Drosophila melanogaster: A case study of a model genomic sequence and its consequences

Michael Ashburner and Casey M. Bergman

References	This article cites 103 articles, 48 of which can be accessed free at: **http://www.genome.org/cgi/content/full/15/12/1661#References** Article cited in: **http://www.genome.org/cgi/content/full/15/12/1661#otherarticles**
Email alerting service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here**

Notes

To subscribe to *Genome Research* go to:
http://www.genome.org/subscriptions/

Perspective

Drosophila melanogaster: A case study of a model genomic sequence and its consequences

Michael Ashburner[2] and Casey M. Bergman[1]

Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, United Kingdom

The sequencing and annotation of the *Drosophila melanogaster* genome, first published in 2000 through collaboration between Celera Genomics and the *Drosophila* Genome Projects, has provided a number of important contributions to genome research. By demonstrating the utility of methods such as whole-genome shotgun sequencing and genome annotation by a community "jamboree," the *Drosophila* genome established the precedents for the current paradigm used by most genome projects. Subsequent releases of the initial genome sequence have been improved by the Berkeley *Drosophila* Genome Project and annotated by FlyBase, the *Drosophila* community database, providing one of the highest-quality genome sequences and annotations for any organism. We discuss the impact of the growing number of genome sequences now available in the genus on current *Drosophila* research, and some of the biological questions that these resources will enable to be solved in the future.

It is almost 100 years since William Castle introduced *Drosophila melanogaster* to the pleasures and rigors of biological research (Castle 1906). Four major phases of *Drosophila* research can, perhaps, be distinguished. The period ~1910–1940, of classical genetic analysis was a period of rapid development in which most of the major principles of classical genetics were established: the chromosome theory of heredity, the nature of genetic linkage and genetic maps, the genetic behavior of chromosome aberrations, the induction of gene and chromosome mutations by radiation, the discovery of mitotic recombination, and so on. This was followed by a long period, ~1940–1968, of growth but relative sterility, a period in which many of the best minds in genetics turned their attention to microbes and phage. The period from, roughly, ~1968–2000 was a renaissance, witnessed by many molecular biologists moving into the field, creating an analytical, rather than descriptive, study of development and behavior. This metamorphosis was fueled by many major technical advances within the field, for example, the invention of in situ hybridization, of the *P*-element-based transformation technology, of powerful methods for clonal analysis, the discovery of potent chemical mutagens, and by the extraordinary external advances in molecular biology. New generations of researchers selected *Drosophila* as a model organism for the study of fundamental problems in biology. From 2000, fly research has matured into its fourth period: the genome era, for, on March 24, 2000 the first release of the "complete" genomic sequence of *Drosophila melanogaster* was published, timed to coincide with that year's annual fly meeting in Pittsburgh. Five years into the post-genomic era we can begin to ask: What have we learned and what may lie ahead?

The genome

Prior to 1998, two groups, the Berkley *Drosophila* Genome Project and the European *Drosophila* Genome Project, were beginning to sequence the genome of *D. melanogaster* by the tried and tested way of sequencing a minimal tiling path of clones (cosmids, P1 clones, and BACs) chosen from physical maps of the genome (Hartl et al. 1992; Madueno et al. 1995; Kimmerly et al. 1996; Hoskins et al. 2000). That changed on May 12, 1998, when Craig Venter invited Gerry Rubin to participate in an attempt to sequence the genome of this fly by whole-genome shotgun sequencing (WGS), a method untried and untested for anything larger than a bacterial genome of one or a few megabases in size. There was considerable skepticism in the community that WGS would succeed for a large and complex genome with much repetitive DNA (see Green's 1997 riposte to Weber and Myers' 1997 paper in this journal). It was a leap of faith that the combination of the new capillary sequencing machines, of very careful construction of clone libraries, and of software (then not yet written) would allow the 120-Mb euchromatic genome of *D. melanogaster* to be assembled. By September 1999 this faith had been justified: A WGS assembly of the euchromatic portion of the fly genome had been achieved. This proof-of-principle for a metazoan WGS was the first landmark contribution of the fly genome project.

At that time, only one metazoan genome, that of *Caenorhabditis elegans*, had been sequenced and annotated (The *C. elegans* Sequencing Consortium 1998). Experience in genome annotation, both as a technical problem and as a community problem, was sparse. The "annotation jamboree," hosted by Celera in November 1999, was important, not only for what it did, but for how it did it—an intimate and intensive collaboration of software engineers and of biologists drawn from the community working together both to build gene models and to annotate these with functional information using the then fledgling Gene Ontology (http://www.geneontology.org/) (Ashburner et al. 2000; Lewis 2005). This is a model that has been followed by many other communities. Also unusual (although not unique, vide *C. elegans*) to this day was the very close association between the sequencing and annotation groups and FlyBase (http://www.flybase.org/), the community database for Drosophilists. The act of analyzing the fly genome sequence, therefore, was novel in a second respect: It introduced new community methods for genome annotation and curation.

