

## **Drosophila White Paper 2003**

August 13, 2003

*Explanatory Note: The first Drosophila White Paper was written in 1999. Revisions to this document were made in 2000 and the final version was published as the Drosophila White Paper 2001*

<http://flybase.bio.indiana.edu/docs/news/announcements/drosboard/Whitepaper2001.html>

*In 2003, the Drosophila Board of Directors voted 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. A draft prepared by the Board was circulated to the community-at-large through FlyBase and directed email. With the input of the community included, a revised version was submitted for formal approval by the Drosophila Board. The final version will be provided to the Trans-NIH Genomics Resources Group, a committee including representatives of the various NIH institutes that oversees broad resource and infrastructure initiatives for genome research. It will also be available as a resource to other agencies and interested parties to inform them of recent progress and priorities of the Drosophila research community. This Drosophila White Paper 2003 and a summary of community input are posted on FlyBase at:*

<http://flybase.bio.indiana.edu/docs/news/announcements/drosboard/>

The contributions of *Drosophila* as a model system for understanding basic biological mechanisms are even more evident today than in the previous years. This is in large part due to the advances in genomic technologies, which when combined with the powerful genetic manipulations possible with *Drosophila*, allow researchers to dissect complex biological problems that could not have been successfully approached in the past. In addition, the translations of *Drosophila* research to other arenas, including studies of human population dynamics, development and disease mechanisms continue to yield impressive successes. To name a few recent examples, we note how the signaling pathway for dorsal/ventral pattern formation in *Drosophila* embryos has quite unexpectedly provided a crucial paradigm for signaling in human inflammation and innate immunity. Counterparts for over 70% of human disease genes are found in *Drosophila* and many of these fly genes are being extensively studied. Hence the number of examples showing that *Drosophila* can serve as an excellent disease model continues to increase. Indeed, a growing number of researchers studying human biology are collaborating with *Drosophila* researchers to apply the powerful genetics of *Drosophila* to understand the mechanisms of Huntington disease, Parkinson disease, spinocerebellar ataxia, early onset Alzheimer disease, and other genetic disorders. Overall, more than 60% of human genes have homologs in *Drosophila*. Thus most cellular and developmental processes are functionally conserved. Key insights have been gained in recent years into the genetic and cellular mechanisms of processes such as neurodegeneration, vasculogenesis, stem cell determination, cell and tissue polarity, signal transduction, growth control and organogenesis. Models proposed for the function of many newly described mammalian proteins are based on the mutant phenotypes associated with the *Drosophila* homologues and *Drosophila* is now widely used for *in vivo* functional analyses that are difficult to carry out in vertebrates. The availability of *Drosophila* genomic sequences and its integration into well-studied biology of flies have provided a boost to the power of comparative genomics. A recent example is how the identification of genes that play a role in

malaria transmission is relying heavily on comparisons of the genomes of *Anopheles* and *Drosophila*.

Studies of *Drosophila* have provided fertile testing ground for new approaches in genomic research. Continued and even greater success relies on the maintenance or expansion of key projects and facilities and on 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:

- High quality finishing sequence of the euchromatin of *Drosophila melanogaster*
- Reannotation of the euchromatin (Release 3.1)
- An expanding library of complete cDNAs
- An expanding collection of mutant strains with transposable element insertions in newly annotated genes
- Progress toward the goal of complete coverage of the genome with chromosomal deficiencies
- Progress on a heterochromatin genome project
- Development of RNA-interference technologies for cultured cells and flies
- Transcriptional profiling of the complete life cycle and many tissue types
- Database development to integrate genome and genetic resources for *Drosophila melanogaster*
- Sequence and partial assembly of the euchromatin of *Drosophila pseudoobscura*

There is overwhelming agreement that the following three resources must be supported to serve the entire community of *Drosophila* researchers.

- 1) A well-funded stock center with a carrying capacity of at least 20,000 strains. This number takes into account current efforts to accumulate at least one mutant allele for every gene, deficiencies that provide extensive coverage of the genome, and the lines being generated by the ongoing gene disruption projects. The Bloomington Stock Center, which is serving the community extremely well, can accommodate this immediate goal if it is provided adequate funding.

It is important to note that the community anticipates a need to house 10,000 - 20,000 additional strains in the near future. This number includes having at least two different mutant alleles of each gene, a refined set of molecularly mapped deficiencies and duplications (particularly needed for mapping X chromosomal genes), and sets of widely used transgenic marker strains for inducible gene expression or protein trapping. Given current ongoing efforts to generate these strains, well-characterized collections should be available to the community in three to five years. This expansion will require either a significant expansion of the physical facilities and personnel at the Bloomington Stock Center, or the identification of a second national facility.

- 2) Expanded and improved electronic databases to capture and organize *Drosophila* data, and integrate the information with databases used by other research communities. It is essential to support efforts that can keep pace with the enormous rate and increasing complexity of data being generated by *Drosophila* researchers, including up-to-date gene annotations and the characterization of mutant phenotypes, RNA and protein expression profiles, 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) A molecular stock center that would provide the community with fair and equal access to key molecular resources at affordable costs. These resources include commonly used vectors, cDNA and genomic libraries and quality controlled cDNA or oligo-based microarrays and genomic tiling arrays. Reliance on commercial companies to provide microarrays may not be an adequate long-term solution as it limits the widespread use and data distribution of important technologies and information. We believe that a molecular center that could generate and distribute these reagents, particularly cDNA and genomic arrays, and serve as a technological advice center would do much to advance the use of functional genomics by individual investigators. Finally, we point out that a well-run molecular stock center would be cost effective for grant dollars and could serve multiple research communities.

