Drosophila Board White Paper 2009

Explanatory Note: The first Drosophila White Paper was written in 1999. Revisions to this document were made in 2001, 2003, 2005 and 2007. The most recent past version is:
http://flybase.bio.indiana.edu/static_pages/news/whitepapers/DrosBoardWP2007.pdf
At its 2009 meeting, the Drosophila Board of Directors decided to write a new White Paper to include progress made in the preceding two years and to assess current and future needs of the Drosophila research community. This draft was prepared by the Board, made available to the entire Drosophila research community, and modified according to feedback received from community members.

The fruit fly, *Drosophila*, continues to occupy a central place in biomedical research. Our understanding of the basic principles of genetics, including the nature of the gene, genetic linkage, meiotic chromosome segregation, and recombination, all arose from studies in *Drosophila*. When recombinant DNA technology was developed in the 1970s, *Drosophila* DNA was among the first to be cloned and characterized, leading to pioneering studies that linked molecular lesions in the genome with mutant phenotypes in a multicellular animal. Over the past several decades, research using *Drosophila* has paved the way for our understanding of the central regulatory pathways that control animal development. Many of the signaling systems discovered through this research, such as Notch, Wnt, and hedgehog, are now recognized as central contributing factors for major human diseases, including cancer, cardiovascular diseases, and neurological disorders. Similarly, *Drosophila* research has defined many fundamental biological processes that directly impact human health, including vasculogenesis, the innate immune response, stem cell determination and maintenance, cell and tissue polarity, growth control, pattern formation, circadian rhythms, learning and memory, neural pathfinding, and synaptic transmission. *Drosophila* also serves as the closest genetic model for the major insect vectors of disease, such as *Anopheles gambiae* (malaria), *Aedes aegypti* (dengue fever, yellow fever), and *Culex pipiens* (West Nile fever), as well as many major agriculturally important insects, including pollinators such as honeybees and pests that include many species of beetles and aphids. *Drosophila* provides an excellent model for understanding the genetic basis of complex traits, providing insight into the importance of gene-gene and gene-environment interactions, and identifying genes and pathways relevant to orthologous complex traits in humans. In addition, the genus *Drosophila* has been a key model system for understanding population biology, the molecular basis of speciation, and evolution.

A unique defining feature of *Drosophila* is its combination of rapid and facile genetics with a complex body plan and major organs and tissues that reflect the fundamental physiological, behavioral, and metabolic pathways in humans. Current technology allows researchers to manipulate the fly genome at a level of precision that exceeds that of any other multicellular genetic model system, from exact base changes by gene targeting to molecularly-defined chromosomal deficiencies and duplications. Single copy transposon insertions have long been routine in *Drosophila*, most recently with the added advantage of being able to target insertions to precise locations in the genome. With the genome sequences of multiple *Drosophila* species now at hand, the fruit fly also provides the best system for conducting studies of evolutionarily conserved regulatory networks, providing an ideal model for systems biology.

Studies of *Drosophila* have provided fertile testing ground for new approaches in genomic research and continue to have a significant impact on biomedical research. Maintaining and expanding this tradition relies on the recognition by the scientific community and by the NIH that *Drosophila* remains central to our understanding of human biology and the origins of disease, and requires the support of key projects and facilities as well as the development of new

technologies. To this end, the *Drosophila* research community has identified current bottlenecks to rapid progress and defined its most critical priorities for the next two years. We begin by first noting recent achievements that have been the most important for the community-at-large:

- Completion of the *Drosophila melanogaster* genome (Release_5) through refinement in the sequencing of some highly repetitive regions dispersed in euchromatin, assembly of telomeric sequence on the *4*th and *X* chromosomes as well as significant progress toward sequence finishing and assembly of 15 Mb of the moderately repetitive portion of the heterochromatin.
- Updates to the *Drosophila melanogaster* gene annotation set (Release_5.22 as of 11/04/09).
- Insights gained into gene and genome organization and evolution through the whole genome shotgun sequencing, assembly, alignment and annotation of the euchromatin of eleven additional *Drosophila* species: *simulans, sechellia, yakuba, erecta*, *ananassae*, *pseudoobscura, persimilis*, *willistoni, mojavensis, virilis, and grimshawi.*
- An expanding library of complete cDNAs and derivatives, including an expanding library of open reading frame (ORF) clones in a recombinational cloning vector.
- An expanding collection of mutant strains with transposable element insertions or point mutations disrupting over 60% of the approximately 15,000 annotated genes.
- Ten-fold expansion of the number of *Drosophila* cell lines available for study.
- Expanded chromosome deletion collections providing near complete (>98%) genome coverage and finer subdivision of the genome with deletion breakpoints mapped to the sequence.
- Continued successful use of RNA-interference (RNAi) in cultured cells, as well as development of related technologies and resources for cell screening and verification.
- Development of RNAi technologies for whole animals, including expanding libraries for tissue-specific *in vivo* RNAi for the vast majority of fly genes.
- Continued improvement of genetic techniques such as targeted gene disruption and the ability to integrate large fragments of genomic DNA into flies.
- Production and distribution of GFP-based protein traps and enhancer traps in 900 genes.
- Development of *phi*C31 integrase-mediated site-specific integration of transgenes to minimize position effects and reliably integrate DNA into the genome.
- Genomic libraries for *phi*C31 integrase-mediated site-specific integration of large DNA fragments allowing rescue of almost any *Drosophila* mutation.
- Transcriptional profiling of the complete life cycle and many tissue types.
- Progress toward genome-wide tiling arrays and next generation sequencing platforms for comprehensive transcriptional profiling and genome-wide protein binding site mapping by ChIP.
- Database development to integrate genome and genetic resources for *Drosophila*.
- Expanding international stock resources, with over 100,000 publicly available stocks.

These achievements have been accomplished through collaboration within the research community, to recognize and prioritize its most pressing needs. In addition, none of these projects could have been undertaken or completed without funding support provided, in whole or in part, by the NIH. Further progress in *Drosophila* research depends upon a continuation of this crucial collaboration. This White Paper represents an updated view of the most important priorities for near term future needs of the community.

There is overwhelming agreement that two broad areas need to be supported and expanded to serve the *Drosophila* research community in the upcoming years. These are **(I) basic**

community resources, consisting of *Drosophila* stock centers, electronic databases, and the molecular stock center, and **(II) research support for functional analysis of the *Drosophila* genome, including characterization of temporal and spatial expression patterns for all *Drosophila* genes and proteins.** These broad areas are described in detail below.

I. Basic Resources that Serve the *Drosophila* Community

A) Stock centers that provide a comprehensive range of genetically defined stocks at affordable costs are essential.

i) *D. melanogaster* strains: Unfettered access to stocks is a cornerstone of *Drosophila* research. Saving, sharing and reusing stocks maximize the yield of past and future investments in *Drosophila*, allowing researchers to build effectively on past achievements and stay at the forefront of biological research. Despite extensive efforts, cryopreservation of *Drosophila* stocks has not proven to be a practical technology and research stocks must be kept as living cultures. Consolidation of stock maintenance and distribution at dedicated centers is the most effective way to provide rapid and universal access to these indispensable research materials. Expansion of worldwide stock capacity over most of the last ten years has allowed stock centers to keep pace with the growth of *Drosophila* research, but the ability of existing infrastructure to respond to future community needs is limited. The recent closing of a major European stock center has brought these limitations into sharp focus at a time when powerful new tools are being generated at an unprecedented pace.

As described in the "Loss-of-function mutations" paragraph below, the fly community is ready to create multiple large-scale sets of versatile new stocks to support the functional analysis of the *Drosophila* genome. These stocks would be created for use by the whole community and should be housed in a public facility. Using guidelines that assure that only the most-needed stocks will be maintained, we estimate that a capacity of 50,000 stocks in the U.S. within the next five years is required to meet minimum community needs. Increased productivity and future technical developments are likely to necessitate additional capacity, however, and we anticipate that stockkeeping capacity within the U.S. will need to increase to 100,000 stocks to support *Drosophila* researchers for the foreseeable future.

We therefore consider investment in increased stock center capacity as our highest priority for NIH infrastructure funding. The only *D. melanogaster* stock center in the U.S. that accepts new stocks is the Bloomington *Drosophila* Stock Center (BDSC). In its current facility at Indiana University, the BDSC can house a maximum of 35,000 strains. That capacity will be filled in the next 3-5 years, largely by expansion from ongoing NIH-sponsored resource development projects, even after removal of several thousand strains whose usefulness has been largely superseded by newer stocks. While other solutions are possible, the most straightforward means of increasing stockkeeping capacity is expansion of the BDSC. The construction of new space or renovation of existing space at Indiana University would provide a long-term solution for the *Drosophila* community to the problem of inadequate stock center capacity.

ii) Other *Drosophila* species: The sequencing of 11 new species continues to drive demand for stocks of the twelve sequenced species and their relatives from the San Diego Stock Center at UCSD (SDSpSC). The SDSpSC currently maintains approximately 1,900 different stocks representing about 250 species, an increase of over 20% in the last year. Two hundred of these are newly created transgenic stocks in eight species. As additional genetically marked and transgenic stocks of these and other species are generated, the number of stocks will double in the next two to three years. While the SDSpSC's space and infrastructure are adequate to

accommodate the increase, the Center is already understaffed. Thus, at the very time when the role of the SDSpSC is even more central to the research community, insufficient staffing is compromising its function.