The "complete" sequence of the genome of *D. melanogaster* we have today is not that released in March 2000 (Adams et al. 2000; Myers et al. 2000; Rubin et al. 2000). Since Release 1, there

have been three subsequent genome releases (Celniker et al. 2002) (Release 4 was in April 2004; Release 5 is planned to be the final release of the genome sequence); each release is improved in quality, with the correction of errors, both of sequence and assembly, the closure of physical and sequence gaps (only 23 now remain in the Release 4 euchromatin), and the correct assembly of repetitive sequences (this has been the responsibility of the Berkeley *Drosophila* Genome Project; http://www.fruitfly.org/). In addition, a separate project, the *Drosophila* Heterochromatin Genome Project, has been funded to sequence the complex heterochromatic sequences of the telomeres and pericentromeric chromosome regions (http://www.dhgp.org/). FlyBase has been responsible for keeping the annotation of the genome up to date. A major effort by about 10 FlyBase annotators resulted in a complete revision of all gene models and other genome features, based on the first "finished" (Release 3) sequence; this was published in a series of papers in a special issue of *Genome Biology* in December 2002 (http://www.genomebiology.com/Drosophila). Revision of gene models and other features subsequently is an ongoing, reiterative, task being done by FlyBase (Drysdale et al. 2005).

Immediate lessons

Before November 1999 there had been decades of debate as to the number of protein-coding genes in *D. melanogaster*. That debate then stopped: it is ~14,000. Some, for example, Hild et al. (2003), have argued that the number of protein-coding genes had been seriously underestimated (perhaps by as many as 2000 protein-coding genes) by the original annotation. A careful experimental evaluation of these "missed" gene models shows few of them to be real; many are simply new exons of genes already known or predicted (see Yandell et al. 2005). Before December 2002, the abundance and diversity of the transposable elements in the genome of *D. melanogaster* was unknown: The first attempt at their annotation (Kaminker et al. 2002) gave numbers of 1572 elements in 93 families; a more recent analysis using improved methods and including additional families (such as the enigmatic *INE-1* element) (Locke et al. 1999), indicates that the Release 4 "euchromatin" (an operational definition for the assembled chromosome arms including the first few megabases of the pericentromeric heterochromatin) has 6013 elements in 127 families (Quesneville et al. 2005; http://dynagen.ijm.jussieu.fr/repet/dmel4/index.html).

Added benefits

In addition to revealing the parts list of the *Drosophila* genome, the completed sequence of *D. melanogaster* has changed the practice of *Drosophila* genetics and led to many unexpected discoveries. Having the genome has enormously accelerated—by a factor of at least 10—the time required to clone a particular gene of interest; this tedious task is no longer rate limiting or essential for biological discovery. The large, and growing, collection of inserted transposons used for gene disruption (mostly *P*-elements, but also *hobo*, *Minos*, and *piggyBac*) can now be mapped precisely to the genome sequence, rather than to a 50–100-kb interval by in situ hybridization to polytene chromosomes. About 65% of the genes of *D. melanogaster* have been disrupted by at least one transposon insertion (Bellen et al. 2004; Thibault et al. 2004; Venken and Bellen 2005). With advances in *P*-element technology, this has led to methods for the construction of deletions whose limits are known with base-pair accuracy, and to attempts

to cover the entire genome with a minimal tiling path of deletions (Parks et al. 2004; Ryder et al. 2004). Single nucleotide polymorphisms between the sequenced strain and others have led to the construction of several SNP maps, which enormously help the mapping of, for example, EMS-induced point mutations (Berger et al. 2001; Hoskins et al. 2001; Martin et al. 2001). The genome sequence has also greatly facilitated the recovery of EMS-induced mutations in selected gene regions using the method of tilling (Winkler et al. 2005).

The completion of the fly genome in 2000 coincided with great advances in genomic technology that have revolutionized our abilities to study transcription, protein binding to specific DNA sequences, and genetic variation at the molecular level. We can now make microarrays for expression profiling, either targeted to all known or predicted coding regions or against whole-genome tiling paths of high resolution (e.g., the INDAC resource; see http://www.indac.net/); we can now map the binding sites of chromatin-associated proteins to the genome at high resolution, using either DamID (Orian et al. 2003; Sun et al. 2003; Bianchi-Frias et al. 2004) or chromatin immunoprecipitation (chIP) (MacAlpine et al. 2004; Birch-Machin et al. 2005); we can now conduct genome-scale surveys for polymorphisms using high-throughput PCR strategies (Glinka et al. 2003; Orengo and Aguade 2004), and effectively re-sequence other genomes of the same species, using tiling paths of oligonucleotides (http://www.dpgp.org/). Genome resources have also revolutionized the genetic studies of complex traits in *Drosophila* (Pletcher et al. 2002; Harbison et al. 2005).

The task of obtaining one full-length cDNA from each fly gene is not only facilitated by the genomic sequence (Stapleton et al. 2002a,b), it helps enormously in refining gene models (Misra et al. 2002). We can now look forward to the day when each gene is represented by one full-length cDNA (in a versatile vector that will allow it to be shuttled to a variety of useful constructs) (http://www.fruitfly.org/EST/; S. Celniker, pers. comm.), and perhaps even to the availability of full-length cDNAs from every alternatively spliced transcript. Likewise, the genomic sequence has enabled the design of antisense RNA reagents that are now allowing large-scale, systematic RNAi screens for gene function in tissue culture cells (Bettencourt-Dias et al. 2004; Boutros et al. 2004).

The proper study of the genome is the genome itself. Quite unexpected properties of genomes have come from following this edict. Many individual examples of tandemly repeated genes had been known from work prior to the genome. But it was only the analysis of a 2.9-Mb trial sequence (Ashburner et al. 1999) and of the genome itself in 2000 that showed just how common this is, and the extent to which some protein families (e.g., of serine proteases) had expanded by duplication. Similarly, nested genes were first discovered in flies (Henikoff et al. 1986), but these were thought to be rare exceptions: They are not. More than 7% of the genes in the *D. melanogaster* genome are nested (Ashburner et al. 1999; Misra et. al 2002), and flies have at least a dozen examples of nests within nests. mRNAs that do not encode proteins also appear to be more common than previously thought (Tupy et al. 2005), an observation that may help to explain the phenomenon of "intergenic" transcription (Hild et al. 2003; Stolc et al. 2004).

