Annu. Rev. Genomics Hum. Genet. 2003. 4:89–117
doi: 10.1146/annurev.genom.4.070802.110323

First published online as a Review in Advance on June 4, 2003

THE DROSOPHILA MELANOGASTER GENOME

Susan E. Celniker[1] and Gerald M. Rubin[1,2]

[1]Berkeley Drosophila Genome Project, Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, California 94720; email: celniker@bdgp.lbl.gov

[2]Howard Hughes Medical Institute, Department of Molecular and Cellular Biology, University of California, Berkeley, California 94720; email: gerry@fruitfly.org

Key Words cDNAs, gene annotation, gene disruption, comparative genomics, gene expression

■ **Abstract** Drosophila's importance as a model organism made it an obvious choice to be among the first genomes sequenced, and the Release 1 sequence of the euchromatic portion of the genome was published in March 2000. This accomplishment demonstrated that a whole genome shotgun (WGS) strategy could produce a reliable metazoan genome sequence. Despite the attention to sequencing methods, the nucleotide sequence is just the starting point for genome-wide analyses; at a minimum, the genome sequence must be interpreted using expressed sequence tag (EST) and complementary DNA (cDNA) evidence and computational tools to identify genes and predict the structures of their RNA and protein products. The functions of these products and the manner in which their expression and activities are controlled must then be assessed—a much more challenging task with no clear endpoint that requires a wide variety of experimental and computational methods. We first review the current state of the *Drosophila melanogaster* genome sequence and its structural annotation and then briefly summarize some promising approaches that are being taken to achieve an initial functional annotation.

INTRODUCTION

As originally conceived, the role of model organisms in the Genome Project was to test methods of large-scale sequence determination and to develop high-throughput analyses of the resulting sequence data. This is an opportune time to assess the role the Drosophila Genome Project has played in meeting these goals and to review the current state of genomic analysis in this classic model organism. The euchromatic portion of the Drosophila genome now exceeds the community quality standards for finished sequence, although we are only in the early stages of determining the sequence and structure of the centric heterochromatin—a difficult task that has not been accomplished for any genome. With the aide of extensive expressed sequence tag (EST) and complementary DNA (cDNA) sequence data, we have a reasonably accurate estimate of the gene number and overall gene structure of most protein-coding genes in the genome. We are less certain of the extent

of alternative splicing or the number and types of nonprotein-coding transcripts. Moreover, the surprising complexity of gene organization revealed by these early analyses makes it clear that the structural annotation of metazoan genomes will depend greatly on experimental data for some time. The goals of structural annotation can be clearly defined and a wealth of comparative sequence data is becoming available to facilitate this task. Functional annotation—determining the biological roles of encoded transcripts and proteins and how their expression and activities are regulated—is a more diverse, open-ended undertaking that will involve more members of the research community than the determination and annotation of the DNA sequence. Although much of this work will be done one or a few genes at a time, some can be accomplished by systematic, genomic-wide approaches.

GENOME SEQUENCE

The chromosomes of *Drosophila melanogaster* are shown in Figure 1. The 180-megabase (Mb) genome is roughly two-thirds euchromatic and one-third heterochromatic. The euchromatin contains about 98% of the protein-coding genes in the genome; the heterochromatin is largely composed of simple sequence repeats. This satellite DNA cannot be stably cloned, limiting analysis of the heterochromatin by standard genomic methods. For this reason we discuss the euchromatin and heterochromatin separately.

Euchromatin

Celera Genomics and the Berkeley Drosophila Genome Project (BDGP) collaboratively performed a whole genome shotgun (WGS) assembly and generated a sequence of the euchromatin (1, 116) superseding a traditional P1- and bacterial artificial chromosome (BAC)-based sequencing effort that produced 29 Mb of sequence (approximately 25%) by the end of 1999. The first assembly (*WGS1*) used only plasmid and BAC-paired end sequences, and the second added BAC- and P1-based finished and draft sequences [see table 3 of (1) for details]. This joint assembly was submitted to GenBank as Release 1. This sequence contained many gaps and regions of low sequence quality.

In October 2000, Release 2 corrected errors in the order and orientation of small scaffolds present in Release 1 and filled a few hundred small sequence gaps. Celniker et al. (26) evaluated the quality of the Release 2 assembly by comparing it to the Release 3 high quality euchromatic sequence (see below), attributing any differences between them to sequence errors in the WGS assembly; the Release 2 sequence represents over 97.7% of the finished sequence. In autosomal regions of unique sequence, the error rate of Release 2 was one in 20,000 base pairs (bp). However, the error rate on the X chromosome was higher due to lower sequence coverage, and repetitive sequences in this assembly were only draft quality.

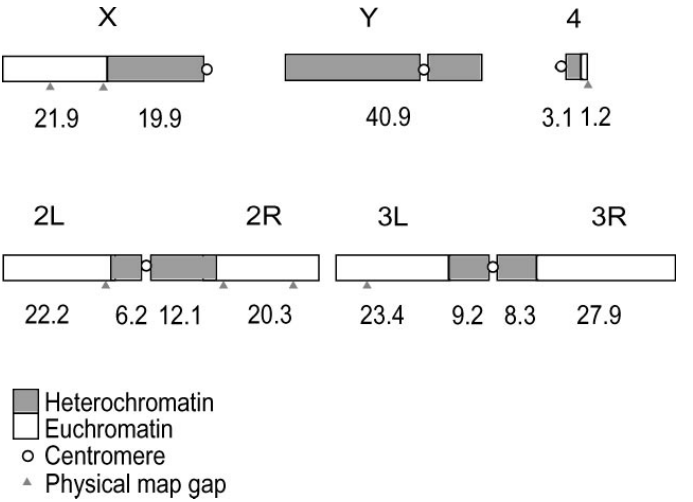


Figure 1 Chromosome structure of *Drosophila melanogaster*. The sex chromosomes X and Y, the small chromosome 4, and the left and right arms of chromosomes 2 (2L, 2R) and 3 (3L, 3R) are shown [adapted from (64)]. The numbers given below the chromosomes correspond to their lengths in megabases (Mb). The euchromatic portions of the chromosome arms (*white*) correspond to the Release 3 euchromatic sequence described in Celniker et al. (26). The lengths of the heterochromatic portions of the chromosome arms (*gray*) are estimated from measurements of mitotic chromosomes (170). The length of the heterochromatin on the X chromosome is polymorphic among strains and can comprise from one third to one half the length of the mitotic chromosome. Cytogenetic experiments show that Release 3 euchromatic sequence extends into the centric heterochromatin by approximately 2.1 Mb (64). Seven physical map gaps (*triangles*) correspond to regions of tandem repeats that are either not clonable in BACs or cannot be assembled. In addition to the physical map gaps there are 37 sequence gaps (data not shown) (26).

