DNA 甲基化基因家族，基因组中约有 1.6% 的胞嘧啶被甲基化，重复序列区高度甲基化。与基他昆虫不同的是，基因内含子区甲基化高于外显子区。为了适应长距离迁飞，蝗虫进化出一套高效的能量储存和代谢的机制，其主要能源物质为脂类，基因组中与脂类运输和抗氧化保护以及脂质降解有关的基因家族显著扩增，如基因组中 *perilipins*、*fatty-acid-binding protein*、*Prdx*6*s*、*sigma GST*、*enoyl-CoA hydratase*、*acetyl-CoA acyltransferase* 2 等基因出现多拷贝。蝗虫基因组中 *OBPs*、*ORs*、*GRs*、*IRs* 等基因家族出现显著的扩增，可能与其食性很广有关，同时 *UGTs* 和 *carboxyl/choline esterases* 基因家族也出现显著扩增，以帮助其降解不同食物中的化学成分。

3.3.10 家蝇基因组

家蝇 *Musca domestica* 是生活中常见的昆虫，幼虫以动物排泄物等为食，成虫能够携带 100 多种病原菌，对人类和动物的健康带来极大的威胁，其基因组测序于 2014 年完成。家蝇基因组大小 691 Mb，重复序列含量较高，Scaffold 数为 20487，N50 值为 226 kb，基因组注释获得蛋白质编码基因 14180 个。在家蝇基因组中共发现 771 与免疫相关的基因，具有完整的 Toll、imd、JAK/STAT 和 JNK 免疫通路，这与家蝇长期生活在富含动物病原体腐烂性环境有关。先后从基因组找到 146 个 P450s、11 个 P450 pseudogenes、33 个 GSTs、92 个脂酶基因，显示家蝇基因组中解毒代谢相关的基因家族出现了明显扩张，以应对生境中各种有害物质。家蝇基因组中 *CysLGIC* 超基因家族具有 23 个基因，为抗药性研究和农药新靶点开发提供了参考。家蝇的味觉受体基因家族显著出现扩增，推测与家蝇需要通过味觉来识别不同的有害物质有关（Scott *et al.*，2014）。

3.3.11 南极蠓基因组

南极蠓 *Belgica antarctica* 是唯一生活在南极的一种地方性昆虫，需要适应极端温度、结冰、脱水、渗透压平衡、紫外线辐射以及环境产生的其他各种选择压力，其基因组测序于 2014 年完成。南极蠓基因组大小 89.6 Mb，是目前最小的昆虫基因组。其 Contig 序列为 5003 条，N50 值为 98.2 kb。虽然拼接质量不高，CEGMA 基因组评估和比较基因组学研究表明南极蠓的基因组数据可以用于后续数据分析，预测得到蛋白质编码基因 13517 个。相比于其他昆虫，重复序列含量的大幅减少，内含子长度变短，这是其南极蠓基因组明显变小的主要原因。通过基因组个体杂合度分析发现，由于其基因组比较小，南极蠓受到的选择压力非常大，因此杂合度相对其他昆虫低。基因家族分析显示南极蠓 OBP 基因出现明显的收缩，推测与其生活环境、食物相对单一，活动范围也较小等习性有关（Kelley *et al.*，2014）。

3.3.12 褐飞虱基因组

褐飞虱 *Nilapavata lugens* 是水稻上的重要害虫，具有迁飞习性和翅二型现象，其基因组测序完成于 2014 年。作者采用 HiSeq2000 测序技术，利用单对交配纯化 13 代的褐飞虱，使用与小菜蛾相似的测序策略，得到了共 1.14 Gb 的褐飞虱基因组序列，基因组 Scaffold N50 为 356.6 kb，注释得到 27571 个蛋白编码基因。通过对褐飞虱和其它 14 个节肢动物基因组的比较分析，发现褐飞虱等半翅目的 3 个物种基因数目、特异基因数目都比

