Proposal for the Sequencing of *Drosophila yakuba* and *D. simulans*.

David J. Begun and Charles H. Langley, University of California - Davis

Overview

Comparative genome sequencing has the greatest impact on biology when the targeted genomes impinge directly on analysis or interpretation of the human genome or the genome of a genetic model system. Comparative genomics may also shed light on the genetic and evolutionary mechanisms that determine genome organization and composition. The most obvious benefit of comparative genomics has been the discovery of conserved putative functional elements present in each of two distantly related genomes. However, comparisons between distantly related genomes are biased towards identifying only those functional elements that evolve very slowly. Alternatively, comparisons between more recently diverged genomes provide quantitatively critical elements in the analysis of population genomic variation and a clearer view of the mechanisms causing genome evolution. In this white paper we argue that determining the genome sequences of *Drosophila simulans* and *D. yakuba* will greatly facilitate two fundamental goals of genomics research: inferring the mutational and evolutionary mechanisms underlying genome divergence and investigating the causes and consequences of population genomic polymorphism within species.

The scientific value of determining the genome sequence of *two* additional Drosophila species is noted several times in the white paper. However, we will also mention it here because it is such an important point. The outgroup (*yakuba* in this case) allows divergence between sister taxa (*melanogaster* and *simulans*) to be "polarized," i.e. assigned to a particular lineage. The outgroup also permits reliable polarization of polymorphisms segregating within *melanogaster* populations. The ability to infer the ancestral state of a nucleotide is vital for sophisticated analyses of genomic divergence and polymorphism that seek to reveal evolutionary mechanism.

In summary, the genome sequences of *Drosophila yakuba* and *D. simulans* will lead to important advances in our understanding of the mechanisms responsible for the organization and composition of the *melanogaster* genome and in our understanding of the evolutionary processes controlling divergence of the *melanogaster* genome from that of other Drosophila. Furthermore, *simulans* and *yakuba* genome sequences would open up new areas of research in the genetic and developmental basis of species differences. The *simulans* and *yakuba* genomes would lay the groundwork for whole genome approaches to the study of molecular and phenotypic population variation within the *melanogaster* model system. Advances in the analysis and interpretation of *melanogaster* population variation will have direct impacts on the study of human variation. Finally, the *yakuba* genome will add an important dimension to annotation of the functionally important yet rapidly evolving component of the *melanogaster* genome.

Justification

Genomic analysis of closely related species is the key to understanding the mechanistic basis of genome evolution because such genomes permit reliable inferences of the histories of individual mutations that have fixed in the recent past. Insights into mechanisms of genome divergence are also enhanced by phylogenetically informed analyses of at least three species. Such analyses allow one to tease apart evolutionary changes on each of the two ingroup lineages, which is particularly important given that evolutionary processes often differ even between closely related species. These considerations, and others provided later in the white paper, motivate our proposal to determine the genome sequence of two Drosophila species, *D. simulans* and *D. yakuba*. Drosophila of the *melanogaster* subgroup have played a central role in biology. The virtues of the best known species, *D. melanogaster*, in genetics and developmental biology need not be recited here. *D. melanogaster* and its close relatives have also played a central role in evolutionary biology and population genetics. Many fundamental principles of population genetics have been discovered through the study of *melanogaster* and its relatives. To cite just one example, the relationship between crossing-over and DNA polymorphism and its possible interpretation was first noted in research on *melanogaster* over 10 years ago (Aguade *et al*. 1989; Begun and Aquadro 1992). A similar relationship has recently been recently reported in humans (Nachman 2001, IHGSC, 2001, Venter, *et al*. 2001). Species differences have also been studied to great effect in

this group of flies - more is probably known about the genetic basis of hybrid infertility and inviability in the *melanogaster* subgroup than in any other group of organisms (Hollocher 1998).

D. simulans and *D. melanogaster* shared a common ancestor roughly 2-3 million years ago (Figure 1). *D. simulans* has several properties that make it extraordinarily useful for comparative genomics. Like *melanogaster*, *simulans* evolved in Africa but is now cosmopolitan and found throughout the world in association with humans; it is similar to *melanogaster* in terms of generation time and ease of culture. With only one exception, a single, large inversion on chromosome 3R, there are no cytologically detectable karyotype differences between these species. The euchromatic portion of the *simulans* genome is smaller because of reduced transposable element copy number (Dowsett and Young 1982). Over 100 mutant *simulans* stocks are currently available from the Tucson Stock Center. Inbred lines of *D. simulans* are available and are easily generated. P-element transformation can be used in *D. simulans* (Scavarda and Hartl 1984). The *simulans* and *melanogaster* sequences are generally easily aligned, which will make assembly of the *simulans* genome relatively simple. *D. melanogaster* and *D. simulans* can produce both F1 and backcross hybrids (Sturtevant 1920; Davis, *et al*.1996, Barbash, *et al*. 2000), opening up the possibility of deploying genetic tools from *melanogaster* in the investigation of species differences and incompatibilities.