B) Expanded and improved electronic databases to capture and organize *Drosophila* data, and integrate the information with other databases used by the research community. It is essential to support efforts that can keep pace with the enormous acquisition rate and increasing complexity of data being generated by *Drosophila* researchers. These include the sequence of eleven new *Drosophila* species, re-sequencing and deep phenotyping of hundreds of wild-derived inbred *D. melanogaster* strains, up-to-date gene annotations, the characterization of mutant phenotypes, RNA and protein expression profiles, and interacting gene, protein, RNA and small molecule networks. These efforts must also include effectively linking *Drosophila* databases with those of other organisms, including other well-established model systems and emerging systems for genome research. Not only will this development promote more rapid progress in *Drosophila* research, it should also significantly enhance progress in functional genomics overall by promoting crosstalk among scientists working in different fields. Up-to-date and well-organized electronic databases are essential conduits to translate information from fly research to other areas of study that can impact human health, including the study of human biology, genetic disease and biomedicine, cellular responses to infectious pathogens, and Dipteran disease vectors.

C) Continued support for a molecular stock center that provides the community with fair and equal access to an expanding set of key molecular resources at affordable costs. The *Drosophila* Genomics Resources Center (DGRC) serves the community by collecting, maintaining and distributing valuable reagents that are utilized by labs throughout the world. Currently the DGRC houses an inventory of over 1,000,000 cDNA clones, transformation vectors, and clones in yeast as well as collections of vectors, full-length cDNA clones, EST clones, and genomic libraries. The DGRC also carries 108 cell culture lines including embryonic lines from *D. melanogaster* and other *Drosophila* species, imaginal disc cell lines, and those derived from the central nervous system. Acquisition of these resources is possible through cooperation with large-scale projects, such as the Berkeley *Drosophila* Genome Project, as well as donations from individual labs that have generated collections of clones or developed new vectors, and donations from groups that have created new cell lines or wish to share existing unique cell lines. It is important to maintain a reliable, central molecular repository that is able to expeditiously distribute key reagents to the scientific community as it can relieve individual labs of this responsibility and afford the end user with a dependable timeline for receiving materials. A central repository also ensures that these valuable resources are not degraded or lost, and provides technical guidance and ready access to reliable, relevant protocols. In addition, the importance of a molecular stock center is magnified by NIH guidelines that require investigators to make materials widely available.

II. Research Support for Functional Analysis of the *Drosophila* Genome.

A) Genetic resources. The most powerful advantage of *Drosophila* as a model system lies in the wide repertoire of genetic manipulations possible. Below we list the major current and future needs of the *Drosophila* community in continuing to support the goal of complete functional analysis of the *Drosophila* genome.

i) Loss-of-function mutations**:** Central to all genetic studies in *Drosophila* is the ready availability of loss of function mutations in all genes, including insertion, deletion, point mutation and RNAi knock-down lines. The Genome Disruption Project (GDP) has tagged 60% of annotated genes

with P-element, piggyBac and most recently, Minos insertions. Minos provides a broader spectrum of insertion sites, improving the yield of tagged genes. Also, insertions of a new Minos vector, MIMIC, can be modified by Recombination Mediated Cassette Exchange (RMCE) to allow tagging *in vivo* with any DNA element. This new strategy goes beyond generating mutations in protein coding genes. For example, it makes possible the generation of protein trap lines to reveal the temporal and spatial expression patterns and subcellular localization of thousands of proteins *in vivo* (see section C below). It also provides novel access to control sequences, structural DNAs, small RNA genes and the entire ensemble of currently unknown genetic elements. This new tool encompasses numerous applications that impinge on every aspect of fly research. In another approach, a first-generation collection of RNAi knock-down lines directed at all annotated genes has become available. Subsequently, technological improvements have been developed that result in more reliable and effective knock-down of any gene in any tissue, and second-generation collections of lines are now being generated. We strongly support continued NIH funding for insertional mutagenesis that allows RMCE tagging, for centralized RNAi screening, and for distribution of validated resources to the community. We encourage new funding opportunities for the development of RNAi resources in transgenic flies. In addition, creating collections of mutants that carry defined mutations on FRT-bearing chromosomes for thousands of genes represents a valuable step toward completing the functional analysis of the entire *Drosophila* genome.

