Proposal for Drosophila as a Model System for Comparative Genomics

Andrew Clark, ac347@cornell.edu
Greg Gibson, ggibson@unity.ncsu.edu
Thom Kaufman, kaufman@bio.indiana.edu
Bryant McAllister, bryant-mcallister@uiowa.edu
Eugene Myers, gene@eecs.berkeley.edu
Patrick O'Grady, ogrady@amnh.org

June 9, 2003

Introduction

The challenge of obtaining a complete annotation of functional genes and regulatory elements in the genomes of higher organisms remains a rate-limiting step to biological discovery. Recently, impressive progress has been reported based on comparative analysis of genome sequences of related species (Boffelli et al. 2003, Kellis et al. 2003). These studies underscore the pressing need to establish a model system for developing and applying sophisticated comparative algorithms directed at problems of whole genome annotation. Many attributes of genomes are shared across species, so there is great potential in applying interspecific contrasts to identify genomic features with varying degrees of conservation. For example, the use of comparative information will accelerate the verification of gene coding regions and the discovery of regulatory regions. To date, this has mostly been accomplished by relatively simplistic application of large-scale alignments using MultiPIP or VISTA. It seems time to develop a model system for designing next-generation computational tools that will result in a far more powerful application of comparative approaches. Indeed, NHGRI's mammalian ENCODE project requires the development of such tools. This development is best carried out in systems with smaller genomes that nonetheless share many features with mammalian systems. The central argument of this white paper is that Drosophila species provide an ideal target for inculcating and accelerating such advances in comparative genomics leading not only to a complete annotation of the *D. melanogaster* but also the human genome.

There is little question that comparisons among different species provide insight into those regions whose sequence conservation reflects evolutionarily conserved function. Traditional wisdom has been that divergent pairwise comparisons (*e.g., D. melanogaster vs. D. pseudoobscura*) were appropriate for this task (Bergman et al. 2002). However, recent analysis (Boffelli et al. 2003) suggests that multiple comparisons among species at various levels of divergence or "phylogenetic shadowing" may provide better resolution by (a) maximizing evolutionary change across all regions of the genome, and (b) minimizing alignment problems resulting from saturation in more rapidly evolving regions. Consider, for example, an alignment of the human and mouse genomes. Those regions that are unchanged are likely to be under significant functional constraint. Sequences that are difficult to align due to saturation (upstream regions, for example) are still critical to the proper function of the genome and may be under constraint in a much shallower time scale. The idea of phylogenetic shadowing is to obtain genomic sequence from multiple species at varying levels of divergence. By combining data from the various species comparisons, one should have enough mutational hits to identify what is and is not significantly conserved. In this case the identification and alignment of both slowly and rapidly evolving orthologs is straightforward because divergence at multiple time points has been sampled.

We advocate a modification of phylogenetic shadowing to annotate genome data. The "ladder and constellation approach," where the ladder rungs are the various points of divergence and the constellations are clusters of species attaching to these points, improves upon phylogenetic shadowing by offering multiple points of comparison (constellations) from the same divergence point (ladder). More sophisticated models can be applied that contrast rates of substitution at synonymous vs. nonsynonymous sites, or that contrast upstream substitutions that implicate changes in transcription factor binding, DNA conformation, or other potentially important regulatory features. The approach advocated here ameliorates the problem of "saturation" at synonymous sites by sampling from multiple species, including a breadth of phylogenetic diversity. Thus, the power to obtain unambiguous counts of substitutions keeps the analysis clean, while statistical power is gained by adding additional species (Anisimova et al. 2001, 2002).

These results make a compelling case for the complete sequencing of constellations of closely related species selected on the basis of phylogenetic information, suitability of species, and potential use to the broader community. We propose that by considering full genomic sequences from multiple species, it will be possible to develop annotation methods that make full use of the latent information in patterns of interspecific divergence. The proposed goals are to:

1. Use *D. melanogaster* and select taxa in the genus Drosophila as a model system to learn rules for optimizing the selection of species and to guide the development of computational tools for full annotation of the complex genomes.

2. Combine evolutionary information with genetic analysis in a representative series of Drosophila species to make biological sense of whole genome sequence data.
3. Characterize the rates and patterns (and by inference, the selective forces) underlying evolutionary divergence, which may in turn lead to speciation, across the genomes of multiple species.
4. Explore the relationships between genotype, developmental pattern and process, and the resultant phenotype of higher eukaryotes.
5. Study the emergence and loss of new pathways, the gain and loss of complexity of pathways, and the changes in the regulatory network of complex genomes.

Opportunities in Bioinformatics Realized by Comparative Genomics

A primary thrust of this project would be to establish a close collaboration between biologists and bioinformaticists to develop the next generation of computational tools for automated annotation and analysis of multiple genomes in an evolutionary context. The special issues that arise in bioinformatics in the context of comparative genomics include:

- Assembly – Are there unique features of collections of sequence data among related organisms that allow more efficient and accurate genome assembly?
- Gene finding – Once the alignment is in hand for several species, what are the best algorithms for gene finding and for annotating the structural features of genes?
- Regulatory sequences – How accurately can regulatory regions be annotated based on the alignments of gene regions from related taxa?
- Transposable elements– What can a full annotation of appearance, disappearance, expansion and contraction of transposable element families reveal about the evolution of mobile elements and consequences for genome evolution?
- Small RNAs – What is the role of micro RNAs in gene expression? Can these be more readily detected via comparative genomic analysis?
- Adaptive evolution – Is it feasible to make inferences about adaptive evolution across the entire genome?
- Functional validation – How might the potential value of an evolutionary change in gene or genome structure on biological function be rigorously assessed?

In summary, the completion of genome sequences of multiple Drosophila species is motivated by: 1) the need to develop tools for annotation by incisive comparative analysis; 2) the desire to fully annotate and validate the model organism, *Drosophila melanogaster*; and 3) to stimulate development of efficient bioinformatics approaches that can be applied to elucidate genome function and evolution beyond model systems to all higher organisms including humans.