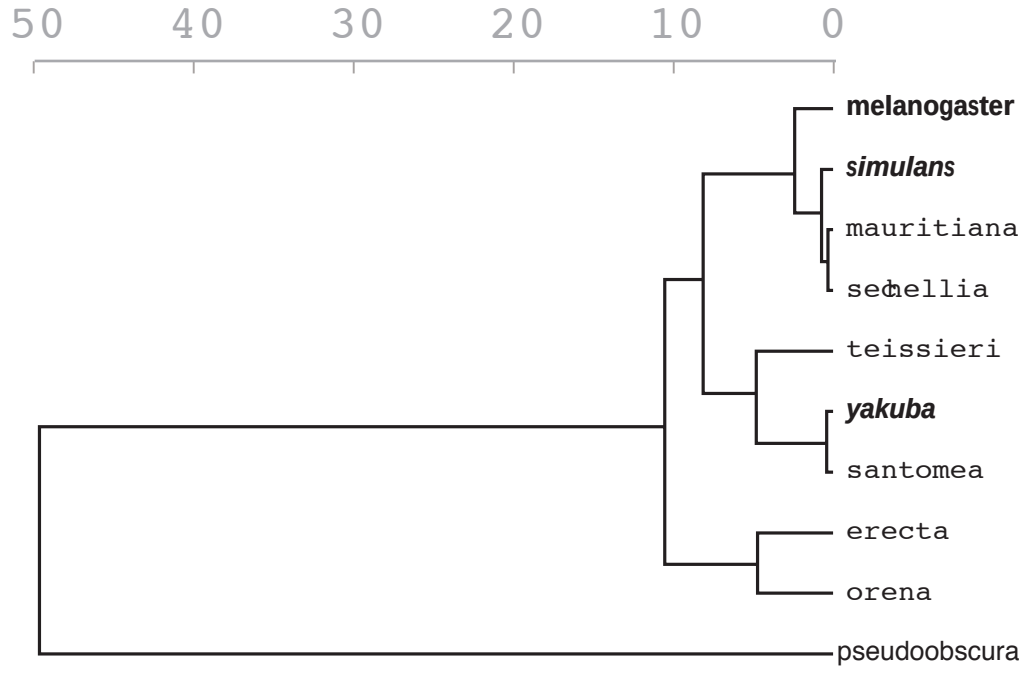

Figure 1. Phylogeny of the *melanogaster* subgroup, redrawn from Powell (1997).

The usefulness of *simulans* in evolutionary genomics is further strengthened by the existence of two closely related species, *D. mauritiana* and *D. sechellia*. These three species are referred to as the *simulans* clade. Members of the *simulans* clade are phenotypically diverged for several traits (see Table 1), yet are partially interfertile with each other. P-element transformation can be used in *D. mauritiana* (True *et al*. 1996). Thus, interspecific genetics is possible, and in fact has been used to investigate both the genetic and population genetic basis of phenotypic evolution and species incompatibilities (e.g., True *et al*. 1996a,b; Laurie *et al*. 1997; Jones 1998; Macdonald and Goldstein 1998). As we note below, a *simulans* genome sequence would provide a major impetus to such research.

Drosophila yakuba is an outgroup species relative to *melanogaster* and *simulans*, having split from these two species roughly 10 million years ago (see Figure 1). It too, evolved in Africa, and is currently found only on that continent. *D. yakuba* is similar to *melanogaster* and *simulans* in its husbandry. Highly inbred, standard karyotype stocks are available and are easily made.

Sequence divergence between *yakuba* and *melanogaster* is heterogeneous (Martin and Meyerowitz 1986). In many genomic regions, a nucleotide in *yakuba* will be identical to the homologous nucleotide in *melanogaster* and/or *simulans* (e.g., Akashi 1996; Begun and Whitley 2000). However, in some genomic regions, alignment of certain sequences (e.g., introns, rapidly evolving exons) in the two species is difficult (e.g., Tsaur and Wu 1997); PCR primers designed from the *melanogaster* reference sequence have a failure rate of about 50% in many *yakuba* genomic regions. The recent discovery of *D. santomea* (Lachaise *et al*. 2000) significantly increases the evolutionary and population genetic interest of *yakuba*. These two species show several morphological differences yet are partially interfertile. Thus, this species pair provides another opportunity to investigate the genetic basis of species differences and incompatibilities in a system in which the infrastructure of the *melanogaster* research community can be deployed to great advantage (Coyne *et al*. 2002, Llopart *et al*. 2002). Thus, while *erecta* or *teissieri* might be comparable to *yakuba* from a sequence divergence perspective (*e.g*., Takano-Shimizu 2001), there are strong conceptual advantages for *yakuba*.

Genome sequences of *melanogaster*, *simulans* and *yakuba* would provide a uniquely valuable resource. First, evolutionary analysis of several protein-coding regions in these species has already revealed their value for generating and testing population genetic hypotheses of genomic divergence (and polymorphism). Extension of these approaches to a genomic scale will definitely provide unprecedented insights into mechanisms of genome divergence.

Second, it is likely that the *yakuba* genome will enhance the *melanogaster* annotation in a manner distinct from, yet complementary to the enhancement expected from the *pseudoobscura* genome sequence. For example, significantly conserved regions in *melanogaster* vs. *pseudoobscura* comparisons may often reflect functional constraints, but they may also result from low mutations rates in some genomic regions or the stochastic nature of the substitution process. However, conserved *melanogaster* vs. *pseudoobscura* sequence elements which are also conserved in *yakuba* (and *simulans*) may be more attractive targets for functional analysis. Furthermore, functionally important sequences can evolve. Thus, sequences conserved in *melanogaster* vs. *pseudoobscura* comparisons will only be a subset of functionally important *melanogaster* sequences. This suggests that the *yakuba* genome sequence will identify several important functional *melanogaster* elements (*e.g*., small, rapidly evolving proteins) that would not be revealed simply by comparison of *melanogaster* and *pseudoobscura* genomes.

Third, the *simulans* and *yakuba* genome sequences would dramatically facilitate investigation of genetic basis of interspecific phenotypic differences. The possibility of cataloguing with high confidence virtually all genomic changes that have fixed in *melanogaster* vs. *simulans* opens up to direct experimentation the evolutionary and functional causes and consequences of such changes, and takes advantage of the power of *melanogaster* genetics, genome resources, and annotation. Moreover, the availability of these two genome sequences would dramatically accelerate genetic analyses of phenotypic divergence of *simulans* vs. *mauritiana*, *simulans* vs. *sechellia*, *mauritiana* vs. *sechellia* and *yakuba* vs. *santomea*. Thus, these two new sequences would thrust the *melanogaster* subgroup to the forefront of evolutionary developmental genetics. Tools that would be developed for integrating phylogenetically informed, multiple species comparisons with a well-annotated model system in Drosophila would be directly applicable to similar research questions in human and primate biology.