The analysis of the genome of *D. melanogaster* has led to the insight that this genome is far more complex than we had imagined. In flies, as in other species (Cohen et al. 2000; Caron et al. 2001; Roy et al. 2002), the sequence has allowed us to observe

that the genome is organized into large gene-expression neighborhoods, within which even unrelated genes tend to be coexpressed (Boutanaev et al. 2002; Spellman and Rubin 2002; Parisi et al. 2004; Stolc et al. 2004; Belyakin et al. 2005; Thygesen and Zwinderman 2005). The existence of gene-expression neighborhoods suggests coadapted genomic regions that may be related to chromatin domains that may be preserved as syntenic regions during evolution (see below). The strongest evidence for gene-expression neighborhoods appears to come from genes expressed in the male testis (Boutanaev et al. 2002; Parisi et al. 2004), a genomic organization that may be necessary to facilitate proper gene expression during the final stages of sperm development in a highly condensed chromatin environment. The distribution of intergenic lengths in the compact *D. melanogaster* genome has also been shown to be nonrandom: Genes with complex regulation have long intergenic regions (Nelson et al. 2004). These observations suggest that relationships between genome structure and gene regulation are encoded in the fly genome sequence.

It is no coincidence that perhaps the greatest recent breakthrough in our understanding of gene regulation has come after the completion of genomic sequences of key eukaryotes like *Drosophila*: the discovery of the vast array of microRNAs (miRNAs) and their functions. In fact, the genome sequence of *D. melanogaster* helped reveal the fundamental hairpin structure of pre-miRNAs from mature miRNA expressed sequences (Lagos-Quintana et al. 2001). In turn, this detailed understanding of miRNA structure has allowed their genome-wide prediction in *Drosophila* (Lai et al. 2003). The genome sequence has also been critical for the prediction of miRNA targets (Enright et al. 2003; Stark et al. 2003; Rajewsky and Socci 2004; Brennecke et al. 2005; Burgler and Macdonald 2005).

There is still much to do

We hope not too many scientists will think that all the fun is over with *Drosophila*, and turn to the study of the *Trichoplax* or *Loxodonta* genomes. There remains much to discover, and many resources are now available to catalyze discovery by individual research groups (Matthews et al. 2005), who will remain the bedrock of the *Drosophila* community in the post-genomic era (Gilbert 1991). Large-scale projects to catalog functional elements in the genome sequence could be integrated and distributed through a *Drosophila* ENCODE project (http://rana.lbl.gov/drosophila/dencode.html), which would capitalize on the tradition of resource-sharing among Drosophilists, and serve as a model for community-driven, comprehensive genome annotation in higher eukaryotes.

The genomic sequences of a further 11 species of *Drosophila* (http://species.flybase.net/) will provide a rich source of data for expanding on lessons learned from the *D. melanogaster* genome. *Drosophila* genome sequences may, in fact, continue to push advances in WGS and comparative assembly techniques by providing in *D. melanogaster* a "finished" reference genome. The impact of finishing on genome assembly, annotation, and biological inference can now be evaluated to direct future strategies for genome sequencing projects (Myers et al. 2000; Benos et al. 2001; Celniker et al. 2002).

Heterochromatin has long been recognized as a major, yet mysterious, component of most metazoan genomes. We have already learned much about its molecular nature from studies with *Drosophila* (Dimitri et al. 2005). We know, for example, that much of the complex heterochromatin of *D. melanogaster* is composed of a graveyard of decaying, often nested, transposable elements with a sprinkling of protein-coding genes (Hoskins et al. 2002; Dimitri et al. 2003). We know that its chromatin differs in the spectrum of its proteins (Elgin and Grewal 2003). But, it would be an exaggeration for even the most zealous "heterochromatist" to claim that we have anything approaching a full understanding of either the structure or function of this important genome component. The completion of the sequence of *D. melanogaster* now requires the sequence of the complex heterochromatin (we except the 36 Mb of simple sequence satellite sequences) and new methods for its analysis.

Straightforward in principle, but demanding in practice, is the challenge to discover "functions" for all of the genes. The Gene Ontology has provided not only a structured language to describe gene "function," but also tools for the prediction of gene function. Yet no scientist should be satisfied for long with only predicted function. Of the 14,461 predicted protein-coding genes of *D. melanogaster*, only 5402 have known mutant alleles; on the other hand, there are 9875 genes in *D. melanogaster* whose existence is reasonably well attested by classical methods but that have yet to be identified on the sequence (data computed from FlyBase) (A. de Grey, pers. comm.). Linking the wealth of results published in the literature to the genome is absolutely necessary if we are to leverage the depth of our understanding of development, behavior, and evolution in *Drosophila* using the genome sequences. Continued progress toward completion of the gene disruption projects and expression profiling (see above) will prove essential for finding functions for the remaining as-yet-uncharacterized genes.