In the two and a half years since the initial WGS sequence report, the two sequencing centers participating in the BDGP (Lawrence Berkeley National Laboratory Drosophila Genome Sequencing Center and Baylor College of Medicine Human Genome Sequencing Center) carried out a finishing process that closed gaps, improved sequence quality, and validated the assembly. This resulted in Release 3 of the genomic sequence (26). The process was a hybrid between the clone-by-clone and WGS approaches, taking advantage of the expertise and experience at each center. Sequence traces derived from the WGS and draft sequences of individual BACs were assembled into BAC-sized segments. These segments were brought to high quality by standard methods and then joined computationally to reconstruct each chromosome arm.

In Release 3, the 6 euchromatic chromosome arms are represented in 13 scaffolds comprising116.9 Mb. A scaffold is a set of contiguous sequence

contigs, ordered and oriented with respect to one another; within a scaffold, gaps between adjacent contigs are of known size and are spanned by clones (1). There are 37 sequence gaps, all of which lie in regions of repetitive DNA. The Release 3 sequence contains an estimated error rate of less than one error in 100,000 bp, and the overall assembly was verified by comparing it to a physical map of fingerprinted BAC clones. Hoskins et al. (64) used BAC-based fluorescence in situ hybridization (FISH) analysis to correlate the genomic sequence with the cytogenetic map and found that the Release 3 euchromatic sequence extends into the centric heterochromatin on each chromosome arm.

Using improved WGS sequence assembly algorithms, two additional assemblies containing only the WGS plasmid and BAC paired end sequences used in *WGS1* were generated in March 2001 (*WGS2*) and July 2002 (*WGS3*); 99% of the Drosophila euchromatin was assembled accurately in *WGS3*, with virtually no global errors and few local order and orientation errors (26). Except for a larger number of gaps, overall sequence quality in *WGS3* approaches the National Human Genome Research Institute (NHGRI) standard for finished sequence, illustrating the potential of WGS assembly at high sequence coverage (12X).

Heterochromatin

Heterochromatin comprises about 30% of the fly genome, approximately 59 Mb in the female and 100 Mb in the male (1, 170). The centric region of every major chromosome is heterochromatic, as is most of chromosome 4 and all of the Y (Figure 1). The transition between heterochromatin and euchromatin is gradual; the density of transposable elements in the euchromatin increases continuously toward the centromeres (1, 72) (Figure 2). The heterochromatin contains satellite sequences, middle repetitive elements (e.g., transposons), and some single-copy DNA. Satellite sequences comprise approximately two-thirds of the heterochromatin. Cytogenetic maps of the heterochromatin were made, and in situ hybridization was used to map the distribution of the different satellite DNAs (49, 50, 125, 126). For example, more than 70% of the DNA in the centric heterochromatin of chromosome 2 is composed of five simple repeated sequences (102). Satellite DNA is inefficiently recovered and unstable, even in small insert plasmid vectors (101, 160), limiting the use of standard genomic methods to study the heterochromatin. The third of the heterochromatin not consisting of satellite DNA can be cloned in plasmid vectors; the *WGS3* includes a 20.7-Mb draft-quality assembly of nonsatellite heterochromatin, measured as the part of the *WGS3* that extends beyond, or otherwise does not align to, the Release 3 euchromatic sequence. The *WGS3* heterochromatic sequence is represented in 2597 scaffolds, including portions of five scaffolds that extend into the Release 3 euchromatic sequences of the chromosome arms. The scaffolds range in length from 1 kilobase (kb) to 712 kb and include 1170 sequence gaps that account for 3.7 Mb (18%) of the 20.7-Mb sequence.