其他昆虫多，显示出半翅目物种的基因扩张现象。褐飞虱的 OR 和 GR 基因家族收缩，这与褐飞虱只以水稻韧皮汁液为食的严格单食性特性相符；研究还发现褐飞虱中解毒和消化相关基因存在着基因丢失现象，如 P450、GST 基因数目很少，淀粉降解必须的 alpha-淀粉酶缺失，几丁质合成酶 CHS2 缺失，这些特点也可能与褐飞虱专一食性有关；褐飞虱与真菌 YLS 和细菌 *A. nilaparvatae* 组成了共生系统，通过对真菌 YLS 和细菌 *A. nilaparvatae* 测序并组装注释，分析三者的共生关系，发现褐飞虱缺少 10 种必需氨基酸合成能力，而在 YLS 中能找到对应的氨基酸合成基因；还发现 YLS 能够利用尿酸，跟褐飞虱共同形成了氮素循环的完整途径；YLS 能合成酵母甾醇中间产物，褐飞虱参与利用酵母甾醇中间产物进一步合成胆固醇，从而形成完整的胆固醇合成途径；YLS 和褐飞虱在维生素生物合成途径上都有缺陷，但 *A. nilaparvatae* 带有完整的维生素 B 合成途径，可能为褐飞虱提供维生素（Xue，*et al.*，2014）。

3.3.13　臭虫基因组

臭虫 *Cimex lectularius* 是与人类健康密切相关的皮外寄生物，其基因组于 2016 年完成。作者首先臭虫对经过 6 代近交纯化，然后采用二代 Illumina Solexa 平台测序，基因组大小为 650.47 Mb，拼接得到 1402 条 scaffold 序列，scaffold N50 为 7.17 Mb，MAKER 软件预测和手工注释共获得 14220 个蛋白质编码基因。基因组分析表明，为了适应臭虫独特的生态环境和生活习性，很多基因或基因家族出现了丢失或扩张。与臭虫专性寄生习性相关，在黑暗环境生存使得 CRY1 与 JET 感光基因退化，气味受体、味觉受体、离子受体等化学感受基因以及免疫通路相关基因均出现了显著的基因家族收缩；臭虫的专性吸血习性使得其唾液蛋白家族扩增，以阻止在吸食过程中的寄主血液凝固，水通道蛋白（AQP）的扩增可以快速去除血液中大量的水分；臭虫具有皮下受精交配习性，在基因组中节肢弹性蛋白基因大量扩增，使得雌虫可以最大限度地免于交配产生的创伤或修复创伤。臭虫抗药性发展迅速，基因组分析发现臭虫的电压门控钠通道基因出现了多个点突变使得靶标不敏感；差异表达分析发现 P450、羧酸酯酶、谷胱甘肽-S-转移酶等代谢酶基因的表达增强，ABC 转运蛋白基因家族扩增，CPR 家族基因扩增等均是造成了臭虫日趋严重抗性的原因。通过微生物和寄主分析，发现了臭虫与其体内 walbacia 菌形成营养共生关系，在臭虫基因组发现了 805 个潜在的水平转移基因。臭虫基因组使得从分子机制水平研究和解释臭虫的寄生习性、嗜血习性、抗药性等科学问题成为可能，为研究吸血昆虫、共生关系以及寄生行为等提供了新的模式材料（Benoit，*et al.*，2016）。

