A White Paper Requesting BAC Library Construction: Drosophila as a Model for Comparative Genomics

Submitted by Therese Ann Markow
Dept. of Ecology and Evolutionary Biology, BSW 310
University of Arizona, Tucson, AZ 85721
Office: (520) 621 3323
Email: tmarkow@arl.arizona.edu

Bryant F. McAllister
Dept. Biological Sciences, 138 BB
University of Iowa, Iowa City, IA 52242
Office (319)335-2604
Email: bryant-mcallister@uiowa.edu

Thomas Kaufman
Dept. of Biology
Indiana University, Bloomington, IN 47405
Office (812) 855-3033
kaufman@bio.indiana.edu

On behalf of the Tucson *Drosophila* Species Stock Center

Introduction

Genome sequences from a wide variety of eukaryotes are accumulating at an astounding pace and the informatics and research communities are facing a diversity of problems annotating these genomes and validating the functional roles of the annotated sequences. For example, most predicted coding sequences are not associated with any known function (Adams et al. 2000). For those well-defined genes that do have known biological functions, annotation of sequences important for *cis*-regulation is still in its infancy (Ohler et al. 2002). Furthermore, networks of gene interactions are even more poorly understood (Halfon & Michelson 2002).

One recognized mechanism for genome-wide functional annotation and validation is the use of cross-species comparative analyses (Bergman et al., 2003; Boffelli et al. 2003). This White Paper requests the construction of BAC genomic libraries for a set of species in the genus Drosophila that will facilitate comparative studies designed to **1) provide sequencing resources for comparative annotation of the *D. melanogaster* genomic sequence, and 2) provide genomic resources for experimental investigation of gene function throughout the genus Drosophila.**

Importance of the Organism: A Model for Comparative Genomics

Research over the past century has established *D. melanogaster* as the premier model system for understanding basic processes important to genetics, developmental biology, neurobiology, and medicine. However, this species is only a single member of a diverse genus known to contain approximately 2000 species. The genus *Drosophila* has proven itself as an unprecedented model for comparative experimental research with many species having been the focus of studies of genome evolution, morphological variation, physiological adaptation, behavioral evolution, ecological specialization, phylogenetic systematics, and species differentiation (reviewed in, Powell 1997). In fact, no other group has such a well-defined phylogeny and an extensive literature on genetics, genome content, evolutionary history, ecology, and behavior.

Many resources are available for studies capitalizing on the diversity within the genus *Drosophila*. About 250 species, and many wild type and mutant lines within these species, are maintained in culture at the NSF-supported Tucson *Drosophila* Species Stock Center http://stockcenter.arl.arizona.edu. The Tucson Stock Center has experienced a nearly two-fold increase in orders for diverse species over the last three years, a trend that reflects the needs and research directions of the genetics community. These lines are a resource for use of universal transformation vectors (Horn & Wimmer 2000) in reciprocal transformation studies of gene function among species within the genus *Drosophila*. Furthermore, the current finishing efforts on the genome sequence of *D. melanogaster* (Celniker et al. 2002) and the recent release of a first assembly of the genome sequence of *D. pseudoobscura* (Human Genome Sequencing Center 2003), will further stimulate comparative studies utilizing the diversity within the genus *Drosophila*. It is anticipated that the genus *Drosophila* will continue to be utilized as a model for research in comparative genomics through the development of computational techniques, through the use of experimental comparative approaches in functional validation, and through the use of phenotypic diversity in gene pathway discovery.

Use of the BAC Libraries

The greatest advances in experimental comparative genomics using *Drosophila* species will result from the development of resources that enable functional analysis at the sequence level in any member of the genus, and the requested BAC libraries are essential to the realization of this goal. We propose the construction of BAC libraries for 20 species representing a broad spectrum of phylogenetic diversity (Figure 1) within the genus *Drosophila*. These species can be divided into three broad

classes based on their relation to *D. melanogaster*: (1) closely related species in the *melanogaster* group, (2) other members of the subgenus Sophophora in the *obscura* and *willistoni* groups and (3) more distantly related species in the subgenus Drosophila.

