Pergamon

Insect Biochemistry and Molecular Biology 32 (2002) 1193–1197

Insect Biochemistry and Molecular Biology

www.elsevier.com/locate/ibmb

Culicine genomics

Dennis L. Knudson [a,*], Susan E. Brown [a], David W. Severson [b]

[a] *Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO 80523, USA*
[b] *Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA*

Received 10 February 2002; received in revised form 28 March 2002; accepted 2 April 2002

Keywords: Aedes aegypti; Culicidae; Diptera; Genome; Genomics; Linkage map; Physical map; Whole genome sequence

1. Introduction

In our recent review on genetic and physical mapping in mosquitoes, we discussed the future of mosquito genomics (Severson et al., 2001). We stated that reagents and methodologies were in place for fine-scale mapping in Culicidae and for map-based positional cloning in *Aedes aegypti* and *Anopheles gambiae*. We reported that an *An. gambiae* genome project had been initiated and predicted that there would be considerable gene information developed from the whole genome sequence of the paradigm dipteran, *Drosophila melanogaster*. We were confident in our prediction that the genomes of one or more mosquitoes would be sequenced *in toto*. Little did we know just how quickly culicid genomics would progress.

2. What is the current status of whole genome sequencing in dipterans?

In 1999–2000, the Berkeley Drosophila Genome Project (BDGP), a consortium of the Drosophila Genome Center, and Celera Genomics collaborated to sequence the *D. melanogaster* genome (Adams et al., 2000) (see also www.fruitfly.org and www.celera.com). A whole-genome shotgun approach was used by Celera Genomics to produce a ~12× genome sequence coverage. The BDGP produced the scaffold for the assembly with their BAC-based physical map of a tiled path of overlapping BAC clones spanning the euchromatic portion of the genome and 1.5× shotgun sequence of each BAC clone in the tiling path. The 'scaffold sequence' assisted in the assembly of the 12× whole-genome shotgun data to produce Release 1.0 of the genomic sequence in March 2000 (Adams et al., 2000). Release 2.0 of the genomic sequence was made public in October 2000 where ~330 of the gaps in the Release 1.0 sequence were filled and some annotations were corrected or added by Celera and BDGP, and the scaffolds were orientated left–right across the chromosome. BDGP is currently finishing the genomic sequence to high quality (phase 3, i.e., finished with no gaps, see www.ncbi.nlm.nih.gov/HTGS and the third release of annotated whole genome sequence of *Drosophila melanogaster* is expected in March 2002 (www.fruitfly.org)).

In June 2001, the National Institute of Human Genome Research (NHGRI) announced its interest in receiving applications or a 'Request for Applications' (RFA) to obtain sequence coverage and an initial assembly of the genome of *Drosophila pseudoobscura* (grants.nih.gov/grants/guide/rfa-files/RFA-HG-01-003.html). By using a comparative approach, the biological significance of conserved regions between *D. pseudoobscura* and *D. melanogaster* would be identified. The underlying impetus for this comparative approach is simple. If organisms are sequenced with the purpose of understanding genes and gene function to illuminate basic biology and disease principles, then our ability to make such predictions from sequence data ought to be very good. Unfortunately, our gene-finding, predictive skill is only moderately successful. In contrast, comparative methods that examine sequence from two or more appropriately selected organisms can be very illuminating. The evolutionary distance between the two organisms to be compared is an important criterion in the selection of the second organism. For example, a comparative sequence analysis of very closely related species would likely reveal sequence variation that is informative for speci-

* Corresponding author. Tel.: +1-970 491-7255; fax: +1-970 491-4888.
E-mail address: Dennis.Knudson@colostate.edu (D.L. Knudson).

PII: S0965-1748(02)00082-6

ation. In contrast, a comparative sequence analysis of more diverged species would reveal conserved regions that are not only coding sequences for genes, but also identify important functional genome elements involved in transcriptional regulation (i.e., promoters and enhancers). *Drosophila pseudoobscura* was chosen as the second *Drosophila* species to be sequenced because *D. pseudoobscura* is estimated to have diverged from *D. melanogaster* 50–60 million years ago (Myr) and conserved functional elements would likely be identified. This time scale should allow the sequence of functional elements to drift, but the element should still be identifiable as one with a particular function. This comparative approach is not only satisfying intellectually, but the data generated should be exceptionally instructive and provide considerable insight into genome structure.

