Drosophila ENCODE (Encyclopedia of DNA Elements) White Paper

General Rationale: The sequencing of the human genome and those of major genetic model organisms has greatly facilitated biomedical research and will ultimately improve human health. However, maximizing this benefit crucially depends on our ability to interpret all of the information encoded within genome sequences, a daunting task that will require extensive experimental and computational analysis [1]. Recognizing this considerable challenge, the NIH has launched a pilot ENCODE project to comprehensively explore 1% of the human genome [2, 3]. Moving towards a 100% human ENCODE project requires testing our abilities to map features onto the more manageably-sized genomes of model organisms [4]. The tractable genomes of model organisms will help us develop full literacy in "reading" genomes. Genome literacy in turn will provide a framework for understanding how complex biological systems work. Developing a comprehensive map of known, and discovering unknown, features in the *Drosophila melanogaster* genome was the focus of a recent workshop in Bethesda, Maryland [5], and resulted in this white paper.

Drosophila offers a variety of attractive features well suited for a model organism ENCODE project. The sequence of the euchromatic portion of the *Drosophila* genome is virtually complete and is of very high quality [6], and finished heterochromatic sequence is being generated [7]. The structural arrangement of *Drosophila* and human genes is similar, and both organisms employ extensive splicing of pre-mRNAs to generate greater protein diversity [8]. Additionally, >70% of human disease associated genes have *Drosophila* homologs [9], making this genome an important translational tool for human health research. The *Drosophila* genome has good baseline annotation [10] and an established comprehensive database [11] of curated functional information. The impending whole genome sequences from 11 additional species of *Drosophila* will complement the *Drosophila melanogaster* sequence and annotation. These 12 sequences offer a unique combined resource for exploiting comparative techniques [12] and

improving the discovery and annotation of conserved functional sequence elements. These sequences will also provide important insights into genome evolution. Importantly, the sequenced *Drosophila* species form a ladder of genetic divergence comparable to that found across mammalian genomes [13, 14]. Understanding the structure of these genomes will greatly assist efforts towards enriching human genome annotation by comparative mammalian genomics.

The tractability of *Drosophila* genetics and the many shared features of *Drosophila* and humans (*e.g.* a complex organ-based body plan) and will provide for inexpensive functional validation of ENCODE annotations, using a suite of tools including classical genetics, transgenic expression, high throughput point mutation detection and RNAi [15-18], as well as emerging proteomic tools that can be applied to annotation [19]. One of the primary tools in the human ENCODE arsenal will be the study of allelic diversity and population structure. Mapping of quantitative traits will depend on methods for putting together large datasets including gene expression and regulatory polymorphisms. *Drosophila* has long served as the preeminent model for the development of population and quantitative genetic approaches. These methods are increasingly important for finding variation associated with human disease [20, 21]. As rigorous tests of human DNA element function is often not feasible and/or ethical, model organisms will continue to be critical for a full scale human ENCODE project.

A large and engaged Drosophila research community will facilitate DNA element discovery and validation. The *Drosophila* genomics, genetics and population biology communities are populated with talented and well-trained scientists with a strong history of embracing new techniques to uncover basic biological processes. A deep literacy of reading genomes will require these types of skills. The community effort required for a successful ENCODE project will be a challenging sociological exercise. The Drosophila community has a strong tradition of generous data and resource sharing to match its considerable expertise.

General Conclusions: Workshop participants, from the international *Drosophila*, genomics-technology, and funding communities [5] explored how a *Drosophila* ENCODE project would contribute to the discovery and annotation of coding and non-coding genome features; how *Drosophila* genetics could be used to attach functional annotation to already-defined and yet-to-be-discovered features; and how such efforts would contribute to the goals of the human ENCODE project [2, 3]. A *Drosophila* ENCODE project would leverage the work of greater than 1,600 *Drosophila* laboratories worldwide with a major NHGRI investment in the sequencing of a dozen species of *Drosophila*.