Progress, both experimental and computational, in the understanding of regulatory networks in *Drosophila* is dramatic: Indeed, it can be argued that the regulation of A-P and D-V axes formation in early fly development is one of the best (if not the best) understood complex biological system (http://bdtnp.lbl.gov/). The syncytial embryonic environment is also optimal for the decoding of networks based on transcriptional control. Indeed, these networks can even be emulated in vitro (Isalan et al. 2005). Yet, from experimental analyses, we understand in any detail the structure of the regulatory regions of relatively few genes, and the annotation of even this limited set is regrettably incomplete (http://www.flybase.org/annot/dmel_release4.1.txt). Nevertheless, the ability to reconstruct core features of the *Drosophila* segmentation network automatically in silico from annotated regulatory sequences (Fig. 1) suggests that a complete genomic inventory of regulatory elements will have direct impact on gene regulatory network analysis in flies. The genomic sequence will enable great advances here, as computational methods for the prediction of regulatory regions (Berman et al. 2002; Ohler et al. 2002; Rajewsky et al. 2002; Ochoa-Espinosa et al. 2005), particularly using comparative data (Bergman et al. 2002; Berman et al. 2004; Grad et al. 2004; Sinha et al. 2004), improve and as the size of the database of functionally characterized *cis*-regulatory sequences mapped to the genome increases (Lifanov et al. 2003; Bergman et al. 2005). The development of a genome tile microarray in *Drosophila* will also be essential for experimentally enumerating the targets and binding specificities of the vast majority of the >700 predicted transcription factors in the fly genome, for which no regulatory information is currently available. Integrating results of gene networks inferred from multiple genomic and proteomic (e.g., two-hybrid screens) (Giot et al. 2003; Stanyon et al. 2004; Formstecher et al. 2005) data sources

A

B
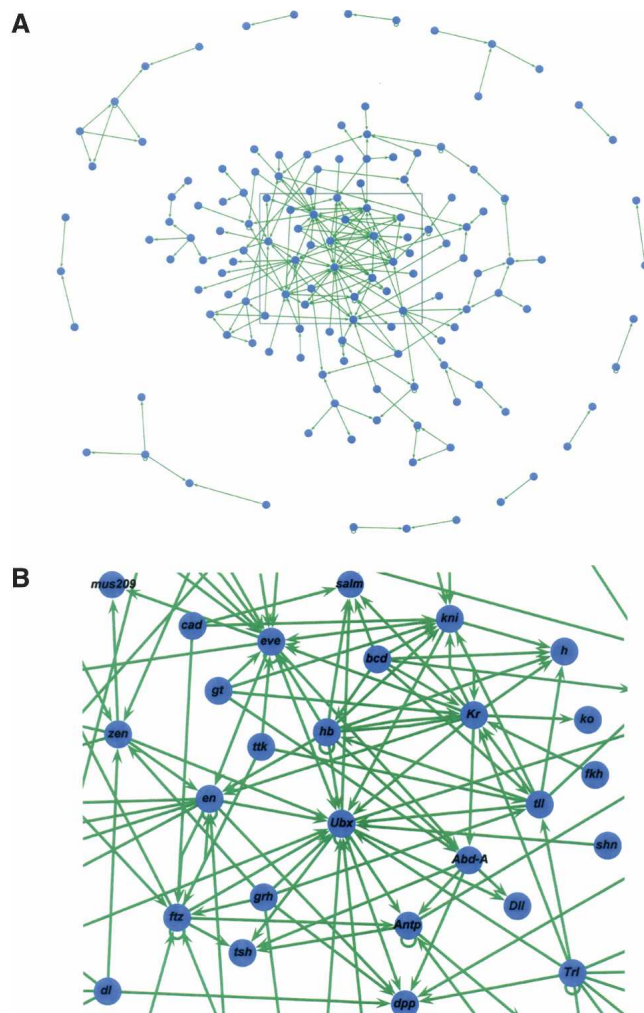
Figure 1. Partial gene regulatory network (GRN) for *Drosophila mela-nogaster* automatically generated from genome annotations. (*A*) GRN determined from entire set of protein–DNA interactions in the *Drosophila* DNase I footprint database (http://www.flyreg.org/) using Cytoscape (Shannon et al. 2003). Note that the majority of currently annotated protein–DNA interactions coalesce into a single interconnected network. (*B*) Sector of GRN shown in panel *A* demonstrating that known transcriptional regulatory interactions can be computed directly from text-based annotation of genomic DNA.

will hopefully expand and link together functional regulatory modules into coherent systems that specify fly biology.

We doubt that the discovery of gene-expression neighborhoods is the last surprise for our understanding of genome structure at a large scale. Here, we believe that comparative data will have much to say. It was about 70 years ago that Sturtevant and Dobzhansky discovered that overlapping inversions can be used for phylogenetic reconstruction (Sturtevant and Dobzhansky 1936), a fact most remarkably used to study the relationships for the >120 species of endemic Hawaiian species (Carson et al. 1992). Now, the high rate of genome restructuring by inversions in *Drosophila* (Ranz et al. 2001) can be used to reveal groups of genes that maintain positional order, perhaps for functional reasons such as coexpression. The relationship between syntenic regions and gene expression neighborhoods is only beginning to emerge (Stolc et al. 2004); however, the genomic material to re-

veal the scope of functional constraints on chromosome restructuring is now available. The rigorous definition of syntenic regions and the development of software to achieve this end are an important proximate goal, but the vagaries of incomplete genome sequences, overlapping inversions, and the possibility of breakpoint re-use will present challenges to reconstructing the complete history of inversions in *Drosophila*.