3. What is the current status of whole genome sequencing in culicidae?

The whole genome sequence of *Anopheles gambiae* is already at an advanced stage (whole genome assembly) and the genome program has an organizational structure similar to the *D. melanogaster* project, that is, there are two components in the overall genome program. A US and European collaboration on the *An. gambiae* genome project seeks to sequence 10,000 ESTs from normalized cDNA libraries made from five different mosquito tissues (50,000 total ESTs), to physically map 2000 genomic DNA bacterial artificial chromosomes (BACs) to polytene chromosome cytogenetic map, to sequence both ends of 25,000 genomic DNA BACs, and to assess local scale synteny between the mosquito and *D. melanogaster* by sequencing 3 Mb of the Adh region (NIH, NIAID, May 2001, AI048846). A second component involves the whole genome shotgun sequencing to a 10× coverage, and this effort has been undertaken by Celera Genomics (NIH, NIAID, August 2001, AI050687). Celera Genomics has deposited raw trace files (raw output files from the sequencing machine that are used in the assembly) in the Trace Archive at The National Center for Biotechnology Information. Currently (January, 2002), there are 4,445,304 traces online from Celera Genomics' the whole genome shotgun sequencing effort (//www.ncbi.nlm.nih.gov/Traces/). Celera Genomics is presently undertaking assembly and annotation of the sequence that it has produced together with scaffold and EST sequence data produced by the international collaboration. The assembled genome sequence is anticipated to be freely available to the research community via GenBank by March, 2002.

4. Why sequence the whole genome of another mosquito and why should the mosquito be *Aedes aegypti*, the yellow fever mosquito?

All mosquitoes are members of the family Culicidae, which is divided into three subfamilies—Culicinae, Anophelinae and Toxorhynchitinae (Knight and Stone, 1977). Toxorhynchitinae species do not blood-feed and species from Culicinae and Anophelinae are principal vectors for mosquito-borne diseases. Arbovirus transmission, including dengue fever and yellow fever, and also, transmission of lymphatic filariasis are associated with Culicinae, and species from Anophelinae are primary vectors for malaria transmission. Considerable research effort over the past decade has been directed toward developing molecular approaches to genetic analyses of selected species within both subfamilies (Severson et al., 2001). These two subfamilies, however, vary significantly taxonomically and in their genomic structure (Knudson et al., 1996; Rai and Black, 1999; Severson et al., 2001). As discussed recently, anophelines are more primitive members of Culicidae based upon their bionomics, morphology, chromosome structure and other related DNA studies (Foley et al., 1998). Hennig indicated that the superfamily Culicoidea likely existed in the Upper Triassic period (215 Myr) (Hennig, 1981) which is earlier than the 106–46 Myr suggested by recent molecular taxonomic data (Foley et al., 1998). The fossil record indicates that the family Culicidae was well evolved by 58–36 Myr (Eocene period) with most of the fossils coming from the Oligocene period (38–26 Myr). Recent molecular data are consistent with the fossil record because *Aedes* and *Culex* exhibited a divergence range of 52-24 Myr (Foley et al., 1998).

The genome size estimate for *An. gambiae* is ca. 260 Mb, while that for *Ae. aegypti* is considerably larger at ca. 780 Mb (Brown et al., 2001; Rao and Rai, 1987; Warren and Crampton, 1991). Most of the size differences between culicine and anopheline mosquitoes have been attributed to the amount and pattern of dispersion of their repetitive DNA sequences found in their genomes. Culicines exhibit a short period interspersion pattern and anophelines a long period interspersion pattern of repetitive DNA (Black and Rai, 1988; Cockburn and Mitchell, 1989; Severson et al., 2001). In short, the genome organization of *An. gambiae* is more like *Drosophila* with long stretches of largely unique sequences followed by large blocks of repetitive DNA. The *Ae. aegypti* genome organization is more similar to that seen in higher vertebrate genomes, where relatively short regions containing repetitive elements alternate with single copy sequences throughout the genome.