Choice of additional Drosophila species for genome sequencing

The small genome size of most Drosophila species, the excellent phylogenetic work that has been done on this genus, and the knowledge base of the biology of drosophilids in general allow us to select a set of species for whole genome sequencing that are optimal with respect to the general problems in annotation cited above. Importantly, it will be possible to perform this multiple species analysis at a fraction of the cost of sequencing a single mammalian genome. Another benefit of using Drosophila as a comparative genomic model is that an active community, working via the Flybase interface, has already greatly assisted the annotation of *D. melanogaster* and *D. pseudoobscura* genomes. Arrangements for similar annotation capability will be made available to streamline the annotation of the additional genomes proposed here (see letter of support from W. Gelbart).

Our recommendations of candidate taxa for whole genome sequencing can be divided into three broad classes based on their relationship to *D. melanogaster* (Figure 1), the main goals and expected benefits of the proposed sequencing, and the level at which sequencing will be undertaken. The first class includes taxa that are closely related to *D. melanogaster* and the *melanogaster* subgroup. These species (*D. simulans*, *D. yakuba, D. erecta,* and *D. ananassae*) represent increasingly divergent lineages that have been independent of *D. melanogaster* for ~5-15 million years. These taxa should be sequenced at the 8x level.

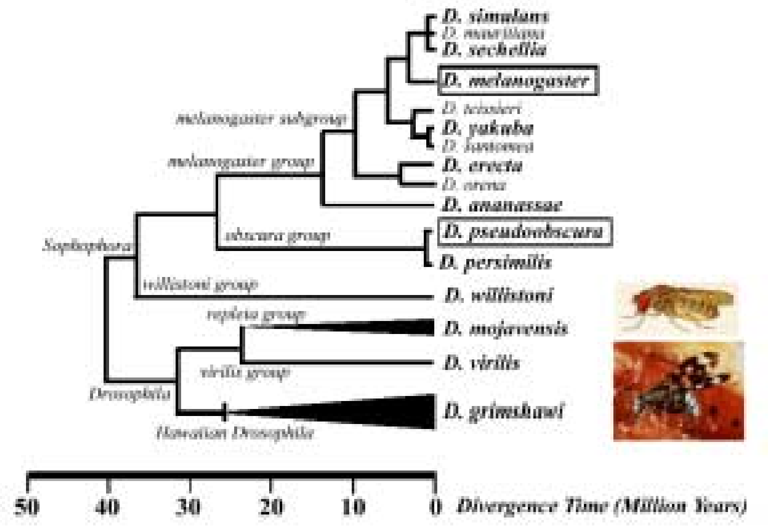

Figure 1. Phylogenetic relationships and estimated divergence times of major lineages in the genus Drosophila. Taxa in bold are proposed for sequencing at this time; boxed taxa have nearly complete genome sequences. Thickened lines in *D. mojavensis* and *D. grimshawi* indicate clades of greater than 100 or 1,000 species, respectively. While the Hawaiian Drosophila lineage diverged from the *virilis*-*repleta* clade ~32 million years ago, this lineage began to rapidly diversify 26 million years ago (hatch mark). Divergence times are based on a linearized *Adh* molecular clock (Russo et al., 1995; reviewed in Powell, 1997).

The addition of these four genomes will also further studies of population and developmental genetics within this group through the discovery of *cis*-regulatory regions and developmental motifs in rapidly evolving genes thought to be involved in speciation. The second class of species includes taxa that are more distantly related to the *melanogaster* group, having been evolving independently for ~35-50 million years. This more divergent set of species (*D. willistoni, D. grimshawi, D. virilis*, and *D. mojavensis*), all of which display behaviors and ecological attributes that are significantly different from those of *D. melanogaster* and each other, will complete the series of species used in the comparative annotation of *D. melanogaster*. These species should also be sequenced at the 8x level. Comparisons with these more distantly related species, which display significant rate heterogeneity with respect to *D. melanogaster,* will identify constrained elements evolving at a variety of rates in the genome. Furthermore, this entire series of species will serve to elucidate new genes, gene functions and gene expression patterns. A better appreciation of whole genome evolution over a 60 million year time frame will result from this work, particularly in the identification of localized synteny, which will yield a better understanding of local neighborhoods of gene regulation. New insights will also be obtained in molecular evolutionary research through an enhanced understanding of the evolution of codon bias and transposable elements from closely related species.

The third class, which includes two species that are closely related to *D. melanogaster* and *D. pseudoobscura* (*D. sechellia* and *D. persimilis*, respectively), should be sequenced at the 3x level. The analysis of these sequence data will be used to further genetic research into the mechanisms of speciation in the *D. melanogaster* and *D. pseudoobscura* model systems, in part by establishing efficient means for developing marker loci throughout the genome and identifying candidate genes in regions of interest. All species being considered here have a high degree of biological interest and have been used for many years as model systems in evolutionary biology, population genetics, and physiology. Whole genome sequences will stimulate diverse biological investigations while uniting them through well annotated and curated databases. In addition, practical issues considered for determining the target species for sequencing included, ease of sequencing, the ability to maintain healthy laboratory cultures, and the possibilities for experimental manipulations, including the generation of interspecific hybrids.

Class One – Closely Related Species

We note that a proposal to sequence the genomes of *D. yakuba* and *D. simulans* was the subject of a previous White Paper (Begun and Langley), and it is our understanding that this project has been approved, so that our proposal leverages those data even further. The additional species recommended for sequencing are listed below with rationalizations for their inclusion. The relative phylogenetic positions of those species chosen is shown in Figure 1 and their genome sizes and other relevant data are given in Table 1 below.

