Pacific Symposium on Biocomputing 7:512-523 (2002)

SHALLOW GENOMICS, PHYLOGENETICS, AND EVOLUTION IN THE FAMILY DROSOPHILIDAE

M. ZILVERSMIT, P. O'GRADY, R. DESALLE
American Museum of Natural History, Department of Invertebrate Zoology, Central Park West @ 79th Street, New York, NY 10024, USA

The effects of the genomic revolution are beginning to be felt in all disciplines of the biological sciences. Evolutionary biology in general, and phylogenetic systematics in particular, are being revolutionized by these advances. The advent of rapid nucleotide sequencing techniques have provided phylogenetic biologists with the tools required to quickly and efficiently generate large amounts of character information. We use family Drosophilidae as a model system to study phylogenetics and genome evolution by combining high throughput sequencing methods from the field genomics and standard phylogenetic methodology. This paper presents preliminary results from this work. Separate data partitions, based on either gene function or linkage group, are compared to a combined analysis of all the data to assess support on phylogenetic trees.

1 Introduction.

The traditional goal of molecular systematics has been to use the character information found in one or a few well-characterized loci to infer the evolutionary history of a selected group of taxa. The assumptions here are twofold; first, that the gene history will accurately represent the history of the taxa sampled and second, that the evolutionary rate of the gene selected will provide ample characters to resolve each node in the phylogeny. Many authors have rightly pointed out that the history of a single gene sequence sampled for a species is not necessarily reflective of the history of that species.[1] Reasons for this discrepancy include horizontal gene transfer, lineage sorting of ancestral polymorphisms, and other phenomena.[2] Furthermore, depending on the divergence times of the taxa being examined, one or a few genes may not contain a sufficient number of characters to robustly reconstruct all nodes in the phylogeny.

A variety of genome projects have successfully employed high throughput methods to rapidly and reliably generate large numbers of sequences.[3,4] As a result, a wealth of characters are now readily available to the molecular systematist. Many systematists have begun to take a "phylogenomic" approach by using genomic information to infer phylogenetic relationships. However, other than prokaryotic and organellar genomes, it is unlikely that most systematists will ever be able perform research on a truly genomic scale. Using a large number of loci, sampled throughout the genome, is a far more powerful way to infer phylogenetic

relationships than currently being employed. Such an approach not only yields a more accurate view of how the genome as a whole is evolving, but will also increase the likelihood that loci evolving at many different rates are sampled, thereby increasing the numbers of characters supporting each node. Even though the sequences of entire genomes are not being compared, this genomic sample approach to systematics can be referred to as shallow genomics.

We are currently taking a similar approach to address phylogenetic relationships within the family Drosophilidae. We have developed a relatively simple and cost-effective system for high throughput gene sequencing. High throughput sequencing is not new, as it is exactly this sort of approach that enabled the existence of the genome projects. The methods used for those projects and in commercial laboratories are beyond the reach of most academic laboratories because of the tremendous cost of reagents, equipment, and labor. We have developed a system that is inexpensive enough to be used by several researchers in the same laboratory while being simple enough to be completed by a single individual. This was accomplished by pooling existing methods and protocols (from publications, Web sites and personal communication) and then combining and modifying them.[5,6] The improvements are enhanced by using a capillary sequencer, a system that greatly increases ease and efficiency of data generation relative to slab gel systems.

1.1 Drosophilidae as a shallow genomic model system

This family is an excellent model system for genomic studies because the entire sequence of *Drosophila melanogaster* is now completed and that of a second drosophilid, *D. pseudoobscura*, is nearing completion. Coupling this impressive amount of sequence data with gene location data derived from in situ hybridization to the polytene chromosomes, we can insure that the entire genome is sampled.