Finally, the *yakuba* and *simulans* sequences are vital for inferring the mutational and population genetic mechanisms underlying within-species genome polymorphism in the *melangaster* model and its sister species, *simulans*. Several important population genetic results have emerged from theoretical and empirical investigations of these species on a gene-by-gene basis. Many of these results were later replicated in other species, including humans (e.g., Nachman 2001). These gene-by-gene data show that expansion of these research strategies to genome scales will have major impacts on our understanding of the population genetic mechanisms shaping variation in animal populations. It is also clear that such approaches developed in the *mel*/*sim*/*yak* system will be directly applicable to the study of human genomic polymorphism.

Selected Research Topics

Genome Size

Genome size is labile, sometimes varying by several orders of magnitude even within a particular group of animals (Li 1997). The evolutionary mechanisms underlying genome size evolution remain obscure, though it is clear that gene-number variation plays only a small role. Our best data on the population genetic processes affecting genome size come from Drosophila. Patterns of insertion and deletion (indel) variation in so-called "dead-on-arrival" retrotransposons suggest that deletions are, on average, larger than insertions and occur more often than insertions (Petrov and Hartl 1998). Furthermore, data from dead-on-arrival retrotransposons have been interpreted as supporting the notion that this "deletion bias" in Drosophila reflects fundamental mutational biases of Drosophila DNA replication rather than the influence of natural selection favoring smaller genomes. The retrotransposon data suggest that in the absence of other forces, we would expect all Drosophila genomes to have the same minimal or optimal size. However, the range of genome size variation is considerable within Drosophila (Powell 1997). Thus, other forces must be at play. For example, the addition of *yakuba* sequences to the phylogenetic analysis of divergence of individual genes in *melanogaster* and *simulans* led to the hypothesis that the *melanogaster* genome has actually been increasing in size in the recent past, perhaps as a consequence of weaker natural selection against slightly deleterious insertion mutations (Akashi 1996).

The availability of complete genome sequences of *simulans* and *yakuba* will provide our first complete, unbiased picture of insertion and deletion evolution over an evolutionary time scale which permits inferences about mechanism. Specifically, the patterns of indel variation along the *melanogaster* vs. *simulans* lineage in different genomic regions or for different categories of sequences can be investigated for the first time, and the importance of transposable element vs. other types of sequence variants for genome size variation can be quantified. Such analyses speak much more strongly to mutational and population genetic mechanisms than do similar approaches applied to distantly related genomes. Importantly, the *yakuba* genome allows for determining the direction of indel evolution in the *melanogaster* vs. *simulans* lineages.

Codon Bias

Codon bias, the non-random use of alternative codons, can result from either mutational biases or from natural selection. Codon bias is a ubiquitous feature of genome organization, as it is found in genomes ranging from *E. coli* and yeast to Arabidopsis, Drosophila and humans (Li 1997; Duret and Mouchiroud 1999). Despite its ubiquity, mechanistic explanations of codon bias in multicellular organisms are still poorly developed. The best data on population genetic causes of codon bias come from analysis of individual genes in *melanogaster* vs. *simulans*, with *yakuba* serving as the outgroup (e.g. Akashi 1995,1996; Begun 2001). Several lines of evidence in *melanogaster* suggest that natural selection plays a role in non-random codon usage (Powell 1997). Codons ending in A or T are hypothesized to have lower fitness than codons ending in G or C (Shields 1988). The *melanogaster* genome appears to be accumulating putative lower fitness codons over time (Akashi 1996). *D. simulans* shows a similar pattern, but the accumulation appears to be occurring at a much slower rate (Begun 2001). A whole genome view of the substitution process in two lineages will greatly enrich our ability to investigate this phenomenon in terms of effects of gene expression, gene length, recombination rates, sequence context, and gene function on patterns of base compositional change. To cite just one example, few data on patterns of nucleotide substitution in sequences other than exons or introns are available today. Data from sites that are not clearly associated with genes are vital for testing hypotheses regarding mutational mechanisms, how mutation patterns vary between genomic regions, and the role of mutation biases in evolution of codon bias and base composition. As was true for analyses of genome size, it is only the phylogenetic analysis of substitutions along individual lineages in closely related species that will allow us to tease apart the various mutational and population genetic processes. Thus, the *yakuba* genome is critical.

Transposable Elements

Many issues concerning the biology of transposable elements, such as the forces maintaining genomic copy number and the factors affecting transposition rates or other mutagenic processes, are inherently genomic in scope (Charlesworth and Langley, 1988; Kaminker *et al*. 2002). The *simulans* and *yakuba* reference sequences will yield a complete comparison with *melanogaster* genomic parasites at an ideal resolution for addressing several outstanding issues, including copy number, sequence divergence, insertion target sequences, host accessory genes (e.g., tRNAs), distributions relative to chromatin structure, and (mitotic and meiotic) recombination rates. While genomic distributions of different families, population polymorphism, and sequence divergence of *melanogaster* transposable elements have been studied, much fewer data are available for closely related species. Human euchromatin is filled with ancient insertions of transposable element sequences (e.g., LINES and Alu) which were fixed in the human lineage over tens of millions of years. In contrast, euchromatic insertions of transposable element sequences typically occur at very low allele frequencies in *melanogaster* populations. Our understanding of such a stark difference in the population genomics of transposable elements will be advanced by genomic sequences of these two close *melanogaster* relatives and the research they will support.

Lineage restricted genes

Preliminary genome comparisons suggest that many Drosophila genes are present in nematodes and mammals (e.g., Rubin *et al*. 2000, IHGSC, 2001). However, a large and interesting set of genes have lineage-restricted distributions within animals (Adams *et al*. 2000; Rubin *et al*. 2000, IHGSC, 2001; Mural *et al*. 2002). Though identifying functions for previously unknown genes present in all or most animals is a major component of comparative genomics, understanding the causes and consequences of recruitment and loss of genes is a complementary and important alternative research goal. Such changes are likely to have played an important role in adaptive evolution and to be the basis of the unique properties of particular species or types of animals. This important aspect of genome evolution and organization can be studied to great effectiveness in Drosophila.