Drosophila ananassae

Drosophila ananassae is a member of the *melanogaster* species group, a clade of about 300 species. The *ananassae* subgroup is basal within the melanogaster group (O'Grady and Kidwell 2002) and should serve as a midpoint for comparisons between the completed *D. melanogaster* and *D. pseudoobscura* genomes. There is a large literature available for this species (Nishinokubi et al., 2003; Tobari et al. 1993; Yamada et al., 2002) and an active research community, particularly in Japan. A characteristic of *D. ananassae* unique among Drosophila species is that males undergo a significant level of recombination (Matsuda et al. 1993; Sato et al., 2000). The data available suggest that the genome of *D. ananassae* is similar in content and size to that of *D. melanogaster* and *D. pseudoobscura* (Powell 1997), making whole genome sequencing technically feasible.

Drosophila erecta

Drosophila erecta is basal within the *melanogaster* subgroup and has increasingly been used as an outgroup for comparisons between various species within this subgroup (Arhontaki et al., 2002; Parsch et al., 2001). Bergman et al. (2003) have generated genomic sequences from 8 regions (ca. 500 kb) from *D. erecta* to determine their utility in aiding the annotation of the *D. melanogaster* genome. They found that adding additional genomic sequences can greatly inform annotation and can aid in the discovery of *cis*-regulatory regions and changes in microsynteny. Some data suggest that the genome of *D. erecta*, while about the same size as that of *D. melanogaster* (Table 1), possesses far fewer satellite repeats, potentially making sequencing of the entire genome of this species easier. Highly inbred lines for this species already exist.

Class Two – Divergent Taxa

The genus Drosophila contains over 2000 species, arranged into about 50 species groups. Perhaps the best studied of these are the *virilis*, *repleta*, and Hawaiian *picture wing species* groups, each of which have adapted to a significantly different ecological niche, offering the opportunity to investigate new gene pathways and functions.

Drosophila willistoni

The *willistoni* species group is a Neotropical lineage within the subgenus *Sophophora*, a larger group containing both the *melanogaster* and *obscura* species groups (Figure 1). In addition to being independent of the *D. melanogaster* lineage for ~30 million years, the rate of evolution in the *willistoni* lineage is much accelerated relative the melanogaster lineage, making this species critical for annotation of more conserved genes. Polytene chromosome polymorphism within *D. willistoni* has been examined extensively, and a particularly active research community, in both the United States and Latin America, exists. *Drosophila willistoni* is a significant choice for sequencing in that it has been implicated in the horizontal transfer of the *P* transposable element into the *D. melanogaster* genome and additional components of these genomes may have also transferred in this manner. The genome size of *D. willistoni* is not significantly larger than that of *D. melanogaster* (Table 1), making sequencing time and effort reasonable. Moreover, a strain is now in the fourth generation of single brother-sister mating to develop an inbred, homokaryotypic strain.

Drosophila virilis

The *virilis* species group is considered as retaining many ancestral characteristics of primitive Drosophila, and therefore, is recognized as an excellent outgroup to the Sophophoran lineage. *Drosophila virilis* is one of the most thoroughly studied Drosophila species, in terms of genetic mutants (Alexander 1976), detailed linkage and physical maps (Gubenko and Evgen'ev 1984), and gene sequences (~100 different single-copy gene sequences in GenBank). Although the total genome size of *D. virilis* is about twice as large as *D. melanogaster*, the euchromatin appears to be only about 20% larger (Bergman et al. 2003). Indications are that the *D. virilis* genome may contain more regions of simple repetitive DNA interspersed in euchromatin than encountered with *D. melanogaster*. At 8x coverage, the resulting assembly of the *D. virilis* sequence may be more fragmented; however, nearly 100 markers are mapped to the polytene chromosomes and these will facilitate organization of the resulting sequence scaffolds. An existing set of mapped P1 clones could be used to further assemble the sequences (Lozovskaya et al. 1993; Vieira et al. 1997). The community of Drosophila researchers interested in heterochromatin has expressed strong support (see letter from S. Elgin and 22 others) in obtaining the genome sequence of *D. virilis*, and the findings resulting from their efforts on this species will increase the understanding of genomes with more complex organizations, such as the human genome. Inversions are not present in most lines of *D. virilis*, and many highly inbred strains are available for sequencing.

Drosophila grimshawi

The Hawaiian Drosophila lineage represents one of the most astounding radiations of morphological and biological complexity. This lineage consists of roughly 1,000 species, all of which are descended from a single ancestral migrant that arrived in the Hawaiian Islands ~26 million years ago (DeSalle 1992; Russo et al. 1995). The historical fragmentation of the Hawaiian Islands makes this group an excellent model for species formation and population genetics. Moreover, Hawaiian drosophilids are generally large in size, often several times larger than *D. melanogaster*, and are characterized by elaborate morphological features (Figure 1). Therefore, a complete genome sequence of this lineage will provide a stimulus for studies of the evolution of development in addition to its usefulness in comparative annotation. The *picture wing* species group has served as an evolutionary model system of speciation, sexual selection, and adaptive radiation for over 40 years, and much is known of its behavior, ecology, and chromosome evolution. Carson (1992) used *D. grimshawi* as a standard to infer phylogenetic relationships of nearly all described picture wing species. *Drosophila grimshawi* is therefore an excellent candidate for obtaining a genome sequence for a representative of the Hawaiian Drosophila. An isochromosomal line, which has been maintained in culture for nearly 40 years, could be used to sequence the *D. grimshawi* genome (Table 1).