Phylogenetic relationships among drosophilid genera based on molecular and some morphological analyses are incongruent.[7,8] The major conflict is in the placement of the endemic Hawaiian Drosophilidae, a lineage consisting of two genera, *Drosophila* and *Scaptomyza*. Grimaldi[9] proposed that the Hawaiian *Drosophila* was actually closely related to a clade of mycophagous genera, the *Hirtodrosophila* genus complex. The Hawaiian *Scaptomyza* formed a clade with the continental *Scaptomyza* and were actually more closely related to the subgenus *Drosophila* than the Hawaiian *Drosophila*. In contrast, molecular data,[7] as well as previous morphological work,[10] suggest that the Hawaiian *Drosophila* and all *Scaptomyza* (Hawaiian and continental) form a clade. This group is closely related to the subgenus *Drosophila*, not the *Hirtodrosophila* genus complex.

Our goals are to (1) test the monophyly of the Hawaiian Drosophilidae and (2) determine the sister group relationships of the Hawaiian *Drosophila* and the genus

Scaptomyza. These analyses will examine congruence between partitions generated from different chromosomes and different functional classes of genes (i.e., transcription factors and enzymes). Although the explicit intention of using shallow genomic methods in our lab is for phylogentics, we expect that the data gathered will reveal information about genome evolution in a number of drosophilid species. Our methods allow us to gather data from unstudied genes and gene regions to examine not only organismal evolution, but molecular and genome evolution as well.

2 Methods

1.1 Primer design and testing

Our primer design takes advantage of the large number of known sequences from *Drosophila*, humans, and other organisms to design large numbers of primers that are functional across large phylogenetic distances. Like the CATS system,[5] we have systematically scanned GenBank for genes and gene regions that might be appropriate for primer design and use. Oligonucleotide primers homologous to *D. melanogaster* genes are designed from each chromosomal segment via GADFLY (Genome Annotation Database of *D. melanogaster*). These sequences, and those from two or more homologous genes from other species, are then input to the CODEHOP web site.[6] CODEHOP produces ortholog blocks of sequence and a series of potential primers using the genetic code and codon bias tables. Optimal primers are synthesized based on the following criteria:

(1) The primers should amplify a fragment between 300-800 bp. Shorter fragments are better if template DNA has been fragmented, while longer targets are more economical, as 800 bp is the current maximum fragment size that can be sequenced with a single primer in both directions.
(2) Target sequences should be variable as to the level of sequence conservation between taxa, representing as complete a range as possible.
(3) Primers are chosen with low to intermediate degeneracy, (less than or equal to 32-fold).

In order to automate the sequencing step, we have incorporated the sequences for the T3 and T7 universal sequencing primers into our oligonucleotides. These universal primer sites are also located on the vector (TOPO pCR4; Invitrogen) used for cloning PCR products should that be necessary. With this addition only this

single set of primers is needed to sequence any number of different genes, either directly or ligated into a vector, and enables significantly greater throughput in less time for virtually no extra cost.

Primers are initially screened for success in PCR, assessing both for effectiveness in amplifying gene fragments from divergent taxa (e.g., multiple diptera, plus sturgeon, teleost and mammal samples) and the number of products produced. Testing on such a broad range of taxa is done to indicate both how well a primer will work with members of the family Drosophilidae, for our immediate test study, and its amplification success among a variety of other taxa for later work. In order to reduce problems due to secondary amplification and paralogy, only those primers producing single bands are used for sequencing.

1.2 Gene fragment amplification, processing and sequencing

PCR products are then purified either using a standard isopropanol precipitation method adapted for microtitre plates or via filtration using SOPE resin (Edge Biosystems) with lab-made resin columns in a microtitre-style column rack. The former method has a significant cost effectiveness advantage and the latter permits superior speed and ease at low cost for a commercial product. The purified PCR products are then used as the template for dye terminator (BigDye; ABI) sequencing reactions using 0.5ul dye terminator mixed with an equal part dye terminator extender buffer (Tris-HCl, pH9) per reaction with only universal sequencing primers. Using such a low volume of dye terminator allows for some of the most significant cost reduction for a product that cannot easily be replaced. The sequencing reactions are also purified by an adapted alcohol precipitation method and resuspended in 5ul ABI deionized formamide and sequenced on an ABI Prism 3700 DNA analyzer.

