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

David J. Begun, djbegun@ucdavis.edu
Charles H. Langley, chlangley@ucdavis.edu

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 a clearer view of the mechanisms causing genome evolution.

Here we propose that genome sequences of *Drosophila yakuba* and *D. simulans*, two species in the *melanogaster* subgroup of Drosophila, would significantly enhance, i) the *melanogaster* annotation, ii) our understanding of the mechanisms responsible for the organization and composition of the *melanogaster* genome, and iii) our understanding of the evolutionary processes controlling divergence of the *melanogaster* genome from that of other Drosophila, iv) investigation of the genetic and developmental basis of species differences. The *simulans* and *yakuba* genomes would also 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 made possible by the *yakuba* and *simulans* genome sequences would have direct impacts on the study of human variation.

Background

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). 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). 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 would 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.
D. simulans has two very close relatives, mauritiana and sechellia. These three species are referred to as the simulans clade. The members of simulans clade are phenotypically diverged for several traits, 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.

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

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 would 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.

The simulans and yakuba genome sequences would impact biology at several levels. First, their phylogenetic relationships and sequence divergence make melanogaster, simulans and yakuba an excellent system for improving the melanogaster annotation and for generating mechanistic explanations for conserved and diverged sequences. For example, most non-protein coding regions of pseudoobscura may be difficult to align with melanogaster - in fact, the expected value of the
pseudoobscura sequence is that for non-protein coding regions there will be two classes of sequence, alignable and unalignable. However, functionally important sequences can evolve. Thus, sequences conserved in the melanogaster vs. pseudoobscura comparison will be a subset of functional sequences. In contrast, melanogaster, simulans and yakuba will show a wide distribution of sequence similarities that will be heterogeneous across genomic regions and different classes of nucleotides. Furthermore, yakuba sequence would allow the identification of conserved, ancestral states and the partitioning of melanogaster-simulans differences into derived substitutions on individual lineages. Such data can be used to identify important sequences which evolve, and to test hypotheses on the causes of such evolution. Furthermore, the simulans and yakuba genome sequences would be a major asset in prioritizing melanogaster vs. pseudoobscura conserved sequence elements for functional analysis. Such conserved elements which are also conserved in yakuba and simulans may be more attractive targets for functional analysis in melanogaster.

Second, the simulans and yakuba genome sequences would shed new light on the mechanisms of genome evolution and on the evolution of interspecific phenotypic differences. The melanogaster subgroup species are sufficiently closely related to each other and to melanogaster that the histories of genomic changes can often be reconstructed with high confidence. This 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. In fact, a likely outcome of simulans and yakuba genome sequences is that the melanogaster subgroup would become a favored model system for evolutionary developmental genetics. More generally, the 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 genome polymorphism in the melanogaster model and its sister species, simulans. As in comparisons between species, the outgroup sequences (simulans and yakuba) would allow the inference of the evolutionarily derived states of melanogaster SNPs. Several important population genetic results have emerged from theoretical and empirical population genetic efforts aimed at investigating 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 would 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 would be directly applicable to the study of human genomic polymorphism.

Annotation

The importance of simulans and yakuba for annotation is a recurring theme of the proposal. Here we draw attention to particular aspects of annotation that would be aided by the yakuba and simulans sequence.

Tools for the initial, automated annotation of large genomes are evolving. Currently, the most reliable and widely used algorithms are based on pairwise comparisons of genomes to identify conserved sequences. However, the annotation of the future will have to be much richer than that attainable by two-species comparisons. For example, a goal of human biology will be to obtain a complete, annotated description of the genomic changes which have fixed along the human lineage versus those that have fixed along the chimpanzee lineage. The mel/sim/yak genome sequences would foster the development of tools and approaches for such annotation. These resources would extend the impressive melanogaster annotation and be directly applicable to human genome annotation. Furthermore, empirical studies have demonstrated the value of multiple taxa (at different levels of divergence) in identification of functionally important sequences (Cliften et al. 2001, Tompa 2001). Theoretical studies suggest that a phylogenetically informed approach, such as that available with mel/sim/yak and pseudoobscura is expected to be more efficient than pairwise approaches for automated annotation (Blanchette and Tompa 2002).

Sequences that are conserved between distantly related genomes may be important. However, many important aspects of genome function, organization and evolution will not be revealed simply through comparison of two distantly related genomes. For example, two well-studied melanogaster
genes, even-skipped and Esterase-6, exhibit conserved features of 5'-flanking regulatory function in spite of extensive sequence divergence in the 5'-flanking region between melanogaster and pseudoobscura (Tamarina et al. 1997; Ludwig et al. 1998, 2000). Such results show that important sequences can evolve, and that simple pairwise similarity between distant relatives may give us an incomplete picture of candidate functional elements. Thus, the sequence of two more closely related Drosophila species would enrich the melanogaster annotation in ways that the pseudoobscura sequence will not, and would support the identification of patterns of "conserved" and "diverged" genomic regions using phylogenetically informed algorithms.