Aedes aegypti is an important vector of viral disease in nature, and arthropod-borne viral diseases have re-emerged as major health problems in most tropical countries. For example, the incidence of dengue fever has

increased greatly over the past 20 years, including its hemorrhagic form with associated human fatalities. Dengue virus (family Flaviviridae) occurs as four serotypes that are biologically transmitted between humans, principally by *Ae. aegypti* mosquitoes. This virus causes a nonspecific febrile illness termed dengue fever. It is also the etiological agent of dengue hemorrhagic fever, a severe and sometimes fatal form of the disease. Dengue is a threat to >2.5 billion people, with an annual incidence in the tens of millions and ~24,000 deaths per year (Anon, 1996; Beaty, 2000). No effective vaccine candidates are available nor soon anticipated for preventing dengue transmission.

It is important to note that dengue is a disease based on interactions among three rather than two different organisms: the human, the virus, and the mosquito vector. Disruption of any of these interactions can prevent disease. Indeed, historically, the greatest successes against dengue have come from controlling the vector. This realization has led to a recent resurgence of research on vector biology and molecular studies of mosquitoes in particular. While complete genome information is readily available for dengue and other mosquito-borne viruses, and most recently for humans, little detailed genome information exists for the third key organism *Ae. aegypti*. An *Ae. aegypti* whole genome sequence, therefore, would represent an opportunity that has not yet been adequately exploited: genomic studies of the world's most important vector of dengue and yellow fever virus.

Aedes aegypti is also an excellent laboratory model organism and, as such, its genetic and physical maps are extensive and more detailed than any other culicine or any other mosquito. In addition, there is a complete collection of recombinant libraries representing a >18-fold genome coverage (Severson and Knudson, unpublished). *Aedes aegypti* is clearly the best choice as the second mosquito for whole genome sequencing (Brown and Knudson, 1997; Brown et al., 1995; Brown et al., 1997; Severson et al., 2001).

In the past year, an *Ae. aegypti* genome project has been proposed and it has received wide support from the scientific community (Severson, Loftus and Knudson, unpublished). The ultimate objectives of this project are to bring contemporary genomics to bear on the study of the yellow fever mosquito, *Ae. aegypti*, such that the complete genome sequence can be elucidated, assembled and annotated. This includes a goal to promote and foster collaboration and intellectual input from the *Ae. aegypti* research community. Indeed, the resources and information generated will be made freely available to the entire research community, in a timely manner. The immediate objective of an *Ae. aegypti* genome project is to develop a body of information, including expressed sequence tag (EST) sequences, genomic sequences, and the physical map locations of large genomic clones that will enhance gene discovery and also provide critical tools for an eventual complete genome sequencing project. The availability of these resources will likely significantly influence the timetable for realization of a complete sequencing project. Importantly, it will immediately result in a body of genome information that will be invaluable to a number of ongoing and anticipated *Ae. aegypti* research projects. It will also provide EST and genomic sequence information that can be used for anticipated gene expression analysis. Since *Ae. aegypti* has an estimated genome size of 780 Mb (Rao and Rai, 1987; Warren and Crampton, 1991), the proposed *Ae. aegypti* genome project is dedicated to developing the resources critical to a whole genome sequencing effort. This project is designed to provide the framework and scaffold for the assembly of *Ae. aegypti* whole genome shotgun sequence data.

5. What level of sequence coverage should be done?

The recent successes with the *D. melanogaster* and other sequencing projects suggest that a whole genome shotgun sequencing approach is not only the most efficient, but also the most likely strategy that will eventually be applied to *Ae. aegypti*. The sheer logistics and major funding required for an *Ae. aegypti* whole genome sequencing effort indicates that this project will need to be performed by an established, large sequencing center(s).

During the assembly of *D. melanogaster*, it was noted that a 6.5× whole genome shotgun sequence coverage could be assembled with a high degree of fidelity with some gaps (Adams et al., 2000; Myers et al., 2000). The recent RFA for the comparative sequencing of *D. pseudoobscrua*, further, suggests that a genome of 150 Mb could be completed for ca. $5 M with a coverage that could be of "sufficiently high quality to form a basis for sequence comparison with the sequenced genome of *D. melanogaster*, and eventually that of yet other species."