The Sequencher software package (Gene Codes) was used to edit raw sequences and create consensus sequences from several clones. Sequences were aligned either by eye or by using a multiple alignment program.[11] PAUP, version 4.0[12] was then used to generate phylogenetic trees using parsimony, likelihood, and distance methods. Other sequence analyses were performed in MacClade.[13]

2.3 Phylogenetic Analyses

PAUP 4.0[13] was used to perform all analyses in this study. Most parsimonious trees (MPTs) were found using an exhaustive search algorithm. Tree characteristics, shown in Figures 1 and 2, are number of MPTs, number of steps on the shortest tree, number of parsimony informative characters (PICs), and the consistency index (CI). Bootstrap proportions[14] and decay indices[15] were used as measures of support

on the phylogeny. Partitioned branch support[16] was calculated using TreeRot[17]. Best estimate data removal indices[18] were also calculated.

3. Results

3.1 Primer Design and Amplification.

To date, we have designed 500 primers, amplifying a range of different functional classes of genes, for loci located on each major linkage group in *Drosophila*

Table 1. Characteristics of the loci used in this study.

Gene/Locus	Size (bp)	PICs	% PIC[1]	Functional Class
16S*	908	66	7.2 / 3.5	mitochondrial ribosomal RNA
28S*	812	29	3.6 / 1.5	nuclear ribosomal RNA
Adh*	771	198	25.7 / 10.5	enzyme
amy	186	4	2.2 / 0.2	enzyme
BcDNA	630	119	18.9 / 6.3	peptide transporter
boss	225	31	13.8 / 1.7	signal transduction
COII*	688	127	18.5 / 6.8	mitochondrial protein coding
dpp	436	59	13.5 / 3.1	TGF-B receptor ligand
dsh	468	20	4.3 / 1.1	segment polarity
esc	374	55	14.7 / 2.9	gene silencing
fkh	201	35	17.4 / 1.9	RNA polymerase II trans. fact.
Gpdh*	770	119	15.5 / 6.3	enzyme
glass	763	48	6.3 / 2.6	RNA polymerase II trans. fact.
kuz	544	2	0.4 / 0.01	enzyme (metalloendopeptidase)
mago	360	9	2.5 / 0.5	germ plasm assembly
pdm2	599	4	0.7 / 0.02	RNA polymerase II trans. fact.
Sod*	439	131	29.8 / 7.0	enzyme
sia	448	56	12.5 / 3.0	zinc finger domain
snf	295	116	39.3 / 6.2	small nuclear ribonucleoprotein
Xdh*	2078	449	21.6 / 23.9	enzyme
PSAP	448	74	16.5 / 3.9	enzyme
26S prot	431	101	23.4 / 5.4	26S proteosome regulatory subunit
CG3869	268	26	9.7 / 1.4	unknown
Total	**13124**	**1878**		

1. Percent parsimony informative characters in each locus / percent parsimony informative characters in entire data matrix. * Primers used prior to this project.

melanogaster. Of these primer pairs 179 have been synthesized and tested. Thus far, we have been able to successfully amplify 125 of the 179 primer pairs we have tested. Primer pairs that fail to give any amplification product are immediately rejected. Of the 125 primer pairs that gave a PCR product, about 2/3 give strong, single band amplification products that we anticipate little difficulty sequencing 16 of which are at a level of completion to be used in the preliminary data matrix (Table 1). These loci are found on the X, second and third chromosomes of *Drosophila melanogaster* and represent a broad range of functional classes of genes.