Drosophila mojavensis

A fourth independent lineage representing distant comparisons with *D. melanogaster* is provided by *D. mojavensis*, a species in the *repleta* group. The *repleta* species group contains over 100 species, all of which are native to the Americas. In addition to being appropriately positioned for comparative bioinformatics, the repleta group provides unique biological aspects for which a genome sequence will facilitate more detailed studies. Three species in particular, *D. hydei*, *D. buzzatii,* and *D. mojavensis*, have been used as genetic and evolutionary model systems and each have a number of inbred lines and, in some cases, libraries. Of these species, *D. mojavensis* emerges as the most appropriate candidate for sequencing because it is highly specialized on its host plants and is in the process of diversification. As such, it can provide unprecedented insights into the genetics of two primary evolutionary processes, speciation and specialization. *Drosophila mojavensis*, unlike *D. melanogaster* and its relatives, easily hybridizes and introgresses with its close relative, *D. arizonae*, facilitating genetic studies (Ruiz et al, 1990; Pantazidis et al 1993). In addition, it is, itself, in the process of speciating, with different geographic host races exhibiting genetic differentiation and reproductive isolation (Markow et al 2002). This species also specializes on particular host cacti endemic to the inhospitable Sonoran Desert, and offers the opportunity to address issues of adaptation to harsh dietary chemicals (Fogleman and Danielson 2002) and to thermal extremes (Stratman and Markow 1998). At this time a large microsatellite library is available (Ross et al 2003) that is being used to construct a genetic map for *D. mojavensis.* A cDNA library has also been constructed for use in producing a microarray. *Drosophila mojavensis* is sufficiently closely related to other repleta species, such as *D. buzzatii*, to take advantage of existing genetic and linkage data to assist in sequence assembly and interpretation. The unique position of *D. mojavensis* will enable investigations into the speciose repleta group.

Speciation Studies in the *D. obscura* and *D. melanogaster* Groups

Drosophila sechellia

Within the *melanogaster* subgroup, the sibling species triad including *D. simulans, D. sechellia*, and *D. mauritiana* has been extensively studied in an attempt to understand the population genetic and other factors that might lead to speciation. *Drosophila simulans*, a human commensal, has the largest distribution within this group. Initially, this species was found in sub-Saharan Africa, along with some other members of the *melanogaster* subgroup (*D. melanogaster* and *D. yakuba*). *Drosophila sechellia* and *D. mauritiana* are endemic to Seychelles and Mauritius, respectively. Moreover, in addition to its restricted distribution and smaller population sizes, *D. sechellia* is also known to breed in the highly toxic fruit of *Morinda citrifolia* (Rubiaceae). Therefore, *D. sechellia* is not only an excellent model for species formation, but ecological adaptation as well, and provides an essential counterpoint to *D. melanogaster* for interpretation of the effect of recent population expansion on genetic diversity.

Drosophila persimilis

Dobzhansky's work on *D. pseudoobscura* and its sister species, *D. persimilis*, has become one of the paradigms of modern population genetics (Anderson et al. 1991; Jones et al. 1981; Popadic and Anderson 1994). With the near completion of the *D. pseudoobscura* genome, additional studies into the population genetic basis of species formation are now possible. This work will serve as an important comparison to similar studies in the melanogaster subgroup. Like *D. sechellia*, we propose a shallow shotgun sequence (3x) of the *D. persimilis* genome, assembled on the *D. pseudoobscura* backbone, to further enable such work. There are several stocks available from the Tucson Stock Center, some of which contain phenotypic mutants (Table 1). The establishment of a highly inbred line suitable for sequencing would be trivial for this species and its genome size makes it quite feasible for sequencing and assembly.

Table 1. Summary of Criteria Used to Assess the Proposed Species

	Genome	No.		Inbred	
Species	Size[a]	Stocks[b]	Genetics[c]	Line[d]	Crossable[e]
D. ananassae	0.38-0.39	15	Y	Y	Y
D. erecta	0.31-0.32	1	N	Y	N
D. willistoni	0.52	20	Y	Y	Y
D. grimshawi	0.47-0.53	5	Y	Y	Y
D. mojavensis	0.38-0.48	30	Y	Y	Y
D. virilis	0.62-0.68	54	Y	Y	Y
D. persimilis	0.34-0.39	14	Y	Y	Y
D. sechellia	0.34-0.35	14	Y	Y	Y

[a] Genome size is in picograms. [b] Number of wild type and mutant stocks available from Bloomington, Tucson, or other laboratories. [c] Indicates that there is a history of genetic crosses with the species and, in many cases mutant lines are available. [d] There is either an inbred line suitable for sequencing at this time or such a line will be available by August 2003. [e] Will form fertile hybrids with closely related species, allowing for interspecific genetic studies.

Sequencing of an outgroup

Although the strategy of phylogenetic shadowing concentrates on a group as a whole, it should be pointed out that for many evolutionary analyses an outgroup species is essential. Within subsets of the species targeted for this proposal, some will provide outgroups for others, *e.g., D. willistoni* can serve as an outgroup for the *melanogaster* species group and members of the subgenus Drosophila can serve as an outgroup to the *willistoni-melanogaster* comparison. For the study as a whole however there would be great value in sequencing a more remote relative. We consider this task outside the domain of this proposal, but suggest two possibilities for future efforts. The first would be a species from another drosophilid genus of such as *Chymomyza* or *Scaptodrosophila*. Both are well outside genus Drosophila and would be suitable outgroups for any evolutionary comparisons within that group. Alternatively, a more distant Acalyptrate Dipteran would bridge the evolutionary distance between the drosophilids and mosquitoes. One possible choice, in view of its taxonomic position, economic importance and research community, would be the medfly *Ceratitis capitata* (family Tephritidae). Data from any of these groups would help bridge the gap between the extant *Anopheles gambiae* mosquito sequence and the body of information that would be obtained from the group of Drosophila species proposed here.

Stimulus to radical advances in bioinformatics

The use of a collection of related genomes to better annotate each individual genome is a primary goal of comparative genomics, but a second, equally exciting goal is to learn more about how genomes evolve. The proposed project using drosophilids provides unprecedented opportunities to accomplish these goals. In both cases, it is essential to sample genome sequences from species that are at various levels of divergence, such that the maximum number of genomic regions can be annotated. Such an arrangement is necessary for genome-wide comparisons, because different parts of a genome evolve at different rates. In this proposal, this “ladder” is represented by *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. willistoni*, and the more distantly related species in the subgenus Drosophila (*D. grimshawi*, *D. mojavensis*, *D. virilis*). From such an evenly distributed set of species one can understand the rate and nature of evolutionary events. Genome sequences in the “ladder” pattern of increasing evolutionary distance define more and more functionally important sequences as differences accumulate over evolutionary time. For example, a rapidly evolving gene or enhancer is more likely to be detected among the near-species, whereas detection of a housekeeping gene will likely be facilitated by the distant-species.