One of the great lessons of the post-genomic era is the added value of comparative sequence data for the functional annotation of model systems such as *Drosophila*. The second genome in the genus, that of *Drosophila pseudoobscura*, was published in January 2005 (Richards et al. 2005) and represents the first stage in the explosion of comparative genomics data being generated currently in flies (http://rana.lbl.gov/drosophila/multipleflies.html), an inevitability heralded by the *D. melanogaster* WGS. As with interpreting all biological systems, the *D. melanogaster* genome must be viewed as the product of evolution; thus, the decoding of functional information in this model genome will be intimately intertwined with knowing the evolutionary history and forces that have produced it. Much, however, remains to be learned about the mechanisms of genome evolution in *Drosophila*: the relative contributions of gene transposition and inversions to gene movement (Ranz et al. 2003b); the gain of lineage-specific genes by retrotransposition and tandem duplication (Betran et al. 2002; Wang et al. 2004); the divergence and proliferation of transposable elements (Kaminker et al. 2002; Lerat et al. 2003; Sanchez-Gracia et al. 2005); the evolution of *cis*-regulatory sequences (Emberly et al. 2003; Phinchongsakuldit et al. 2004; Ludwig et al. 2005; Negre et al. 2005; Sinha and Siggia 2005); the relationship between *cis*-regulatory and transcriptome evolution (Ranz et al. 2003a; Rifkin et al. 2003); and the resolution of the roles that mutation, recombination, genetic drift, population history, and natural selection have jointly played in shaping the genomic landscape we observe today (Andolfatto 2001; Aquadro et al. 2001; Glinka et al. 2003; Orengo and Aguade 2004; Haddrill et al. 2005). By identifying both highly conserved and positively selected sequences to predict and inform function, the integration of evolutionary genomics with classical and forward genetics will continue to propel biological discovery long into the second century of fly research.

Acknowledgments

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Andolfatto, P. 2001. Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Dev.* **11:** 635–641.

Aquadro, C.F., Bauer DuMont, V., and Reed, F.A. 2001. Genome-wide variation in the human and fruitfly: A comparison. *Curr. Opin. Genet. Dev.* **11:** 627–634.

Ashburner, M., Misra, S., Roote, J., Lewis, S.E., Blazej, R., Davis, T.,

Doyle, C., Galle, R., George, R., Harris, N., et al. 1999. An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: The *Adh* region. *Genetics* **153:** 179–219.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene Ontology: Tool for the unification of biology. *Nat. Genet.* **25:** 25–29.

Bellen, H.J., Levis, R.W., Liao, G., He, Y., Carlson, J.W., Tsang, G., Evans-Holm, M., Hiesinger, P.R., Schulze, K.L., Rubin, G.M., et al. 2004. The BDGP gene disruption project: Single transposon insertions associated with 40% of *Drosophila* genes. *Genetics* **167:** 761–781.

Belyakin, S.N., Christophides, G.K., Alekseyenko, A.A., Kriventseva, E.V., Belyaeva, E.S., Nanayev, R.A., Makunin, I.V., Kafatos, F.C., and Zhimulev, I.F. 2005. Genomic analysis of *Drosophila* chromosome underreplication reveals a link between replication control and transcriptional territories. *Proc. Natl. Acad. Sci.* **102:** 8269–8274.

Benos, P.V., Gatt, M.K., Murphy, L., Harris, D., Barrell, B., Ferraz, C., Vidal, S., Brun, C., Demaille, J., Cadieu, E., et al. 2001. From first base: The sequence of the tip of the X chromosome of *Drosophila melanogaster*, a comparison of two sequencing strategies. *Genome Res.* **11:** 710–730.

Berger, J., Suzuki, T., Senti, K.A., Stubbs, J., Schaffner, G., and Dickson, B.J. 2001. Genetic mapping with SNP markers in *Drosophila*. *Nat. Genet.* **29:** 475–481.

Bergman, C.M., Pfeiffer, B.D., Rincon-Limas, D.E., Hoskins, R.A., Gnirke, A., Mungall, C.J., Wang, A.M., Kronmiller, B., Pacleb, J., Park, S., et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3:** research0086.

Bergman, C.M., Carlson, J.W., and Celniker, S.E. 2005. *Drosophila* DNase I footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21:** 1747–1749.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci.* **99:** 757–762.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. 2004. Computational identification of developmental enhancers: Conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol.* **5:** R61.

Betran, E., Thornton, K., and Long, M. 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res.* **12:** 1854–1859.

Bettencourt-Dias, M., Giet, R., Sinka, R., Mazumdar, A., Lock, W.G., Balloux, F., Zafiropoulos, P.J., Yamaguchi, S., Winter, S., Carthew, R.W., et al. 2004. Genome-wide survey of protein kinases required for cell cycle progression. *Nature* **432:** 980–987.

Bianchi-Frias, D., Orian, A., Delrow, J.J., Vazquez, J., Rosales-Nieves, A.E., and Parkhurst, S.M. 2004. Hairy transcriptional repression targets and cofactor recruitment in *Drosophila*. *PLoS Biol.* **2:** E178.

Birch-Machin, I., Gao, S., Huen, D., McGirr, R., White, R.A., and Russell, S. 2005. Genomic analysis of heat-shock factor targets in *Drosophila*. *Genome Biol.* **6:** R63.

Boutanaev, A.M., Kalmykova, A.I., Shevelyov, Y.Y., and Nurminsky, D.I. 2002. Large clusters of co-expressed genes in the *Drosophila* genome. *Nature* **420:** 666–669.

Boutros, M., Kiger, A.A., Armknecht, S., Kerr, K., Hild, M., Koch, B., Haas, S.A., Heidelberg Fly Array Consortium, Paro, R., and Perrimon, N. 2004. Genome-wide RNAi analysis of growth and viability in *Drosophila* cells. *Science* **303:** 832–835.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. 2005. Principles of microRNA-target recognition. *PLoS Biol.* **3:** e85.

Burgler, C. and Macdonald, P.M. 2005. Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomics* **6:** 88.

Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., et al. 2001. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291:** 1289–1292.