3.3.14　地中海实蝇

地中海实蝇 *Ceratitis capitata* 是世界性的入侵害虫，其基因组大小为 479 Mb，基因组测序完成于 2016 年。作者先后采用 454 平台和 Illumina HiSeq2000 平台进行测序，利用单对纯化后的个体 DNA 进行测序以提高数据质量，将 contig N50 从 3.1 kb 提高到 45.8 kb，Scaffold N50 从 29.4 kb 提高到 4.1 Mb。基因组注释获得 14547 个基因，23075 个 CDS。与其它 14 个节肢动物的基因组进行同源分析，确定了 26212 个同源组。地中海实蝇中有 1608 条推定的氨基酸序列没有分到任何同源组内，推测是最近才进化的新基因。利用地中海实蝇的唾液腺多线染色体，通过克隆基因和微卫星序列（Medflymic）的原位杂交，将克隆基因和微卫星序列所在的 43 个 scaffold 定位到 5 条常染色体上（染色体 2－6 号），1 个 scaffold 定位到 X 性染色体上。与黑腹果蝇和家蝇基因组进行比较分析，发现多个基因/基因家族的扩张现象可能导致地中海实蝇较高的适应性和入侵性，包括 *IR* 和 *GR* 味觉受体基因家族、性诱剂受体、细胞色素 *P450* 基因和 *CYP6* 亚家族、免疫系统基因（*Toll* 和 *spätzle* 家族）、*TWDL* 和 *CPLCA* 表皮蛋白家族、水通道蛋白基因以及特异的 *ceratotoxin* 基因。对各基因家族的分析表明，可利用化学感受分子作为种群监测或诱捕的引诱剂或驱避剂，视蛋白 opsin 指导最佳陷阱颜色的选择，RHG 促细胞凋亡基因（*reaper*、*grim*）、精液蛋白 SFP 等用于 SIT 昆虫不育技术（Papanicolaou *et al.*，2016）。

4 昆虫基因组数据库

随着测序技术的突破性发展，海量的生物数据在不断累积，每 14 个月就会增长一倍，如何进行数据的管理、存储、展示、共享，变成了非常迫切的问题（Baxevanis *et al.*，2015，Stephens *et al.*，2015）。为了最大化地体现数据的价值和提高数据的利用率，数据库在管理和维护、共享与

挖掘生物大数据中发挥着重要作用。

依据数据资源分类，生物数据库可以分为三类。第一类是大型综合存储型数据库。这类数据库的特点就是，大而杂地收录了大量的数据，数据之间层次和质量良莠不齐，且仅仅是接近原始版的堆积，更新、修改和管理较为困难，而且数据库比较大，维护的成本很高，主要是发挥数据仓库的作用。这类数据库以美国国家生物技术信息中心（NCBI）、欧洲生物信息研究所（EBI）和日本核酸数据库（DDBJ）国际上公认的三大生物信息数据库为代表，这三个数据库各具特色。第二类是单一类群的基因组数据库。这类数据库是围绕某一个研究类群的基因组数据库，数据量较第一类数据库明显缩小，数据之间的层次和质量比较接近，且质量有所保证，数据也经过了加工，维护者管理起来也比较方便，使用者用起来也可以很快的掌握。VectorBase（Giraldo-Calderon *et al.*，2015）是这类型数据的经典代表，其中收录了与众多与疾病媒介传播有关物种的基因组数据。第三类是小型的单个物种或单一属的物种数据库，围绕单一物种的数据构建数据库，数据质量很高，数据加工很精细，功能很齐全，维护和更新迅速和简便，使用便捷。这类数据库目前有膜翅目数据库 Hymenoptera Genome Database（Munoz-Torres *et al.*，2011）、农业害虫数据库 Agripestbase、小菜蛾数据库（中国）DBM-DB（Tang *et al.*，2014）、小菜蛾数据库（日本）KONAGAbase（Jouraku *et al.*，2013）、帝王蝶数据库 MonarchBase（Zhan *et al.*，2013）、蚜虫数据库 APHIDBASE（Legeai *et al.*，2010）、家蚕数据库（中国）SilkDB（Duan *et al.*，2010，Wang *et al.*，2005）、家蚕数据库（日本）KAIKObase（Shimomura *et al.*，2009）、诗神袖蝶数据库 Heliconius Genome Project、二化螟数据库 ChiloDB（Yin *et al.*，2014）和 WaspAtlas 金小峰数据库（Davies *et al.*，2015）。