Even given our superficial descriptions of genome evolution in *melanogaster* and its relatives, at least two examples of lineage-restricted genes been discovered. The *Sdic* locus codes for a sperm specific axonemal-dynein protein in *melanogaster*. The gene is a chimera which originated through a complex set of rearrangements including a gene-fusion event between the cell-adhesion protein annexin X and a cytoplasmic dynein intermediate chain (Nurminsky *et al*. 1998). Interestingly, the gene is absent in the sister species, *simulans*, suggesting it originated in the very recent past. *Jingwei* is a novel chimeric gene, which originated as a result of the insertion of an *Adh* retrosequence into a duplicated locus in the *melanogaster* subgroup of Drosophila (Long and Langley 1993). It is present in *yakuba* (and *teissieri*) but absent from *melanogaster* and *simulans*. The jingwei protein has no typical Drosophila ADH activity (as one might expect for such a radically altered gene), though its function is not yet known. As was the case for *Sdic*, *jingwei* is expressed in a sex-specific manner.

Given that both of the these examples were discovered serendipitously, systematic searches of closely related genomes will reveal many additional examples and provide us with insights into generalities regarding mechanisms of gains or losses of genes. Investigation of this phenomenon in *melanogaster* and its close relatives would open promising avenues of genetic and developmental analysis, including exploration of novel gene function that take advantage of the tools and resources associated with the *melanogaster* model system. Again, the importance of the outgroup sequence *yakuba* is clear, in that it allows one to determine when differences result from gains vs. losses of genes in *melanogaster* vs. *simulans*. Development of the analytical tools for identifying recently recruited or lost genes in *melanogaster* and its close relatives will be directly applicable to the investigation and annotation of the genomes of humans and other primates.

Principles of function and evolution

The data on novel Drosophila genes are consistent with other types of data suggesting that sexual selection is a major cause of Drosophila evolution. Such data include the rapid evolution of

male genitalia (Liu *et al*. 1996), hybrid male sterility (Palopoli and Wu 1994), sexual behavior (e.g., Wu *et al*. 1995), sperm morphology (Pitnick 1996), reproductive tract-related phenotypes (Patterson 1952; Price 1997, Pitnick 1999; Knowles and Markow 2000), and testis and accessory gland protein sequences (Coulthart and Singh1988; Civetta and Singh 1998; Begun 2000). Such data are tantalizing, though from a genomics point of view, extremely limited in scope. Thus, the notion that the reproduction-related component of the genome is more dynamic than the rest of the genome should be seen as a hypothesis, which can only be properly addressed with genome sequences from *simulans*, *melanogaster* and *yakuba*.

This is just one example of a fundamental biological question about the relationship between protein function and protein evolution. The *mel/sim/yak* sequences would be ideal material for addressing this issue in a complete and rigorous manner. Such analyses would capitalize on several unique features of the *mel/sim/yak* system. First, the high quality *melanogaster* annotation extends easily to *simulans* and *yakuba* and allows for powerful integration of functional and evolutionary description. Second, because the *melanogastger* and *simulans* genomes are closely related, rates of amino acid evolution can be compared with rates of evolution for other types of sites (e.g., intron, intergenic, silent) to provide powerful hypothesis tests without the great complications associated with the uncertainty introduced by extensive divergence of most non-exonic sequences. As the *melanogaster* sequence annotation integrates additional features such as protein domains and structures, analysis of protein evolution in the *mel/sim/yak* system will also become a much richer and deeper arena for analysis of genomic scale protein evolution. Moreover, the *simulans* and *yakuba* genomes will be vital for interpretation of *melanogaster* population protein variation and its causes. Note that although analysis of the *pseudoobscura* genome will permit the partitioning of faster and more slowly evolving proteins, the genome is so distantly related to *melanogaster* that comparison of the two offers little hope of providing a mechanistic understanding of why some proteins or protein domains evolve much more quickly than others. The *yakuba* genome would permit sophisticated analysis of lineage-specific effects on protein evolution and the importance of protein function on such effects.

Finally, though a primary goal of *melanogaster* vs. *pseudoobscura* genome comparison will be identification of conserved sequences that are likely to be functionally important, several aspects of genome function are unlikely to be revealed simply through comparison of two distantly related genomes. For example, comparison of *even-skipped* 5'-flanking regions in *melanogaster* vs. *pseudoobscura* reveals highly conserved function, despite extensive sequence divergence (Ludwig *et al*. 1998, 2000). However, despite functional conservation, several functional *melanogaster* elements are diverged between *melanogaster* and *pseudoobscura*. In fact, two cis-acting *melanogaster* elements, one bicoid-binding site and one hunchback-binding site, are entirely absent from the homologous region of the *pseudoobscura* genome. Such results clearly demonstrate that important sequences can evolve, and that simple pairwise comparisons of distant relatives provides an incomplete picture of candidate functional elements. This pattern of 5'-flanking sequence evolution may be common in Drosophila. Further evidence of the potential importance of a *yakuba* genome can be found in the results of an analysis of eight homologous regions from each of five Drosophila species: *melanogaster*, *erecta*, *pseudoobscura*, *willistoni*, and *littoralis* (Bergman *et al*. 2002). Four of 51 predicted *melanogaster* genes in this study (about 8%) had a homologous copy in *erecta*, but not in *pseudoobscura*, *willistoni* or *littoralis*. The published data are consistent with three different explanations, gene over-prediction, transposition of homologous genes to non-homologous locations, or lineage-restricted genes. However, three of these predicted genes have at least one corresponding *melanogaster* EST, while the fourth contains a large ORF (350 amino acids) and has a *melanogaster* vs. *erecta* Ka/Ks significantly less than one. Given these data and the data cited in the previous section, it appears that a large number of *melanogaster* genes may be restricted to *melanogaster* subgroup species. This suggests that a *yakuba* genome will fill an important gap in the annotation of a significant portion of the *melanogaster* genome. Similarly, rapidly evolving open reading frames (e.g., accessory gland protein genes) are likely to be unrecognizable as homologous in *melanogaster* vs. *pseudoobscura* comparions, but recognizable in *melanogaster* vs. *yakuba* comparisons. Finally, spatial patterns of conserved non-coding sequence clusters in *melanogaster* vs. *pseudoobscura* comparisons are highly correlated with patterns in *melanogaster* vs. *erecta* comparisons for some genomic regions