3.2 Combined Phylogenetic Analysis.

Our phylogenetic results indicate that the endemic Hawaiian Drosophilidae (Hawaiian *Drosophila* plus *Scaptomyza*) are monophyletic. *Scaptomyza* is more

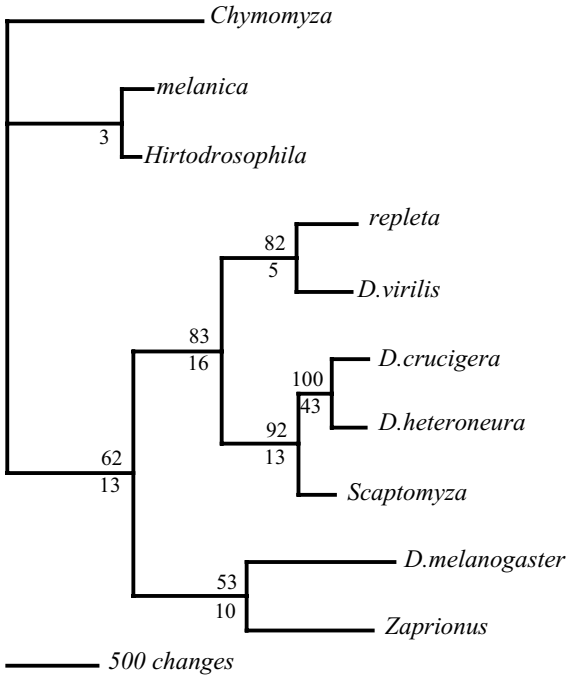

Figure 1. Phylogram showing relationships based on maximum parsimony analysis of 23 molecular partitions (Table 1). 1MPT, 7591 steps, 1879 PICs. Support is indicated above (bootstrap proportions) and below (decay indices) each node.

closely related to the Hawaiian *Drosophila* species *D. crucigera* and *D. heteroneura*, suggesting that the genus *Scaptomyza* may have originated on Hawai'i and subsequently colonized the rest of the world. The Hawaiian Drosophilidae clade is the sister group of two subgenus *Drosophila* species groups, *virilis*