Both the *Drosophila* and human research communities will directly benefit from the establishment of a D-ENCODE central database populated with computational and experimental evidence. Data capture under the D-ENCODE umbrella requires a generic structure that facilitates interoperability within D-ENCODE and external public databases, as well as an accessible framework for adding community generated data.

The human ENCODE project is based on the analysis of 1% of the genome. The *Drosophila* genome, sized at ~5% of the human genome, will provide an important milestone on the road to a whole-genome human ENCODE project. A whole-genome D-ENCODE project will help define the catalogue of elements likely to be encountered in human ENCODE, provide a realistic test of annotation scale-up and finishing, and help with element prioritization for human ENCODE. Additionally, new ideas and approaches are inevitable. For example, chromosomal-scale elements that might escape detection in pilot-scale human genome surveys may be discovered by D-ENCODE.

Much of the infrastructure for an ENCODE project is already in place or is coming online, including:

- 12 sequenced *Drosophila* genomes.
- Pilot re-sequencing of 50 *Drosophila melanogaster* lines.
- Rich EST and cDNA collections.
- *In situ* hybridization database.

- Gene disruption and protein trapping efforts.
- Genomic resource centers.
- Genetic resource centers (for tilling mutation detection, P-element disruption, deletion sets, and RNAi).
- Full genome tiling path arrays (community and commercial).
- Assorted public database repositories.
- A primary database of the *Drosophila* genome annotation, functional data, and literature.

Funding for many of these resources will need to be modestly augmented to support a D-ENCODE project. In particular, molecular and whole organism stock centers with comprehensive sets of reagents are absolutely essential. The provision of adequate stock center capacity is critical for both focused D-ENCODE projects and for the rapid dissemination of reagents to the wider research community to enable rigorous validation of functional annotations.

Genome tiling arrays for comprehensive expression and ChIP-on-chip experiments are currently available through commercial vendors and are coming online in academic labs. Rapid launch of D-ENCODE requires immediate access to cost effective array resources. Array resources for additional species must also be developed to generate orthogonal sequence-conservation and functional-genomic datasets. Such resources are essential for exploring the considerable portion of the genome where there is evidence of evolutionary conservation and/or transcriptional activity but where we have little or no functional knowledge. Creation of affinity-based protein reagents (antibodies and tagged proteins) for capturing the associated DNA elements for ChIP-on-chip studies should also be coordinated and accelerated.

Similarly, lower-level annotation of additional species is required to inform the annotation of *Drosophila melanogaster* and better define the 'rules' for computational genome annotation in metazoans. Additional resources will be required to move beyond first-pass automated gene finding in non-*melanogaster* genomes. In particular, the sequence of additional species of flies is deemed to be important (especially for groups

falling between *Drosophila* and *Anopheles*) as is moderate-coverage cDNA sequencing. Other community groups are organizing these effort and D-ENCODE will benefit.

While community species-specific databases are an ideal home for 'gold-standard' validated D-ENCODE data [22], they should not be the home of e-published working data from D-ENCODE members. Additionally, these data should not be stored and displayed exclusively at lab-based sites because of uncertain long-term stability. Data should instead reside in appropriate, established public repositories. The interoperability of these data sources must be strong. A working group will be established to harmonize data object fields and formats, optimize data synchronization and tracking, and define standards and criteria for inclusion of data in the D-ENCODE project database. Following the long tradition of data and resource sharing among *Drosophila* researchers, there is great enthusiasm for making all verified data simultaneously available to D-ENCODE and non-D-ENCODE researchers.

Infrastructure Priorities:

While a significant advantage of D-ENCODE is the availability of pre-existing infrastructure, some of the grants supporting these resources will be up for renewal as D-ENCODE is coming online. These resources must be maintained to support the project. Novel organizational and database needs require the creation of new infrastructure.