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) Sequencing of a set of complete cDNAs representing the vast majority, if not all of the genes of *Drosophila melanogaster*. The cDNAs will be of enormous use by the community of researchers for gene annotations and expression studies at the level of individual genes or on global scales by microarrays. We understand that NIH has made a 3-year commitment to the BDGP to sequence ~ 5000 new cDNAs with full length ORFs. Together with the previous work, this should provide an estimated 80% coverage. We emphasize the importance of full funding of this project and the need to identify alternative transcripts for many genes to understand the added complexity of multiple gene products.
- 5) Insertion of the complete cDNA set into appropriate vectors for proteome and ribonome studies. 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 large-scale 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.
- 6) Gene disruption for a mutational analysis of the genes of *Drosophila melanogaster*. 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, PCR-based deletion screening, and SNP mapping of point mutants are important to accomplish the functional analysis of the entire genome by mutations.

- 7) Completion of a *Drosophila* heterochromatin genome project. The sequence analysis of heterochromatin remains the major roadblock toward the completion of the genome projects of essentially all multi-cellular organisms. Developing and testing technologies to tackle the challenges of dealing with heterochromatin can best be accomplished in *Drosophila melanogaster* where a variety of experimental tools can be brought to bear on the challenges of dealing with highly repetitive DNAs. In addition, a heterochromatin genome project is necessary to completely understand the informational content and molecular organization of the *Drosophila* genome.
- 8) The sequencing of additional *Drosophila* species. The sequencing of *D. pseudoobscura* has recently been completed and researchers worldwide are reaping the benefits for functional annotation of coding sequences, for prediction of DNA enhancer sequences and RNA *cis*-regulatory sequences and identification of non-coding RNAs. The sequencing of *D. simulans* and *D. yakuba* remain the top priorities for immediate sequencing in the next year. In March 2003, the *Drosophila* Board asked a group of colleagues, with expertise in the areas of ecology, phylogeny, evolutionary biology, developmental biology, and bioinformatics, for advice on the number and identity of species that should be considered top priorities for the next sequencing projects. After careful consideration, the expert group recommended the following eight species, in addition to *D. simulans* and *D. yakuba*, for genome sequencing in the next two years: *D. willistoni*, *D. erecta*, *D. ananassae*, *D. virilis*, *D. grimshawi*, and *D. mohavensis* at 8X coverage; *D. sechellia* and *D. persimilis* at 3X coverage. This proposal received enthusiastic endorsement by the *Drosophila* Board and widespread community support. A White Paper proposal that incorporated community input was submitted to the NHGRI in June 2003 and is currently under review. Applications of the proposed comparative genomics project include improving *D. melanogaster* gene annotations, identification of conserved non-coding and coding regions of genes (including non-coding RNAs), and tracking changes associated with gene and chromosome evolution. Because of the vast knowledge of the phylogeny and biology of the drosophilids, we are confident that the investment in these genome projects will be considered an outstanding success, not only by *Drosophila* researchers, but by all who are interested in comparative genomics and molecular evolution. Beyond the benefits to the *Drosophila* community, this project will lead to the development of bioinformatic tools that can be applied subsequently to the comparison and annotation of larger vertebrate genomes. Improvements in genome sequencing technologies over the last several years have lowered the costs involved considerably to an estimated \$3 million for a genome the size of *D. pseudoobscura*. Thus, the total cost of this project should be a fraction of the cost of sequencing a mammalian genome.
- 9) Capturing spatial expression patterns for all *Drosophila* genes. Particularly powerful is the protein-trap technology using a transposable element with a GFP-containing exon to mark proteins and analyze tissue and sub-cellular distribution of proteins *in vivo*. Support to

generate, maintain and provide these lines to the community is considered a high priority since *in vivo* applications are broad and powerful. Ongoing efforts have also demonstrated the utility of genome-wide analysis of RNA expression patterns using RNA *in situ* hybridization to embryos. Thus far, 2500 genes have been analyzed and these efforts have demonstrated an economy of scale. This analysis should be completed for all genes and extended to other tissues at different stages of the life cycle. The development of sophisticated imaging methods that could capture dynamic expression patterns in multi-dimensions and with sub-cellular resolution will add substantially to the utility of this information.

Below we categorize additional needs of the community that are judged to be best met by R0-1, investigator-initiated efforts or pilot grants, rather than by large project grants.

- 1) An efficient means of cryopreservation of *Drosophila* at any stage of development. There is no question that renewed efforts to develop a suitable cryopreservation technique remains a high priority for *Drosophila* researchers. Successful application would reduce the stress on the national stock center, ensure that valuable genetic resources are not lost and could curtail costs involved in running fly kitchens, and constantly maintaining laboratory stocks in all *Drosophila* labs.
- 2) Continued development of technologies for RNAi in whole flies. RNAi is now being used with high success in cultured cell lines using simple delivery methods. However, efficient delivery in whole flies remains a major challenge.
- 3) Molecular mapping of chromosomal deficiencies and duplications. The community uses chromosomal deletions extensively to map genes of interest and to identify dosage-sensitive modifiers of phenotypes. Currently, an estimated 85 to 91% of the euchromatic portion of the *Drosophila* genome is deleted and subdivided by existing chromosomal deletions. A project to molecularly map the endpoints of the existing set of deletions would be straightforward to carry out. The results would immediately define molecular intervals for mutations of interest and tie cytogenetic breakpoints of these heavily used chromosomes to the genome sequence. This would complement the DrosDel project currently being carried out by a consortium in Europe. In addition, and particularly relevant for analysis of X-linked genes, is the molecular characterization of existing and newly generated genomic duplications.
- 4) Development of new cell lines. Cell lines have found increasing use in *Drosophila* 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.