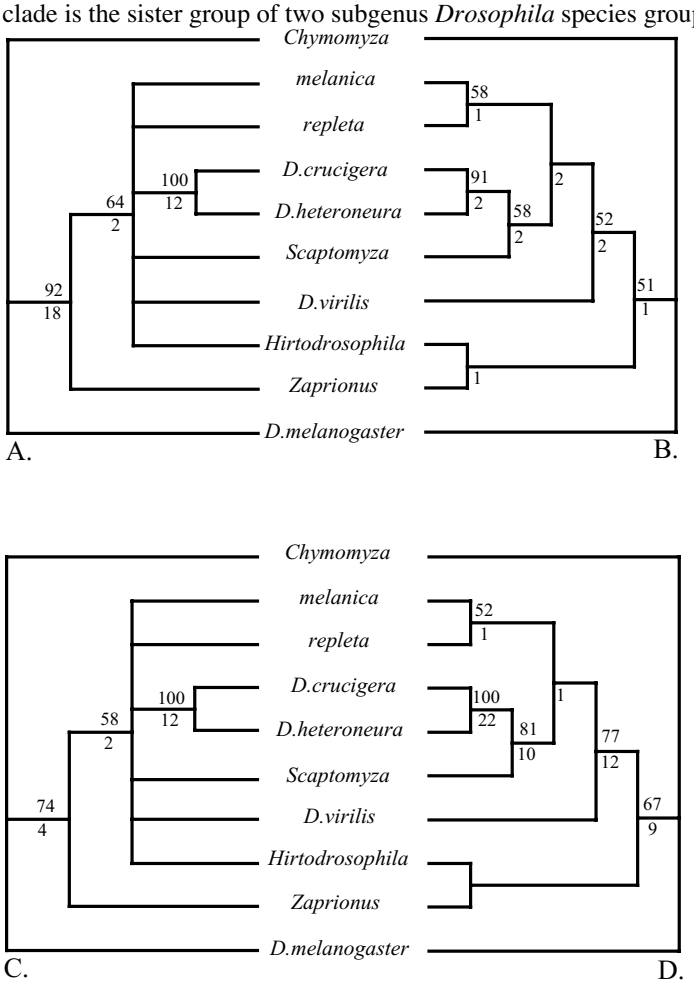

Figure 2A-D. Phylogenetic analysis of several partitions examined in this study. A. Enzyme tree (Adh, amy, Gpdh, kuz, PSAP, Sod, Xdh). 6MPTs, 3539 steps, 977 PICs, CI – 0.760. **B.** Transcription factor tree (BcDNA, boss, dpp, dsh, fkh, glass, mago, pdm2, sia). 1MPT, 2106 steps, 381 PICs, CI = 0.872. **C.** Chromosome 2 tree (26S, dpp, Gpdh, esc, pdm2, kuz, Adh, amy, mago). 10 MPTs, 2247 steps, 551 PICs, CI = 0.823. **D.** Chromosome 3 tree (PSAP, Sod, sia, Xdh, glass, boss, fkh). 1 MPT, 3320 steps, 824 PICs, CI = 0.782. Support is indicated above (bootstrap proportions) and below (decay indices) each node.

(represented by *D. virilis*) and *repleta*. These data suggest that the genus *Drosophila* is not monophyletic with respect to the genera *Scaptomyza* and *Zaprionus*. Interestingly, the subgenus *Sophophora*, represented by *D. melanogaster,* is quite basal within the drosophilid taxa we sampled here. This group may constitute a lineage separate from the genus *Drosophila*, as some previous studies have suggested.[19] Also of interest is that though other forms of statistical support are relatively strong here, best estimate data removal indices[18] are relatively low (1-3 for all nodes with the exception of the *Scaptomyza* -Hawaiian clade, estimated at 5), indicating that still more data is needed to produce the strongest phylogeny.

The 23 molecular loci were analyzed in several partitions. First, we divided the genes based on functional class, either enzymes (Fig. 2A) or transcription factors (Fig. 2B). The enzyme tree (Fig. 2A) is based on the coding regions of seven loci, which contain a total of 977 parsimony informative characters, and is quite unresolved. It does, however, support the sister group relationship of the two Hawaiian *Drosophila* species sampled, *D. crucigera* and *D. heteroneura*, as well as the placement of genera *Hirtodrosophila* and *Zaprionus* as close relatives of the subgenus *Drosophila*. The transcription factor tree (Fig. 2B), which is based on nine loci containing a total of 381 parsimony informative characters, is most similar to the combined analysis tree, although support on this phylogeny is generally not high.

Next we examined the three major linkage groups in the *Drosophila melanogaster* genome. Loci located on the first chromosome yielded a completely unresolved phylogeny (data not shown). Those on the second (Fig. 2C) and third (Fig. 2D) chromosomes, however, were quite resolved. The third chromosome tree (Fig. 2D) was most similar to the combined analysis tree (Fig. 1). Bootstrap proportions at four nodes on this tree were near or above 70%. The major difference was in the placement of several subgenus *Drosophila* species groups, *repleta*, *virilis*, and *melanica*.

4. Discussion

Using a genome-based approach is necessary for the studies of phylogenetics and molecular evolution to reach their full potential. Studying these elements of biology by using a few, carefully selected genes has been a practice born out of the limitations of technologies and techniques as much as out of intellectual paradigms. Now that it is clear that many of these restrictions no longer apply it is time to benefit from genomic data.

4.1 Drosophilidae Phylogeny: Molecular Evidence

Note that seven of the loci used (16s, 28s, Adh, COII, Gpdh, Sod and Xdh) in this preliminary analysis (Table 1) were not developed for this project and have been used in previous studies of drosophilid systematics and evolution. However, these seven loci do not represent a broad range of chromosomal positions or functional classes (they are all ribosomal or mitochondrial genes or code for enzymes). This sampling is biased toward enzyme coding genes, which evolve slowly relative to other types of sequences, such as transcription factors or introns.