1. **Develop an organizational structure for D-ENCODE.** A self-identified core of individuals or labs will be responsible for communication and dissemination of results and methods; working with both internal and external groups. Within the D-ENCODE project, the group will act as the primary project coordinator, establishing the operating procedures and standards for data exchange (syntax, controlled vocabularies, versioning, and unique identifiers), and coordinating the use of cell types, tissues, and developmental stages to ensure the generation of cohesive data sets for meta-analysis. This group will work closely with other D-ENCODE groups (*via* regular conference calls and meetings) to monitor progress, facilitate data sharing, and identifying areas warranting closer examination. This

core will also report progress and concerns to the community-elected *Drosophila* Board, and foster the participation of the *Drosophila* community -- long the source of rich *Drosophila* annotation. To facilitate transfer of expertise and data to human ENCODE and ensure communication of arising human ENCODE needs, D-ENCODE will be in regular contact with human ENCODE. NHGRI project managers will be integral members of the core group. A Scientific Advisory Board will be established. Estimated costs: $250,000/yr (*via* cooperative agreement U01).

2. **Develop a coherent data management structure for D-ENCODE.** A centralized portal for D-ENCODE data is essential for two reasons. It will serve as a community kiosk, a single site where all of the D-ENCODE data (directly anchored to the *Drosophila* genome) will be available for visual comparisons. And, it will provide a location where researchers can collect and retrieve an internally consistent set of data, that correlates to the same released version. Using the shared operational processes and data standards that the coordination group develops, this group will be responsible for providing the data and making it available from a single portal. D-ENCODE encourages the use of pre-existing public databases where possible and will support the development of new databases when required. Estimated costs: $1,000,000/yr (*via* collaborative agreement U01).

3. **Technologies and resources to support discovery and verification of DNA elements.**
   - Ensure adequate and cost effective *Drosophila melanogaster* array resources, particularly tiling path arrays for transcript discovery and ChIP-on-chip studies. Estimated costs: $250,000/yr (Assumes community generation. Additional cost recovery from end users).
   - Accelerated development and distribution of antibody and other affinity reagents for ChIP-on-chip studies. Estimated cost: $500,000/yr.

- Develop comparative genomic tools for multi-genome alignments and within species polymorphism data. Estimated cost: $250,000/yr.
- Develop array resources for additional *Drosophila* species. Estimated cost: $500,000/yr.

4. Technologies and resources to support validation of DNA elements.

- Complete comprehensive genetic libraries of *Drosophila melanogaster* strains with mapped deletions, insertions, or point mutations. Estimated costs: $100,000/yr (stocks. Assumes infrastructural support, mostly space, from host institutions), $250,000/yr (point mutation collections).
- Continue support for systematic RNAi in *Drosophila*. Estimated costs: $500,000/yr (funded until 2007).
- Re-sequencing of *Drosophila melanogaster* lines should be accelerated to make DNA sequence variants available (to be funded outside of D-ENCODE).
- Continue and expand support for distribution of vectors and clones for use as reporters and in functional tests in transgenic *Drosophila*. $100,000/yr.
- Deriving population based methods for inferring function based on allele frequencies: Exploratory work in this area will be important. Estimated costs: $250,000/yr.
- Continued support for systematic in situ hybridization of full-length cDNAs. Estimated costs: $250,000/yr.
- Directed cloning of predicted *cis*-regulatory elements and promoters to support systematic generation of transgenic reporter lines. Estimated cost: $500,000/yr.
- Moderate sampling of gene expression from all 11 non-*melanogaster* *Drosophila* species (e.g. ESTs, MPPS) (to be funded outside of D-ENCODE).

