Mol Genet Genomics
DOI 10.1007/s00438-014-0831-7

ORIGINAL PAPER

OMIGA: Optimized Maker-Based Insect Genome Annotation

Jinding Liu · Huamei Xiao · Shuiqing Huang · Fei Li

Received: 23 May 2013 / Accepted: 17 February 2014

Abstract Insects are one of the largest classes of animals on Earth and constitute more than half of all living species. The i5k initiative has begun sequencing of more than 5,000 insect genomes, which should greatly help in exploring insect resource and pest control. Insect genome annotation remains challenging because many insects have high levels of heterozygosity. To improve the quality of insect genome annotation, we developed a pipeline, named Optimized Maker-Based Insect Genome Annotation (OMIGA), to predict protein-coding genes from insect genomes. We first mapped RNA-Seq reads to genomic scaffolds to determine transcribed regions using Bowtie, and the putative transcripts were assembled using Cufflink. We then selected highly reliable transcripts with intact coding sequences to train de novo gene prediction software, including Augustus. The re-trained software was used to predict genes from insect genomes. Exonerate was used to refine gene structure and to determine near exact exon/intron boundary in the genome. Finally, we used the software Maker to integrate data from RNA-Seq, de novo gene prediction, and protein alignment to produce an official gene set. The OMIGA pipeline was used to annotate the draft genome of an important insect pest, *Chilo suppressalis*, yielding 12,548 genes. Different strategies were compared, which demonstrated that OMIGA had the best performance. In summary, we present a comprehensive pipeline for identifying genes in insect genomes that can be widely used to improve the annotation quality in insects. OMIGA is provided at http://ento.njau.edu.cn/omiga.html.

Keywords Insect genome · OMIGA · RNA-Seq · Re-training · Annotation

Communicated by Q. Xia.

J. Liu · H. Xiao · F. Li (✉)
Department of Entomology, College of Plant Protection, Nanjing Agricultural University, Nanjing 210095, China
e-mail: lifei@njau.edu.cn

J. Liu · S. Huang
College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China

Introduction

As the cost of DNA sequencing has dramatically reduced, an increased number of genome sequencing projects have begun. Entomologists jointly launched the i5k initiative, which intends to sequence the genomes of 5,000 insects (Robinson et al. 2011). Insect genome annotation remains challenging because most insects have high heterozygosity, and their genomes have not been assembled as well as those of model species. This makes it difficult to obtain a list of high-quality genes. Several genome annotation pipelines have been developed and were used for annotating insect genomes such as those of *Drosophila melanogaster* (Adams et al. 2000), *Bombyx mori* (Xia et al. 2004), and *Anopheles gambiae* (Holt et al. 2002). However, these pipelines must be carefully optimized because of the high heterozygosity, and some errors have been found in the annotations of complex genomes (Jouraku et al. 2013). Therefore, it is necessary to develop a pipeline for predicting protein-coding genes from high-heterozygosity insect genomes.

There are several strategies for genome annotation. The most straightforward method is to predict protein-coding genes from genomes using de novo prediction software such as GeneScan (Ramakrishna and Srinivasan 1999; Aggarwal and Ramaswamy 2002), GenomeScan (Yeh et al. 2001), and Augustus (Stanke and Morgenstern

Published online: 09 March 2014

2005). These programs were developed using principles of machine learning, such as hidden Markov models, support vector machines, and neural networks. Recently, RNA-Seq has been widely used to obtain the transcriptomes of species of interest. In contrast to expressed sequence tags (EST), RNA-Seq data can be used to detect very low abundance transcripts because it has deep coverage. Thus, transcriptome data are used to determine the transcribed regions of a genome by mapping raw RNA-Seq reads to genomic scaffolds. Some programs, such as PASA (Haas et al. 2003) and G-Mo.R-Se (http://www.genoscope.cns.fr/externe/gmorse/), have been developed to integrate RNA-Seq data into genome annotation. In addition, many algorithms also consider the alignment of RefSeq proteins with the genomic scaffolds to refine the gene structure. The programs Glean and Maker were developed to integrate three kinds of genetic evidence: de novo gene prediction, gene expression, and protein homology evidence (Elsik et al. 2007; International Silkworm Genome 2008; Richards et al. 2008; Smith et al. 2011a, b; Suen et al. 2011; Zhan et al. 2011; You et al. 2013).