Carson, H.L., Tonzetich, J., and Dopescher, L.T. 1992. Polytene chromosome maps for Hawaiian *Drosophila*. In *Drosophila inversion polymorphism*. (eds. C.B. Krimbas and J.R. Powell), pp. 441–453. CRC Press, Boca Raton, FL.

Castle, W.E. 1906. Inbreeding, cross-breeding and sterility in *Drosophila*. *Science* **23:** 153.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2018.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E., et al. 2002. Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3:** research0079.

Cohen, B.A., Mitra, R.D., Hughes, J.D., and Church, G.M. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **26:** 183–186.

Dimitri, P., Junakovic, N., and Arca, B. 2003. Colonization of heterochromatic genes by transposable elements in *Drosophila*. *Mol. Biol. Evol.* **20:** 503–512.

Dimitri, P., Corradini, N., Rossi, F., and Verni, F. 2005. The paradox of functional heterochromatin. *Bioessays* **27:** 29–41.

Drysdale, R.A., Crosby, M.A., Gelbart, W., Campbell, K., Emmert, D., Matthews, B., Russo, S., Schroeder, A., Smutniak, F., Zhang, P., et al. 2005. FlyBase: Genes and gene models. *Nucleic Acids Res.* **33:** D390–D395.

Elgin, S.C. and Grewal, S.I. 2003. Heterochromatin: Silence is golden. *Curr. Biol.* **13:** R895–R898.

Emberly, E., Rajewsky, N., and Siggia, E.D. 2003. Conservation of regulatory elements between two species of *Drosophila*. *BMC Bioinformatics* **4:** 57.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. 2003. MicroRNA targets in *Drosophila*. *Genome Biol.* **5:** R1.

Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C., et al. 2005. Protein interaction mapping: A *Drosophila* case study. *Genome Res.* **15:** 376–384.

Gilbert, W. 1991. Towards a paradigm shift in biology. *Nature* **349:** 99.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302:** 1727–1736.

Glinka, S., Ometto, L., Mousset, S., Stephan, W., and De Lorenzo, D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: A multi-locus approach. *Genetics* **165:** 1269–1278.

Grad, Y.H., Roth, F.P., Halfon, M.S., and Church, G.M. 2004. Prediction of similarly-acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *D. melanogaster* and *D. pseudoobscura*. *Bioinformatics* **20:** 2738–2750.

Green, P. 1997. Against a whole-genome shotgun. *Genome Res.* **7:** 410–417.

Haddrill, P.R., Thornton, K.R., Charlesworth, B., and Andolfatto, P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* **15:** 790–799.

Harbison, S.T., Chang, S., Kamdar, K.P., and Mackay, T.F. 2005. Quantitative genomics of starvation stress resistance in *Drosophila*. *Genome Biol.* **6:** R36.

Hartl, D.L., Ajioka, J.W., Cai, H., Lohe, A.R., Lozovskaya, E.R., Smoller, D.A., and Duncan, I.W. 1992. Towards a *Drosophila* genome map. *Trends Genet.* **8:** 70–75.

Henikoff, S., Keene, M.A., Fechtel, K., and Fristrom, J.W. 1986. Gene within a gene: Nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. *Cell* **44:** 33–42.

Hild, M., Beckmann, B., Haas, S.A., Koch, B., Solovyev, V., Busold, C., Fellenberg, K., Boutros, M., Vingron, M., Sauer, F., et al. 2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.* **5:** R3.

Hoskins, R.A., Nelson, C.R., Berman, B.P., Laverty, T.R., George, R.A., Ciesiolka, L., Naeemuddin, M., Arenson, A.D., Durbin, J., David, R.G., et al. 2000. A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. *Science* **287:** 2271–2274.

Hoskins, R.A., Phan, A.C., Naeemuddin, M., Mapa, F.A., Ruddy, D.A., Ryan, J.J., Young, L.M., Wells, T., Kopczynski, C., and Ellis, M.C. 2001. Single nucleotide polymorphism markers for genetic mapping in *Drosophila melanogaster*. *Genome Res.* **11:** 1100–1113.

Hoskins, R.A., Smith, C.D., Carlson, J.W., Carvalho, A.B., Halpern, A., Kaminker, J.S., Kennedy, C., Mungall, C.J., Sullivan, B.A., Sutton, G.G., et al. 2002. Heterochromatic sequences in a *Drosophila* whole-genome shotgun assembly. *Genome Biol.* **3:** research0085.

Isalan, M., Lemerle, C., and Serrano, L. 2005. Engineering gene networks to emulate *Drosophila* embryonic pattern formation. *PLoS Biol.* **3:** e64.

Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D.A., Lewis, S.E., Rubin, G.M., et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: A genomics perspective. *Genome Biol.* **3:** research0084.

Kimmerly, W., Stultz, K., Lewis, S., Lewis, K., Lustre, V., Romero, R., Benke, J., Sun, D., Shirley, G., Martin, C., et al. 1996. A P1-based physical map of the *Drosophila* euchromatic genome. *Genome Res.* **6:** 414–430.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853–858.

Lai, E.C., Tomancak, P., Williams, R.W., and Rubin, G.M. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4:** R42.

Lerat, E., Rizzon, C., and Biemont, C. 2003. Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. *Genome Res.* **13:** 1889–1896.

Lewis, S.E. 2005. Gene Ontology: Looking backwards and forwards. *Genome Biol.* **6:** 103.

Lifanov, A.P., Makeev, V.J., Nazina, A.G., and Papatsenko, D.A. 2003. Homotypic regulatory clusters in *Drosophila*. *Genome Res.* **13:** 579–588.