DNA Element Priorities:

1) **Protein-encoding genes.** Some protein-coding genes are still very difficult to identify in the absence of direct experimental evidence; for example, those with short open reading frames or nested within other genes. Additionally, transcriptome complexity due to alternatively spliced primary transcripts is difficult to determine from known sequence rules that regulate mRNA processing. Work to improve gene models will require cycles of computational and experimental work. Computational work will be greatly facilitated by comparative genomics. Cost-effective methods for generating biological evidence will include resources used for multiple projects (*e.g.* full-length cDNAs and high throughput sampling of the transcriptome via massively parallel signature sequencing, the proteome via tandem mass-spectrometry), directed sampling methods (*e.g.* systematic RT-PCR, in situ hybridization, PolII occupancy), and non-biased discovery (*e.g.* protein trapping).
    - Develop gene models for problematic genes. Estimated cost: $500,000/yr (2 R01, U01s).
    - Map alternative splice forms. Estimated cost: $1,000,000/yr (4 R01, U01s).
    - Map transcription start and stop sites. Estimated cost: $500,000/yr (2 R01, U01s).
    - Directed annotation of transposable elements in multiple *Drosophila* species. Estimated cost: $250,000/yr (1 R01, U01)

2) **Non-protein coding genes.** A growing body of evidence indicates that much more of the genome is transcribed than initially thought. Our abilities to identify non-coding RNAs, and to determine their functions are still rudimentary. Consequently both innovative computational and high-throughput experimental methods need to be developed or refined. Importantly, we do not known how much of this transcribed genomic 'dark matter' is in the form of genes that elude annotation due to short ORFs or RNA-based function and how much is due to transcriptional 'noise'. Much work on these genes will be generated in the course of protein-encoding gene model refinement.
    - Develop models for micro-RNAs and antisense transcripts. Estimated costs: $500,000/yr (2 R01, U01s).

- Create a map of all transcribed regions and focus on possible genetic function. Estimated costs: $750,000/yr (3 R01, U01s).
- Deep analysis of 10Mb from several species to examine conserved 'dark matter' expression. Estimated costs: $250,000/yr (1 R01, U01s).

3) **Structural and Regulatory Elements.** Transcriptional regulation in metazoans is complex and experiments aimed at deciphering some of this complexity are a clear goal of a D-ENCODE project. Major difficulties in reading the regulatory code stem from the fact that binding sites for transcriptional regulators are short and degenerate, hampering computational detection. *In vivo* mapping techniques such as ChIP-on-chip or DamID with whole-genome tiling arrays can help to direct analysis to occupied binding sites. When combined with evolutionary comparisons and careful DNA/protein biochemistry, this will lead to more reliable functional binding site identification. The identification and interpretation of functionally significant genomic variants is a main goal of human population genomics and will be essential for personalized medicine. D-ENCODE will provide a testing ground for functional analysis of polymorphic DNA elements. We have a poor genome-wide understanding of structural elements in the DNA sequence, such as origins of replication and scaffold attachment sites. Additionally, sequences may show emergent properties and will become evident as large blocks of structure in a full genome analysis. The discovery of additional novel and mysterious genome features (*e.g.* elements involved in organizing the genome in the 3-D interphase nucleus) is expected from a thorough functional annotation of the whole *Drosophila* genome.
    - Improve our understanding of sequence rules for transcription factor binding. Estimated costs: $500,000/yr (2 R01, U01s).
    - Map *in vivo* occupancy of binding sites for all known transcription factors and the basal transcriptional machinery. Estimated costs: $1,000,000/yr (4 R01, U01s).
    - Determine the functional targets of all transcription factors. Estimated costs: $1,000,000/yr (4 R01, U01s).

- Develop maps of polymorphic gene expression between *Drosophila melanogaster* lines to be used in assays for complex genetic traits. Estimated costs: $500,000/yr (2 R01, U01s).
- Map nuclear architecture onto the genome. Estimated costs: 500,000/yr (2 R01, U01s).
- Map replication origins in diploid and polytene cells. Estimated costs: $250,000/yr (1 R01, U01).