Although these programs have been applied to insect genome annotation, and have been used to successfully obtain a catalog of genes in some insects, there are several major disadvantages in annotating high-heterozygosity insect genomes. (1) De novo gene prediction software must be trained to achieve high accuracy. However, known genes are limited because the genomes of insects are less well studied than those of vertebrate model organisms. Thus, it is difficult to find enough insect genes for training (Robinson and Baumgartner 2011). In this case, the genes from a closely related species were often used as the training set. (2) Normally, insect RNA-Seq data are not as much available as model organisms. Therefore, the transcription evidence is not sufficient for those insects that are not well studied. (3) The N50 length of insect genomic scaffolds is generally not as long as that of model organisms, which reduces the reliability of de novo gene prediction.

Here, to overcome these disadvantages, we developed a computational pipeline, named Optimized Maker-Based Insect Genome Annotation (OMIGA), to identify protein-coding genes from insect genomes with high heterozygosity. In OMIGA, we mixed RNA-Seq data from different developmental stages of insects and mapped them to genomic scaffolds. This helped us to identify transcribed regions in the genome. Next, we extracted ≈1,000 high reliability genes from the assembled RNA-Seq data and used them as the training set to re-train de novo gene prediction software. This significantly increased the accuracy of de novo prediction. The results indicated that OMIGA has better performance than conventional insect genome annotation strategies.

Materials and methods

Data

Genome and transcriptome data from an important rice insect pest, *Chilo suppressalis*, were obtained by genomic sequencing using an Illumina platform (GA II). The whole genome shotgun project has been deposited in DDBJ/EMBL/GenBank under the accession number ANCD00000000. Raw RNA-Seq data are available under the accession number SRA060774. Detailed descriptions of these data have been reported in another manuscript (under review). The genome data were stored in standard Fasta format, whereas RNA-Seq sequencing reads were stored in Fastq format. Invertebrate RefSeq proteins were downloaded from the RefSeq database (Pruitt et al. 2007). The NCBI non-redundant protein database was used for functional annotations.

Identifying repeat sequences

We used the software RepeatMasker to identify known repeat sequences in the genomic scaffolds (Tempel 2012). Novel repeat sequences were predicted by RepeatModeler, which includes two de novo programs, RECON (Bao and Eddy 2002) and RepeatScout (Price et al. 2005). Both programs were used with default parameters.

Mapping RNA-Seq raw data with the genome scaffolds

Bowtie was used to align RNA-Seq raw reads with genomic scaffolds to collect expression data (Langmead et al. 2009). Next, TopHat was used to determine the exon/intron junctions within the genome. Finally, we used Cufflinks to obtain putative transcripts (Trapnell et al. 2009, 2010, 2012). We named these transcripts the Cufflink gene set. This step determined the transcribed regions of the scaffolds. All programs were used with default parameters. The numbers and lengths of transcripts depended on the depth and coverage of RNA-Seq. The transcripts that were expressed only at a specific stage were more difficult to locate. Most transcripts identified by this step were incomplete because of sequencing bias.

