Drosophila Board White Paper 2007
November 2007
Explanatory Note: The first Drosophila White Paper was written in 1999. Revisions to this document were made in 2001, 2003 and 2005.
http://flybase.bio.indiana.edu/static_pages/news/whitepapers/DrosBoardWP2001.pdf
http://flybase.bio.indiana.edu/static_pages/news/whitepapers/DrosBoardWP2003.pdf
http://flybase.bio.indiana.edu/static_pages/news/whitepapers/DrosBoardWP2005.pdf
At our 2007 meeting, the Drosophila Board of Directors decided to write a new White Paper to take stock of the 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, and modified according to feedback received from the Drosophila research community.

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, 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 twenty years, research using *Drosophila* has paved the way for our understanding of the central regulatory pathways that control animal development, culminating with the award of the Nobel Prize in 1995. 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, neural pathfinding, and neurosynaptic function. *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). *Drosophila* is 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 with major organs and tissues that reflect the fundamental physiological and metabolic pathways in humans. Current technology allows the researcher 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*, now with the added advantage of being able to target insertions to precise regions of the genome. With the recent genome sequences of multiple *Drosophila* species, 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. Continued and even greater success 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 maintenance and expansion of key projects and facilities and 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 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 towards sequence finishing and assembly of 15 Mb of the moderately repetitive portion of the heterochromatin.
- Updates to the *Drosophila melanogaster* gene annotation set (Release_5.3).
- 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.
- An expanding collection of mutant strains with transposable element insertions or point mutations disrupting over 50% of the nearly 14,000 annotated genes.
- Ten-fold expansion of the number of *Drosophila* cell lines available for study.
- Expanding chromosome deletion collections providing near complete (96-98%) genome coverage and finer subdivision of the genome with deletion breakpoints mapped to the sequence.
- Development of RNA-interference (RNAi) technologies for cultured cells and whole animals and the establishment of transgenic RNAi libraries targeting more than 85% of genes.
- Continued improvement of genetic techniques such as targeted gene disruption.
- Production and distribution of GFP-based protein traps and enhancer traps in 900 genes.
- Development of *phi*C31 integrase-mediated site-specific integration of trans-genes to minimize position effects and reliably integrate large segments of DNA into the genome.
- Transcriptional profiling of the complete life cycle and many tissue types.
- Progress toward genome-wide tiling arrays for complete transcriptional profiling and genome-wide protein binding site mapping by ChIP-array.
- Database development to integrate genome and genetic resources for *Drosophila*.
- Expanding international stock resources, with over 85,000 publicly available stocks.

These achievements have been accomplished through a collaboration of the research community to recognize and prioritize its most pressing needs, and the funding agencies to provide the resources and coordination necessary to meet these needs. Further progress in *Drosophila* research depends upon a continuation of this most important collaboration. This White Paper represents an updated view of the most important priorities for near term future community resources.

There is overwhelming agreement that the following three resources must continue to be supported and expanded to serve the entire research community.

1) Stock centers that provide a comprehensive range of genetically defined stocks at affordable costs are essential. Our research often depends on the availability of fly stocks that are very difficult and labor intensive to construct. To maximize the yield of research dollars invested in *Drosophila*, the community must be able to save, share and reuse stocks efficiently. Unlike *C. elegans*, zebrafish, and mice, however, cryopreservation of *Drosophila* stocks is not a practical alternative to the maintenance of living cultures. Thus, the stock centers need to maintain live cultures of stocks for public distribution, which is highly labor-intensive. Unfettered access to stocks is one of the fundamental reasons why *Drosophila* research has been able to build effectively upon past achievements and stay at the forefront of biological research. It is of

paramount importance that these resources are EXPANDED as the *Drosophila* community creates new stocks that allow more sophisticated experiments and hence better biology.

Additional US stock center capacity will be required to continue to accommodate the stocks generated by high-throughput community resource projects and high-value stocks resulting from hypothesis-driven research. The existing capacity of 25,000 *D. melanogaster* strains at the Bloomington Stock Center is expected to meet community needs for the next two years only. The success of current efforts to accumulate functionally defined mutant alleles or gene-disruption insertion alleles for every gene (more than 16,800 lines and growing), deficiencies that provide extensive coverage and fine subdivision of the genome (more than 2,100 lines and growing), and other critical tools for controlled manipulation of the *Drosophila* genome (more than 2,600 lines and growing) will saturate funded US stock center capacity at a time when many powerful new tools are being generated. These sets of tools include strains that produce fluorescently-labeled endogenous proteins, RNAi lines, strains for sophisticated mosaic analysis, sequence-defined duplications, and lines for next-generation transgenesis methods. Sets of strains focused on genes with related functions, including components of *Drosophila* models of human disease, will also become available and, as with strains developed specifically for community use, will require public distribution to be widely exploited. We estimate that 15,000 new high-value stocks will become available over the next 5 years, only about 5,000 of which can currently be accommodated in a public stock center. Therefore, additional US stock center capacity must be developed to bring the total to approximately 35,000 strains. Corresponding financial support is critical to reaching that goal.