4) **Chromatin Structure**. Higher order chromatin structure is an important level of organization in eukaryotic genomes, and a critical facet of genome regulation. Epigenetic marks will need to be identified and analyzed functionally to fully annotate the *Drosophila* genome. Additionally, chromatin associated proteins (*e.g.* Polycomb and Trithorax groups) and modified histones (*e.g.* acetylation and methylation) are critical for understanding gene regulation. The sequence-based rules for chromatin structure are not well understood, making direct biological evidence critical.
    - Map DNA accessibility (hypersensitivity sites) to characterize 'open' and 'closed' chromatin regions as well as signatures of promoters, enhancers, chromatin boundaries, and origins of replication. Estimated costs: $500,000/yr (2 R01, U01s).
    - Map the entire histone code of nucleosome modifications and histone variants. Estimated costs: $500,000/yr (2 R01, U01s).
    - Map in vivo occupancy of known chromatin remodeling complexes. Estimated costs: $500,000/yr (2 R01, U01s).
    - Use affinity reagents and proteomics to identify all proteins closely associated with DNA. Estimated costs: $500,000/yr (2 R01, U01s).

Total Cost and Duration: $14,000,000 per year. D-ENCODE will be front loaded and expected to wind-down or be absorbed by ENCODE within 5 yrs. The priority lists for infrastructure and DNA elements are interdependent. Within those lists, numbered headings and bullet points are in order of priority.

Criteria for D-ENCODE Participation:

1) **Genomic Scope.** All D-ENCODE labs or groups of labs will study the entire genome and will focus on the first tier of genome features such as transcription, factor occupancy, chromatin status, and sequence conservation. While D-ENCODE data will be useful for understanding interactions and networks, these are beyond the scope of D-ENCODE. D-ENCODE data will be verified using independent techniques and validated by direct assays for genetic function. Genetic validation can be based on multiple lines of evidence from knockouts, knockdowns, and allele frequencies in populations. The major concentration is on cost effective data generation, rather than technical development.

2) **Coordination.** The core group is expected to be truly community minded and work unselfishly to ensure the success of D-ENCODE. As this will be a cooperative agreement (U01), these PIs are expected to work with major input from NHGRI. Satellite groups working on data generation are strongly encouraged to form meaningful, cross-institutional and inter-disciplinary collaborations (although innovative proposals from individual PIs are also encouraged). At least two groups should tackle each DNA element type to allow for cross verification of data (using SAGE-like and array-based assays for example). These groups will be in close contact with each other and with the core D-ENCODE group and must be comfortable with a congenial arrangement that will maximize startup speed and ultimately help transfer skills, knowledge, and social structure to human ENCODE.

3) **Data Sharing.** There is no doubt that immediate access to genomic sequence data has been a boon to biology, but there are concerns about ensuring that the deposited data is of high quality and that those producing the primary data receive appropriate credit. The human ENCODE pilot has established policies for data sharing and release [2]. D-ENCODE members will further liberalize the open distribution of verified experimental data prior to traditional publication, in the spirit of open-source software or large physical science consortiums. Data will be attributed to the discoverer with a citable 'e-publication' to credit those scientists that devote a substantial portion of their effort to

community projects. Attribution also assures that released data passes a 'comfort level' test of quality. The *Drosophila* ENCODE project will be a model for the organization of collaborative biological science.