Re-training de novo gene prediction software

To obtain high accuracy, de novo gene prediction software must be re-trained before it can be used for insect genome annotation. The best training strategy is to use sufficient genes of the same species as the training dataset (Makarov 2002). To collect enough genes for training, we selected transcripts from the Cufflink genes. The criteria were as follows. (1) The transcripts were annotated as protein genes

using BLASTX with an *E* value <1e^{-5}. (2) The open reading frame (ORF) length of the transcript was more than 80 % of that of its orthologs in the RefSeq sequences. (3) Both BLASTX and TransDecoder identified the same ORF. (4) For better training, genes with a single exon were also included. These genes were used to re-train the prediction software such as Augustus (Stanke et al. 2006) and SNAP (Korf 2004). For GeneMark, more than 10 Mb of genome sequence was used to re-train the software (Lukashin and Borodovsky 1998). The default parameters were used for training.

Aligning protein sequences with the genome

Invertebrate RefSeq protein sequences were aligned with the genome scaffolds using BLASTX. To get more informative alignments, we used the program Exonerate to polish BLAST hits. Exonerate realigns BLAST hit sequences around splice sites. The alignment result has a high level of quality and therefore can be used to determine intron/exon boundary (Slater and Birney 2005).

Evidence integration

We obtained three kinds of genetic evidence by the above analyses. To produce a consensus gene dataset, we integrated all types of evidence using the software Maker with default parameters (Cantarel et al. 2008).

Results

OMIGA pipeline

The OMIGA pipeline is shown in Fig. 1. A shell script was written to execute each step of the pipeline with a configuration file. The parameters can be set flexibly according to the requirements. OMIGA runs on a Linux operating system and requires at least four gigabytes of random-access memory. Perl scripts were used to process intermediate data and convert different file formats. The Perl package (version 5.8.0 or higher) and the Mysql database management system are required. The open source of OMIGA can be obtained from http://ento.njau.edu.cn/omiga.html. The software used in this pipeline is listed in Table 1.

OMIGA can be divided into four steps: mapping RNA-Seq data with the genome, re-training Augustus and SNAP, mapping invertebrate RefSeq proteins to the genome, and integrating genetic evidence using Maker. Genes meeting either of the following criteria were kept: (1) de novo predicted genes with homologs in the RefSeq dataset; (2) a region of >30 % of a de novo predicted gene that could be mapped with RNA-Seq data.

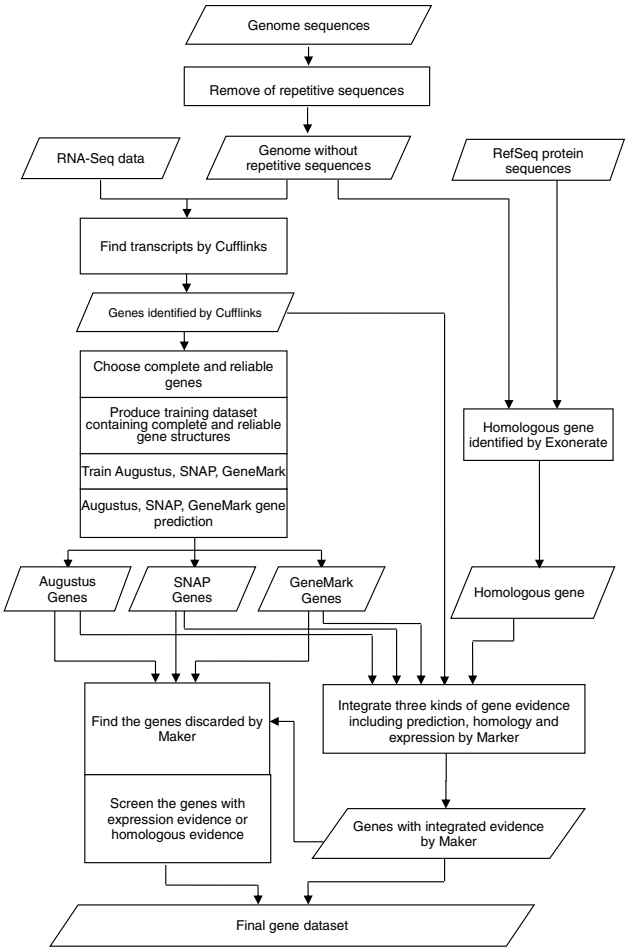

Fig. 1 Optimized Maker-Based Insect Genome Annotation (OMIGA) pipeline. Repeat sequences were removed by RepeatMasker. Cufflinks was used to obtain a reliable gene set for training gene prediction software. Maker was used to integrate different types of genetic evidence and to generate a consensus gene dataset