目前昆虫基因组数据主要存储于大型综合存储型数据库中。NCBI 共收录了 215 个昆虫的基因组拼接数据，Ensemble 上收录了 31 个，这两个公共数据库涵盖了大部分的昆虫基因组数据。由于 NCBI 等大型数据库并不是单一地为昆虫领域服务，主要集中在医学、模式生物领域。目前 NCBI 基本没有针对昆虫基因组数据进行挖掘和数据注释等，仅仅只是数据仓库服务。为此，这么多昆虫基因组研究者纷纷建立了单个类群或单个个体的基因组数据库（表 2），在众多的昆虫基因组数据库，涌现了 2 个综合型的昆虫基因组数据库，i5k workspace@ NAL（Poelchau *et al.*，2015）和 InsectBase（Yin *et al.*，2016）。

4.1 i5k Workspace@NAL

i5k Workspace@ NAL 数据库是由美国农业部主导构建的节肢动物基因组学服务型数据库，共收录昆虫基因组 46 个，数据库提供基因组数据的浏览、下载、数据提交、序列比对、基因组可视化及在线基因组手工注释平台，以及 HMMER、CLUSTAL 两个在线工具（Poelchau *et al.*，2015）。随着 i5k 计划的提出，越来越多的节肢动物基因组被测序。在此背景下，美国农业部相关科学家希望在纷乱无章的测序潮流中推出一套基因组测序、组装、注释、维护、共享的标准化流程和平台，因此构建了 i5k Workspace@ NAL 数据库。然而事与愿违，在目前基因组数据依旧是稀缺资源的环境下，大多数研究人员没有遵从 i5k Workspace@ NAL 提出的共享数据标准。目前，i5k Workspace @ NAL 主要收录了美国农业部主导的一些节肢动物基因组测序数据，其他国家科学几乎没有提交数据。

4.2 InsectBase

InsectBase 昆虫基因组与转录组数据库旨在有效的解决目前昆虫基因组数据库的纷乱杂陈的现状，构建一个综合的全能化的昆虫领域的生物信息数据库，为广大研究者提供方便快捷的后基因组时代基因组、转录组等数据服务和交流合作平台（Yin *et al.*，2016）。

InsectBase 昆虫基因组数据库（http://www.insect-genome.com/）的总数据存储量达 120 G。InsectBase 通过筛选和质量过滤共收集了 155 种昆虫基因组（隶属于 16 个目），其中 61 个基因组具有注释信息（Official Gene Set，OGS），116 个转录组数据，237 个物种的 EST 序列，69 个物种的 7544 条 miRNA 序列，2 个物种的 83262 条 piRNA 序列，构建了 78 个物种的 22536 个信号通路，116 个昆虫的 UTR 序列和 CDS 序列。针对 61 个有 OGS 注释的昆虫，开展了数据挖掘。

InsectBase 对研究较多的 36 个基因家族开展了系统分析，运用 OrthoMCL 直系同源算法发现了 7 个物种中的直系同源基因，共找到 1∶1∶1 直系同源基因 973 个。InsectBase 昆虫基因组数据库提

供序列查询、序列比对、基因组可视化、信号通路和注释、进化分析和进化树构建等功能服务，所有基因数据均可下载。从 PubMed 中下载了 94758 条昆虫研究相关文献，通过数据挖掘，建立了昆虫学领域的关系网络平台 iFacebook，初步实现“基因 - 研究者 - 昆虫物种”等三者之间的关系网络，便于促进学术交流。InsectBase 是综合型的生物信息学数据库，数据种类齐全、功能全面、用户使用方便，有利于昆虫学研究者对基因数据的获得、整理和分析，促进昆虫分子生物学研究。自 2015 年 8 月上线以来，到目前已经累计有来自全世界 86 个国家的研究学者近 10 万次的访问，其中最活跃的当属中国和美国，中国的访问量占到 86. 23% 。