(Bergman *et al*. 2002 Figure 4). Previously collected data show that *melanogaster* vs. *yakuba* nucleotide divergence is roughly the same as *melanogaster* vs. *erecta* divergence (Schmid and Tautz 1997, Schmid and Aquadro 2001; Takano-Shimizu 2001). This suggests that a *yakuba* genome would play a significant role in providing statistical power for identifying conserved non-coding clusters that are functionally significant in *melanogaster*.

The study of gene expression variation provides another example of the central role *mel/sim/yak* genomes would play. The relative importance of amino acid vs. regulatory evolution has been debated for decades, mostly in the absence of good data. Regardless of the experimental strategy used to collect genome-wide gene expression data, complete *yakuba* and *simulans* genomes are required. For example, interpretation of *simulans* expression data based on *melanogaster* expression arrays would require estimates of sequence divergence for each gene, as well as complicated correction algorithms. More properly, the *simulans* genome sequence will permit the design of *simulans* arrays, or arrays which can be used for both species without introducing bias from sequence divergence. Finally, genome wide expression studies in *yakuba*, which are required for interpretation of *melanogaster/simulans* differences, cannot be done without a complete *yakuba* sequence

Evolutionary developmental biology

Providing a developmental genetic explanation for phenotypic variation within and between species is a major goal of biology. Most experimental approaches in "evodevo" ask questions about major morphological differences between distantly related species. However, the *melanogaster* subgroup system offers excellent material and experimental tools for investigating the developmental and evolutionary basis of morphological evolution.

Phenotypic variation among *melanogaster* subgroup species can be studied at three levels (Table 1). First, the very closely related species of the *simulans* clade differ at numerous characters. Many are sexually dimorphic, while others are ecological or behavioral differences such as those associated with host-plant divergence between *simulans* and *sechellia*. It also appears that there has been rapid evolution of traits relating to interactions between male and female reproductive tracts, though the details of the biology of the differences are still obscure (Price 1997). Species of the *simulans* clade are partially interfertile, allowing for genetic analysis of the species differences. The genome sequence of *simulans* and *yakuba* would open up opportunities to describe all nucleotides fixed between *simulans* and *mauritiana*, and to determine their ancestral states. In combination with genetic experiments and the *melanogaster* annotation, such approaches would result in a list of annotated nucleotide changes in candidate genes for phenotypic differences between species. Such candidates could be tested by various experimental approaches, including fine scale genetic mapping which is straightforward given a *simulans* genome sequence, and by transgenic experiments in *simulans*.

Second, *Drosophila yakuba* is very closely related to the recently discovered species, *Drosophila santomea* (Lachaise *et al*. 2000, Cariou *et al*. 2001, Llopart *et al*. 2002; Coyne *et al*. 2002). These two species also show morphological differences and are partially interfertile (see Table 1). The *yakuba* genome sequence would permit rapid progress in understanding the genetic basis of these species differences.

Finally, *melanogaster* and the *simulans* clade species are diverged for many morphological and behavioral characters. Though they produce fertile hybrids only with great difficulty, the genetic tools associated with *melanogaster* can be used to investigate phenotypic differences. Here again the knowledge and experimental power of the Drosophila model system will add enormous value and opportunity. Together the *simulans* and *yakuba* genomes would open up the possibility of true, comparative genomics of phenotypic evolution in the *melanogaster* subgroup.

Hybrid incompatibilities

Many fundamental and practical questions arise from the functional genomic interactions displayed in incompatibilities of species hybrids. Drosophila hybrids have been a rich source of data and ideas about genomic interactions that produce hybrid sterility or inviability (see Table 1). These interspecific genomic incompatibilities play a major role in evolutionary processes leading to speciation. Haldane's rule (in F1 hybrids the heterogametic sex is more likely to be sterile or

inviable) is still not understood at the mechanistic level. The *simulans* and *yakuba* reference sequences would expedite discovery of the types of mutations, genes and developmental pathways commonly involved in *melanogaster* subgroup species incompatibilities. For example, RNAi promises to provide a rapid avenue to the functional genomic analysis within the *melanogaster* subgroup. Complete genome sequences of *simulans* and *yakuba* would not only expedite such comparative analysis but also allow allele-specific targeting in species hybrids. The *simulans* and *yakuba* reference sequences would also play a major role in expediting fine scale interspecific genetics and interpretation of gene expression experiments in *melanogaster* subgroup hybrids.

Interpretation of *melanogaster* variation

The completion of a high quality, human reference sequence is transforming the study of human genomic variation. Just as genomic research in Drosophila has advanced the molecular and developmental core of functional genomics, the investigation of population genomic variation in Drosophila will serve as a rigorous, fertile foundation for development of tools and talent needed to understand human biology and medicine from the population genomic variation perspective. A grant proposal to initiate re-sequencing of 50 independent *D. melanogaster* genomes is currently under review. These genomic polymorphism data (and genetic stocks from which they would be derived) would be a platform for the advancement of population genetics to population genomics. Furthermore, they would serve as a model for the development of human population genomics. Enormous synergies would occur between this re-sequencing project and the sequencing of the *simulans* and *yakuba* genomes.