In addition to having a ladder of species, having a “constellation” of species that are equally diverged provide information in a manner similar to phylogenetic shadowing. Genome sequences in the “constellation” pattern of similar distances in independent evolutionary lineages emphasize features by highighting the intersection of mutually conserved regions that may be functionally important. For evolution, one can better see which elements of genomic organization are essential as opposed to coincidental. In the proposed suite of species, there are two major constellations representing two different levels of divergence: 1) *melanogaster- yakuba-erecta* and 2) *melanogaster-pseudoobscura-willistoni-virilis-*

grimshawi-mojavensis. Finally, two sibling species pairs – *D. sechellia* & *D. simulans*, and *D. persimilis* & *D. pseudoobscura* –provide the opportunity to investigate the mechanisms of speciation, and how this process influences divergence of gene structure and function throughout the genome.

The proposed sequencing program will provide a dataset of unparalleled utility, both in terms of the development of comparative informatics software and biological investigations. Drosophila genomes are compact and feature rich, and, at 1/30th the size of the human and mouse genomes, they are significantly less expensive in terms of computer time for assembly and annotation. This allows investigators to focus on the analysis, rather than merely the logistics of achieving the computation, so that more sensitive and complex computations can be undertaken. The techniques and software that results will be of general utility, and lessons learned and tools derived from this data set can be applied to investigation on plants, mammals, and other comparative efforts. For example, annotation remains the most time consuming aspect of genome biology. The development of broadly applicable tools will be a service to the research community well beyond that provided by a deeper understanding of Drosophila.

By the very nature of large mammalian genomes it will be a number of years before we acquire greater than 15 genomes in the mammalian clade with close to complete coverage. Yet investigation into comparative whole genome techniques at the structural (*e.g.,* assembly), mechanistic (*e.g.,* protein coding genes, structural RNA genes, *cis*-regulation) and adaptive (*e.g.,* positive selection on coding regions) levels should be encouraged. Moreover, without widespread sample datasets, both population theoretical motivated approaches and sequence alignment based bioinformatics methods will lag behind the appearance of mammalian data. Experience shows that it typically takes two to three years for new algorithms and software to appear for any new domain. Thus in a scenario without an early developmental set of metozoan genomes it is possible that the mammalian clade data will not deliver on the clear promise of unraveling many of the aspects of human cellular and organismal biology in a timely fashion.

Annotation methods based on multiple aligned genome sequences can draw on evolutionary information in the form of a broad range of conservation levels to greatly increase accuracy and utility of annotation (Cooper 2003). For a given feature, models can be created for all genomes subject to the constraint that the predictions have the same structure or homology. Consider for example finding protein-coding genes. The most powerful way to find genes comparatively is to train Hidden Markov Model (HMM) predictors (*e.g.*, Genescan) on each genome using available cDNA verified genes as the training set, and then apply each HMM to its genome, subject to the constraint that all predicted proteins whose similarity to each other is consistent with the known evolutionary relationship among the genomes. Note that this approach is symmetric, since a prediction is made for all genomes, not just one target genome. Current methods are not ideal, as they are still based on the idea of masking a reference genome with the segments that are conserved in other genomes. Loss of sequence conservation has also been effectively used, somewhat surprisingly, within closely related species, *i.e.,* phylogenetic shadowing (Boffelli et al. 2003). The proposed data set would allow for the development of methods that simultaneously exploit shadowing and footprinting.

This same idea applies to non-translated genes such as structural RNAs, where the constraint is that all predictions simultaneously have the same or nearly the same secondary structure (Eddy 2002). While just the presence of secondary structure in one genome is insufficient, conservation across several is likely to be informative at certain evolutionary distances. Similarly for core promoters and enhancers, a set of position weight matrices can be created that preserve the number and concentration of sites in the enhancer region, and allow variation in order and orientation across species. It is further conjectured that the very close-in species will be needed for finding these types of elements.

While various nascent software tools have aligned several large genomes, it is hardly the case that this is a solved problem. For example, a recent issue of GENOME BIOLOGY dedicated to the comparison of the human and mouse genomes demonstrates this. All three analytical methods used (Schwartz et al. 2003, Bray et al. 2003, Couronne et al. 2003) employed different heuristics and resulted in somewhat discordant results. Development of better methods is needed and having 6-12 complete and related genomes in a spectrum of divergence will not only improve the pair wise case but may also lead to methods that explicitly take advantage of the known evolutionary relationship among the genomes.

Except for the sibling species pairs created by sequencing *D. sechellia* and *D. persimilis*, we argue for 8x shotgun sequencing in order to have independently assembled sequences that are nearly complete. Whole shotgun assemblers are now available that would reconstruct the euchromatin with perhaps 1,000 or so internal, small gaps and further provide a good picture of the non-repetitive portions of the heterochromatin. With less than 5x shotgun sequencing, one is forced into mapping reads to a known genome in order to assemble the data. In such an approach most of the information about translocations, duplications, and deletions is lost. While individuals who wish to view the genome as an unordered sack of proteins may be content with such a result, anyone interested in genomic evolution, the diaspora of retrotransposons, or *cis*-regulation, to name but a few topics, will find the results inadequate. With 5-6x one can assemble independently, but one will invariably be missing a portion of a gene such as a bit of an exon, core promoter, or enhancer region. If we truly want to provide a dataset that is high value resource to both experimental and computational biologists, it seems wise to let the sequencing pipelines run to 8x.

