Opinion

Beenomes to *Bombyx*: future directions in applied insect genomics

Jay D Evans* and Dawn Gundersen-Rindal†

Addresses: *USDA-ARS Bee Research Lab and †USDA-ARS Insect Biocontrol Lab, Beltsville, MD 20705, USA.

Correspondence: Jay D Evans. E-mail: evansj@ba.ars.usda.gov

Published: 26 February 2003

Genome ***Biology*** 2003, **4**:107

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/3/107

Abstract

The recent sequencing of the *Anopheles gambiae* genome showcases the genetic breadth of insects and a trend towards sequencing organisms directly involved with human welfare. We describe traits in other insect species that make them important candidates for genomics projects, and review several recent workshops aimed at uniting researchers working with insect species to efficiently address problems in medicine, biotechnology, and agriculture.

The recent sequencing of the *Anopheles gambiae* genome [1] is a watershed event in genomics for two reasons. First, this species is of sufficient phylogenetic distance from the previously sequenced *Drosophila melanogaster* to provide the best view to date of changes in genome organization and composition across the insects. The 250 million-year spread between these species, abetted by a high rate of sequence evolution, allows genomic comparisons over an evolutionary time-scale equal to that between humans and fish [2], larger by one-third than that between humans and chickens. Although this is a fraction of the distance covered by insects as a whole, it allows new tests of inferences drawn from *Drosophila* about gene function in insects in general.

The second reason that the *Anopheles gambiae* genome is a landmark is that *Anopheles* is the first animal to be sequenced, other than ourselves, whose actions have a strong direct impact on human lives. In the near future such 'applied' genomic projects will probably become the norm, as agencies involved with human health and agriculture develop plans to sequence key pests and beneficial species. This trend is particularly evident in insect genomics. The next two species in the insect genome queue, the honey bee (*Apis mellifera*) and silkworm moth (*Bombyx mori*), were selected in part because of their longstanding use in agriculture. Other insect candidates, including another mosquito (*Aedes aegypti*), the medfly (*Ceratitis capitata*), and flour beetle (*Tribolium castaneum*), also have longstanding histories of research driven by their impacts on humans. In this article, we discuss criteria that might be used to evaluate the candidacy of various insect taxa for whole-genome sequencing. Specifically, we compare and contrast genome size, current genetic knowledge, species diversity, and the human impact of insects from 11 different insect orders and suggest how scientists and funders could use these criteria to help justify and prioritize future sequencing efforts. In addition, we briefly summarize recent scientific workshops aimed at integrating scientists and research programs focused on questions concerning basic and applied genomics in non-traditional insect species.

Genome sequencing criteria in insects

Given limited time and funding, robust criteria must be developed by which to weigh insect species as new sequencing candidates. One obvious goal is taxonomic breadth, and the eventual completion of full genome sequences from members representing the three major insect clades (Figure 1 [3]) will be an essential contribution to comparative genomics. Taxonomic breadth by itself is not a sufficient criterion for comparing sequencing candidates, however. Full genome sequences from multiple species of *Drosophila*, for example, can complement each other by clarifying gene function and organization in this well-studied genus, and by extrapolation in insects in general. Furthermore, multiple candidates within the same insect order may warrant

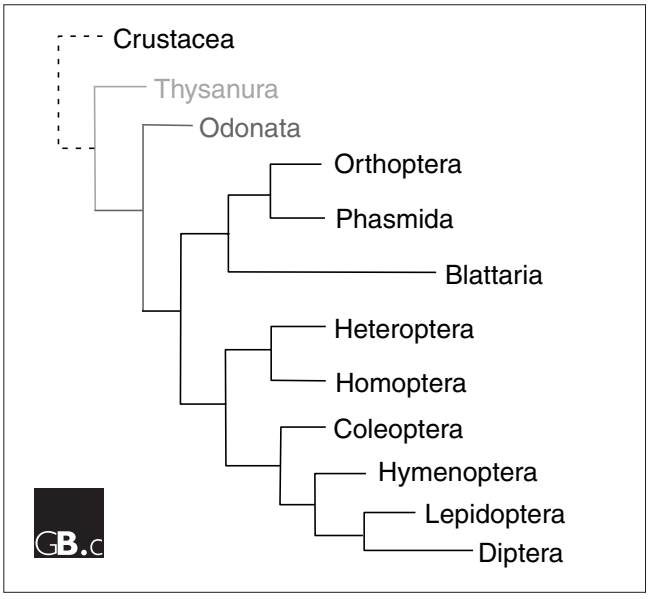

Figure 1
Phylogenetic tree of insect orders, after Wheeler *et al.* [3]. Light gray, Archaeognatha (primitive wingless insects); Dark gray, Paleoptera (primitive winged insects); Black, Neoptera (higher insects). Crustacea are shown as an arthropod outgroup. Thysanura include silverfish; Odonata, dragonflies; Orthoptera, grasshoppers and crickets; Phasmida, stick insects; Blattaria, roaches; Heteroptera, true bugs; Homoptera, aphids, scales and tree hoppers; Coleoptera, beetles; Hymenoptera, ants, bees and wasps; Lepidoptera, moths and butterflies; and Diptera, flies.

sequencing on the basis of other criteria: for example, within the order Diptera are mosquitoes, *Drosophila* and the economically important tephritid fly *Ceratitis capitata*.