表 2　昆虫基因组数据库统计

Table 2　Insects genome database statistics

数据库 Database		生物种数 Species	Genome with Gene Sets	链接 URL
NCBI (GeneBank & Refseq)		215	112	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
WaspAtlas		1	1	http: //waspatlas. com/
InsectBase		138	61	http: //www. insect - genome. com

5　总结与展望

随着测序费用的急剧下降，昆虫基因组测序计划如雨后春笋般地涌现。由于昆虫基因组杂合度高导致的拼接困难等问题，在 2020 年前完成 5000 种昆虫测序的目标也许很难实现，但随着技术的进步，这些困难最终会得到彻底解决。对 948 种昆虫基因组大小进行统计分析，结果显示平均大小为 1. 15 Gb，按 1000 美元完成人基因组 (3 Gb) 测序来计算，完成一个昆虫基因组的测序仅需不到 400 美元。相信在不久的将来，昆虫基因

组测序和重测序将成为日常实验设计的一部分。

组学数据的大量积累，将会对昆虫学研究起巨大的推动作用。首先，系统生物学的研究思路将占据昆虫分子生物学研究的高地，研究人员不仅仅将基因组作为数据仓库在使用，而且可以从组学角度寻找重要科学问题的答案，才是功能基因组学研究时代的突破性飞跃。其次，生物数据的积累对生物信息学提出了更高的要求。目前，数据分析工作主要依赖于公司的技术人员完成，但是常规的通用分析流程将越来越不能胜任具有针对性的数据分析需求，生物信息学技术将如同上世纪 90 年代末的分子生物学技术一样，成为每一个实验室的重要技术平台。因此，昆虫学研究中应该注重培养既懂昆虫学问题也熟悉生物信息学分析的两栖人才。最后，基因组重测序、转录组、蛋白组和代谢组等将成为功能基因组时代的四驾马车，将 DNA、RNA、蛋白质和代谢产物4 个不同层次的大数据充分整合，是功能基因组时代的重要研究手段。

在昆虫基因组学研究中，还应当注意和明确的是，数据和技术应该为科学问题服务。昆虫基因组数据的大量堆积，数据质量良莠不齐，需要提高和发展；技术层面上的问题重重，需要实现突破。他山之石，可以攻玉。昆虫基因组研究可以并应当借鉴医学研究领域的领先技术和思路，但技术的突破和数据的提高，应该紧密围绕昆虫科学问题，服务于害虫防治和益虫利用的最终目标。

参考文献(References)

Adams MD, Celniker SE, Holt RA, *et al.* The genome sequence of Drosophila melanogaster [J]. *Science*, 2000, 287 (5461): 2185 – 95.

Allen JE, Majoros WH, Pertea M, *et al.* JIGSAW, GeneZilla, and GlimmerHMM: Puzzling out the features of human genes in the ENCODE regions [J]. *Genome Biol.*, 2006, 7 (S9): 1 – 13.

Bao Z, Eddy SR. Automated de novo identification of repeat sequence families in sequenced genomes [J]. *Genome Res.*, 2002, 12 (8): 1269 – 1276.

Baxevanis AD, Bateman A. The importance of biological databases in biological discovery [J]. *Curr Protoc Bioinformatics*, 2015, 50111 – 50118.

Bellaousov S, Reuter JS, Seetin MG, *et al.* RNAstructure: Web servers for RNA secondary structure prediction and analysis [J]. *Nucleic Acids Res.*, 2013, 41 (Web Server issue): W471 – 474.

Benoit JB, Adelman ZN, Reinhardt K, *et al.* Unique features of a global human ectoparasite identified through sequencing of the bed bug genome [J]. *Nat. Commun.*, 2016, 710165.

Butler J, MacCallum I, Kleber M, *et al.* ALLPATHS: De novo assembly of whole – genome shotgun microreads [J]. *Genome Res.*, 2008, 18 (5): 810 – 820.