Biological Rationale for the Project

Functional annotation is at least as important as structural annotation. Functional annotation begins with definition of the molecular, cellular, and organismal roles of each and every gene in a genome, but extends much more broadly to include: understanding of the ecological and phylogenetic context within which genes function; assembling the structure of biological pathways and networks of interaction through which the parts assemble into a whole; and investigating the relationship between genetic and phenotypic variation both within and between species. As the 1/30th scale model for annotation of the human or mouse genome, Drosophila provides exceptional opportunities for biological interpretation of genomic diversity. In this section, we describe how whole genome sequencing of the selected Drosophila species will advance biomedical research through its impact on functional annotation of the human genome, the *Anopheles* genome, and of course the genome of *D. melanogaster*.

The rate-limiting step in functional annotation of genes has always been the availability of mutant stocks. No vertebrate species offers a capability approaching that of *D. melanogaster* for rapid and efficient validation of functional predictions in relation to individual genes, or for forward genetic generation of phenotypes for unannotated genes. Deficiency stocks will soon cover the entire genome, and P-element mutations are expected to soon exist for more than 80% of all fly genes (Spradling et al. 1999; Peter et al. 2002). Numerous mutant stocks also exist for several other fly species, and modern transposon-mediated transformation methods as well as RNAi techniques enable genetic analysis for each of the other Drosophila species (Atkinson and James 2002; Giordano et al. 2002; Brown et al. 1999). As genome comparisons identify genes that are either absent or significantly altered in particular lineages, the molecular genetic tools of Drosophila will facilitate direct assessment of the particular genes with both gain- and loss-of function assays. Microarray tools and proteomic methods are in place that will allow comparative analysis of gene and protein interactions in related species, and the wide range of cell and developmental biological tools for hypothesis-driven analysis of gene function in Drosophila need not be described in detail here. The majority of human disease-associated genes have orthologs in the Drosophila genome (Rubin et al. 2000), and the ability to compare function and interactions in a variety of species contexts will undoubtedly further enhance the attractiveness of Drosophila for annotating human, not just fly, genes.

Comparative sequence analysis has been recognized as a core element of the human genome project since its inception, and is now embraced by the commitment to sequencing of almost a dozen chordate genomes in the next few years. These species will cover the range of phylogenetic organization from primate through mammal, and tetrapod to primitive vertebrates. Useful as these sequence data will be, there is no question that comprehensive functional dissection will be required to make sense of it. Furthermore, there are more complex questions that cannot be addressed purely with bioinformatic tools, including: 1) how can function be ascribed to the one third of the predicted proteome that simply returns “function unknown” query results; 2) what methods will enable characterization of functional divergence of the conserved fraction of genes; 3) to what extent are developmental and physiological pathways conserved across species, and what empirical approaches can be used to detect such conservation; 4) what are the effects of structural factors such as chromosomal location, heterochromatin, transposable element movement, and nucleotide content bias on gene function and evolution; and 5) how has recent speciation followed by rapid population expansion and migration shaped the genetic risk factors that contribute to complex diseases such as diabetes, heart disease, psychological disorders and neurological decline with age. As described above, the genus Drosophila provides an obvious model for addressing each of these questions, and the various species have been selected with these questions in mind.

A considerable investment has been made in sequencing of the genomes of malaria-transmitting mosquitoes. Like other insects that directly impact public health as carriers of disease, the best hope for functionally annotating and potentially manipulating the genomes of these species lies in detailed understanding of Drosophila genetics (Zdobnov et al. 2002). Whole genome sequencing will facilitate this process, for example by allowing mutational analysis of homologs and transgenic experimentation. It is also sure to promote renewed interest in Drosophila as a model for host-parasite interactions. The interaction between the bacterium Wolbachia and *D. melanogaster* is well known (*e.g.,* Rand et al. 2001), but various Drosophila species including those included in this proposal are also host to nematodes and/or form symbiotic relationships with a range of microbes.

The biological basis of species differences

In addition to providing detailed annotation for the *D. melanogaster* genome, this project would identify genes and genomic features that are unusually highly diverged, potentially because of natural selection for diverse function. Critical to this analysis are sequences with sufficient divergence to detect a signal, yet not so divergent that multiple mutations have obscured the actual signal. Both likelihood and Bayesian methods have been devised to estimate parameters of explicit evolutionary models for both coding and noncoding regions in a series of increasingly divergent species. These methods are able to identify the genes that exhibit strongly selected patterns, implying that these genes exhibit evidence for directional selection and hence functional differentiation between species (*e.g.,* Yang and Nielsen 2000, 2002, Yang et al. 2000, see Swanson et al. 2001 for a specific example). The *D. melanogaster - D. pseudoobscura* species pair is already of an evolutionary distance where synonymous sites in coding regions are saturated for mutations (*e.g.,* Table 2 of Bergman et al. 2003). However, with a series of additional intermediate sequences, the stage is set for the genome-wide application of likelihood analysis of genome features that carry adaptive attributes. In short, this will be a remarkable resource for understanding the nature of adaptive differences between closely related species.

Impact on the Drosophila research community

Worldwide, there are well over one thousand active Drosophila research groups that are in a position to capitalize immediately on the data obtained from the proposed project. This results in a total volume of research annually that is similar to the mouse community, and dwarfs the magnitude of research conducted on any other model organism. Many of the groups already use comparative data to guide their annotation of gene structure, and it is noteworthy that the Tucson Species Stock Center has seen a four-fold increase in orders for *D. pseudoobscura* since the draft genome of this species was released two months ago. While the bulk of fly research is focused on *D. melanogaster*, > 200 labs actively employ other species in their research on the evolution of development, the biology of diverse species, and population/quantitative genetics. Indeed, evolutionary and developmental biologists are making a substantial contribution to genome research using Drosophila (Reinke and White 2002). Moreover, a computational research community is already in place to ensure the rapid development of new annotation approaches and interpretative frameworks (via Flybase, etc.). As the human genome project increasingly focuses on the genetics of complex diseases, there is every expectation that Drosophila will continue to drive theoretical developments and understanding of the forces that shape genetic and phenotypic variation. The increasing number of medical school-based labs that are working on Drosophila species further attests to the medical relevance of this group of organisms. Survey of symposium and workshop titles over the past five years also testifies to dramatic expansion of research interest from a focus on development to all aspects of cell biology, neurobiology and behavior, physiology and biochemistry, and chromatin structure, such that Drosophila is now truly a broad biomedical research organism. Further interest in the genetics of adaptation and divergence relating to all of these aspects of organismal biology will be guaranteed by the generation of sequence data from species with a range of ecological and phylogenetic divergence.