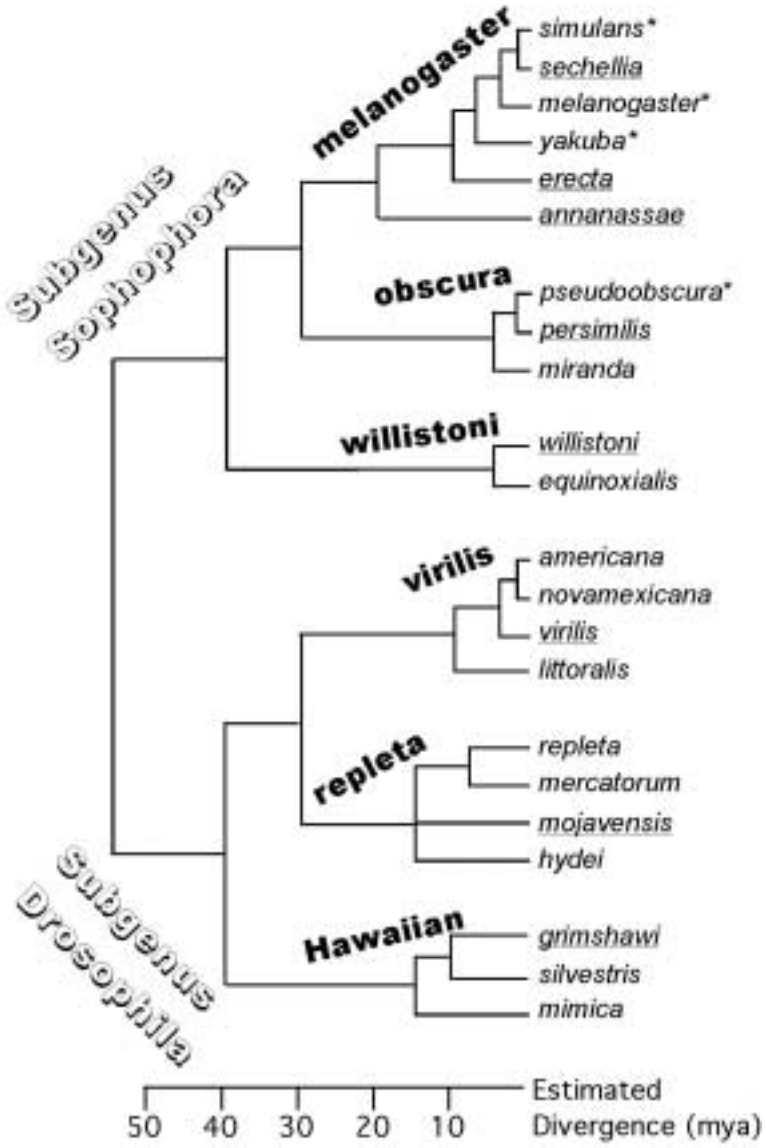

Figure 1. Phylogenetic relationships among the species selected for BAC library construction. Libraries are already available for *D. melanogaster* and *D. pseudoobscura*, which are included for reference. An asterisk beside a species name indicates its genome has been selected for sequencing. Underlining indicates the species has been proposed to have its genome sequenced. The major species groups targeted by this request and the two subgenera are identified. Estimates of divergence times based on Russo et al. 1995 and Powell 1997.

BAC libraries are a critical component of large-scale sequencing projects. White Papers requesting genome sequences for ten Drosophila species for which BAC libraries are also requested have been submitted and the motivating questions for the genome sequences are detailed in the sequencing requests (Begun and Langley 2003; Clark et al. 2003). By creating an additional set of ten libraries from species that are closely related to the targets of genome sequencing efforts, we will greatly facilitate experimental genomics research in these groups. Classical *Drosophila* genetics has relied on saturation mutagenesis in *D. melanogaster* as a means for relating genotype and phenotype and identifying genetic pathways. A tangible benefit of BAC libraries will be to broadly expand the use of existing phenotypic diversity within the genus as a means of comparative analysis of gene function and pathway discovery. Below we describe the species for which BAC libraries are requested.