Here, we use four criteria, from among the many possible, to compare the merits of insects from the 11 different insect orders shown in Figure 1. First, we use genome size based on estimates in the Animal Genome Size Database [4] as a predictor of direct sequencing costs. We estimate a mean genome size for each order, weighted at the level of family (for example, the numerous estimates of genome size in the fly family Drosophilidae were averaged and used as a single data point). This correlates well with estimates for the smallest genome in each order, and arguably is a more relevant estimate of the genome size of potential candidate species. We assume that sequencing costs increase linearly with genome size, given that any economy of scale achieved in sequencing larger genomes is likely to be mitigated by a need for higher sequence redundancy prior to assembly of large genomes.

Next, we evaluate the current level of genetic research for different organisms, because this is both an aid to sequence assembly and annotation and reflects the number of scientists who would be likely to benefit from a complete sequence. Our surrogate marker for the level of genetic research is the number of protein sequences present in GenBank [5] as of 15 December 2002 (excluding protein entries for species whose genome sequences are known in full, namely *Anopheles gambiae* and *Drosophila melanogaster*). We then compare species diversity across orders [6], because the strongest inferences in terms of gene function and synteny are likely to occur within insect orders, and orders with greater diversity are likely to have greater ecological and economic importance as well as a larger community of researchers. Finally, we estimate the direct human impact of specific insect orders. For this, we counted the number of papers about each insect order referenced in the CAB Abstracts Database [7], an international abstract service for agricultural and applied sciences, from 1993-2002. We then derive a composite score for each insect order by ranking GenBank records, species diversity, and relative human impact, and dividing this value by the mean genome size (Table 1).

Our results indicate strengths in all criteria for the holometabolous insect orders (those with complete metamorphosis) - Hymenoptera, Diptera, Lepidoptera and Coleoptera - as predicted previously [8] (see Figure 1 legend for the common names of insects in these orders). Coleoptera are the most speciose worldwide, but have slightly lower economic impact than Lepidoptera, Hymenoptera, and Diptera. Beyond the holometabolous insects, the order Homoptera stands out for having species with generally small genomes and great economic and agricultural importance. The representatives from the primitive insect orders Thysanura and Odonata, while valuable from the standpoint of phylogenetic breadth, fare poorly compared to other orders using our criteria.

The emphasis in these criteria on insects with recognized human impact and ecological importance is not meant to negate the value of model insect species as sequencing candidates. Model insect genomes can provide general insights into biological mechanisms, gene structure and function, and the conserved evolutionary processes that select for certain genetic traits. Thus model organisms, as illustrated by species of *Drosophila*, yield invaluable insights for all insect genomes. And as a final caveat, we should emphasize that although we present several ways to compare the merits of different insect groups, we do not mean to infer that these are the only criteria useful for such decisions. (Our views are our own and need not reflect the opinions of our agency or the US government.)

Recent insect genome collaborations, and progress

Several recent workshops have been held with a specific focus on insect genomics and its applications. The Comparative Insect Genomics Workshop (Washington DC, USA; October 2001; sponsored by the US Department of Agriculture) was the first international meeting of scientists from academia, private industry, and government with the

Table 1

Insect orders evaluated for sequencing priority

Insect order	Genome size	GenBank records	Species diversity	Human impact	Composite score
Thysanura					
Odonata					
Orthoptera					
Phasmida					
Blattaria					
Heteroptera					
Homoptera					
Coleoptera					
Hymenoptera					
Lepidoptera					
Diptera					

Insect orders are listed according to the phylogenetic tree shown in Figure 1 and show relative genome size (weighted mean, with size corresponding to the area of the circle). Also shown are the number of protein records in GenBank, number of worldwide species, and human impact estimate (see text for further details). In each case the proportion of the circle that is black indicates the number relative to the order with the highest value for each measure (the highest ranking order being shown with a filled circle). The last column shows a composite ranking of orders assuming that equal weight is given to each of these criteria.

purpose of addressing and promoting the broad field of insect genomics. Discussions at this meeting focused on current approaches for analyzing and comparing genomes, the evaluation of candidate insects for genome sequencing, and ways to coordinate genomic efforts and ensure public access to materials and datasets. Leaders from the fruitfly, nematode, plant, and microbial genomics communities discussed the evolution of their own genome initiatives, and offered critiques of impending projects in insects. Because *Drosophila*-associated projects have served as models for all insect genomicists, there was substantial discussion of how new insect projects might benefit, and might benefit from, studies involving *Drosophila*. FlyBase [9], a key database for *Drosophila* genetics, forms one venue for comparative analyses in insects that is already widely used by those working on other insect species. Similar resources available through the US National Institutes of Health [10] and the Gene Ontology Consortium [11] were also identified as being key to generating testable inferences for new genome sequences.