Applying the OMIGA pipeline to find genes in an insect genome

We sequenced the genome of the rice insect pest *C. suppressalis* using Illumina Solexa, yielding 670 Mb of draft genome data. Unfortunately, this assembly of this draft genome was of poor quality because of high heterozygosity. The scaffold N50 was only 5.2 kb. Moreover, only 13 known *C. suppressalis* genes were available in the GenBank database, which was not sufficient to train de novo prediction software. To overcome these disadvantages, we applied the OMIGA pipeline to identify protein-coding genes in the *C. suppressalis* genome. We screened 222 single-exon transcripts and 612 multiple-exon transcripts to train Augustus and SNAP. All of these transcripts had intact ORFs. After integrating the homology and expression evidence, 12,548 protein-coding genes were identified in the draft genome of *C. suppressalis*.

Table 1 Software packages used in the OMIGA pipeline

Classification	Software package	Functional description
Repetitive sequence identification	RepeatMasker	Identify the repetitive sequences of the genome
	RepeatModeler/RECON/RepeatScout	De novo predict the repetitive sequence in the genome and establish the repetitive sequence library
Identification expression gene	Bowtie	Map the short reads to the genome
	TopHat	Identify the splicing junctions
	Cufflinks	Determine transcript structure
Gene prediction	Augustus	De novo predict protein-coding genes from the genome. The accuracy relies on the re-training
	SNAP	De novo predict protein-coding genes from the genome. Re-training is necessary. The accuracy is relatively low in predicting long intron genes
	GeneMark	De novo predict protein-coding genes from the genome. Use >10 M genome of the species specific for self-training. Re-training is necessary. The accuracy is relative low in predicting long intron genes
Identification protein homologous gene	Exonerate	Global sequence alignment tool, use protein sequence to align genome to find the conserved domain of homologous genes
Gene synthesis	Maker	Integrate gene evidence of various aspects to produce consistent gene set
Others	TransDecoder	Identify the coding region on nucleotide sequences
	Blast	Sequence alignment tool
	Samtool	Process alignment results
	Fastx_toolkit	Process short sequence file
	BioPerl	Perl module

Table 2 Comparison of results of four annotation strategies

Strategy	Training set	Expression evidence	Total genes	Nr homologous genes ($1e^{-5}$)	N50
Strategy I	Monarch butterfly gene	None	6,034	5,102	862
Strategy II	Monarch butterfly gene	Cufflinks genes	11,067	9,062	1,103
Strategy III	Cufflinks identification gene	None	7,993	6,544	1,086
OMIGA	Cufflinks complete gene	Cufflinks genes	12,548	10,221	1,283

Assessing the OMIGA pipeline

To evaluate the performance of OMIGA, we designed three additional strategies to annotate the draft genome of *C. suppressalis*.

Strategy I: We used 1,000 Monarch butterfly genes (including 168 single-exon genes) to train Augustus and SNAP (Zhan and Reppert 2013). To re-train GeneMark, we used 20 Mb of Monarch butterfly genome sequence as the training dataset. The final gene set was obtained by integrating de novo gene prediction and protein alignment evidence.

Strategy II: Monarch butterfly genes were used as the training set. The RNA-Seq data was mapped with genomic scaffolds to determine the transcribed regions. The final gene set was obtained by integrating de novo gene prediction, protein alignment, and RNA-Seq data.

Strategy III: We selected 834 *C. suppressalis* genes from the Cufflink gene set as the training set. To train GeneMark, we used 20 Mb of *C. suppressalis* genome sequence as the training dataset. The final gene set was obtained by integrating de novo gene prediction and protein alignment evidence.