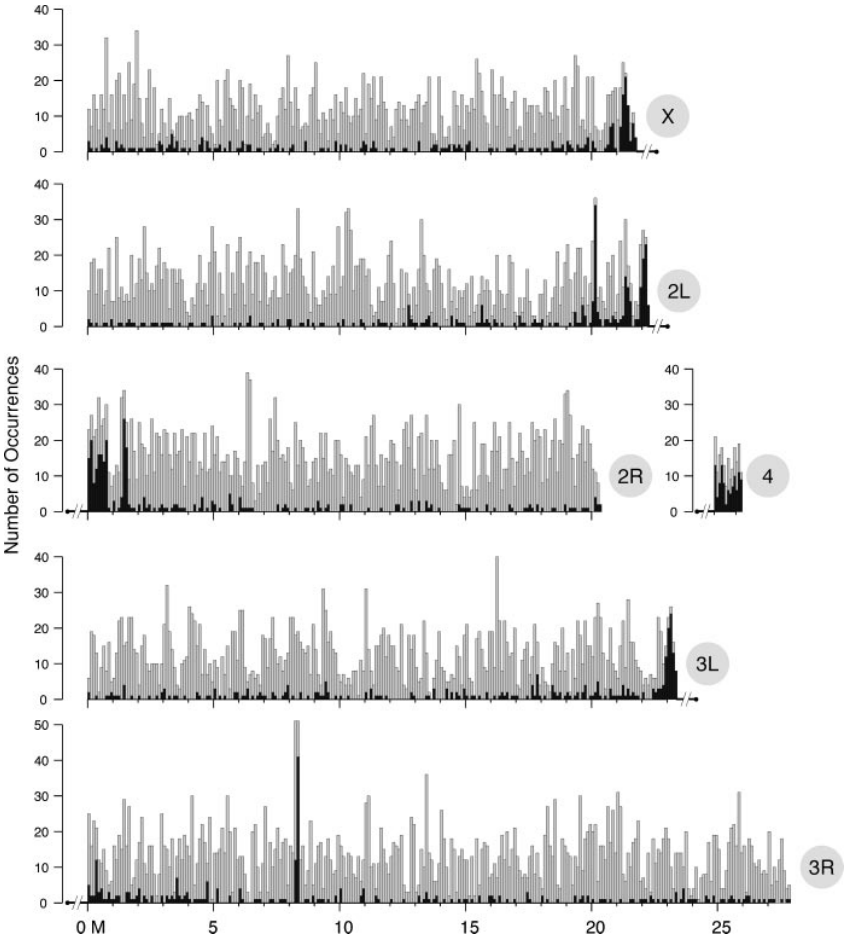


Figure 2 Distribution of protein-coding genes and transposable elements in the *Drosophila melanogaster* euchromatic genome. Each chromosome arm is represented by a black horizontal line with a circle indicating its centromere. The number of transposable elements (*black*) and protein-coding genes (*gray*) is shown for 100-kilobase (kb) windows along each chromosome arm. Although there is local variation, gene density is rather uniform along the chromosome arms, whereas transposable elements are highly enriched at the centromere-proximal regions of each arm. A scale in megabases (Mbs) is shown at the bottom of the figure.

EST AND cDNA RESOURCES

Over 260,000 *Drosophila melanogaster* ESTs have been sequenced from a number of developmental stages and tissues (Table 1). Nearly all of these are 5′-ESTs from cDNA libraries constructed to maximize the percentage of full-length clones. ESTs

TABLE 1 *Drosophila melanogaster* EST collections

Library name	Source	# of ESTs	Year	Reference
RE normalized embryo pFlc-1	Embryo	60,235	2002	(155)
RH normalized head pFlc-1	Adult head	55,406	2002	(155)
LD embryo BlueScript	Embryo	37,744	2000	(138, 155)
GH head pOT2	Adult head	24,598	2000	(138, 155)
AT adult testes pOTB7	Adult testes	23,087	2002	(155)
SD Schneider L2 cell culture pOT2	Schneider cell line	21,132	2000	(138, 155)
GM ovary pOT2	Ovary	11,482	2000	(138, 155)
LP larval-early pupal pOT2	Larvae, pupae	9,381	2000	(138, 155)
D adult testis library	Adult testes	7,298	2000	(2)
ESG01	Salivary glands[a]	4,688	2003	(57)
HL head BlueScript	Adult head	3,183	2000	(138, 155)
CK embryo BlueScript	Embryonic rough ER	1,653	1998	(77)
8–12 hr Embryonic cDNA library	Embryonic 8–12 hr	1,124	2002	(147)
	Total	261,011		

[a]Glands were isolated from mixed stage animals including those that were collected at 6, 18, 20, 22, and 24 h after puparium formation at 18°C.

provide valuable evidence for the existence of genes and their structures. The BDGP also used computational analysis of these ESTs to select a nonredundant set of putative full-length cDNAs for sequencing, the Drosophila Gene Collection (DGC). Initially, cDNAs were selected by choosing the 5′-most clone resulting from *inter se* clustering of ESTs (138). Once the Release 1 annotated genomic sequence was available, cDNAs were selected by aligning ESTs to the genome sequence, followed by a comparison of the aligned ESTs to one another and the predicted gene models; this selection strategy more reliably identified full-length cDNA clones (155).

Full-length cDNAs provide the most accurate evidence available for determining the intron/exon structure of genes and for detecting alternative splicing. They also provide an essential resource for proteomics and functional analyses. cDNAs of more than 3,500 individual genes have been sequenced by members of the Drosophila research community; additionally, the BDGP generated high-quality full-insert sequences for an overlapping set of 8,921 clones in the DGC (154). The annotated Release 3 genomic sequence allows a rigorous quality control assessment of these sequenced cDNAs. Comparison to the genomic sequence detected errors caused by reverse transcriptase, clones that do not contain a complete open reading frame (ORF), and other artifacts of library construction. Through this

process, cDNA clones containing complete and accurate ORFs for 5,300 unique genes, representing 40% of all predicted Drosophila genes, were identified.

In the course of annotating the Release 3 genomic sequence, curators identified ESTs representing 2,013 new clones that were added to the cDNA sequencing pipeline. These include 309 clones chosen to replace previous clones containing truncated ORFs, 543 clones for genes that were not previously represented in the DGC, and 833 clones that represent alternative splice forms. Obtaining a sequenced cDNA for each major splice variant of each gene will require additional EST sequencing and directed library screening with probes designed from the annotated genomic sequence.

STRUCTURAL ANNOTATION

The objective of the initial annotations of the euchromatic portion of the genome was to obtain accurate gene structures, which are necessary to predict the Drosophila proteome. The Release 1 sequence was annotated largely during a two-week period in November 1999 as part of an "Annotation Jamboree," an unprecedented collaborative effort between bioinformatics experts and biologists from the private and public sector. This initial annotation reported 13,601 genes (1). Attempting to assess the accuracy of Release 1 annotations, Karlin et al. (73) compared conceptual translations to a dataset of 1,049 polypeptide sequences existing in SWISS-PROT prior to the 1999 publication of the annotated Release 1 sequence, identifying many apparent annotation errors. Misra et al. (108) reexamined these cases and concluded that most resulted from strain polymorphisms or sequence errors in the SWISS-PROT entries. Release 1 was an imperfect annotation based on an imperfect sequence, but it has been an invaluable tool both for the Drosophila research community and those interested in comparative genomics (85, 140, 164).

Release 2 of the annotated sequence was generated in October 2000, but included only minor changes to the sequence and annotation. In contrast, Release 3 represents a major update of the sequence and its annotation. Keeping the annotation of the Drosophila genome up to date is a core responsibility of FlyBase (45), the community database. A dozen FlyBase curators carried out the Release 3 annotation effort (108) over a nine-month period. The curators used a comprehensive set of guidelines for integrating computational analyses, cDNA data, and protein alignments into updated annotations to enhance the consistency of the effort. Traceable evidence for each gene model was stored and is publicly accessible from FlyBase. Several new open source software tools were built to support this effort. Data relevant to each model were visualized and evaluated using Apollo (93), a sequence annotation editor. An integrated computational pipeline and database (115) were used to generate the required computational analyses and store the annotations and their supporting evidence. However, the most significant reasons for the improved quality of the Release 3 annotations are the availability of cDNA

and EST sequences aligning to 78% of predicted genes and more time devoted to annotation.