Cantarel BL, Korf I, Robb SM, *et al.* MAKER: An easy – to – use annotation pipeline designed for emerging model organism genomes [J]. *Genome Res.*, 2008, 18 (1): 188 – 196.

Chen Y, Liu YS, Zeng JG, *et al.* Progresses on plant genome sequencing profile [J]. *Life Science Research Feb.*, 2014 (1): 66 – 74.

Consortium HGS. Insights into social insects from the genome of the honeybee *Apis mellifera* [J]. *Nature*, 2006, 443 (7114): 931.

Davies NJ, Tauber E. WaspAtlas: A Nasonia vitripennis gene database and analysis platform [J]. *Database* (*Oxford*), 2015.

Duan J, Li R, Cheng D, *et al.* SilkDB v2.0: A platform for silkworm (*Bombyx mori*) genome biology [J]. *Nucleic Acids Res.*, 2010, 38 (Database issue): 453 – 456.

Edgar RC, Myers EW. PILER: Identification and classification of genomic repeats [J]. *Bioinformatics*, 2005, 21 (Suppl): 152 – 158.

Elsik CG, Mackey AJ, Reese JT, *et al.* Creating a honey bee consensus gene set [J]. *Genome Biol.*, 2007, 8 (1): R13.

Friedlander MR, Chen W, Adamidi C, *et al.* Discovering microRNAs from deep sequencing data using miRDeep [J]. *Nat. Biotechnol.*, 2008, 26 (4): 407 – 415.

Giraldo – Calderon GI, Emrich SJ, MacCallum RM, *et al.* VectorBase: An updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases [J]. *Nucleic Acids Res.*, 2015, 43 (Database issue): 707 – 713.

Heather JM, ChainB. The sequence of sequencers: The history of sequencing DNA [J]. *Genomics*, 2016, 107 (1): 1 – 8.

Jouraku A, Yamamoto K, Kuwazaki S, *et al.* KONAGAbase: A genomic and transcriptomic database for the diamondback moth, *Plutella xylostella* [J]. *BMC Genomics*, 2013: 14464.

Kelley JL, Peyton JT, Fiston – Lavier AS, *et al.* Compact genome of the Antarctic midge is likely an adaptation to an extreme environment [J]. *Nat. Commun.*, 2014, 54611.

Kirkness EF, Haas BJ, Sun W, *et al.* Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle [J]. *Proceedings of the National Academy of Sciences*, 2010, 107 (27): 12168 – 12173.

Kozomara A, Griffiths – Jones S. miRBase: Annotating high confidence microRNAs using deep sequencing data [J]. *Nucleic Acids Res.*, 2014, 42 (Database issue): 68 – 73.

Legeai F, Shigenobu S, Gauthier JP, *et al.* AphidBase: A centralized bioinformatic resource for annotation of the pea aphid genome [J]. *Insect Mol. Biol.*, 2010, 19 (Suppl): 25 – 12.

Liu JD, Improvement of Insect Genome Annotation Method and Analysis of Two Insect Genomes [D]. Nanjing Agricultural University, 2014.

Liu Q, Mackey AJ, Roos DS, *et al.* Evigan: A hidden variable model

for integrating gene evidence for eukaryotic gene prediction [J]. *Bioinformatics*, 2008, 24 (5): 597 – 605.

Luo R, Liu B, Xie Y, *et al.* SOAPdenovo2: An empirically improved memory – efficient short – read de novo assembler [J]. *Gigascience*, 2012, 1 (1): 18.

Maxam AM, Gilbert W. A new method for sequencing DNA [J]. *Proc. Natl. Acad Sci. USA*, 1977, 74 (2): 560 – 564.

Miller JR, Delcher AL, Koren S, *et al.* Aggressive assembly of pyrosequencing reads with mates [J]. *Bioinformatics*, 2008, 24 (24): 2818 – 2824.