The sequencing of 12 new species has driven increasing demands for stocks of the twelve sequenced species and their relatives from the Tucson Stock Center. Tucson currently maintains approximately 1500 different stocks of about 250 species. Given the envisioned acquisition of fresh wild type and newly created genetically marked stocks of these other species, this number will double in the next two to three years. While Tucson's space and infrastructure are adequate to accommodate the increase, the Center is understaffed. It is also worth noting that several groups are attempting to manipulate many of the 12 fly species and that it is anticipated that sophisticated genetic manipulations may soon be possible in these species. Hence, the role of the Tucson Stock Center will become even more important in the near future and an increase in support is truly important.

2) 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, including the sequence of eleven new *Drosophila* species, up-to-date gene annotations and 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 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 human research.

3) 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. These resources include commonly used vectors, full-length cDNA clones, EST clones, cell lines, genomic libraries and microarrays. A well-run molecular stock center is cost effective

for grant dollars, serves multiple research communities and plays a catalytic role by making available resources that might otherwise remain closely held. Moreover, the importance of a molecular stock center is magnified by new NIH guidelines requiring investigators to make materials available through such centers.

In addition to the resources described above, certain research projects that require large infrastructures and investments over several years must be in place to realize the full potential of *Drosophila* as a model system for functional and comparative genomics. Several of these projects are ongoing, use existing technologies, and require adequate funding for their successful completion. Others are projects that require the development of new technologies.

The research community considers the following high priority projects.

4) Functional analysis of the *Drosophila* genome. The most powerful advantage of *Drosophila* as a model system lies in the wide repertoire of genetic manipulations possible. Key to all genetic approaches is the ready availability of loss of function mutations in all genes. An ongoing NIH-funded project will provide for the generation and sequencing of nearly 10,000 unique *P*-element insertions for an anticipated 75% coverage of the annotated genes. Because many genes will be refractory to mutagenesis by transposable elements, alternatives to *P*-element gene disruption techniques should also be considered a high priority. Developing technologies such as TILLING, homologous gene replacement, PCR-based deletion screening, and SNP mapping of point mutants are important to accomplish the functional analysis of the entire genome by mutations. RNAi screening is another powerful approach for functional analysis of the *Drosophila* genome, both in tissue culture and *in vivo*. The value of a centralized facility has already become clear from the experience of the NIGMS-supported *Drosophila* RNAi Screening Center (DRSC). Over 7,000 genes have been linked to a phenotype in one or more assays developed in *Drosophila* cells. Important improvements include: replacing ~5% of the dsRNAs in the existing library that have off-target effects, generating dsRNA libraries targeting specific classes of genes, improving screen automation, data acquisition and data normalization, and integration of the DRSC database with existing ones to enable cross validation and high confidence references for data mining purposes. We encourage continued support for centralized RNAi screening, as well as distribution of validated RNAi resources to the community. In addition, conditional expression of hairpin constructs *in vivo*, known as tissue-specific RNAi, makes it now possible to disrupt the activity of single genes with exquisite spatial and temporal resolution. The construction and distribution of libraries of transgenic RNAi lines, targeted to specific regions of the genome to ensure consistent results, will be an important resource for the community.

5) Capturing temporal and spatial expression patterns for all *Drosophila* genes and proteins. Documenting the expression of all transcripts and proteins at single cell resolution ultimately will be essential to fully understand the structure and function of the *Drosophila* genome. Over 5,000 genes have been analyzed by *in situ* hybridization to embryos, and this analysis should be completed for the remaining genes and extended to other tissues at different stages of the life cycle. New attention should be focused on *in situ* mapping the expression patterns of Drosophila RNA genes, including those encoding piRNAs. Protein-trap technology, i.e., the modification of endogenous genes to produce GFP fusion proteins *in vivo*, has been shown to provide accurate information on developmental expression and subcellular localization. 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. 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 remain invaluable tools for expression profiling, biochemical analyses, and are synergistic with protein traps. Speeding the production of

antibodies against large numbers of *Drosophila* proteins is a high priority. A pilot project should be funded to prove that a centralized production facility can 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 expression information for electronic distribution should also be expanded. Currently, databases strive to contain information of sufficient quality to allow candidate genes of possible interest to particular investigators to be identified. The value of such databases will increase with each improvement in resolution and breadth of coverage. Acquiring higher resolution, and more biologically meaningful expression data remains heavily dependent on the quality of the tissue preparations, and on expert knowledge. 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.