We have examined a number of partitions, defined based on linkage group or functional class, in addition to the combined data matrix. Such comparisons highlight the benefits of shallow genomics and combined phylogenetic analysis by showing that a tree based on a single gene, gene class, or linkage group does not generate a phylogeny as resolved and robust as the entire data matrix analyzed simultaneously. A genomic-based sample may be more effective at reconstructing the phylogenetic relationships than an analysis producing a tree based on the congruencies or conflicts of just a few genes.

We divided the genes based on functional class and analyzed two partitions, enzymes (Fig. 2A) and transcription factors (Fig. 2B). While the enzyme tree is mostly unresolved at the tips, it is well supported at the base. In contrast, the transcription factor tree shows the most support for nodes at the tips. This is consistent with the notion that transcription factors are more rapidly evolving relative to enzyme coding genes. These results suggest that, when approaching a specific phylogenetic question, it may be most effective to select a gene based on function. Evaluation by linkage group (Fig. 2C-D) shows a similar example of a group of slow versus fast genes. Thus using a combination, and representative sampling, of all gene aids in producing a tree that has both better resolution and support, and is based more closely on the overall evolution of the taxa.

4.2 Future Directions: Interpreting and Evaluating Conclusions from Shallow Genomic Data

One of the limitations to generating a large number of gene sequences via shallow genomic methods is that there are not many algorithms that allow one to easily analyze these disparate data fully. For example, the program TreeRot.2b[17] which is widely available can only calculate the partitioned branch support for up to 12 data sets. Partitioned branch support data was only efficiently produced for the shallow genomic data after the program was altered by the creator to process up to 30 partitions. Note also that the data removal indices mentioned above are only listed as best estimates. It is necessary to list them tentatively because the methods and algorithms for properly calculating this measure of support between 23 data sets does not exist. These are just a few examples of how technical and algorithmic systems available are inadequate for large data sets. Now that this volume of data

can be produced, it is up to those in the field of informatics and computer science to meet the challenge presented by it.

4.3 Additional Data on Genome Evolution

It has been mentioned above that we expected our methods to reveal multiple indications and artifacts of the evolution of the genome studied. Although our work is still in its preliminary stages we have already come across a number of interesting phenomena using previously unstudied genes and gene regions. Each of these is a paper unto itself, but we will summarize three examples to highlight the information that can be revealed with the survey methods we describe above.

The first example concerns the evolution of CAG repeats. The glass gene encodes a specific RNA polymerase II transcription factor which has been implicated in photoreceptor determination in *Drosophila melanogaster*.[20] We have examined a portion of exon four in several members of the genus *Drosophila* as part of the genus level project described above. The glass gene of *D. micromelanica*, a species in the *melanica* group, contains a large polyglutamine (CAG) repeat region not been found in any other species yet examined. This region is interesting in that it may be used as a model to better understand the molecular basis of how repeat regions form. Such work has applications beyond genome evolution in *Drosophila*, and can be used in the study of a number of diseases since polyglutamine repeats, particularly CAG repeats, have been implicated in at least eight human neurological disorders. These diseases include Huntington's Disease for which *Drosophila* is now one of the emerging model systems.[21,22]

In the course of examining our loci we have also come across an interesting case of horizontal gene transfer. In the 26s proteosome encoding region there is a 60 base pair insertion in three of the taxa in our analysis. BLAST[23] searches revealed that these inserts are possibly viral in origin. Whether these represent excised viruses, some sort of coevolving polydnavirus, or a third option remains to be seen and is a very interesting question under study in our laboratory.

The third element of interest found was an unusually long non-coding region in the CG3869 gene fragment, with significant phylogenetic uses beyond our higher level study. This gene was discovered only during the genome project and its function is still unknown. Finding introns in *Drosophila* genes and designing primers for them is not especially difficult, yet it is not generally worth the effort as *Drosophila* introns tend to be rather small. The CG3869 non-coding region, by contrast, is roughly 500bp and shows a surprising amount of variation, including a number of indels and microsatellites. This region has already proven to be a useful in species and population level studies done in our lab and the primers are in high demand from those in other labs who want to do the same.