Recognizing the success of the completed and ongoing dipteran genome projects, several working groups have formed to develop and promote genome projects in new insect groups. Within the Hymenoptera, an international genomics effort has been emerging for several years around the honey bee, arguably the best studied and economically most important member of this group. Propelled in part by the Comparative Insect Genomics Workshop, a successful funding white paper was submitted to the US National Human Genome Research Institute for honey bee genome sequencing (now nearing completion) at the Baylor College of Medicine Genome Center. A more recent Honey Bee Biotechnology Workshop (Sapporo, Japan; July 2002) focused both on details of this genome project and on independent genomics efforts. New applications of functional genomic techniques described there foretell the many ways researchers will use genomic data to answer basic and applied questions in this species. As one example, two lab groups discussed the successful application of RNA interference methods in honey bee embryos and brains.

Within the Lepidoptera, an international genomics effort has centered on the economically important silkworm moth [12], for which a completed genome sequence is expected in 2004. The recent International Workshop of Lepidopteran Genomics (Tsukuba, Japan; September 2002) focused on key aspects of this genome project, most notably the integration of large-insert libraries, expressed sequence tags (ESTs), and applications of transgenic technologies. The International Lepidopteran Genome Project [13] has been charged with applying new technologies to compare the genomes of a growing list of agriculturally important moths and butterflies. Among these, the crop-feeding heliothine moths have long been appreciated as significant genome candidates. One privately funded genome project in this group, involving the tobacco budworm *Heliothis virescens,* is apparently complete but remains inaccessible to the public. By contrast the *Bombyx mori* project [12] and projects involving additional heliothine species are expected to be carried out with full public access.

Although no formal gatherings have been held to date, working groups representing additional insect orders (for example within the Coleoptera and Homoptera) continue to develop within the insect genomics research community. Well-defined and concerted research efforts, combined with advancing technologies and access to post-genomic tools and data, will speed advances in these taxa. As one example, functional studies using RNA interference and related methods are now feasible for all insect species, using orthologs identified through matches with current genome projects. Additionally, newly available large-insert libraries, for example those available through [14], can be used to begin testing for synteny and structure in diverse insect genomes. Finally, comparative genomics databases from flies, moths, and bees will undoubtedly be used to inform other genomics projects.

In conclusion, the field of insect genomics is experiencing an exceptional year that should invigorate insect genetic studies. The outbreak of genome sequences is also likely to impact genetic studies more broadly. New estimates suggest that 61% and 66% of protein coding sequences from *Drosophila* and *Anopheles*, respectively, have known orthologs in non-insect genomes (human, mouse, *Arabidopsis*, worm, yeast, zebrafish, rat and rice [15,16]). This upward trend (only 20-30% of *Drosophila* genes were identified as having non-insect matches two years ago [17]) is certain to continue with incoming genome data for bees, moths, and their relatives. Researchers studying insect genomes can look forward to using these shared traits to better address general problems in medicine, biotechnology, agriculture, and evolutionary biology.

References

1. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, *et al.*: **The genome sequence of the malaria mosquito *Anopheles gambiae*.** *Science* 2002, **298:**129-149.
2. Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, *et al.*: **Immunity-related genes and gene families in *Anopheles gambiae*.** *Science* 2002, **298:**159-165.
3. Wheeler WC, Whiting M, Wheeler QD, Carpenter JM: **The phylogeny of the extant hexapod orders.** *Cladistics* 2001, **17:**113-169.
4. Gregory T: **Animal Genome Size Database** [http://www.genomesize.com]
5. **GenBank** [http://ncbi.nih.gov/Genbank]
6. Borror DJ, Triplehorn CA, Johnson NF: *An Introduction to the Study of Insects.* Sixth edn. Philadelphia,: Saunders College; 1989.
7. **CAB Abstracts** [http://www.cabi-publishing.org/Products/Database/Abstracts/Index.asp]
8. Kaufman TC, Severson DW, Robinson GE: **The *Anopheles* genome and comparative insect genomics.** *Science* 2002, **298:**97-98.
9. **FlyBase** [http://flybase.bio.indiana.edu]
10. **National Center for Biotechnology Information** [www.ncbi.nlm.nih.gov]
11. **Gene Ontology Consortium** [http://www.geneontology.org]
12. **International Lepidopteran Genome Project** [http://www.ab.a.u-tokyo.ac.jp/lep-genome]
13. **SilkBase** [http://www.ab.a.u-tokyo.ac.jp/silkbase]
14. **GENEFinder Resource** [http://hbz.tamu.edu]
15. Gilbert DG: **euGenes: a eukaryote genome information system.** *Nucleic Acids Res* 2002, **30:**145-148.
16. **euGenes** (July 2002 update) [http://iubio.bio.indiana.edu:8089]
17. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, *et al*: **Comparative genomics of the eukaryotes.** *Science* 2000, **287:**2204-2215.