Locke, J., Howard, L.T., Aippersbach, N., Podemski, L., and Hodgetts, R.B. 1999. The characterization of *DINE-1*, a short, interspersed repetitive element present on chromosome and in the centric heterochromatin of *Drosophila melanogaster*. *Chromosoma* **108:** 356–366.

Ludwig, M.Z., Palsson, A., Alekseeva, E., Bergman, C.M., Nathan, J., and Kreitman, M. 2005. Functional evolution of a *cis*-regulatory module. *PLoS Biol.* **3:** e93.

MacAlpine, D.M., Rodriguez, H.K., and Bell, S.P. 2004. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes & Dev.* **18:** 3094–3105.

Madueno, E., Papagiannakis, G., Rimmington, G., Saunders, R.D.C., Savakis, C., Siden-Kiamos, I., Skavdis, G., Spanos, L., Trenear, J., Adam, P., et al. 1995. A physical map of the X chromosome of *Drosophila melanogaster*: Cosmid contigs and sequence tagged sites. *Genetics* **139:** 1631–1647.

Martin, S.G., Dobi, K.C., and St Johnston, D. 2001. A rapid method to map mutations in *Drosophila*. *Genome Biol.* **2:** research0036.

Matthews, K.A., Kaufman, T.C., and Gelbart, W.M. 2005. Research resources for *Drosophila*: The expanding universe. *Nat. Rev. Genet.* **6:** 179–193.

Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E., et al. 2002. Annotation of the *Drosophila melanogaster* euchromatic genome: A systematic review. *Genome Biol.* **3:** research0083.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., et al. 2000. A whole-genome assembly of *Drosophila*. *Science* **287:** 2196–2204.

Negre, B., Casillas, S., Suzanne, M., Sanchez-Herrero, E., Akam, M., Nefedov, M., Barbadilla, A., De Jong, P.J., and Ruiz, A. 2005. Conservation of regulatory sequences and gene expression patterns in the disintegrating *Drosophila* Hox gene complex. *Genome Res.* **15:** 692–700.

Nelson, C.E., Hersh, B.M., and Carroll, S.B. 2004. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5:** R25.

Ochoa-Espinosa, A., Yucel, G., Kaplan, L., Pare, A., Pura, N., Oberstein, A., Papatsenko, D., and Small, S. 2005. The role of binding site cluster strength in Bicoid-dependent patterning in *Drosophila*. *Proc. Natl. Acad. Sci.* **102:** 4960–4965.

Ohler, U., Liao, G.C., Niemann, H., and Rubin, G.M. 2002. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3:** research0087.

Orengo, D.J. and Aguade, M. 2004. Detecting the footprint of positive selection in a European population of *Drosophila melanogaster*: Multilocus pattern of variation and distance to coding regions. *Genetics* **167:** 1759–1766.

Orian, A., van Steensel, B., Delrow, J., Bussemaker, H.J., Li, L., Sawado, T., Williams, E., Loo, L.W., Cowley, S.M., Yost, C., et al. 2003. Genomic binding by the *Drosophila* Myc, Max, Mad/Mnt transcription factor network. *Genes & Dev.* **17:** 1101–1114.

Parisi, M., Nuttall, R., Edwards, P., Minor, J., Naiman, D., Lu, J., Doctolero, M., Vainer, M., Chan, C., Malley, J., et al. 2004. A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. *Genome Biol.* **5:** R40.

Parks, A.L., Cook, K.R., Belvin, M., Dompe, N.A., Fawcett, R., Huppert, K., Tan, L.R., Winter, C.G., Bogart, K.P., Deal, J.E., et al. 2004. Systematic generation of high-resolution deletion coverage of the *Drosophila melanogaster* genome. *Nat. Genet.* **36:** 288–292.

Phinchongsakuldit, J., MacArthur, S., and Brookfield, J.F. 2004.

Evolution of developmental genes: Molecular microevolution of enhancer sequences at the *Ubx* locus in *Drosophila* and its impact on developmental phenotypes. *Mol. Biol. Evol.* **21:** 348–363.

Pletcher, S.D., Macdonald, S.J., Marguerie, R., Certa, U., Stearns, S.C., Goldstein, D.B., and Partridge, L. 2002. Genome-wide transcript profiles in aging and calorically restricted *Drosophila melanogaster*. *Curr. Biol.* **12:** 712–723.

Quesneville, H., Bergman, C.M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. 2005. Combined evidence annotation of transposable elements in genome sequences. *PLoS Comp. Biol.* **1:** 166–165.

Rajewsky, N. and Socci, N.D. 2004. Computational identification of microRNA targets. *Dev. Biol.* **267:** 529–535.

Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E.D. 2002. Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics* **3:** 30.

Ranz, J.M., Casals, F., and Ruiz, A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* **11:** 230–239.

Ranz, J.M., Castillo-Davis, C.I., Meiklejohn, C.D., and Hartl, D.L. 2003a. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* **300:** 1742–1745.

Ranz, J.M., Gonzalez, J., Casals, F., and Ruiz, A. 2003b. Low occurrence of gene transposition events during the evolution of the genus *Drosophila*. *Evol. Int. J. Org. Evol.* **57:** 1325–1335.

Richards, S., Liu, Y., Bettencourt, B.R., Hradecky, P., Letovsky, S., Nielsen, R., Thornton, K., Hubisz, M.J., Chen, R., Meisel, R.P., et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15:** 1–18.

Rifkin, S.A., Kim, J., and White, K.P. 2003. Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* **33:** 138–144.