It is important to note that the completed large genome (>100 Mb) sequencing efforts that have successfully employed shotgun sequencing have also depended heavily on extensive sequence tagged site, physical and genetic map information, and end-sequences from large insert genomic DNA clones (e.g., cosmids, BACs or YACs) for correct sequence assembly and efficient gap closure. Indeed, there is no precedent for the initiation of a shotgun sequencing effort of this scale without these preliminary data. While these resources will not be needed for *D. pseudoobscrua* sequence assembly because of its high similarity to *D. melanogaster*, these resources will definitely be needed for *Ae. aegypti* because it does not have a drosophilid or anopheline

genome organization and it exhibits a collection of repetitive elements that are not well represented in *An. gambiae* (Knudson, unpublished). Since the evolutionary distance for *Aedes* and *Anopheles* is ~27–62 Myr (Foley et al., 1998), which is a scale similar to the two drosophilids (~50–60 Myr), comparative genomics between *Ae. aegypti* and *An. gambiae* should be extremely informative. The comparative genomes between aedine and anopheline mosquitoes and drosophlids which diverged ~46–106 Myr (Foley et al., 1998) will also provide a different perspective on the chromosomal structure and evolution in dipterans. For example, a recent comparative bioinformatic analysis of *An. gambiae* ESTs and genomic sequences versus the *D. melanogaster* whole genome sequence has yielded a tantalizing glimpse of dipteran evolution (Bolshakov et al., 2002). Even though the amino acid level sequence divergence between orthologous genes ranged from 26–97%, orthologs were still readily discernable. Gross chromosomal similarity was also observed such that the five chromosomal arms of *An. gambiae* could be assigned to its *Drosophila* cognate with a high level assurance. While chromosomal similarity was seen, gene order was shuffled with the length of syntenic regions being relatively small. Clearly, the extent of microsynteny between these two dipterans will be forthcoming once *An. gambiae* is assembled (Bolshakov et al., 2002).

Using the recent *D. pseudoobscrua* RFA as guide (~$0.033/base), the estimated cost for an *Ae. aegypti* whole genome shotgun sequence would be ~$26 M at today's prices for a similar level of genome coverage (6.5×). The case for doing only a 5× coverage is also strong because this level of genome coverage will likely yield a >95% contig coverage of the whole genome and it would cost ca. $6 M less today. Clearly, this estimate is high as sequencing costs continue to drop with improving technology.

6. What will we expect from the whole genome sequence of *Aedes aegypti*?

The availability of the complete and annotated *D. melanogaster* genome, the soon to be released, assembled whole genome sequence for *An. gambiae* and a whole genome sequence for *Ae. aegypti* would constitute important resources for gene discovery and comparisons in the two most important vector mosquito species. Moreover, a number of currently important scientific questions can be addressed more efficiently by comparative genomic approaches. For example, genetic studies in the laboratory have demonstrated that arbovirus susceptibility in *Ae. aegypti* is, in part, determined by multiple gene effects (Bosio et al., 1998), and the general genome locations for quantitative trait loci (QTL) influencing both dengue virus midgut infection and midgut escape have been identified (Bosio et al., 2000). *Ae. aegypti* provides an excellent model system for investigating genetic interactions between the vector mosquito and parasitic nematodes responsible for human lymphatic filariasis such as *Brugia malayi* (MacDonald, 1962). Indeed, QTL have been identified that determine *B. malayi* midgut penetration by microfilariae (Beerntsen et al., 1995) as well as their successful development to the vertebrate infective L3 stage (Severson et al., 1994). *Ae. aegypti* also provides an excellent model system for investigating genetic interactions between the vector mosquito and a malaria parasite such as *Plasmodium gallinaceum* (Kilama and Craig, 1969). Again, QTL have been identified that determine *Ae. aegypti* susceptibility to *P. gallinaceum* (Severson et al., 1995b). Contemporary genomics provide us with essential tools for investigating, at the most fundamental level, parasite–host interactions thereby facilitating development of new approaches for preventing these diseases.