Validation of Annotation Using Comparative Genomics

Gene structures can be validated and improved using comparative sequence analysis in the absence of full-length cDNA sequence. In a pilot study, Bergman et al. (9) analyzed gene conservation at eight genomic regions from five Drosophila species (*D. melanogaster*, *D. erecta*, *D. pseudoobscura*, *D. willistoni*, and *D. littoralis*) covering more than 500 kb of the *D. melanogaster* genome. The four other species were chosen to cover a range of divergence times 6–15, 46, 53, and 61–65 million years, respectively (128). All *D. melanogaster* genes (and 78–82% of coding exons) identified in divergent species such as *D. pseudoobscura* showed evidence of functional constraint.

More distant organisms were also examined. The lineages of *D. melanogaster* and *Anopheles gambiae* diverged about 250 million years ago. Comparing the genomes and proteomes of these two diptera (13, 171) reveals that almost half of the genes in both genomes are orthologous, showing an average sequence identity of about 56%. This level of sequence identity is slightly lower than that observed between the orthologues of *Fugu rubripes* and *Homo sapiens*, which diverged about 450 million years ago, indicating that the two insects diverged much faster than the two vertebrates. Some conservation in gene order can be observed in the fly and mosquito genomes, but at a lower level than observed between vertebrate species. Analysis of aligned fly and mosquito sequences reveals that orthologous genes have retained only half of their intron/exon structure, indicating that intron gains or losses occurred at a rate of about one per gene per 125 million years.

Approximately 20% of the predicted Drosophila proteins do not show similarity to proteins encoded by the genomes of organisms from other phyla. Although the number of proteins in a public database such as TrEMBL increased exponentially in the time between the Release 1 and Release 3 annotation in 2000 and 2002, respectively, the increased size of the protein datasets resulted in only a 14% increase in the number of fly genes that produce proteins with similarity to other proteins. This does not include a comparison to the proteins of *A. gambiae* (63), the only other dipteran with a complete genome sequence, because Release 3 was frozen before the *A. gambiae* data was available. We expect many of the Drosophila proteins that now appear to be species specific will show sequence similarity to Anopheles proteins.

Gene Number

The number of genes reported in Release 3 (13,676) is essentially unchanged from the Release 2 annotation (see Table 2). However, revisions in the annotations result in changes to the sequence of 45% of predicted proteins in the genome and in structural changes to 85% of gene models. This is largely due to improved

TABLE 2 Annotations in Release 3

Description[a]	Euchromatin 116.8 Mb	12 Mb of heterochromatin
Protein-coding genes	13,379	297
tRNA genes	290	0
microRNA genes	23	ND
snRNA genes	28	ND
snoRNA genes	28	ND
Pseudogenes	17	ND
Misc. noncoding RNA	38	2
rRNA genes	—	6
Transposons	1,572	ND
Total protein-coding genes	13,379	297
Total length of euchromatin/heterochromatin	116.8 Mb	12 Mb
Exons	60,897	1109
Protein-coding exons[b]	54,934	999
Length of genome in exons	27.8 Mb (24%)	382 kb (3%)
Introns	48,257	803
Genes with 5′ UTR	10,227 (76%)	152 (51%)
Transcripts with 5′ UTR	14,707 (81%)	215 (57%)
Average 5′ UTR length	265 nucleotides	217 nt
Genes with 3′ UTR	9,646 (72%)	119 (40%)
Transcripts with 3′ UTR	14,012 (77%)	172 (46%)
Average 3′ UTR length	442 nucleotides	311 nt
Average ratio of length of CDS/transcript[c]	0.75	0.79
Total protein-coding transcripts	18,106	379
Genes with alternative transcripts	2,729 (20%)	49 (16%)
Average number of transcripts per alternatively spliced gene	2.75	2.8
Total number alternative transcripts	4,743	88
Unique peptides	15,848	351
Gene-prediction data only	815	89
BLASTX/TBLASTX homologies	10,996	167
ESTs and DGC cDNA sequencing reads	10,406	134
GenBank accessions	3,104	22
ARGS (RefSeq)	795	0
Error report submissions	825	7
Full-insert DGC cDNAs	9,297	58

[a]Abbreviations: UTR, untranslated region; CDS, (protein)-coding sequence; R2, Release 2; R3, Release 3; ND, not determined. All statistics are for protein-coding genes only. The numbers reflect the FlyBase annotation database of November 25, 2002.

[b]Any exon containing CDS, even if the majority of the exon is UTR.

[c]The length of the coding region divided by the length of the entire protein coding transcript, averaged over all protein-coding transcripts.

annotation of 3′ and 5′ untranslated regions (UTRs); 70% of genes now have annotated UTR sequences. The evidence supporting the individual gene models has also improved: 78% of the gene models are supported by EST or cDNA sequence and 82% are supported by Basic Local Alignment Search Tool (BLASTX or TBLASTX) homology (see Table 2). The increased EST coverage has revealed many additional cases of alternative splicing and 20% of the Release 3 genes have annotated alternative transcripts. Transposable elements and nonprotein-coding RNAs were systematically annotated for the first time in Release 3.

After publishing the Release 1 sequence and annotations, two groups published analyses that suggested that the number of genes reported was significantly underestimated. Andrews et al. analyzed 3,000 testis-derived EST sequences (2). Two hundred failed to align using BLAST to any other ESTs or to Release 1 annotations. Andrews et al. speculated that these represented new genes, or at least unannotated exons. However, the Release 3 annotation contains only 38 genes that are supported solely by testis-specific ESTs; of those, only six were previously unannotated. Therefore, it seems likely that Andrews et al. identified unannotated exons, particularly 5′ UTR exons, and not entire genes.

Gopal et al. undertook a computational analysis that identified plausible genes, regardless of other evidence, using the gene-finding program Genscan, and then translated these to identify protein motifs that might support the existence of those genes (56). They reported 1,042 novel genes that were not included in the Release 2 annotation. The 781 genes that mapped to the euchromatin were compared to the Release 3 annotations. Most of these are now represented by annotations in Release 3; however, no corresponding annotations exist for the remaining 261 genes and some may be real genes that were missed in the Release 3 annotation.

Transposable Elements

Release 3 is the first sequence of the genome that accurately represents repetitive sequences; a careful analysis of the transposable element families became feasible. Kaminker et al. (72) identified more than 1,500 full or partial transposable elements, comprising nearly 4% of the euchromatin. These elements belong to more than 90 distinct families of transposable elements that vary in copy number from one to 146. Transposable elements are preferentially found in intragenic regions, often nested within other transposons.