Scheduling for the proposed work

Because this one proposal entails multiple species sequences, it is useful to consider the order in which the proposed work is done. This ordering is based on the expectation of the immediate gain to be accrued from each successive genome sequence. The next Drosophila genome to be done should be *D. willistoni*, because this sequence would produce the largest initial impact in resolving ambiguities in *D. melanogaster* annotations even with the current software tools. The second set of genomes to do would be *D. erecta* and *D. ananassae*, as these would provide a series needed to establish a powerful, genome-wide phylogenetic shadowing approach, and would stimulate rapid development of bioinformatics tools for comparative genomic annotation. Third would be *D. virilis*, to reach further out from all other Drosophila, serving to annotate the most slowly evolving developmental features. Fourth would be to perform the lighter shotgun coverage of *sechellia* and *persimilis*, because by this time the informatics tools for annotation would be ready to capitalize on the two additional closely related species, and the evolutionary genetic community would be primed to use genome-wide comparisons of close sibling species to investigate the genetic basis for differences between them. Finally would come *D. grimshawi* and *D. mojavensis*, not because they are the least important, but rather because by this time these research communities would be fully prepared to take optimal advantage of this new resource. Note that the small size of the Drosophila genomes means that the entire span of time before acquiring all eight genomes should be between 12 and 20 months.

Conclusion

The core thesis of this proposal is that the genus Drosophila provides a superb model system for comparative genomics. Dollar for dollar, sequencing of ten Drosophila species will provide more information than sequencing of a single large vertebrate, because it will facilitate the development of computational tools for using comparative genomics in both structural and functional annotation. Furthermore, the Drosophila research community is large and the tools of Drosophila genetics are well tuned, so that any inferences of function derived from the novel bioinformatics tools described here will rapidly face the harsh test of direct biological validation. The species to be studied parallel the diversity seen at three levels of importance to understanding the genetic basis of human variation, namely primate, mammal, and vertebrate. The sequence resources will greatly enhance fundamental experimental research into the mechanisms of speciation, the genetics of adaptation, genotype-phenotype mapping, chromatin organization, and of course developmental and biochemical mechanisms. Thus, we can envisage these sequences being the substrate that leads to the emergence of a myriad of computational and empirical strategies for making sense of sequence diversity. In addition to theoretical and bioinformatics advances, the interplay between the comparative genomic analysis of these sequences and evolutionary inferences to be drawn from them are likely to generate a wealth of proposed hypotheses in the fields of embryogenesis, development, neuronal molecular biology, innate immunity and life span. These will face immediate test through the excellent molecular biology tools available in *D. melanogaster*. This approach has particular relevance to human genetic disorders, and there is a growing effort to use Drosophila as a direct model for human genetic diseases (Rubin et al. 2000). This project is of importance to the public health mission of the NIH both through the development of a powerful infrastructure of methods for comparative genomics, and also by further empowering the Drosophila model system to make primary discoveries about gene function and gene regulation relevant to human disease.

References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185-2195.

Alexander ML. 1976. The genetics of *Drosophila virilis*. Pp, 1365-1427, In: *The Genetics and Biology of Drosophila*. Vol. 1c, Ashburner and Novitski (eds.), Academic Press.

Anderson WW, Arnold J, Baldwin DG, Beckenbach AT, Brown CJ, Bryant SH, Coyne JA, et al., 1991. Four decades of inversion polymorphism in *Drosophila pseudoobscura*. Proc. Natl. Acad. Sci. 88: 10367-10371.

Anisimova M, Bielawski JP, Yang Z 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585-1592

Anisimova M, Bielawski JP, Yang Z. 2002. Accuracy and power of Bayes prediction of amino acid sites under positive selection. Mol Biol Evol. 19:950-958.

Arhontaki K, Eliopoulos E, Goulielmos G, Kastanis P, Tsakas S, Loukas M, Ayala F. 2002. Functional constraints of the Cu,Zn superoxide dismutase in species of the *Drosophila melanogaster* subgroup and phylogenetic analysis. J Mol Evol. 55:745-756.

Atkinson PW, James AA 2002. Germline transformants spreading out to many insect species. Adv. Genet. 47: 49-86.

Barker, JSF, and Starmer, WT. 1982. *Ecological Genetics and Evolution: The Cactus-Yeast-Drosophila Model System.* Academic Press, NY.

Barker, JSF, Starmer, WT, and MacIntyre, RJ. 1990. *Ecological and Evolutionary Genetics of Drosophila*. Plenum Press, NY.

Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, Stapleton M, Wan K, George RA, De Jong PJ, Botas J, Rubin GM, Celniker SE. 2003. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biol. 3:Research0086-6.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299:1391-1394.

Bray N, Dubchak I, Pachter L 2003. AVID: A global alignment program. Genome Res 13:97-102.

Brown SJ, Mahaffey J, Lorenzen M, Denell RE, Mahaffey JW 1999. Using RNAi to investigate orthologous homeotic gene function during development of distantly related insects. Evol Dev 1: 11-15.

Carson, H. L. 1992. Inversions in Hawaiian Drosophila. Pp. 407-440 In: Krimbas, CB and Powell, JR (eds) *Drosophila Inversion Polymorphism*. CRC Press, Boca Raton, FL.