6) Production of comprehensive cDNA resources. cDNA sequences for the majority, if not all, of the genes of *Drosophila melanogaster* will be of enormous use for gene annotations and expression studies at the level of individual genes or on global scales using microarrays. Ongoing efforts to obtain and sequence full-length cDNAs should be supported. In addition, the insertion of the complete cDNA set into appropriate vectors for proteome and ribonome studies is a high priority. Such studies may include analysis of protein-protein, DNA-protein and RNA-protein interactions. In addition to these studies, the complete cDNA set could be used as a tool for the production of antibodies against Drosophila proteins. Well-characterized cDNAs, which have been corrected for amplification-mediated mutations, need to be placed in vectors that can be manipulated for various proteomics applications. This would allow these tools to be efficiently produced and made available to the community at reasonable costs.

7) Functional annotation of *Drosophila* genomes. Thanks to four separate National Human Genome Research Institute (NHGRI) funded initiatives, the sequence of 11 additional species of *Drosophila* is now complete. 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 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. In addition, projects aimed at sequencing EST's 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. 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. A panel of sequenced reference strains will galvanize community efforts for whole genome association studies of *Drosophila* complex traits of relevance to human health and adaptive evolution.

8) Completion of the mapping, sequencing, and annotation of *Drosophila melanogaster* heterochromatin. The difficulty of analyzing 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 annotation of existing sequences, including the demonstration that

~3% of all *Drosophila* protein-coding genes reside in heterochromatin. However, much of the existing sequence is unmapped and unfinished, and reliable annotations require more complete information. The NHGRI has generously funded the *Drosophila* Heterochromatin Genome Project (DHGP), and we encourage continued support for this project as well as other investigations of heterochromatin sequence and function.

Below we categorize additional high-priority needs of the community that may be best met by R01, investigator-initiated efforts or pilot grants, rather than by large project grants.

- **Development of new methodologies that broaden the scope of the use of RNAi in *Drosophila* cells and whole animals.** In particular, the application of RNAi to primary tissue culture cells will facilitate the design of novel cell-based assays that reflect complex *in vivo* biological processes (i.e., axonal outgrowth, muscle differentiation). In addition, any technological advances that aim to improve the efficiency and miniaturization of the high-throughput RNAi approach (i.e., dsRNA chips) are clearly needed in order to improve the reliability and affordability of the current technology. Improvement of methods to deliver RNAi to whole animals, especially embryos, is also needed. Finally the construction and distribution of validated resources for RNAi screening will greatly expand access to this technology.

- **Development of new cell lines and molecular characterization of existing cell lines.** Cell lines have found increasing use in *Drosophila* research, but only a limited number of *Drosophila* cell lines are available. In particular, there is a need for tissue-specific cell lines that could be used in RNAi screens (for example epithelial cells to screen for genes involved in epithelial cell polarity), and for cell-cell interaction studies (i.e. cell lines that fail to express a certain signaling pathway). Having access to a diverse set of cell lines should facilitate the biochemical purification and analysis of molecular complexes and would complement whole organism approaches.

- **Development of methods to understand the evolution of gene function.** It is important to understand the functional evolution of genetic pathways, not just sequence evolution. This requires support to develop tools for the sequenced non-*melanogaster* species such as gene replacement and transformation that have been successfully used in *D. melanogaster*.

- **Generation of a well-characterized collection of conditional (ts lethal) mutants.** Such a collection would be of real value to the *Drosophila* community for several reasons. First, the majority of available lethal mutations are embryonic lethal; and thus, studying post-embryonic development using these mutants is extremely difficult. Second, even for those lethals that do die later in development, potential embryonic defects can be masked by stores of protein or RNA deposited into the oocyte by heterozygous mothers. In such cases, it is necessary to make germline clones, however, if the protein is also required for germline development, such eggs may stop developing or they may be disorganized and therefore difficult to analyze. Third, even in the best cases, conditional mutants are required to determine the precise temporal requirement of a gene product. One of the best ways to address all of the above limitations and move the field forward is to isolate conditional mutants in as many genes as possible.

- **Developing an efficient means of cryopreservation of *Drosophila* at any stage of development.** This has long been a high priority for *Drosophila* researchers. Existing methods allow preservation of embryos but are not robust and have had no impact on *Drosophila* stock keeping. A practical cryopreservation procedure would reduce the stress on the stock centers, ensure that valuable genetic resources are not lost and

would curtail costs common to every *Drosophila* lab associated with long-term stock maintenance.