The genomic variation in present-day populations (humans or Drosophila) is the end result of complex processes of mutation, recombination, natural selection and demography occurring over several hundred thousand years. Many concepts and tools for analyzing sequence divergence, polymorphism and their association with phenotypes arose within the Drosophila population genetics community (*e.g.*, Hartl and Clark 1997). Virtually all of the fly data were from *melanogaster*, *simulans*, or *yakuba*. For example, the introduction of quantitative contrasts of DNA polymorphism within species to divergence between species occurred fifteen years ago in Drosophila (Hudson *et al*. 1987), and was followed by several advances that are routinely applied to population genetic data from humans to *Arabidopsis*. Gene based (as opposed to genomic) comparisons of DNA sequence polymorphism and divergence in *melanogaster*, *simulans* and *yakuba* have yielded clear evidence of the effects of selection on different scales, and involving different mechanisms. When such approaches are extended to the genomic scale, new tools, discoveries and annotation will follow. For example, genes or other features with atypical contrasts of levels of polymorphism and divergence will be candidates for further investigation of potentially interesting biology and medicine. New methods and questions arising from complete genomic comparisons will lead to exciting advances. The interpretation of population variation in genomic expression patterns (RNA and protein) must be linked to genomic polymorphism, which in turn must be linked to genomic divergence between closely related species.

Value

The sequences of *simulans* and *yakuba* are exceptional values in terms of effort and cost. Assembly of a *simulans* random shotgun into a high quality reference sequence can be built on the *melanogaster* reference sequence. Similarly a *yakuba* shotgun should assemble into large contigs based on synteny (and high similarity) with *melanogaster*. BAC libraries of both genomes will be available for finishing, if resources allow. The straightforward nature of data collection and genome assembly minimizes per-base and per-genome costs. Given the large fraction of *simulans* and *yakuba* regions which would be alignable to *melanogaster*, these two sequences would be exceptional values in terms of cost "per- annotated- base." For example, virtually every base in *simulans* can be aligned to the *melanogaster* reference. Each diverged base ($>5x10^6$ bp) as well as the few unalignable regions, would be of value in characterizing mutational and substitutional processes causing genomic variation.

Communities

This project would be built on the *melanogaster* model and would be aimed at the broad research goal of understanding genetic variation (in humans, as well as in *melanogaster* and its relatives). Thus, the project would intersect several "communities." First, the white paper originates from the Drosophila population genetics community. This community is positioned between the Drosophila molecular and developmental community, and the rigorous theoretical population genetics community. The interests and productivity of the Drosophila population genetics community over the last decade make *simulans* and *yakuba* its clear choice. An attachment documents the broad support in that community including the choice of these two particular species. The larger Drosophila research community is a second constituency for the project. The annotated *melanogaster* reference sequence has already transformed Drosophila biology. The *simulans* and *yakuba* sequences would provide a new dimension to that central resource. Finally, these two genome sequences would be essential components in the extension of the *melanogaster* model to the general development of comparative and population genomics. As many of the tools and ideas which would be developed in *melanogaster* population genomics would also greatly enhance human biomedical research and medicine, the broad biological community focused on genomic variation will find value in these resources.

Literature

Adams, M.D., *etal*., 2000 The genome sequence of *Drosophila melanogaster*. *Science* **287**:2185-2195.

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* **144**:1297-1307.

Ashburner, M., 1989 Drosophila: A Laboratory Handbook and Manual.

Barbash, D.A., Roote, J., Ashburner, M., 2000 The *Drosophila melanogaster* hybrid male rescue gene causes inviability in male and female species hybrids. *Genetics* **154**:1747-1771.

Begun, D. J,, Aquadro, C. F., 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rate in *D. melanogaster*. *Nature* **356**: 519-520.

Begun, D.J., Whitley, P., 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. *Proc. Natl. Acad. Sci. USA* **97**:5960-5965.

Begun, D.J., 2001 The frequency distribution of nucleotide variation in *Drosophila simulans*. *Molec. Biol. Evol.* **18**:1343-1352.

Bergman, C.M., Pfeiffer, B.D., Rincon-Limas, D.E., Hoskins, R.A., Gnirke, A., Mungall, C.J., Wang, A.M., Kronmiller, B., Pacleb, J., Park, S., Stapleton, M., Wan, K., George, R.A., de Jong, P.J., Botas, J., Rubin, G.M., Celniker, S.E., 2002 Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biology* **3**:1-20.

Cariou, M.L., Silvain, J.F., Daubin, V., Da Lage, J.L., Lachaise, D., 2001 Divergence between *Drosophila santomea* and allopatric or sympatric populations of *D.yakuba* using paralogous amylase genes and migration scenarios along the Cameroon volcanic line. *Molec. Ecol.* **10**:649-660.

Charlesworth, B., Langley, C.H., 1989 The population genetics of Drosophila transposable elements. *A. Rev. Genet.* **23**:251-287.

Civetta, A., Singh, R.S., 1998 Sex-related genes, directional sexual selection, and speciation. *Molec. Biol. Evol.* **15(7)**:901-909.

Cobb, M., Burnet, B., Connolly, K., 1988 Sexual isolation and courtship behavior in *Drosophila simulans*, *Drosophila mauritiana*, and their interspecific hybrids. *Behav. Genet.* **18**:211-225.

Coulthart, M.B., Singh, R.S., 1988 Low genic variation in male-reproductive-tract proteins of *Drosophila melanogaster* and *Drosophila simulans*. *Molec. Biol. Evol.* **5**:167-181.

Coyne, J.A., 1996a Genetics of a difference in male cuticular hydrocarbons between two sibling species, *Drosophila simulans* and *D. sechellia*. *Genetics* **143**:1689-1698.

Coyne, J.A., 1996b Genetics of differences in pheromonal hydrocarbons between *Drosophila melanogaster* and *D. simulans*. *Genetics* **143**:353-364.

Coyne, J.A., 1996c Genetics of sexual isolation in male hybrids of *Drosophila simulans* and *D. mauritiana*. *Genet. Res.* **68**:211-220.