Culicine chromosome evolution largely reflects whole arm rearrangements (Matthews and Munstermann, 1994) and comparative DNA-based genetic maps have been constructed for a number of species including *Aedes albopictus* (Severson et al., 1995a), *Armigeres subalbatus* (Ferdig et al., 1998), *Culex pipiens* (Mori et al., 1999), *Culex tritaeniorhynchus* (Mori et al., 2001), *Aedes triseriatus* (Anderson et al., 2001), and *Aedes togoi* (Mori and Severson, unpublished data). These studies demonstrate that once a gene location is determined in *Ae. aegypti*, the orthologous locus in other culicine species can be predicted with a high likelihood. Once the *Ae. aegypti* whole genome sequence is completed, it will provide an immediate springboard for rapid genetic and evolutionary investigations with other culicine mosquitoes.

7. Conclusion

The *Ae. aegypti* whole genome sequence will cost ~$20,000,000 at today's prices. This represents a significant resource investment and it will require financial input from a number of diverse, concerned sources. We propose that a 5× *Ae. aegypti* whole genome shotgun sequencing effort be initiated with the goal of elucidating the vector–parasite interaction so that strategies can be developed to control effectively mosquito-vectored diseases.

Acknowledgements

This work was funded in part by National Institutes of Health Grants RO1 AI33127, RO1 AI34337, PO1 AI45123, and the Colorado Agricultural Experiment Station. We also acknowledge the Berkeley Drosophila

Genome Project, the consortium of the Drosophila Genome Center, Celera Genomics, and National Center for Biotechnology Information for their continuing efforts to make dipteran genomics information readily available to the scientific community.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287, 2185–2195.

Anderson, J.R., Grimstad, P.R., Severson, D.W., 2001. Chromosomal evolution among six mosquito species (Diptera: Culicidae) based on shared restriction fragment length polymorphisms. Mol. Phylogenetics Evol 20, 316–321.

Anon 1996. The World Health Report 1996—Fighting Disease, Fostering Development. The World Health Organization, Geneva pp. 143.

Beaty, B.J., 2000. Genetic manipulation of vectors: A potential novel approach for control of vector-borne diseases. Proc Natl Acad Sci USA 97, 10295–10297.

Beerntsen, B.T., Severson, D.W., Klinkhammer, J.A., Kassner, V.A., Christensen, B.M., 1995. *Aedes aegypti*: A quantitative trait locus (QTL) influencing filarial worm intensity is linked to QTL for susceptibility to other mosquito-borne pathogens. Exp. Parasit 81, 355–362.

Black, W.C. IV, Rai, K.S., 1988. Genome evolution in mosquitoes: intraspecific and interspecific variation in repetitive DNA amounts and organization. Genetic Research 51, 185–196.

Bolshakov, V.N., Topalis, P., Blass, C., Kokoza, E., della Torre, A., Kafatos, F.C., Louis, C., 2002. A comparative genomic analysis of two distant diptera, the fruit fly, *Drosophila melanogaster*, and the malaria mosquito, *Anopheles gambiae*. Genome Res 12, 57–66.

Bosio, C.F., Beaty, B.J., Black, W.C.I., 1998. Quantitative genetics of vector competence for dengue-2 virus in *Aedes aegypti*. Am. J. Trop. Med. Hyg. 59, 965–970.

Bosio, C.F., Fulton, R.E., Salasek, M.L., Beaty, B.J., Black, W.C.I., 2000. Quantitative trait loci that control vector competence for dengue-2 virus in the mosquito *Aedes aegypti*. Genetics 156, 687–698.

Brown, S.E., Knudson, D.L., 1997. FISH landmarks for *Aedes aegypti* chromosomes. Insect Mol. Biol 6, 197–202.

Brown, S.E., Menninger, J., Difillipantonio, M., Beaty, B.J., Ward, D.C., Knudson, D.L., 1995. Toward a physical map of *Aedes aegypti*. Insect Mol. Biol 4, 161–167.

Brown, S.E., Severson, D.W., Knudson, D.L., 1997. Correlation of the genetic linkage and FISH physical maps of the *Aedes aegypti* mosquito. In: International Plant and Animal Genome V Conference, Scherago International Inc., Town and Country Hotel, San Diego, California.