Compared with these three strategies, OMIGA had the best performance in terms of gene number, the number of homologous genes, and N50. Strategies I and II used the genes of a closely related species as the training set. Strategy II was inferior to OMIGA and strategy I was worse than strategy III, suggesting that genes from the same species should be used as the training set to improve accuracy. Strategy II and OMIGA both used RNA-Seq data to integrate transcription evidence. These two strategies were better than strategies III and I, respectively (Table 2).

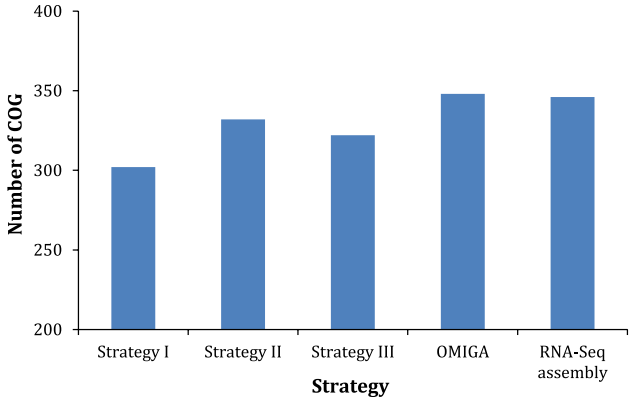

Fig. 2 Numbers of COGs identified by different annotation strategies. CEGMA reported 458 core conserved gene groups (COGs). We used these COGs to evaluate the completeness of the different gene annotation methods. The results indicated that the OMIGA pipeline has the best performance

CEGMA evaluation

CEGMA is a software that has been widely used for evaluating gene space of the assembled genome. There are 458 highly conserved clusters of orthologous groups (COGs) in eukaryotic genomes (Parra et al. 2007). We used CEGMA to identify COGs in the annotation results by different strategies. The numbers of COGs found by strategies I, II, and III and OMIGA were 302, 332, 322, and 348, respectively (Fig. 2). At least 65 % of COGs could be found by either approach. OMIGA identified more COGs than the other three strategies, suggesting it had a better performance. The transcriptome that was directly assembled from RNA-Seq data contained 346 COGs. However, the transcript lengths were shorter than the genes identified by the OMIGA pipeline, demonstrating that identifying genes from the genome is a better method than RNA-Seq assembly alone.

Discussion

An increasing number of insect genomes have been sequenced, and insect genome annotation has become an important and challenging task. Here, we developed a pipeline, named OMIGA, to identify protein-coding genes from insect genomes. The final gene set was obtained by integrating de novo prediction, RNA-Seq data, and homology evidence. Assessment analysis indicated that the OMIGA pipeline performed better than the other three strategies used. This pipeline can be widely used for insect genome annotation to obtain a catalog of highly reliable genes.

Each type of evidence used in genome annotation has advantages and disadvantages. In theory, it is possible to obtain all genes by de novo prediction, but this method has a high false-positive rate. Conserved genes can be found by protein alignment, but it is difficult to identify novel genes from an insect genome using this method. RNA-Seq data have the highest reliability, but do not effectively detect genes expressed at low levels. Therefore, integrating these different types of evidence to produce a final gene set is the best strategy.

Considering the high complexity of insect genomes, we specifically re-trained de novo prediction software. Because most insect pests are not well studied, we selected hundreds of protein-coding genes from the RNA-Seq assembly as the training set. This dramatically increased the accuracy of de novo prediction.