Coyne, J.A., Kim, S.Y., Chang, A.S., Lachaise, D., Elwyn, S., 2002 Sexual isolation between two sibling species with overlapping ranges: *Drosophila santomea* and *Drosophila yakuba*. *Evolution* 56: 2424-2434.

Davis, A.W., Roote, J., Morley, T., Sawamura, K., Herrmann, S., Ashburner, M., 1996 Rescue of hybrid sterility in crosses between *D. melanogaster* and *D. simulans*. *Nature* **380**:157-159.

Demetriades, M.C., Thackeray, J.R., Kyriacou, C.P., 1999 Courtship song rhythms in *Drosophila yakuba*. *Anim. Behav.* **57**:379-386.

Dowsett, A.P., Young, M.W., 1982 Differing levels of dispersed repetitive DNA among closely related species of Drosophila. *Proc. Natl. Acad. Sci. USA* **79**:4570-4574.

Duret, L., Mouchiroud, D., 1999 Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc. Natl. Acad. Sci. USA* **96**:4482-4487.

Eslin, P., Prevost, G., 1998 Hemocyte load and immune resistance to *Asobara tabida* are correlated in species of the *Drosophila melanogaster* subgroup. *J. Insect Physiol.* **44**:807-816.

Hartl, D.L., Clark, A.G., 1997 *Principles of Population Genetics*. Sinauer, Sunderland, Mass.

Hoffmann, A.A., Hercus, M., Dagher, H., 1998 Population dynamics of the *Wolbachia* infection causing cytoplasmic incompatibility in *Drosophila melanogaster*. *Genetics* **148**:221-231.

Hollocher , H., 1998 Reproductive isolation in Drosophila: how close are we to untangling the genetics of speciation? *Current Opinion in Genetics & Development* **8**:709-714.

Hudson, R.R., Kreitman, M., Aguade, M., 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**:153-159.

International Human Genome Sequencing Consortium, 2001 Initial sequencing and analysis of the human genome. *Nature* **409**:860-921.

Joly, D., Bazin, C., Zeng, L.W., Singh, R.S., 1997 Genetic basis of sperm and testis length differences and epistatic effect on hybrid inviability and sperm motility between *Drosophila simulans* and *D. sechellia*. *Heredity* **78**:354-362.

Joly, D., Bressac, C., 1994 Sperm length in Drosophilidae (Diptera): estimation by testis and receptacle lengths. *Int. J. Insect Morph. Embryol.* **23**:85-92.

Jones, C.D., 2001 The genetic basis of larval resistance to a host plant toxin in *Drosophila sechellia*. *Genet. Res.* **78**:225-233.

Jones, C. D., 1998 The genetic basis of *Drosophila sechellia*'s resistance to a host plant toxin. *Genetics*. **149**:1899-1908.

Kaminker, J.S., Kronmiller, B., Bergman, C., Carlson, J., Svirskas, R., Patel, S., Frise, E., Lewis, S., Ashburner, M., Rubin, G., Celniker, S., 2002 A genome-wide analysis of transposable elements. *A. Dros. Res. Conf.* **43**:324C.

Kimura, K., Kidwell, M.G., 1994 Differences in P element population dynamics between the sibling species *Drosophila melanogaster* and *Drosophila simulans*. *Genet. Res.* **63**:27-38.

Knowles, L.L., Markow, T.A., 2001 Sexually antagonistic coevolution of a postmating-prezygotic reproductive character in desert Drosophila. *Proc. Natl. Acad. Sci. USA* **98**:8692-8696.

Lachaise, D., Harry, M., Solignac, M., Lemeunier, F., Benassi, V., Cariou, M.L., 2000 Evolutionary novelties in islands: *Drosophila santomea*, a new *melanogaster* sister species from Sao Tome. *Proc. R. Soc. Biol. Sci.* **267**:1487-1495.

Laurie, C.C., True, J.R., Liu, J., Mercer, J.M., 1997 An introgression analysis of quantitative trait loci that contribute to a morphological difference between *Drosophila simulans* and *D. mauritiana*. *Genetics* **145**:339-348.

Li, WH., 1997 *Molecular Evolution*. Sinauer, Sunderland, Mass.

Liu, J., J.M. Mercer, L.F. Stam, G.C. Gibson, Z.B. Zeng, Laurie, C.C., 1996 Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *D. mauritiana*. *Genetics* **142**:1129-1145.

Llopart, A., Elwyn, S., Lachaise, D., Coyne, J., 2002 Genetics of a difference in pigmentation between *Drosophila yakuba* and *Drosophila yakuba*. *Evolution* 56: 2262-2277.

Long, M., Langley, C.H., 1993 Natural selection and the origin of *jingwei*, a chimeric processed functional gene in Drosophila. *Science* **260**:91-95.

Ludwig, M.Z., Bergman, C., Patel, N.H., Kreitman, M., 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**:564-567.

Ludwig, M., Patel, N., Kreitman, M., 1998 Functional analysis of *eve stripe 2* enhancer evolution in Drosophila: rules governing conservation and change. *Development* **125**:949-958.

Macdonald, S.J., Goldstein, D.B., 1999 A quantitative genetic analysis of male sexual traits distinguishing the sibling species *Drosophila simulans* and *D. sechellia*. *Genetics* **153**:1683-1699.

Martin, C.H., Meyerowitz, E.M., 1986 Characterization of the boundaries between adjacent rapidly and slowly evolving genomic regions in Drosophila. *Proc. Natl. Acad. Sci. USA* **83**:8654-8658.

Nachman ,M.W. 2001 Single nucleotide polymorphisms and recombination rate in humans, *Trends in Genetics* **17**:481-485.

Nurminsky, D.I., Nurminskaya, M.V., De Aguiar, D., Hartl, D.L., 1998 Selective sweep of a newly evolved sperm-specific gene in Drosophila. *Nature* **396**:572-575.