Brown, S.E., Severson, D.W., Smith, L.A., Knudson, D.L., 2001. Integration of the *Aedes aegypti* mosquito genetic linkage and physical maps. Genetics 157, 1299–1305.

Cockburn, A.F., Mitchell, S.E., 1989. Repetitive DNA interspersion patterns in Diptera. Arch. Insect Biochem. Physiol 10, 105–113.

Ferdig, M.T., Taft, A.S., Severson, D.W., Christensen, B.M., 1998. Development of a comparative genetic linkage map for *Armigeres subalbatus* using *Aedes aegypti* RFLP markers. Genome Res 8, 41–47.

Foley, D.H., Bryan, J.H., Yeates, D., Saul, A., 1998. Evolution and systematics of *Anopheles*: insights from a molecular phylogeny of Australasian mosquitoes. Mol Phylog Evol 9, 262–275.

Hennig, W., 1981. In: Pont, A.C. (Ed.), Insect Phylogeny. Wiley, Chichester, UK.

Kilama, W.L., Craig, G.B. Jr., 1969. Monofactorial inheritence of susceptibility to *Plasmodium gallinaceum* in *Aedes aegypti*. Ann. Trop. Med. Parasitol 63, 419–432.

Knight, K.L., Stone, A., 1977. A Catalog of the Mosquitoes of the World. Entomological Society of America, College Park, MD pp. 611.

Knudson, D.L., Zheng, L., Gordon, S.W., Brown, S.E., Kafatos, F.C., 1996. Genome organization of vectors. In: Beaty, B., Marquardt, W.C. (Eds.), The Biology of Disease Vectors. University Press of Colorado, Niwot, pp. 175–214 pp. 175-214.

MacDonald, W.W., 1962. The genetic basis of susceptibility to infection with semi-periodic *Brugia malayi* in *Aedes aegypti*. Ann. Trop. Med. Parasitol 56, 373–382.

Matthews, T.C., Munstermann, L.E., 1994. Chromosomal repatterning and linkage group conservation in mosquito karyotic evolution. Evolution 48, 146–154.

Mori, A., Severson, D.W., Christensen, B.M., 1999. Comparative linkage maps for the mosquitoes (*Culex pipiens* and *Aedes aegypti*) based on common RFLP loci. J. Heredity 90, 160–164.

Mori, A., Tomita, T., Hidoh, O., Kono, Y., Severson, D.W., 2001. Comparative linkage map development and identification of an autosomal locus for insensitive acetylcholinesterase-mediated insecticide resistance in *Culex tritaeniorhynchus*. Insect Mol. Biol 10, 197–203.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., et al. 2000. A whole-genome assembly of *Drosophila*. Science 287, 2196–2204.

Rai, K.S., Black, W.C.I., 1999. Mosquito genomes: structure, organization and evolution. Adv. Genetics 41, 1–33.

Rao, P.N., Rai, K.S., 1987. Inter and intraspecific variation in nuclear DNA content in *Aedes* mosquitoes. Heredity 59, 253–258.

Severson, D.W., Brown, S.E., Knudson, D.L., 2001. Genetic and physical mapping in mosquitoes: molecular approaches. Ann. Rev. Entomol 46, 183–219.

Severson, D.W., Mori, A., Kassner, V.A., Christensen, B.M., 1995a. Comparative linkage maps for the mosquitoes, *Aedes albopictus* and *Ae. aegypti* based upon common RFLP loci. Insect Mol. Biol 4, 41–45.

Severson, D.W., Mori, A., Zhang, Y., Christensen, B.M., 1994. Chromosomal mapping of two loci affecting filarial worm susceptibility in *Aedes aegypti*. Insect Mol. Biol 3, 67–72.

Severson, D.W., Thathy, V., Mori, A., Zhang, Y., Christensen, B.M., 1995b. Restriction fragment length polymorphism mapping of quantitative trait loci for malaria parasite susceptibility in the mosquito *Aedes aegypti*. Genetics 139, 1711–1717.

Warren, A.M., Crampton, J.M., 1991. The *Aedes aegypti* genome: complexity and organization. Genetical Research 58, 225–232.