ii) RNAi screening *in vivo*: Conditional expression of hairpin constructs *in vivo*, known as tissue-specific RNAi, has made it possible to disrupt the activity of single genes with exquisite spatial and temporal resolution. The construction and distribution of libraries of transgenic RNAi lines, which can be targeted to specific regions of the genome to ensure consistent results, is an important resource for the community. We encourage continuing support for development of tissue-specific RNAi and related technologies and resources, including robust systems for RNAi in the germline as well as support for maintenance and distribution of *in vivo* RNAi resources to the community.

iii) RNAi screening in cells: The continued value of a centralized facility for conducting RNAi screens in cultured cells is clear from the experience of the NIGMS-supported *Drosophila* RNAi Screening Center (DRSC). Important improvements include: full-genome dsRNA libraries designed using current rules for minimizing off-target effects and current gene annotations (coding and non-coding genes); availability of dsRNA libraries targeting specific classes of genes; improved image-based screens and analysis; primary cell screening; new and modified cell lines; RNAi "rescue" with *D. pseudoobscura* and *D. persimilis* fosmids (making use of resources generated for genome sequencing), ongoing production of a library for over-expression screening (derived from the community cDNA collection), and production of reagents for both loss and gain of function microRNAs and non-coding RNAs. The utility of RNAi screen results is evident in the large number of publications on individual screens and also in recent bioinformatics analyses based on full-genome RNAi datasets in the DRSC database. The community supports continued funding of the DRSC and further development of new cell screening technologies (including new cell lines and new methods for limiting false positive and false negative results), and for the distribution of data and resources to the community.

iv) cDNA resources: Comprehensive cDNA sequences for *D. melanogaster* will be of enormous use for gene annotations and expression studies, at the level of individual genes or on a genome-wide scale using microarrays. Ongoing efforts to obtain and sequence full-length cDNAs should be supported. These, in turn, can be used to generate high quality libraries of expression-ready cDNA clones that represent the full complement of *Drosophila* protein-coding genes. The insertion of these cDNAs into appropriate vectors for proteome and ribonome

studies is a high priority. Currently 10,000 expression-ready sequence-verified constructs for 5,000 genes have been produced. Approximately 10,000 expression clones have been made and are being used for expression studies in tissue culture and in flies. These resources are being used to generate a protein-protein interaction map of *Drosophila* and will facilitate the analysis of DNA-protein and RNA-protein interactions. In addition to these studies, the complete cDNA set provides a basis for the production of antibodies against *Drosophila* proteins, which represents a high-priority need of the community.

B. Functional annotation of *Drosophila* genomes.

i) Sequencing of additional genomes: Thanks to four separate National Human Genome Research Institute (NHGRI) funded initiatives, the sequence of 11 additional species of *Drosophila* is now complete. Although an important accomplishment, this work needs to be extended to obtain high quality finished genome sequences for the *melanogaster* group species (*D. simulans*, *D. sechellia*, *D. mauritiana*, *D. yakuba*, *D. santomea*, and *D. erecta*). These new data will continue to present an unparalleled opportunity for rapid progress in a range of areas including (1) using comparative sequence analysis to improve the annotations of *D. melanogaster*, (2) understanding genome evolution including the functional evolution of genetic pathways, (3) describing variation at a genome-wide scale, (4) identifying non-coding genes and regulatory elements, and (5) investigating differences between recently diverged species that produce interfertile hybrids. To fully realize the potential of this unique resource, continuing support is needed for assembling, aligning and annotating these genomes.

ii) Additional resources for sequenced genomes: There is also widespread agreement that the community would be well served by having at least one good genomic library available for each of the 12 sequenced *Drosophila* species. The choice that has emerged is a P[acman] BAC library with 40 kb +/- 5 kb insert size, aiming for ~12X coverage. These genomic clones will be used for finishing the genome sequence, allowing the rescue of mutants in different species, and providing evidence for RNAi specificity in other species. They will also allow tagging of genes to determine gene expression patterns and numerous other applications. In addition, projects aimed at sequencing ESTs and cDNA clones for selected species will be invaluable for refining annotations and for developing resources to leverage the new sequence information, such as species-specific microarrays, and high-density SNP genotyping methods for speciation studies. Finally, with NextGen sequencing technologies, upgrading the sequences of the already-sequenced genomes at relatively low cost to fill in gaps and extend long-range contiguity would be valuable to researchers studying these species or using them for comparative analysis.