It is the accumulation of data on individual changes like those above that will allow people using the shallow genome techniques to examine the finer aspects of genome evolution, beyond the coarser level of just gene placement, duplication or elimination

5 Conclusion

High throughput sequencing has had a significant impact on large scale genome sequencing projects.[24] Shallow genome sequencing techniques target a subset of the total genome, making genomics much more tractable for evolutionary and population genetic studies. Determination of sequences from non-model system taxa (or even studies of polymorphism within model systems) will not only prove effective in addressing the questions above, but will also lead to a better understanding of the genes involved in human development and disease.[25] A phylogenetic approach allows for a more complete understanding of the architecture of genomic change that occurs over evolutionary time. The genomic information we collect in the next few decades will not only aid us in reconstructing phylogeny, but will also address a wide range of questions pertinent to how genomes evolve.

Acknowledgements

The first and second authors contributed equally to the content of this paper. Additional help came from Jessica Chen, Jeremy Lynch, Julian Stark, Ilya Temkin and Jake Wintermute. Dr. Michael Sorenson graciously wrote a new version of TreeRot.2 (version "c") to be compatible with our large data matrix. We also would like to thank two anonymous reviewers for their comments and suggestions.

References

1. D. R. Maddison, *Syst. Zool.* **40**:315-340 (1991).
2. J.B. Clark, W.P. Maddison and M.G. Kidwell. *Mol. Biol. Evol.* **11**:40-50 (1994).
3. G. M. Rubin, *et al.*, *Science* **287**: 2222-2224 (2000).
4 S. A. Chervitz, *et al.*, *Science* **282**:2022-2027 (1998).
5. L. A. Lyons, M. M. Raymond, and S. J. O'Brien, *Animal Biotechnology* **5**:103-111 (1997).
6. T. M. Rose, *et al.*, *Nucleic Acids Res.* **26**:1628-1635 (1998).
7. R. DeSalle and D. Grimaldi, *Ann. Rev. Ecol. Syst.* **22**:447-475 (1991).

8. R. DeSalle and D. Grimaldi, *J. Hered.* **83**:182-188 (1992).
9. D. Grimaldi, *Bull. Am. Mus. Nat. Hist.* **197:**1-139 (1990).
10. L. H. Throckmorton, *Univ. Tex. Publs.* **6615**:335-396. (1966)
11. W. C. Wheeler and D. Gladstein, American Museum of Natural History, New York (1993).
12. D. L. Swofford, Sinauer Press, Washington (2000).
13. W. P. Maddison and D. R. Maddison, Sinauer Press, Washington (1992).
14. J. Felsenstein, *Evolution* **39**: 783-791 (1985).
15. K. Bremer, *Evolution* **42**:795-803 (1988)
16. R. Baker and R. DeSalle, *Syst. Biol.* **46**:654-673 (1997).
17. M .D. Sorenson, TreeRot, version 2, Boston University, Boston MA (1999).
18. J. Gatesy, P.M. O'Grady, and R.H. Baker, *Cladistics*, **15**:271-313 (1999)
19. L. H. Throckmorton, in *Handbook of Genetics: Invertebrates of Genetic Interest,* R. C. King, Ed. (Plenum, New York, 1975).
20. K. Moses , M.C. Ellis and G.M. Rubin., *Nature* **340**:531-536 (1989)
21. G. R Jackson, *et al*., *Neuron* **21**, 633-642 (1998).
22. Mitchell, A. , *Nature* **395**, 841 (1998)
23. S. F. Altschul, *et al., Nucleic Acids Res.* **25**:3389-3402 (1997)
24. J. C. Venter, *et al., Science* **280**:1540-1542 (1998).
25. A. R. Templeton, *Ann. Rev. Ecol. Syst.* **30**:23-49 (1999).