melanogaster group: The large number of researchers using *D. melanogaster* as an experimental model and the findings resulting from their activities creates the impetus for further

annotation of its genome sequence. A series of species, all of which are placed in the *melanogaster* species group of the subgenus Sophophora, but are successively more distantly related to *D. melanogaster* are proposed as targets of BAC library construction. These include, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. annanassae*. In addition to the usefulness of these species in annotating the genome of *D. melanogaster*, this set of species from the *melanogaster* group provides a framework for identifying and testing targets of adaptive evolution. Genome sequences have been requested for all of these species (Clark et al. 2003), and BAC libraries will be a necessary component of finishing efforts within these species. Widely available libraries will also provide resources for obtaining segments of genomes for experimental tests of hypotheses arising from the bioinformatics community.

obscura and willistoni groups: The subgenus Sophophora is a diverse group which contains the *melanogaster* species group and two other important species groups, *obscura* and *willistoni*. A representative of the *obscura* species group, *D. pseudoobscura*, is already in the late stages of having its genome sequenced. This request proposes construction of BAC libraries from *D. persimilis* and *D. miranda*, which are both important species within the *obscura* group. Genomic resources (BACs and sequence) for *D. persimilis* will facilitate studies of rapidly evolving differences between closely related species, because of its close relationship with *D. pseudoobscura*. *D. miranda* is an important model for studies of dosage compensation due to the presence of a unique chromosomal rearrangement that has resulted in the formation neo-sex chromosomes (Bone & Kuroda 1996). BAC libraries are also requested for two more divergent species in the subgenus *Sophophora*, *D. willistoni* and *D. equinoxialis*, of the *willistoni* species group. The genome of *D. willistoni* is a candidate for sequencing, and an additional BAC library of *D. equinoxialis* will enable further study within the group. All of the proposed species are members of well-studied groups, with genetic mutants and physical and/or linkage maps. Because these species are strategically placed at increasing levels of divergence relative to *D. melanogaster*, they will facilitate sequence-level investigations within these groups.

The genus Drosophila is divided into two major lineages, the subgenus Sophophora (which contains the *melanogaster*, *obscura* and *willistoni* groups) and the subgenus Drosophila. Estimated divergence between these subgenera is roughly 40-60 million years (Powell 1997). Interestingly, this is the same timeframe as the origin and diversification of the ancestral primate lineage (Goodman 1999). We propose the development of libraries for three groups within the subgenus Drosophila, the *virilis* and *repleta* species groups and the Hawaiian Drosophila lineage. Although this only represents a portion of the diversity within the subgenus *Drosophila*, each of these groups has salient points supporting their selection.

virilis group: The *virilis* species group contains *D. virilis*, a widely-used model for comparative analysis of gene function. Over 75 different *D. virilis* gene sequences have been obtained in an attempt to identify conserved gene regions and, by inference, reveal regions of functional importance. Many of these genes have been transformed into *D. melanogaster* to examine functional equivalence of the sequences. We propose BAC libraries construction for *D. virilis*, *D. littoralis*, *D. americana*, and *D. novamexicana*. A BAC library of *D. virilis* will facilitate sequencing efforts in this species. Furthermore, the total genome size of *D. virilis* is more than twice as large as species in the melanogaster group, and initial sequencing studies reveal the presence of simple repetitive sequences within the euchromatin of *D. virilis* (Bergman et al. 2002). The closely related species *D. littoralis* appears to have a smaller genome with less repetitive DNA (Bergman et al. 2002). Therefore, these species represent a good model for examining mechanisms affecting the evolution of genome size. Finally, *D. americana* and *D. novamexicana* are proposed as candidates for BAC library construction because they are highly interfertile and can be used for quantitative trait mapping (Wittkopp et al. 2003), further facilitating gene identification and functional validation.