1. Collins, F.S., E.D. Green, A.E. Guttmacher, and M.S. Guyer, *A vision for the future of genomics research*. Nature, 2003. **422**(6934): p. 835-47.
2. *The ENCODE Project: ENCyclopedia Of DNA Elements*. http://www.genome.gov/10005107
3. *The ENCODE (ENCyclopedia Of DNA Elements) Project*. Science, 2004. **306**(5696): p. 636-40.
4. *National Advisory Council for Human Genome Research*. http://www.genome.gov/11509849
5. *Drosophila D-ENCODE workshop participants*. http://rana.lbl.gov/drosophila/dencode.html
6. Celniker, S.E., D.A. Wheeler, B. Kronmiller, J.W. Carlson, A. Halpern, S. Patel, M. Adams, M. Champe, S.P. Dugan, E. Frise, A. Hodgson, R.A. George, R.A. Hoskins, T. Laverty, D.M. Muzny, C.R. Nelson, J.M. Pacleb, S. Park, B.D. Pfeiffer, S. Richards, E.J. Sodergren, R. Svirskas, P.E. Tabor, K. Wan, M. Stapleton, G.G. Sutton, C. Venter, G. Weinstock, S.E. Scherer, E.W. Myers, R.A. Gibbs, and G.M. Rubin, *Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence*. Genome Biol, 2002. **3**(12): p. RESEARCH0079.
7. Hoskins, R.A., C.D. Smith, J.W. Carlson, A.B. Carvalho, A. Halpern, J.S. Kaminker, C. Kennedy, C.J. Mungall, B.A. Sullivan, G.G. Sutton, J.C. Yasuhara, B.T. Wakimoto, E.W. Myers, S.E. Celniker, G.M. Rubin, and G.H. Karpen, *Heterochromatic sequences in a Drosophila whole-genome shotgun assembly*. Genome Biol, 2002. **3**(12): p. RESEARCH0085.
8. Graveley, B.R., *Alternative splicing: increasing diversity in the proteomic world*. Trends Genet, 2001. **17**(2): p. 100-7.
9. Fortini, M.E., M.P. Skupski, M.S. Boguski, and I.K. Hariharan, *A survey of human disease gene counterparts in the Drosophila genome*. J Cell Biol, 2000. **150**(2): p. F23-30.
10. Misra, S., M.A. Crosby, C.J. Mungall, B.B. Matthews, K.S. Campbell, P. Hradecky, Y. Huang, J.S. Kaminker, G.H. Millburn, S.E. Prochnik, C.D. Smith, J.L. Tupy, E.J. Whitfied, L. Bayraktaroglu, B.P. Berman, B.R. Bettencourt, S.E. Celniker, A.D. de Grey, R.A. Drysdale, N.L. Harris, J. Richter, S. Russo, A.J. Schroeder, S.Q. Shu, M. Stapleton, C. Yamada, M. Ashburner, W.M. Gelbart, G.M. Rubin, and S.E. Lewis, *Annotation of the Drosophila melanogaster*

euchromatic genome: a systematic review. Genome Biol, 2002. **3**(12): p. RESEARCH0083.
11. FlyBase, *The FlyBase database of the Drosophila genome projects and community literature*. Nucleic Acids Res, 2003. **31**(1): p. 172-5.
12. Miller, W., K.D. Makova, A. Nekrutenko, and R.C. Hardison, *Comparative genomics*. Annu Rev Genomics Hum Genet, 2004. **5**: p. 15-56.
13. *Status of sequencing proposals*. http://www.genome.gov/10002154
14. *Assembly/Alignment/Annotation of 12 Drosophila Genomes*. http://rana.lbl.gov/drosophila/multipleflies.html
15. *The Bloomington Drosophila Stock Center*. http://fly.bio.indiana.edu/
16. *Drosophila Genomics Resource Center*. http://dgrc.cgb.indiana.edu/
17. *Drosophila RNAi Screening Center*. http://flyrnai.org/
18. *Drosophila TILLING Project*. http://tilling.fhcrc.org:9366/fly/
19. Mann, M. and A. Pandey, *Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases*. Trends Biochem Sci, 2001. **26**(1): p. 54-61.
20. Mackay, T.F., *The genetic architecture of quantitative traits: lessons from Drosophila*. Curr Opin Genet Dev, 2004. **14**(3): p. 253-7.
21. Aquadro, C.F., V. Bauer DuMont, and F.A. Reed, *Genome-wide variation in the human and fruitfly: a comparison*. Curr Opin Genet Dev, 2001. **11**(6): p. 627-34.
22. *FlyBase, a database of the Drosophila genome*. http://flybase.bio.indiana.edu/