Cooper GM, Brudno M, Green ED, Batzoglou S., Sidow A. 2003. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Res. 13: 818-820.

Couronne O, Poliakov A, Bray N, Ishkhanov T, Ryaboy D, Rubin E, Pachter L, Dubchak I. 2003. Strategies and tools for whole-genome alignments. Genome Res 13:73-80.

DeSalle, R. 1992. The origin and possible time of divergence of the Hawaiian Drosophilidae: evidence from DNA sequences. Mol. Biol. Evol. 9: 905-916.

Eddy SR. 2002. Computational genomics of noncoding RNA genes. Cell 109: 137-140.

Fogleman, JC. 1982. The role of volitiles in the ecology of cactophilic Drosophila. Pp191-208. In: Barker, JSF, and Starmer, WT. *Ecological Genetics and Evolution: The Cactus-Yeast-Drosophila Model System.* Academic Press, NY.

Fogelman, J.C. and Danielson, P.B. 2002. Chemical interactions in the cactus-microorganism-*Drosophila* model system of the Sonoran Desert. Am. Zool. 41:877-889.

Giordano E, Rendina R, Peluso I, Furia M 2002 RNAi triggered by symmetrically transcribed transgenes in *Drosophila melanogaster*. Genetics 160: 637-648.

Gubenko, I.S., Evgen'ev, M.B. 1984 Cytological and linkage maps of *Drosophila virilis* chromosomes. Genetica 65:127-139.

Jones JS, Bryant SH, Lewontin RC, Moore JA, Prout T 1981. Gene flow and the geographical distribution of a molecular polymorphism in *Drosophila pseudoobscura*. Genetics 98:157-178.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241-254.

Markow, T.A., Castrezana, S., Pfeiler, E. 2002. Flies across the water: Genetic differentiation and reproductive isolation in allopatric desert *Drosophila*. Evolution 56:546-552.

Matsuda, M., Sato, H., Tobari, Y.N. 1993. Crossing over in males. In "*Drosophila ananassae. Genetical and biological aspects*." Ed. Tobari,Y.N. Japan Scientific Societies Press.

Nishinokubi I, Shimoda M, Kako K, Sakai T, Fukamizu A, Ishida N. 2003. Highly conserved *Drosophila ananassae timeless* gene functions as a clock component in *Drosophila melanogaster*. Gene 307:183-190.

O'Grady, PM, Kidwell, MG 2002. Phylogeny of the subgenus Sophophora (Diptera:Drosophilidae) based on combined analysis of nuclear and mitochondrial sequences. Mol. Phylogenet. Evol. 22: 442-453.

Pantazidis, A.C., Galanopoulos, V.K., Zouros, E. 1993. An autosomal factor from *Drosophila arizonae* restores normal spermatogenesis in *Drosophila mojavensis* males carrying the *Drosophila arizonae* Y chromosome. Genetics 134: 309-318.

Parsch J, Meiklejohn CD, Hauschteck-Jungen E, Hunziker P, Hartl DL. 2001. Molecular evolution of the *ocnus* and *janus* genes in the *Drosophila melanogaster* species subgroup. Mol Biol Evol. 18:801-811.

Peter, A., Schottler, P., Werner, M., Beinert, N., Dowe, G., Burkert, P., Mourkioti, F., et al., 2002. Mapping and identification of essential gene functions on the X chromosome of Drosophila. EMBO Rep 3: 34-38.

Popadic, A., Anderson, W.W. 1994. The history of a genetic system. Proc. Natl. Acad. Sci. USA 91: 6819-6823.

Powell, J. R. 1997. Progress and Prospects in Evolutionary Biology: The Drosophila Model. Oxford University Press, New York.

Rand DM, Clark AG, Kann LM. 2001. Sexually antagonistic cytonuclear fitness interactions in *Drosophila melanogaster.* Genetics 159:173-187.

Reinke V and White KP 2002. Developmental genomics approaches in model organisms. Annu. Rev. Genomics Hum. Genet. 3: 153-178.

Ross, C.L., Dyer, K.A., Eerez, T. Miller, S.J. Jaenike, J., and Markow, T.A. 2003. Rapid divergence of microsatellite abundance among species of Drosophila. Mol. Biol. Evol, May 30 Epub, PMID: 12777536

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, et al., 2000 Comparative genomics of the eukaryotes. Science 287: 2205-2215.

Ruiz, A., Wasserman, M.W. and W.B. Heed. 1990. Evolution in the mojavensis cluster of cactophilic *Drosophila* with descriptions of two new species. Jour. Hered. 81:30-42.

Russo, CAM, Takezaki N, and Nei M 1995. Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. 12(3):391-404.

Sato H, Goni B, Matsuda M, Tobari YN. 2000. A site specific increase in recombination in *Drosophila ananassae*. Genes Genet Syst. 75:41-47.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W 2003. Human-mouse alignments with BLASTZ. Genome Res. 13:103-107.

Spradling A, Stern D, Beaton A, Rhem E, Laverty T, Mozden N, Misra S, and Rubin GM 1999. The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. Genetics 153: 135-177.

Stratman, R. and Markow, T.A. 1998. Heat tolerance of Sonoran Desert *Drosophila*. Functional Ecology 8:965-970.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci U S A. 98:7375-7379.

Tobari YN 1993 *Drosophila ananassae*. Genetical and biological aspects. Japan Scientific Societies Press

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32-43.

Yamada H, Matsuda M, Oguma Y. 2002. Genetics of sexual isolation based on courtship song between two sympatric species: *Drosophila ananassae* and *D. pallidosa*. Genetica 116:225-237.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32-43.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908-917.

Yang, Z., Nielsen, R., Goldman, N., Pedersen, A.M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Lai Z, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, Yandell M, Collins FH, Ribeiro J, Gelbart WM, Kafatos FC, Bork P 2002. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science 298:149-159.