Overlapping Gene Structures

One unanticipated finding from the reannotation of the euchromatic genome was the number of unusual overlapping gene structures. Figure 3 shows some examples.

Approximately 15% of annotated genes (2,054) involve the overlap of messenger RNAs (mRNAs) on opposite strands. Complementary sequences between distinct RNAs from overlapping genes on opposite strands were reported in eukaryotes and are implicated in regulating gene expression [for reviews see

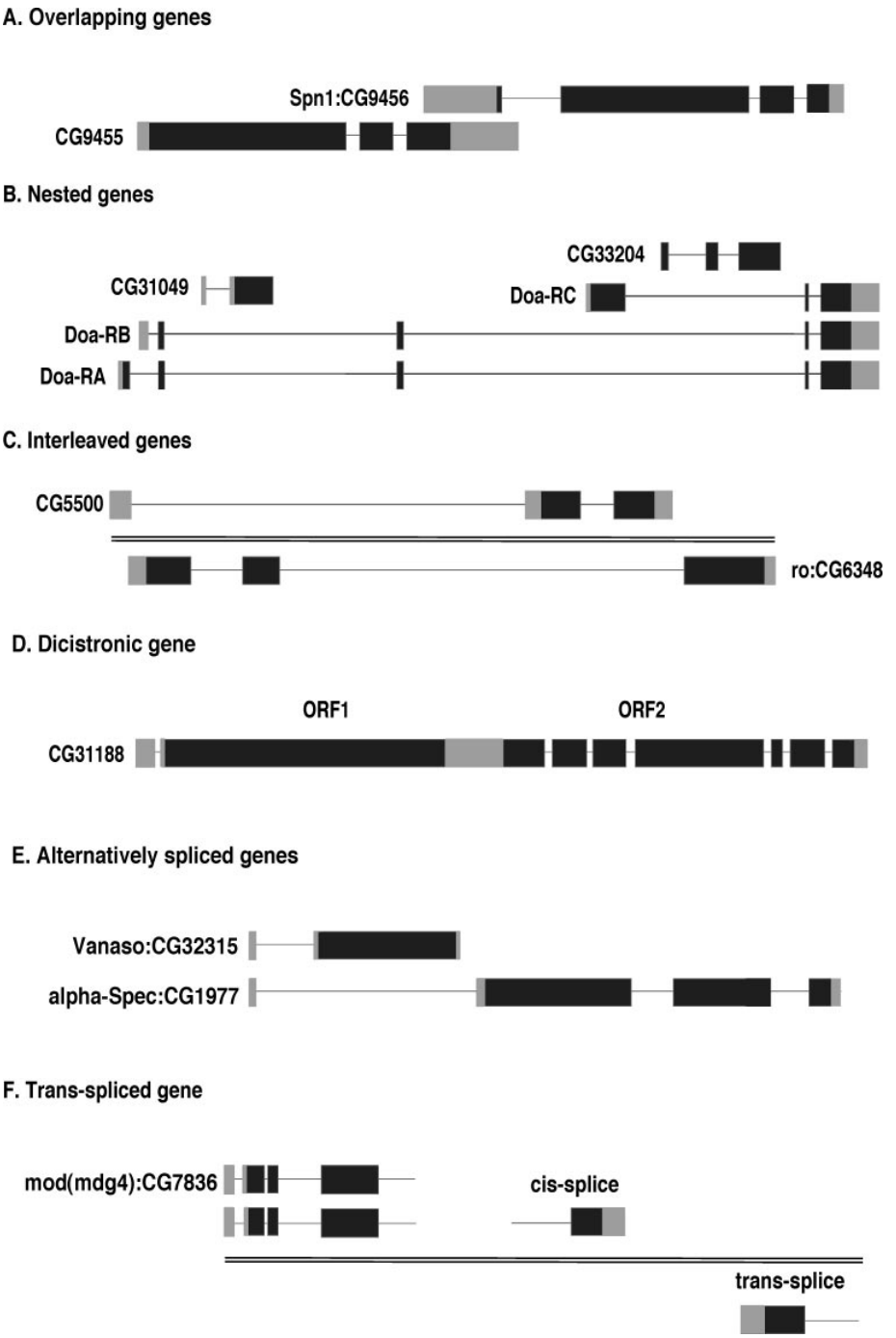
(79, 163)]. For example, the complementary sequence shared between *Dopa decarboxylase* and *CG10561* is involved in regulating levels of these transcripts (150). The large number of identified overlaps raises the possibility that antisense interactions may be a common mechanism for regulating gene expression in Drosophila. Misra et al. also annotated more than 60 instances of overlapping genes on the same strand. In some cases, the 3′ UTR of the upstream gene extends past the putative translation start of the downstream gene (Figure 3*A*).

Approximately 7.6% of Release 3 genes (1038) are included within the introns of other genes (Figure 3*B*). Of the 879 nested protein-coding genes, the majority (574) are transcribed from the opposite strand of the surrounding gene. Interleaved exons and introns (Figure 3*C*) of 26 gene pairs were identified. Transposable elements were found inserted into the introns of 431 genes.

Alternative Splicing

Dicistronic transcripts (Figure 3*D*) have been reported in Drosophila (3, 21, 58, 99, 117, 123, 144, 165). Thirty-one dicistronic transcripts were documented in Release 3, 12 with only one cDNA as evidence. Genes were identified as dicistronic if they contained nonoverlapping coding regions within a single processed mRNA, coding sequences larger than 50 amino acids, and similarity to known proteins. Misra et al. identified 17 additional putative dicistronic genes for which the evidence was less complete. There is evidence supporting alternative monocistronic transcript(s) for the upstream, downstream, or both coding sequences for many predicted dicistronic genes (31/48). In some cases the dicistronic form may be less prevalent than the monocistronic forms; the dicistronic Adh + Adhr transcript is only 5% as abundant as Adh monocistronic transcripts (21). Translating the second coding sequence of a dicistronic transcript requires that translation initiation occur at an internal site. There are two proposed mechanisms for the initiation of internal translation: partial disassembly of the ribosome at the termination of translation of the first coding sequence followed by continued scanning by the 40 S ribosomal subunit (89), and translation initiation using an internal ribosome entry site (124).