iii) Genome-wide variation in *D. melanogaster*: A high priority for further annotating the *Drosophila* genome will be to obtain high quality whole genome sequences of a large number of *D. melanogaster* inbred reference strains. Understanding the effects of natural single nucleotide and copy number variants on a wide range of complex phenotypes, including variation in gene expression, will add a more subtle dimension to genome annotation that will complement other functional studies. Whole genome association studies of *Drosophila* complex traits using several hundred wild-derived strains is an efficient method of genome annotation, particularly for traits, such as behaviors, that are difficult to quantify precisely in high throughput assays. This strategy is unbiased, and includes non-coding genome regions as well as protein coding regions. These complex traits are directly relevant to human health and adaptive evolution.

iv) Genome-scale analysis of DNA elements: DNA element characterizations of great importance to the community include the identification of all sequence-based functional

elements associated with both protein coding and non-protein coding transcribed sequences, characterization of transcription factor binding sites throughout the genome, identification of other binding sites for chromosomal proteins, and locations of various types of epigenetic modifications, origins of DNA replication, and other structural features of the *D.* melanogaster genome. We endorse the value of genome-scale analysis of functional DNA elements in *D. melanogaster* and urge continued funding for such efforts.

v) Completion of the mapping, sequencing, and annotation of *D. melanogaster* heterochromatin: The difficulty of assembling heterochromatin remains the major roadblock toward the completion of genome projects in most multicellular organisms. Mapping, sequencing, and annotation of heterochromatin is essential for genome-wide analyses, such as mapping the distributions of transcription factors and chromatin components, non-protein coding RNAs, and RNAi-mediated gene disruption screens. In addition, elucidating heterochromatin organization is key to understanding the epigenetic regulation of gene expression, with immediate implications in developmental biology and medicine. Important information about the composition and organization of *Drosophila* heterochromatin has been generated through detailed assembly of existing middle repetitive sequences from whole genome shotgun with targeted finishing using BAC-based strategies. Annotation of these revealed that ~3% of all *Drosophila* protein-coding genes reside in heterochromatin. However, much of the satellite sequence is unmapped and unfinished, and reliable annotations require more complete information. While one of the roadblocks has been the availability of techniques to assemble the highly repeated sequences of heterochromatin, as such capability emerges from new sequencing technologies, we urge funding of the application of these technologies to the assembly and annotation of the heterochromatin of *D. melanogaster.*

C) Capturing temporal and spatial expression patterns for all *Drosophila* genes and proteins. Documenting the expression of all transcripts and proteins at single cell resolution will be essential to fully understand the structure and function of the *Drosophila* genome. Although the spatial expression pattern of over 7,500 genes has been determined by *in situ* hybridization to embryos, this effort needs to be completed for the remaining genes, extended to other stages in the life cycle, and done in a few key mutant backgrounds. New attention should be focused on *in situ* mapping of the expression patterns of *Drosophila* RNA genes, including those encoding the expanding number of small RNA classes, such as piRNAs.

Protein-trap technology, which allows the modification of endogenous genes to produce GFP fusion proteins *in vivo*, has been shown to provide accurate information on normal expression patterns and subcellular localization. An ideal approach toward this goal involves the establishment of a large collection of modifiable protein-trap strains that can be converted through RCME to any marker, including markers that can be used for live imaging, Chip-seq, immunoprecipitation, transmission electron microscopy or *in vivo* protein inactivation. New opportunities should be exploited to generate large sets of such fusion genes *in vitro* by recombineering, and to introduce them along with sufficient flanking DNA into specific sites supporting faithful expression using *phi*C31-mediated swapping. Genomic libraries in P[acman] are now available that should greatly facilitate the generation of these fusion transgenes. Support for the generation, maintenance and distribution of these lines to the community is a high priority, due to their versatility and widespread value.

Antibodies represent a high priority for future development. They continue to provide an essential tool for expression profiling, biochemical analyses, and are synergistic with protein traps and labeled transgenes described above. Speeding the production of antibodies against large numbers of *Drosophila* proteins is essential. A pilot project should be funded to prove that

a centralized production facility could economically generate a significant panel of high quality monoclonal and polyclonal antibodies against important classes of proteins. Support to maintain and distribute expression reagents directly to the research community remains essential.

Efforts to record and systematize protein expression patterns for electronic distribution should also be expanded. The value of such databases will increase with each improvement in resolution and breadth of coverage. Projects that combine biological expertise with sophisticated imaging methods that can capture dynamic multi-channel expression patterns in four dimensions, and with sub-cellular resolution, should be given high priority and supported for at least a few key tissues.