Roy, P.J., Stuart, J.M., Lund, J., and Kim, S.K. 2002. Chromosomal clustering of muscle-expressed genes in *Caenorhabditis elegans*. *Nature* **418:** 975–979.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., et al. 2000. Comparative genomics of the eukaryotes. *Science* **287:** 2204–2215.

Ryder, E., Blows, F., Ashburner, M., Bautista-Llacer, R., Coulson, D., Drummond, J., Webster, J., Gubb, D., Gunton, N., Johnson, G., et al. 2004. The DrosDel collection: A set of *P*-element insertions for generating custom chromosomal aberrations in *Drosophila melanogaster*. *Genetics* **167:** 797–813.

Sanchez-Gracia, A., Maside, X., and Charlesworth, B. 2005. High rate of horizontal transfer of transposable elements in *Drosophila*. *Trends Genet.* **21:** 200–203.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13:** 2498–2504.

Sinha, S. and Siggia, E.D. 2005. Sequence turnover and tandem repeats in *cis*-regulatory modules in *Drosophila*. *Mol. Biol. Evol.* **22:** 874–885.

Sinha, S., Schroeder, M.D., Unnerstall, U., Gaul, U., and Siggia, E.D. 2004. Cross-species comparison significantly improves genome-wide prediction of *cis*-regulatory modules in *Drosophila*. *BMC Bioinformatics* **5:** 129.

Spellman, P.T. and Rubin, G.M. 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1:** 5.

Stanyon, C.A., Liu, G., Mangiola, B.A., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J., and Finley Jr., R.L. 2004. A *Drosophila* protein-interaction map centered on cell-cycle regulators. *Genome Biol.* **5:** R96.

Stapleton, M., Carlson, J., Brokstein, P., Yu, C., Champe, M., George, R., Guarin, H., Kronmiller, B., Pacleb, J., Park, S., et al. 2002a. A *Drosophila* full-length cDNA resource. *Genome Biol.* **3:** research0080.

Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., et al. 2002b. The *Drosophila* Gene Collection: Identification of putative full-length cDNAs for 70% of *D. melanogaster* genes. *Genome Res.* **12:** 1294–1300.

Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. 2003. Identification of *Drosophila* MicroRNA targets. *PLoS Biol.* **1:** E60.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E., et al. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306:** 655–660.

Sturtevant, A.H. and Dobzhansky, T. 1936. Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use

in the study of the history of the species. *Proc. Natl. Acad. Sci.* **22:** 448–450.

Sun, L.V., Chen, L., Greil, F., Negre, N., Li, T.R., Cavalli, G., Zhao, H., Van Steensel, B., and White, K.P. 2003. Protein–DNA interaction mapping using genomic tiling path microarrays in *Drosophila*. *Proc. Natl. Acad. Sci.* **100:** 9428–9433.

Thibault, S.T., Singer, M.A., Miyazaki, W.Y., Milash, B., Dompe, N.A., Singh, C.M., Buchholz, R., Demsky, M., Fawcett, R., Francis-Lang, H.L., et al. 2004. A complementary transposon tool kit for *Drosophila melanogaster* using *P* and *piggyBac*. *Nat. Genet.* **36:** 283–287.

Thygesen, H.H. and Zwinderman, A.H. 2005. Modelling the correlation between the activities of adjacent genes in *Drosophila*. *BMC Bioinformatics* **6:** 10.

Tupy, J.L., Bailey, A.M., Dailey, D., Evans-Holm, M., Siebel, C.W., Misra, S., Celniker, S.E., and Rubin, G.M. 2005. Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **102:** 5495–5500.

Venken, K.J. and Bellen, H.J. 2005. Emerging technologies for gene manipulation in *Drosophila melanogaster*. *Nat. Rev. Genet.* **6:** 167–178.

Wang, W., Yu, H., and Long, M. 2004. Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species. *Nat. Genet.* **36:** 523–527.

Weber, J.L. and Myers, E.W. 1997. Human whole-genome shotgun sequencing. *Genome Res.* **7:** 401–409.

Winkler, S., Schwabedissen, A., Backasch, D., Bokel, C., Seidel, C., Bonisch, S., Furthauer, M., Kuhrs, A., Cobreros, L., Brand, M., et al. 2005. Target-selected mutant screen by TILLING in *Drosophila*.

Genome Res. **15:** 718–723.

Yandell, M., Bailey, A.M., Misra, S., Shu, S., Wiel, C., Evans-Holm, M., Celniker, S.E., and Rubin, G.M. 2005. A computational and experimental approach to validating annotations and gene predictions in the *Drosophila melanogaster* genome. *Proc. Natl. Acad. Sci.* **102:** 1566–1571.

Web site references

http://bdtnp.lbl.gov/; Berkeley *Drosophila* Transcriptional Network Project.

http://rana.lbl.gov/drosophila/dencode.html; D-encode workshop.

http://rana.lbl.gov/drosophila/multipleflies.html/; Assembly/Annotation/Alignment of 12 *Drosophila* genomes.

http://species.flybase.net/; *Drosophila* species genome Web page.

http://www.dhgp.org/; *Drosophila* Heterochromatin Genome Project.

http://www.dpgp.org/; *Drosophila* Population Genomics Project.

http://www.flybase.org/; FlyBase.

http://www.flyreg.org/; *Drosophila* DNase I footprint Database.

http://www.fruitfly.org/; Berkeley *Drosophila* Genome Project.

http://www.geneontology.org/; Gene Ontology Consortium.

http://www.genomebiology.com/Drosophila/; Release 3 *Drosophila* genome papers.

http://www.indac.net/; International *Drosophila* array consortium.