Misra et al. annotated alternative splice forms for 2,729 Drosophila genes. The majority of alternatively spliced genes (65%) encode two or more protein products, indicating that alternative splicing generates considerable protein diversity in Drosophila. The other 35% differ only in their 5′ UTRs. Alternative splicing can produce two or more distinct nonoverlapping protein products from a single pre-mRNA species; 12 such cases have been identified so far in Drosophila (for example, *Vanaso* and *Spec*) (Figure 3*E*). The mRNAs produced most commonly share 5′ UTR sequences, but may also share 3′ UTR sequences. Alternative splicing reaches an extreme in the case of the *Down syndrome cell adhesion molecule* (*Dscam*) and *modifier of mdg4* [*mod* (*mdg4*)] genes. *Dscam* encodes an axon guidance receptor. The *Dscam* gene contains 95 alternative exons that are organized into four clusters of 12, 48, 33, and 2 exons each, with the potential to express



38,016 different mRNAs (27, 143). *mod(mdg4)* has been implicated in a range of processes including chromatin insulator functions (8) and apoptosis (61). Twenty-nine distinct transcripts, sharing 5′ exons alternatively spliced to different 3′ exons, are produced from this gene (38, 80). Eight of these transcripts are generated by a trans-splicing (Figure 3*F*) mechanism, using 3′ exons encoded on the opposite strand (38, 80, 108).

Pseudogenes

The number of pseudogenes reported in Drosophila is substantially smaller than in *C. elegans* (39, 112). Twelve previously identified pseudogenes and five new

Figure 3 Complex gene structures. (*A*) Overlapping genes. The 3′ end of *CG9455* overlaps the 5′ end of *Serine protease inhibitor 1* (*Spn1:CG9456*), sharing 408 base pairs of sequence. The translation start site of *Spn1* is 219 bp beyond the translation stop of *CG9455* such that the genes encode distinct proteins. *CG9455* expression is not detected in embryos, and cDNAs from the gene are derived from the testis library, suggesting the gene may be adult specific. Spn1 is expressed in the eye primordium in embryos. Their nonoverlapping expression patterns suggest that transcription of the genes is independent. (*B*) Nested genes. In this example, two genes, *CG31049* and *CG33204*, map to the second and third introns, respectively, of gene *Darkener of apricot* (*Doa*). *Doa* is an example of an alternatively spliced gene with four distinct 5′ exons capable of encoding proteins sharing the carboxy terminus but having distinct amino termini. (*C*) Interleaved genes are transcribed on opposite strands from the same genomic region. Their exons do not overlap but map to an intron of the gene on the complementary strand. In this example, the last two exons of *CG5500* map to the first intron of *rough* (*ro:CG6348*) and the last two exons of *ro* map to the first intron of *CG5500*. (*D*) Dicistronic genes. *CG31188* is an example of a dicistronic gene. A single full-length cDNA contains two open reading frames (ORFs), ORF1 and ORF2, separated by in-frame stop codons. ORF1 is similar to predicted genes in human, rat, and worm. ORF2 shares sequence similarity to *prolyl-4-hydroxylases*. (*E*) Alternatively spliced genes. *Vanaso* and *α-Spectrin* are examples of genes that share an untranslated 5′ exon and no other exons so they encode distinct proteins. Other types of alternatively spliced genes include those that encode identical proteins but have distinct 5′ noncoding sequences and those that encode related but not identical proteins and have some distinct coding exons (see *B*). (*F*) Trans-spliced genes. *modifier of mdg4* [*mod*(*mdg4*):*CG7836*] is currently the only example of a trans-spliced gene in Drosophila. cDNA sequence revealed a gene with four common 5′ exons that are spliced to one of at least 27 possible 3′ exons encoded either on the same (one of 19 shown) or antiparallel DNA strand (one of 8 shown). Exons are shown as boxes, with the ORF in black and the untranslated regions (UTRs) in grey; introns are represented by horizontal lines. In (*E*) and (*F*), the chromosomal DNA is indicated by two closely spaced horizontal lines to emphasize that the two transcripts shown lie on opposite strands.

pseudogenes—four histones and one lectin (*CR31541*)—were annotated in Release 3. Of these 17 pseudogenes, 15 originated by recombination and contain intact introns, one (*Mgstl-Psi*) originated by retrotransposition and lacks introns, and one is too degenerate to classify definitively. No attempt was made to catalogue newly retrotransposed pseudogenes or to annotate the 117 pseudogenes identified by Echols et al. (39) for Release 3, but a list of putative pseudogenes is available at (52). WormBase (157, 168) currently reports 392 pseudogenes in *C. elegans*. It is likely that a subset of the genes identified as protein-coding genes in Release 3 are actually pseudogenes.

RNA Editing

In evaluating the quality of the DGCr1 and DGCr2 cDNAs described above, Stapleton et al. (154) identified 30 genes whose translations do not match the Release 3 predicted proteins and that are consistent with RNA editing. RNA editing generates nucleotide diversity beyond that directly encoded by the genome. Adenosine deaminase (ADAR) targets double-stranded regions of nuclear-encoded RNAs and catalyzes the deamination of adenosine (A) to inosine (I) [reviewed in (7)]. Inosine mimics guanosine (G) in its base-pairing properties, and the translational machinery of the cell interprets I as G. In this way, an A-to-I conversion in the mRNA can alter the genetic information and, consequently, the protein structure. Null mutations in the single ADAR gene in Drosophila (*dADAR*) suggest that pre-mRNA editing modifies adult behavior by altering signaling components in the nervous system (103, 121). Among the mRNAs edited in Drosophila are those encoded by *cacophony* (a calcium channel gene) (148), *paralytic* (a sodium channel gene) (133), and *GluCla* (a chloride channel gene) (145), all of which contain multiple editing sites in their coding sequences. Additional cDNA or EST data is required to distinguish RNA editing from reverse transcriptase errors or strain polymorphisms in many putative additional cases of RNA editing (154). Nevertheless, the combination of a highly accurate genomic sequence and extensive cDNA sequence resources, now available for Drosophila, should allow a comprehensive analysis of the extent of RNA editing as a regulatory mechanism.