repleta group: The *repleta* species group is an extremely diverse clade of more than 100 species, many of which are cactophilic. A set of species representing the diversity within the *repleta* group has been selected for BAC library construction. These include *D. mojavensis, D. repleta, D. hydei,* and *D. mercatorum*. An extensive amount of information has been obtained on the ecology and reproductive biology of *D. mojavensis* and close relatives. The genome of this species is the requested focus of genome sequencing efforts for the *repleta* group. Other species in the group were selected based on representation of strategic lineages within the group, and the fact that each has unique biological aspects. Comparative genomic analyses have already focused on *D. repleta*, studies of heterochromatin and the Y chromosome have focused on *D. hydei*, and existence of parthenogenesis has been documented in *D. mercatorum*.

Hawaiian lineage: The endemic Hawaiian Drosophila, with approximately 1000 species, is one of the most astounding adaptive radiations and has served as a model for the study of biological diversity. These species offer unique opportunities to examine the genetic basis of extreme morphological and behavioral differences. Two species in the picture-wing species group, *D. grimshawi* and *D. silvestris*, and one species in the modified mouthpart species group, *D. mimica*, are proposed as inroads for genome–level analysis of the Hawaiian *Drosophila*.

The set of species for which BAC libraries are requested will further advance Drosophila as a model system for studying basic mechanisms that are important for genetics, developmental biology, neurobiology, and medicine. Although there are many unforeseen research areas that will derive from genomic resources, there are important problems that will immediately be amenable to study by using the diversity within the genus. For example, lifespan of flies in the subgenus Drosophila can be an order of magnitude longer than *D. melanogaster*, presenting unique opportunities for studies of aging. Species in the *obscura* and *virilis* groups generally have geographic ranges in temperate climates, thus providing opportunities to examine genes controlling the unique physiological demands of these environments. Alternative genome arrangements are represented by this spectrum of diversity, and the impact of this variation on domains of gene expression will be amenable to analysis.

Research Communities

Each of the selected species is currently the subject of active efforts and also the selected species groups represent the primary foci of active research using Drosophila. This assessment is supported by the published literature, sequence submissions, and requests for stocks from the Tucson Stock Center. In addition, the Tucson Stock Center has held the Drosophila Species Workshop in each of the past two years and these groups have served as the focal point of the workshop. Enrollment in the workshop has reached capacity within the few days following its announcement, thus this interest is a measure of the enthusiasm that exists in utilizing the existing diversity within the genus.

Status of Sequencing Requests

The sequence of the *D. melanogaster* genome is in its final stages of finishing; however, it remains to be seen what level of annotation for this sequence can be realized. Sequences of genomes from other Drosophila species will facilitate this annotation. The first assembly of the genome of *D. pseudoobscura* is now available from the Human Genome Sequencing Center at Baylor. Sequencing the genomes of *D. simulans* and *D. yakuba* has been listed as "High Priority" by NHGRI. A White Paper requesting genome sequences of *D. sechellia, D. erecta, D. annanassae, D. persimilis, D. willistoni, D. littoralis, D. mojavensis* and *D. grimshawi* was submitted to NHGRI concurrently with this request (Clark et al. 2003).

Strain Selection

To ensure integrity of the strains used for DNA isolation, a "quality control" checklist will be applied to each. Strain identification of appropriate species characteristics will be ascertained through

morphological and sequence analyses. Each strain will have been inbred at least twelve generations. This is most important for the libraries slated for sequencing, because inbreeding will minimize segregating variation that could confound assembly. Chromosomal analysis of polytene chromosomes will be used to ensure chromosomes are homosequential. At least one line from all of these species is currently available from the Tucson Stock Center. In some cases, however, a line may be acquired by the Tucson Stock Center from an independent investigator's lab, because the line is already inbred. In these cases, the strain will be deposited in the stock center, accessed into the permanent collection and identified as the source of the DNA.