Mural, R.J., *et al.*, 2002 The comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296**:1661-1671.

Patterson, J.T., Stone, W.S., 1952 *Evolution in the genus Drosophila*. Macmillan, New York.

Petrov, D.A., Hartl, D.L., 1998 High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Molec. Biol. Evol.* **15**:293-302.

Pitnick, S., Markow, T., Spicer, G.S., 1999 Evolution of multiple kinds of female sperm-storage organs in Drosophila. *Evolution* **53**:1804-1822.

Poinsot, D., Bourtzis, K., Markakis, G., Savakis, C., Mercot, H., 1998 *Wolbachia* transfer from *Drosophila melanogaster* into *D. simulans*: host effect and cytoplasmic incompatibility relationships. *Genetics* **150**:227-237.

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

Price, C.S., 1997 Conspecific sperm precedence in Drosophila. *Nature* **388**:663-666.

Price, C.S., Kim, C.H., Gronlund, C.J., Coyne, J.A., 2001 Cryptic reproductive isolation in the *Drosophila simulans* species complex. *Evolution* **55**:81-92.

Orr, H. A., Presgraves, D. C., 2000 Speciation by postzygotic isolation: forces, genes and molecules. *BioEssays* **22**:1085-1094.

Rubin, G. M., *et al.*, 2000 Comparative genomics of the eukaryotes. *Science* **287**:2204-2215.

Scavarda, N.J., Hartl, D.L., 1984 Interspecific DNA transformation in Drosophila. *Proc. Natl. Acad. Sci. USA* **81**:7515-7519.

Schmid, K.J., Aquadro, C.F., 2001 The Evolutionary Analysis of "Orphans" From the Drosophila Genome Identifies Rapidly Diverging and Incorrectly Annotated Genes. *Genetics* **159**:589-598.

Stern, D.L., 1998 A role of Ultrabithorax in morphological differences between Drosophila species. *Nature* **396**:463-466.

Sturtevant, A.H., 1929 Contributions to the genetics of *Drosophila simulans* and *Drosophila melanogaster*. *Publs Carnegie Instn* **399**:1-62.

Sucena, E., Stern, D.L., 2000 Special feature: divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc. Natl. Acad. Sci. USA* **97**:4530-4534.

Takano-Shimizu, T., 2000 Genetic screens for factors involved in the notum bristle loss of interspecific hybrids between *Drosophila melanogaster* and *D. simulans*. *Genetics* **156**:269-282.

Takano, T.S., 1998 Loss of notum macrochaetae as an interspecific hybrid anomaly between *Drosophila melanogaster and D. simulans*. *Genetics* **149**:1435-1450.

Takano, T.S., 1998 Rate variation of DNA sequence evolution in the Drosophila lineages. *Genetics* **149**:959-970.

Takano-Shimizu, T., 2001 Local Changes in GC/AT substitution biases and in crossover frequencies on Drosophila Chromosomes. *Molec. Biol. Evol.* **18**:606-619.

Ting, C.T., Tsaur, S.C., Wu, M.L., Wu, C.I., 1998 A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* **282**:1501-1504.

True, J.R., Liu, J., Stam, L.F., Zeng, Z.B., Laurie, C.C., 1997 Quantitative genetic analysis of divergence in male secondary sexual traits between *Drosophila simulans* and *Drosophila mauritiana*. *Evolution* **51**:816-832.

True, J.R., Mercer, J.M., Laurie, C.C., 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. *Genetics* **142**:507-523.

True, J.R., Weir, B.S., Laurie, C.C., 1996 A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila mauritiana* chromosomes into *Drosophila simulans*. *Genetics* **142**:819-837.

Tsaur, S.C., Wu, CI., 1997 Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of Drosophila. *Mol. Biol. Evol.* **14**:544-549.

Venter, *et al.* 2001 The Sequence of the Human Genome. *Science* **291**:1304-1351.

Wheeler, D.A., Kyriacou, C.P., Greenacre, N.L., Yu, Q., Rutila, J.E., Rosbash, M., Hall, J.C., 1991 Molecular transfer of a species-specific behavior from *Drosophila simulans* to *Drosophila melanogaster*. *Science* **251**:1082-1085.

Table 1. Differences among *melanogaster* subgroup species

melanogaster* vs. *simulans	
Cheek width	Sturtevant 1929
Eye size	"
Chorion filament	"
Wing size	"
Maxillary palp bristles	Ashburner 1989
Pupation location	"
Trichome pattern	Stern 1998
Susceptibility to Wolbachia	Hoffman *et al*. 1998
Encapsulation of parasitoids	Eslin and Prevost 194 1998
P-element activity	Kimura and Kidwell 1994
Ethanol tolerance	Mercot *et al*. 1994
Sex -related:	
Male genitalia	Sturtevant 1920
Sperm length	Joly and Bressac 1994, Joly *et al*. 1997
Courtship song	Wheeler 1991
Recombination rate	True 1997b
Incompatibility:	
Viability rescue	Barbash *et al*. 2000
Fertility rescue	Davis *et al*. 1996
Sexual isolation	Sturtevant 1920
Hybrid PNS patterning	Takano 1998, 2000
***simulans* clade**	
Sex-related	
Male genitalia	Coyne 1996a, Lauire *et al*. 1997, True *et al*. 1997a, Macdonald and Goldstein 1999
Pheromones	Coyne 1996b
Courtship behavior	Cobb *et al*. 1988
Larval morphology	Sucena and Stern 2000
Sperm length	Joly *et al*. 1997
Seminal receptacle	Joly and Bressac 1994
Incompatibilities	True *et al*. 1996a, Ting et al. 1998
Recombination rate	True *et al*. 1996b
Host plant use	Jones 1998, 2001
yakuba* vs. *santomea	
Pigmentation	Lachaise *et al*. 2000
Incompatibility	Lachaise *et al*. 2000, Cariou *et al*. 2001