Noncoding RNAs

Almost all gene prediction programs assume that genes encode proteins, and noncoding RNA (ncRNA) genes [reviewed in (95)] have gone largely unnoticed even in the era of complete genomic sequences [reviewed in (40, 105)]. Experimental evidence such as cDNAs must exist to identify ncRNA genes. Drosophila ncRNAs were discovered as part of the standard molecular characterization of transcripts produced by loci in the bithorax complex. Lewis (91) proposed that the complex contained a protein-coding gene for each genetic function he identified. Molecular analysis of the transcripts encoded by the *bithoraxoid* (*bxd*) gene showed that,

although alternatively spliced and polyadenylated, none of the splice products encodes a protein. The functions of these transcripts and others that map to the bithorax complex remain obscure (33, 96, 141).

Three spliced and polyadenylated ncRNAs (two nuclear and one cytoplasmic) (65) are transcribed from the 93D, or *hsr*-omega (heat-shock RNA-omega) locus of Drosophila, which is essential for viability in flies. These transcripts have been proposed to function in transport, transcript turnover, and in monitoring the translational machinery of the cell; they specifically affect the transcription and stability of the heat shock–induced hsp70 and alpha-beta transcripts [reviewed in (83)]. The genomic structure of this locus is highly conserved in *D. pseudoobscura* and *D. hydei*, although the nucleotide sequence has diverged.

Studies of dosage compensation in Drosophila provide additional evidence for functional ncRNAs. The *roX1* and *roX2* RNAs are male-specific spliced ncRNAs produced by and associated with the dosage-compensated male X chromosome (106). Functional ncRNAs were found in a wide range of organisms, including human, and a database of such ncRNAs was compiled (42, 43) that currently contains over 20 ncRNA families. Aligning spliced DGC cDNAs that do not contain a significant ORF to the genomic sequence revealed 27 new ncRNA genes. Some of these may be antisense genes; such genes have been reported in other organisms (146); others may encode very small peptides. Further experiments will verify the existence of these interesting genes and determine their function.

MicroRNAs (miRNAs) are a large family of ncRNAs 21–22 nucleotides in length whose functions are unknown (81, 86, 87). All 23 of the known Drosophila miRNAs (81) were annotated in Release 3. Targets of miRNAs are not well characterized, but several Drosophila miRNAs are perfectly complementary to several classes of sequence motifs that mediate negative posttranscriptional regulation (82). In addition to miRNAs, eukaryotic cells contain a complex population of transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), and small nuclear RNAs (snRNAs). Two hundred and ninety tRNA, 28 snoRNA, and 28 snRNA genes were annotated in the Release 3 genome (see Table 2). snoRNAs function as ribonucleoproteins in preribosomal RNA processing reactions and also guide methylation and pseudouridylation of ribosomal RNA. snRNAs are confined to the nucleus, where many are involved in splicing or other RNA processing reactions [reviewed in (44)].

Annotation of Heterochromatin

Hoskins et al. (64) annotated the 12 Mb of heterochromatin that assembled in sequence scaffolds larger than 40 kb in the *WGS3* sequence. This analysis predicted six nonprotein-coding genes and 297 protein-coding genes, including 30 previously known heterochromatic genes. Many regions of similarity to known transposable elements were also identified. This initial annotation of *WGS3* heterochromatic sequences substantially increases the number of predicted genes in

heterochromatin. It also provides a new resource for functional studies of phenomena that differentiate heterochromatin from euchromatin, such as replication timing, chromatin structure, and transcriptional silencing or position effect variegation. Finally, the *WGS3* may facilitate the characterization of *cis*-acting functional elements such as telomeres, chromosome pairing sites, and centromeres.

Previously characterized heterochromatic genes encode products involved in diverse cellular functions. Examples include *light* (post-Golgi protein trafficking), *concertina* (G-protein subunit), *Nipped-B* (morphogenesis), *rolled* (MAP kinase), *poly ADP-ribose polymerase* (chromatin structure), and *bobbed* (ribosomal RNA). These genes were localized to the cytogenetic map through genetic analysis of chromosomal rearrangements and by FISH (49, 66, 78). The genomic structures of several of these genes were determined, and they differ significantly from those of euchromatic genes in that introns and regulatory regions contain clusters of transposable elements (both partial and complete), and some introns are hundreds of kilobases in length (12, 14, 37, 135, 162).

Molecular identification of genes on the Y chromosome of *Drosophila melanogaster* is difficult because the entire chromosome is heterochromatic. Approximately 80% of Y chromosome DNA is composed of nine simple repeated sequences, including (AAGAC)n (8 Mb), (AAGAG)n (7 Mb), and (AATAT)n (6 Mb) (102). In addition, the Y chromosome contains one of two ribosomal DNA loci (*bobbed*) and six genes essential for male fertility (*kl-1*, *kl-2*, *kl-3*, *kl-5*, *ks-1*, and *ks-2*) (22, 75); consistent with this specialization for male fertility, X/0 Drosophila males are sterile but otherwise completely normal (19). The kl-2, kl-3 and kl-5 fertility factors all encode dynein-heavy chains (25, 51). Ten additional Y-linked genes were identified [reviewed in (24)] and there will likely be more found.

FUNCTIONAL ANNOTATION

The promise of genomics has been that knowing all genes in the genome will lead to an understanding of the organism's biology. Some current challenges in Drosophila biology can be addressed using genome-wide approaches, including the development and application of systematic methods to characterize the thousands of genes of unknown function, detection and characterization of non-protein-coding transcripts, the description of gene expression patterns, and the detection of *cis*-acting DNA sequences that play important functional roles in chromosome structure, DNA replication, and control of gene activity.

Genetic Analysis

Drosophila has been the premier metazoan genetic system since Thomas Hunt Morgan began work on this fly in 1910 (110). The genome is easily manipulated and the small number of chromosomes facilitates gene mapping. Balancer chromosomes (113) have been generated that allow recessive lethal mutations to be carried from one generation to the next without selection. Polytene chromosomes

in the salivary glands, discovered in the 1930's, provided a physical map for gene mapping (20, 120). Deficiency collections spanning the genome identify regions of haplo-insufficiency and regions containing genes necessary for viability (94). More modern methods exploiting engineered chromosomes (139, 151), including two-element misexpression systems such as UAS/Gal-4 (17), are used to generate gain-of-function phenotypes. Site-specific recombination can generate mitotic clones of entire chromosome arms (54, 169) or more precise localized rearrangements (88, 158).