Munoz – Torres MC, Reese JT, Childers CP, *et al.* Hymenoptera Genome Database: Integrated community resources for insect species of the order Hymenoptera [J]. *Nucleic Acids Res.*, 2011, 39 (Database issue): 658 – 662.

Ouzounis C A, Valencia A. Early bioinformatics: The birth of a discipline—a personal view [J]. *Bioinformatics*, 2003, 19 (17): 2176 – 2190.

Pang KC, Stephen S, Dinger ME, *et al.* RNAdb 2. 0—An expanded database of mammalian non – coding RNAs [J]. *Nucleic Acids Res.*, 2007, 35 (Database issue): 178 – 182.

Papanicolaou A, Schetelig MF, Arensburger P, *et al.* The whole genome sequence of the Mediterranean fruit fly, *Ceratitis capitata* (Wiedemann), reveals insights into the biology and adaptive evolution of a highly invasive pest species [J]. *Genome Biol.*, 2016, 17 (1): 192.

Parra G, Bradnam K, Korf I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes [J]. *Bioinformatics*, 2007, 23 (9): 1061 – 1067.

Poelchau M, Childers C, Moore G, *et al.* The i5k Workspace@ NAL—enabling genomic data access, visualization and curation of arthropod genomes [J]. *Nucleic Acids Res*, 2015, 43 (Database issue): 714 – 719.

Price AL, JonesNC, Pevzner PA. De novo identification of repeat families in large genomes [J]. *Bioinformatics*, 2005, 21 (Suppl.): 351 – 358.

Richards S, Murali SC. Best Practices in Insect Genome Sequencing: What Works and What Doesn't [J]. *Curr. Opin. Insect. Sci.*, 2015, 71 – 77.

Robinson GE, Hackett KJ, Purcell – Miramontes M, *et al.* Creating a buzz about insect genomes [J]. *Science*, 2011, 331 (6023): 1386 – 1386.

Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase [J]. *J. Mol. Biol.*, 1975, 94 (3): 441 – 448.

Sanger F, Air GM, Barrell BG, *et al.* Nucleotide sequence of bacteriophage phi X174 DNA [J]. *Nature*, 1977, 265 (5596): 687 – 695.

Scott JG, Warren WC, Beukeboom LW, *et al.* Genome of the house fly, *Musca domestica* L., a global vector of diseases with adaptations to a septic environment [J]. *Genome Biol.*, 2014, 15 (10): 466.

Shimomura M, Minami H, Suetsugu Y, *et al.* KAIKObase: An integrated silkworm genome database and data mining tool [J]. *BMC Genomics*, 2009, 10486.

Simao FA, Waterhouse RM, Ioannidis P, *et al.* BUSCO: Assessing genome assembly and annotation completeness with single – copy orthologs [J]. *Bioinformatics*, 2015, 31 (19): 3210 – 3212.

Simpson JT, Wong K, Jackman SD, *et al.* ABySS: A parallel assembler for short read sequence data [J]. *Genome Res.*, 2009, 19 (6): 1117 – 1123.

Stephens ZD, Lee SY, Faghri F, *et al.* Big Data: Astronomical or Genomical? [J]. *PLoS Biol.*, 2015, 13 (7): e1002195.

Tang W, Yu L, He W, *et al.* DBM – DB: The diamondback moth genome database [J]. *Database* (*Oxford*), 2014.

Tarailo – Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences [J]. *Curr. Protoc. Bioinformatics*, 2009, Chapter 4Unit 4 10.

Venter JC, Adams MD, Myers EW, *et al.* The sequence of the human genome [J]. *Science*, 2001, 291 (5507): 1304 – 1351.

Wajid B, Serpedin E. Review of general algorithmic features for genome assemblers for next generation sequencers [J]. *Genomics Proteomics Bioinformatics*, 2012, 10 (2): 58 – 73.