Because many insects have a highly complex genome, we recommend the following aspects must be considered in sequencing an insect genome:

1. To obtain as many transcripts as possible, it is better to use a mixed sample for RNA-Seq sequencing. This provides much more transcription data for subsequent gene prediction.
2. It is important to construct a training dataset containing genes from the same species, which can be obtained from the RNA-Seq data.
3. Many insect genomes have high heterozygosity. Thus, it is difficult to obtain a high N50 in the genome assembly. We found that the OMIGA pipeline can still find reliable genes from an incomplete genome. Thus, we suggest that pursuing a higher N50 at high cost is unnecessary.

Acknowledgments This work was supported by the National Basic Research Program of China (2012CB114100), the National High Technology Research and Development Program ("863" Program) of China (2012AA101505), the National Science Foundation of China (31171843, 31301691), and the Jiangsu Science Foundation for Distinguished Young Scholars (BK2012028).

Conflict of interest The authors declare that they have no conflicts of interest.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE,

Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferriera S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, Worley KC, Woodage T, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC (2000) The genome sequence of *Drosophila melanogaster*. Science 287(5461):2185–2195

Aggarwal G, Ramaswamy R (2002) Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. J Biosci 27(1 Suppl 1):7–14

Bao Z, Eddy SR (2002) Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res 12(8):1269–1276. doi:10.1101/gr.88502

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18(1):188–196. doi:10.1101/gr.6743907

Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM (2007) Creating a honey bee consensus gene set. Genome Biol 8(1):R13. doi:10.1186/gb-2007-8-1-r13

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31(19):5654–5666

Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O'Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q, Zhao S, Zhu SC, Zhimulev I, Coluzzi M, della Torre A, Roth CW, Louis C, Kalush F, Mural RJ, Myers EW, Adams MD, Smith HO, Broder S, Gardner MJ, Fraser CM, Birney E, Bork P, Brey PT, Venter JC, Weissenbach J, Kafatos FC, Collins FH, Hoffman SL (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298(5591):129–149. doi:10.1126/science.1076181

International Silkworm Genome C (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. Insect Biochem Mol Biol 38(12):1036–1045. doi:10.1016/j.ibmb.2008.11.004

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) KONAGA base: a genomic and transcriptomic database for the diamondback moth, *Plutella xylostella*. BMC Genomics 14:464. doi:10.1186/1471-2164-14-464

Korf I (2004) Gene finding in novel genomes. BMC Bioinform 5:59. doi:10.1186/1471-2105-5-59

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25. doi:10.1186/gb-2009-10-3-r25

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26(4):1107–1115

Makarov V (2002) Computer programs for eukaryotic gene prediction. Brief Bioinform 3(2):195–199

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23(9):1061–1067. doi:10.1093/bioinformatics/btm071

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1):i351–i358. doi:10.1093/bioinformatics/bti1018

Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35(Database issue):D61–65. doi:10.1093/nar/gkl842

Ramakrishna R, Srinivasan R (1999) Gene identification in bacterial and organellar genomes using GeneScan. Comput Chem 23(2):165–174

Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beeman RW, Gibbs R, Bucher G, Friedrich M, Grimmelikhuijzen CJ, Klingler M, Lorenzen M, Roth S, Schroder R, Tautz D, Zdobnov EM, Muzny D, Attaway T, Bell S, Buhay CJ, Chandrabose MN, Chavez D, Clerk-Blankenburg KP, Cree A, Dao M, Davis C, Chacko J, Dinh H, Dugan-Rocha S, Fowler G, Garner TT, Garnes J, Gnirke A, Hawes A, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Jackson L, Kovar C, Kowis A, Lee S, Lewis LR, Margolis J, Morgan M, Nazareth LV, Nguyen N, Okwuonu G, Parker D, Ruiz SJ, Santibanez J, Savard J, Scherer SE, Schneider B, Sodergren E, Vattahil S, Villasana D, White CS, Wright R, Park Y, Lord J, Oppert B, Brown S, Wang L, Weinstock G, Liu Y, Worley K, Elsik CG, Reese JT, Elhaik E, Landan G, Graur D, Arensburger P, Atkinson P, Beidler J, Demuth JP, Drury DW, Du YZ, Fujiwara H, Maselli V, Osanai M, Robertson HM, Tu Z, Wang JJ, Wang S, Song H, Zhang L, Werner D, Stanke M, Morgenstern B, Solovyev V, Kosarev P, Brown G, Chen HC, Ermolaeva O, Hlavina W, Kapustin Y, Kiryutin B, Kitts P, Maglott D, Pruitt K, Sapojnikov V, Souvorov A, Mackey AJ, Waterhouse RM, Wyder S, Kriventseva EV, Kadowaki T, Bork P, Aranda M, Bao R, Beermann A, Berns N, Bolognesi R, Bonneton F, Bopp D, Butts T, Chaumot A, Denell RE, Ferrier DE, Gordon CM, Jindra M, Lan Q, Lattorff HM, Laudet V, von Levetsow C, Liu Z, Lutz R, Lynch JA, da Fonseca RN, Posnien N, Reuter R, Schinko JB, Schmitt C, Schoppmeier M, Shippy TD, Simonnet F, Marques-Souza H, Tomoyasu Y, Trauner J, Van der Zee M, Vervoort M, Wittkopp N, Wimmer EA, Yang X, Jones AK, Sattelle DB, Ebert PR, Nelson D, Scott JG, Muthukrishnan S, Kramer KJ, Arakane Y, Zhu Q, Hogenkamp D, Dixit R, Jiang H, Zou Z, Marshall J, Elpidina E, Vinokurov K, Oppert C, Evans J, Lu Z, Zhao P, Sumathipala N, Altincicek B, Vilcinskas A, Williams M, Hultmark D, Hetru C, Hauser F, Cazzamali G, Williamson M, Li B, Tanaka Y, Predel R, Neupert S, Schachtner J, Verleyen P, Raible F, Walden KK, Angeli S, Foret S, Schuetz S, Maleszka R, Miller SC, Grossmann D (2008)

The genome of the model beetle and pest *Tribolium castaneum*. Nature 452(7190):949–955. doi:10.1038/nature06784

Robinson BE, Baumgartner J (2011) Cultivating a demand for clean cookstoves. Science 334(6063):1636–1637. doi:10.1126/science.334.6063.1636-b

Robinson GE, Hackett KJ, Purcell-Miramontes M, Brown SJ, Evans JD, Goldsmith MR, Lawson D, Okamuro J, Robertson HM, Schneider DJ (2011) Creating a buzz about insect genomes. Science 331(6023):1386. doi:10.1126/science.3316023.1386

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinform 6:31. doi:10.1186/1471-2105-6-31

Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, Fave MJ, Fernandes V, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Helmkampf M, Holley JA, Hu H, Viniegra AS, Johnson BR, Johnson RM, Khila A, Kim JW, Laird J, Mathis KA, Moeller JA, Munoz-Torres MC, Murphy MC, Nakamura R, Nigam S, Overson RP, Placek JE, Rajakumar R, Reese JT, Robertson HM, Smith CR, Suarez AV, Suen G, Suhr EL, Tao S, Torres CW, van Wilgenburg E, Viljakainen L, Walden KK, Wild AL, Yandell M, Yorke JA, Tsutsui ND (2011a) Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*). Proc Natl Acad Sci USA 108(14):5673–5678. doi:10.1073/pnas.1008617108

Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, Fave MJ, Fernandes V, Gibson JD, Graur D, Gronenberg W, Grubbs KJ, Hagen DE, Viniegra AS, Johnson BR, Johnson RM, Khila A, Kim JW, Mathis KA, Munoz-Torres MC, Murphy MC, Mustard JA, Nakamura R, Niehuis O, Nigam S, Overson RP, Placek JE, Rajakumar R, Reese JT, Suen G, Tao S, Torres CW, Tsutsui ND, Viljakainen L, Wolschin F, Gadau J (2011b) Draft genome of the red harvester ant *Pogonomyrmex barbatus*. Proc Natl Acad Sci USA 108(14):5667–5672. doi:10.1073/pnas.1007901108