**Genome Size**

Genome size varies by orders of magnitude across animals (Li 1997). The evolutionary mechanisms underlying genome size evolution remain obscure, though it seems clear that gene-number variation plays little role.

Our best data on the population genetic processes affecting genome size come from patterns of small insertion and deletion variation in (presumably unconstrained) Drosophila pseudogenes. These data suggest that (on average) deletions are larger than insertions and occur more often (Petrov and Hartl 1998). The pseudogene data have been interpreted as support for 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. Thus, 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 even within Drosophila (Powell 1997). Thus, other forces must be at play. For example, phylogenetic analysis of individual genes in melanogaster, simulans, and yakuba has 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 would 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 insertion and deletion 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.

**Codon Bias**

Codon bias is the non-random use of alternative codons. This pattern 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, simulans and yakuba. 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). Analysis of genes in melanogaster, simulans and yakuba revealed that melanogaster genes have been accumulating putative lower fitness codons (Akashi 1996). D. simulans genes show 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 would 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 determining the role of mutational and population genetic processes in evolution of codon bias and base composition.
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 would 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. 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 would 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 between closely related species have 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 explorations of novel gene function that take advantage of the tools and resources associated with the melanogaster model system. Development of the analytical tools for identifying recently recruited or lost genes in melanogaster and its close relatives would 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 Singh 1988; 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 melanogaster 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 would also become a much richer and deeper arena for analysis of genomic scale protein evolution. Moreover, the simulans and yakuba genomes will play an important role in the interpretation of melanogaster population protein variation and its causes. Note that although analysis of the pseudoobsccura 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. Finally, the yakuba genome permits sophisticated analysis of lineage-specific effects on protein evolution and the importance of protein function on such effects.

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 would 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.
Table 1. Differences among melanogaster subgroup species

**melanogaster vs. simulans**

<table>
<thead>
<tr>
<th>Character</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheek width</td>
<td>Sturtevant 1929</td>
</tr>
<tr>
<td>Eye size</td>
<td>&quot;</td>
</tr>
<tr>
<td>Chorion filament</td>
<td>&quot;</td>
</tr>
<tr>
<td>Wing size</td>
<td>&quot;</td>
</tr>
<tr>
<td>Maxillary palp bristles</td>
<td>Ashburner 1989</td>
</tr>
<tr>
<td>Pupation location</td>
<td>&quot;</td>
</tr>
<tr>
<td>Trichome pattern</td>
<td>Stern 1998</td>
</tr>
<tr>
<td>Susceptibility to Wolbachia</td>
<td>Hoffman <em>et al.</em> 1998</td>
</tr>
<tr>
<td>Encapsulation of parasitoids</td>
<td>Eslin and Prevost 1998</td>
</tr>
<tr>
<td>P-element activity</td>
<td>Kimura and Kidwell 1994</td>
</tr>
<tr>
<td>Ethanol tolerance</td>
<td>Mercot <em>et al.</em> 1994</td>
</tr>
</tbody>
</table>

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: Babash *et al.* 2000
- Fertility rescue: Davis *et al.* 1996
- Sexual isolation: Sturtevant 1920

**simulans clade**

Sex-related:
- 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

Second, *Drosophila yakuba* is very closely related to the recently discovered species, *Drosophila santomea* (Lachaise *et al.* 2000, Cariou *et al.* 2001). 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 *simulans* are diverged for several 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
**Hybrid incompatibilities**

Many fundamental and practical questions arise from the functional genomic interactions displayed in the 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 (Orr and Presgraves 2000). 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 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 mapping and interpretation of gene expression experiments in *melanogaster* subgroup hybrids.

**Interpretation of melanogaster variation**

The eventual completion of a high quality, human reference sequence will transform the study of human genomic variation. This prospect creates a hope that the integration of population genomic variation data and human genomic annotation will reveal the mutations contributing to common complex diseases, such as mental illness, cancer, diabetes, heart disease and stroke. 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 and fertile proving ground for tools and talent needed to understand human biology and medicine from the population genomic variation perspective. A “white paper” proposing the resequencing of the euchromatic portions of 50 entire *Drosophila melanogaster* genomes has been circulated (a grant proposal is expected in October 2002). 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 resequencing 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 (Hartl and Clark 1997). Virtually all of the 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 would minimize 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 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.

A second constituency for this project would be the larger Drosophila research community. The annotated melanogaster reference sequence has already transformed Drosophila biology. The simulans and yakuba sequences would provide a new dimension to that central resource and would add great value, perhaps comparable and definitely complementary to that of the pseudoobscura sequence.

Finally, these two sequences would be essential components in the extension of the melanogaster model to comparative and population genomics. As many of the tools and ideas which would be developed in melanogaster population genomics would greatly enhance human biomedical research and medicine, the broad biological community focused on genomic variation would find significant value in these resources.

Literature


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.


Takano-Shimizu, T., 2001  Local Changes in GC/AT substitution biases and in crossover


Tompa, M., 2001 Identifying Functional Elements by Comparative DNA Sequence Analysis Genome Res. **11**: 1143-1144.