Wang J, Xia Q, He X, *et al.* SilkDB: A knowledgebase for silkworm biology and genomics [J]. *Nucleic Acids Res.*, 2005, 33 (Database issue): 399 – 402.

Wang X, Fang X, Yang P, *et al.* The locust genome provides insight into swarm formation and long – distance flight [J]. *Nat. Commun.*, 2014: 52957.

Werren JH, Richards S, Desjardins CA, *et al.* Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species [J]. *Science*, 2010, 327 (5963): 343 – 348.

Xia Q, Zhou Z, Lu C, *et al.* A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*) [J]. *Science*, 2004, 306 (5703): 1937 – 1940.

Xiao JH, Yue Z, Jia LY, *et al.* Obligate mutualism within a host drives the extreme specialization of a fig wasp genome [J]. *Genome Biol.*, 2013, 14 (12): R141.

Xu Y, Wang X, Yang J, *et al.* PASA—a program for automated protein NMR backbone signal assignment by pattern – filtering approach [J]. *J. Biomol. NMR*, 2006, 34 (1): 41 – 56.

Xu Z, Wang H. LTR_ FINDER: An efficient tool for the prediction of full – length LTR retrotransposons [J]. *Nucleic Acids Res.*, 2007, 35 (Web Server issue): 265 – 268.

Xue C, Li F, He T, *et al.* Classification of real and pseudo microRNA precursors using local structure – sequence features and support vector machine [J]. *BMC Bioinformatics*, 2005: 6310.

Xue J, Zhou X, Zhang CX, *et al.* Genomes of the rice pest brown planthopper and its endosymbionts reveal complex complementary contributions for host adaptation [J]. *Genome Biol.*, 2014, 15 (12): 521.

Yin C, Liu Y, Liu J, *et al.* ChiloDB: A genomic and transcriptome database for an important rice insect pest *Chilo suppressalis* [J]. *Database* (*Oxford*), 2014.

Yin C, Shen G, Guo D, *et al.* InsectBase: A resource for insect genomes and transcriptomes [J]. *Nucleic Acids Res.*, 2016,

44 (D1): 801 – 807.

You M, Yue Z, He W, *et al*. A heterozygous moth genome provides insights into herbivory and detoxification [J]. *Nature Genetics*, 2013, 45 (2): 220 – 225.

Zhan S, Merlin C, Boore J L, *et al*. The monarch butterfly genome yields insights into long – distance migration [J]. *Cell*, 2011, 147 (5): 1171 – 1185.

Zhan S, Reppert S M. MonarchBase: The monarch butterfly genome database [J]. *Nucleic Acids Res.*, 2013, 41 (Database issue): 758 – 763.

Zhang CX, Current research status and prospects of genomes of insects important to agriculture in China [J]. *Scientia Agricultura Sinica*, 2015 (17): 3454 – 3462.

Zhao Y, Li H, Fang S, *et al*. NONCODE 2016: An informative and valuable data source of long non – coding RNAs [J]. *Nucleic Acids Res.*, 2016, 44 (D1): 203 – 208.

Chen Y, Liu YS, Zeng JG. Progresses on plant genome Sequencing profile [J]. *Life Science Research*, 2014, 18 (1): 66 – 74. [陈勇, 柳亦松, 曾建国. 植物基因组测序的研究进展 [J]. 生命科学研究, 2014, 18 (1): 66 – 74]

Liu JD. Improlement of insect genome annotation method and analysis of two insect geomes [D]. Nanjing Agriculture University, 2014. [刘金定. 昆虫基因组注释方法改进及两种昆虫基因组分析 [D]. 南京农业大学, 2014]

Zhang CX. Current research status and prospects of genomes of insect important to agriculture in China [J]. *Scientia Agricuttura Sinica*, 48 (17): 3454 – 3462. [张传溪. 中国农业昆虫基因组学研究概况与展望 [J]. 中国农业科学, 2015, 48 (17): 3454 – 3462]