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33(Web Server issue):W465–467. doi:10.1093/nar/gki458

Stanke M, Schoffmann O, Morgenstern B, Waack S (2006) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinform 7:62. doi:10.1186/1471-2105-7-62

Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, Denas O, Elhaik E, Fave MJ, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Harkins TT, Helmkampf M, Hu H, Johnson BR, Kim J, Marsh SE, Moeller JA, Munoz-Torres MC, Murphy MC, Naughton MC, Nigam S, Overson R, Rajakumar R, Reese JT, Scott JJ, Smith CR, Tao S, Tsutsui ND, Viljakainen L, Wissler L, Yandell MD, Zimmer F, Taylor J, Slater SC, Clifton SW, Warren WC, Elsik CG, Smith CD, Weinstock GM, Gerardo NM, Currie CR (2011) The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. PLoS Genet 7(2):e1002007. doi:10.1371/journal.pgen.1002007

Tempel S (2012) Using and understanding RepeatMasker. Methods Mol Biol 859:29–51. doi:10.1007/978-1-61779-603-6_2

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111. doi:10.1093/bioinformatics/btp120

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28(5):511–515. doi:10.1038/nbt.1621

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protocol 7(3):562–578. doi:10.1038/nprot.2012.016

Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, Zhao P, Zha X, Cheng T, Chai C, Pan G, Xu J, Liu C, Lin Y, Qian J, Hou Y, Wu Z, Li G, Pan M, Li C, Shen Y, Lan X, Yuan L, Li T, Xu H, Yang G, Wan Y, Zhu Y, Yu M, Shen W, Wu D, Xiang Z, Yu J, Wang J, Li R, Shi J, Li H, Li G, Su J, Wang X, Li G, Zhang Z, Wu Q, Li J, Zhang Q, Wei N, Xu J, Sun H, Dong L, Liu D, Zhao S, Zhao X, Meng Q, Lan F, Huang X, Li Y, Fang L, Li C, Li D, Sun Y, Zhang Z, Yang Z, Huang Y, Xi Y, Qi Q, He D, Huang H, Zhang X, Wang Z, Li W, Cao Y, Yu Y, Yu H, Li J, Ye J, Chen H, Zhou Y, Liu B, Wang J, Ye J, Ji H, Li S, Ni P, Zhang J, Zhang Y, Zheng H, Mao B, Wang W, Ye C, Li S, Wang J, Wong GK, Yang H, Biology Analysis G (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306(5703):1937–1940. doi:10.1126/science.1102210

Yeh RF, Lim LP, Burge CB (2001) Computational inference of homologous gene structures in the human genome. Genome Res 11(5):803–816. doi:10.1101/gr.175701

You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, Douglas CJ, Bai J, Wang P, Cui K, Huang S, Li X, Zhou Q, Wu Z, Chen Q, Liu C, Wang B, Xu X, Lu C, Hu M, Davey JW, Smith SM, Chen M, Xia X, Tang W, Ke F, Zheng D, Hu Y, Song F, You Y, Ma X, Peng L, Zheng Y, Liang Y, Chen Y, Yu L, Zhang Y, Liu Y, Li G, Fang L, Li J, Zhou X, Luo Y, Gou C, Wang J, Yang H (2013) A heterozygous moth genome provides insights into herbivory and detoxification. Nat Genet 45(2):220–225. doi:10.1038/ng.2524

Zhan S, Reppert SM (2013) MonarchBase: the monarch butterfly genome database. Nucleic Acids Res 41(Database issue):D758–763. doi:10.1093/nar/gks1057

Zhan S, Merlin C, Boore JL, Reppert SM (2011) The monarch butterfly genome yields insights into long-distance migration. Cell 147(5):1171–1185. doi:10.1016/j.cell.2011.09.052