Genome Sizes

Total amount of nuclear DNA is variable within the genus Drosophila. Genome content for *D. melanogaster* is estimated at 0.35 pg, whereas genome contents of *D. virilis* and *D. hydei* are estimated at greater than twice this size. The full list of estimated sizes is provided in Table 1. It should be noted, however, that the 116.9 Mb euchromatic genome of *D. melanogaster* is only about half of total genome content. Although *D. virilis* and *D. littoralis* are estimated at having about twice the DNA content of *D. melanogaster*, initial large-scale sequencing suggests the euchromatic genomes of these species may be only 20% larger than *D. melanogaster* (Bergman et al., 2003), indicating that the clonable and sequenceable portion of these genomes is less than 150 Mb.

DNA Sources

Large-scale collection of embryos is possible for most of the species. Established methods for preparation DNA from agarose-embeded embryos can be applied in these cases. For some species, especially the Hawaiian Drosophila, embryo collection may not be a feasible means for obtaining DNA. However, the large size of these flies should facilitate the development of alternative strategies to obtain high quality chromosomal DNA.

Library Specifications

Each library should consist of a total of 18,432 clones arrayed in 384-well microtiter dishes (48 total). This number of clones will facilitate the arraying of clone DNA on 22 x 22 cm filters, and will provide sufficient coverage for each genome. At least two methods of DNA shearing should be used in the construction of each library. Due to small relative genome size of Drosophila species, it may be feasible to construct libraries containing a mix (maybe 3:2) of BAC clones with 150-kb inserts and fosmid clones with 50-kb inserts. For a large genome, such as *D. virilis*, this strategy will produce a genomic library containing ~6x redundancy of the entire nuclear genome and ~12x redundancy of the euchromatic genome. Even greater redundancy would be achieved for species with smaller genome sizes. This level of redundancy would be sufficient for genome sequencing, chromosome walking and contig construction.

Time Frame

There is a great deal of interest in this project, and it can be initiated as soon as feasible. For some of the species, inbred lines are already available. Additional generations of inbreeding and quality control checks will be necessary for some of the species.

Library Development, Characterization, and Dissemination

The Tucson *Drosophila* Species Stock Center at the University of Arizona is prepared to coordinate the BAC library production, characterization and curation as part of its service to the community. One of the NHGRI BAC Library Production Centers is the Arizona Genomics Institute, directed by Dr. Rodney Wing, located at the University of Arizona, in the same building as the Stock

Center. The Tucson Stock Center would seek long-term funding to maintain and distribute the libraries on a cost –recovery basis.

References

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185-2195.

Begun DJ, Langley CH, 2003. Proposal for the sequencing of *Drosophila yakuba* and *D. simulans*. White Paper to NHGRI.

Bergman CM, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. Genome Biology 3:research0086.1-0086.2

Boffelli D, et al. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391-1394.

Bone, JF, Kuroda MI. 1996. Dosage compensation regulatory proteins and the evolution of sex chromosomes in Drosophila. Geneticcs 144:705-713.

Celniker SE, et al. 2002. Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biology 3:research0079.1-0079.14.

Clark A, Gibson G, Kaufman T, McAllister B, Myers G, O'Grady P. 2003. Proposal for Drosophila as a model system for comparative genomics. White Paper to NHGRI.

Goodman M. 1999. The gnomic record of humankind's evolutionary roots. Am J Hum Genet 64:31-39.

Halfon MS, Michelson AM. 2002. Exploring genetic regulatory networks in metazoan development: methods and models. Physiol Genomics 10:131-143.

Horn C, Wimmer EA. 2000. A versatile set for animal transgenesis. Dev Genes Evol 210:630-637.

Ohler U, Liao G-c, Niemann H, Rubin GM. 2002. Computational analysis of core promoters in the Drosophila genome. Genome Biology 3:research0087.1-0087.12.

Powell JR. 1997. *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press.

Russo CAM, Takezaki N, Nei M. 1995 Molecular phylogeny and divergence time of Drosophilid species. Mol Biol Evol 12:391-404.

Wittkopp PJ, Williams BL, Selegue JE, Carroll SB. 2003. Drosophila pigmentation evolution: divergent genotypes underlying convergent phenotypes. Proc Natl Acad Sci 100:1808-1813.