The most valuable mutations in defining gene function are generally those with loss-of-function phenotypes that result in lethality or a visible defect. Of the 13,676 Release 3 genes, approximately 15% have a molecularly defined mutation (45). Nearly all these mutations were originally detected based on the phenotype they produced. However, fewer than one-third of Drosophila genes mutate to lethality or to an easily recognized visible phenotype (107). For this reason, methods that rely on molecular, rather than phenotypic, detection of mutations must play a role in elucidating the function of remaining genes. A collection of mutants generated by such methods will facilitate more difficult phenotypic assays, such as those affecting behavior, learning, and memory, as well as the construction of appropriate double mutants to test hypotheses of genetic redundancy.

Morgan and his students identified the earliest viable loss-of-function mutants, which were spontaneous mutants [reviewed in (159)]. Muller (114) pioneered the use of X-rays to induce double-strand breaks; these successfully generated large deletions and chromosomal rearrangements. The widespread use of chemical mutagens was introduced in the 1960s (92). More recently, gene inactivation by transposable element insertion has played an increasingly important role. The current genome-wide gene disruption project (32, 152, 153) has identified 5,500 genes (40%) associated with molecularly defined *P* element insertions, but not all are known to disrupt gene function. The project's status can be found at Reference 130.

In addition to random mutagenic approaches used to generate null mutations in Drosophila, a number of targeted mutagenic approaches were recently developed, including homologous recombination (136, 137), targeted deletions (67), and chromosomal cleavage (11). RNA interference (RNAi) (74) works exceptionally well in Drosophila cultured cells (30), and researchers are developing methods to generate targeted RNAi in transgenic animals engineered to express double-stranded RNA in specific cells or during specific developmental stages [see (53, 84) for examples]. These methods will be especially useful in trying to understand the distinct role(s) of alternatively spliced transcripts (28).

Gene Expression Patterns

Two large-scale methods were successfully used to determine gene expression patterns, RNA in situ hybridization (127) and DNA microarrays (23, 97, 100, 134). The well-established approach of whole-mount RNA in situ hybridization determines precise spatial gene expression patterns (60, 70), and several groups have used it to examine gene expression in high-throughput fashion (77, 147, 161).

RNA in situ hybridization, despite its use of fixed tissues, provides an overview of developmental changes in gene expression patterns when a large number of differently staged specimens are examined. Tomancak et al. (161) examined 2,179 genes by in situ hybridization to fixed Drosophila embryos as a first step in cataloging a comprehensive set of gene expression patterns in Drosophila embryogenesis (129). Nearly two thirds of the genes assayed display dynamic expression patterns that were documented by extensive digital photomicrography of individual embryos. The photomicrographs were annotated using controlled vocabularies to describe anatomical structures and then organized into a developmental hierarchy. Nearly all the annotated expression patterns are distinct. Additionally, Tomancak et al. found that the RNA transcripts of about 1% of genes show clear subcellular localization.

Microarray data have been used to determine wild-type gene expression profiles during development (4, 161, 167), in particular tissues such as testis (2), mesoderm (48), glial cells (41), and imaginal discs (76). Microarrays were also used to identify new gene members of developmental regulatory pathways. Stathopolous et al. (156) identified targets of the Dorsal regulatory pathway by determining gene expression profiles using *Toll* and *pipe* mutants. Microarrays also revealed gene expression patterns during aging (69), infection (16, 35, 36), starvation, and in conditions of elevated sugar consumption (172). Similar methods identified 93 putative targets of the Drosophila *orthodenticle* gene that are also regulated by the human orthologue Otx2 when expressed in flies (109).

Microarray profiles provide a quantitative overview of the relative changes in each gene's expression level over time, but suffer from several limitations. In multicellular organisms, cell differentiation results in tissue complexity that whole-animal microarray analysis cannot document. It is a formidable task to isolate mRNA from every tissue at different developmental stages, measure gene expression, and assign expression indices to recreate the entire developmental expression pattern. Moreover, the quantitative comparison of expression levels for a given gene, or among different genes, in multiple experiments is technically difficult (29, 142).

A major limitation of both whole mount in situ hybridization and microarray analysis is destruction of the animal under study. Methods that allow gene expression to be monitored in living organisms will be of enormous utility for several reasons: (*a*) They allow more dynamic views of gene expression; (*b*) they enable powerful genetic screens focused on particular structures or tissues, modeled on the classic screens that used cuticle preparations (118) or antibody staining (47); and (*c*) they enable sorting of cells that express a given gene for subsequent biochemical analyses, such as microarray analysis of gene expression. These objectives can be accomplished using transgenic Drosophila expressing gene fusions; such constructs are generated using either in vitro constructed promoter fusions (98) or in vivo gene trapping by insertional mutagenesis. Chia and colleagues (111) pioneered a protein trap approach in which full-length endogenous proteins are expressed as green fluorescent protein (GFP) fusion proteins from their endogenous promoters. They generated several hundred independent lines and, in the

case of known molecules, showed that the subcellular distribution of the chimera mimics that of the wild-type endogenous protein. Cooley and colleagues took this approach to identify genes important for oogenesis (31) using a high-throughput sorting strategy to screen embryos that express GFP.

Studies of gene expression have also revealed unexpected coregulation of neighboring genes. Spellman & Rubin (149) found that over 20% of the genes in the Drosophila genome fall into groups of 10–30 loosely clustered genes that show correlated expression across a wide range of experimental conditions. The data do not reveal the mechanism(s) responsible for the observed similarities in expression of adjacent genes, but the findings are most consistent with regulation at the level of chromatin structure based on the finding that the regions containing similarly expressed genes are quite large—125 kb, on average. Using EST data, Boutanaev et al. (15) found similar clustering of coexpressed, nonhomologous genes, in particular, among genes expressed in the testes.

Identification of *Cis*-Acting Sequences Controlling Gene Expression

Seventy-five percent of the euchromatic portion of the Drosophila genome does not encode exons; determining the function of these DNA sequences is a major challenge for the future. Significant progress has been made in identifying sequences that interact with the transcriptional machinery to ensure proper temporal and spatial gene expression. The basal promoter sequence determines the site of transcription initiation. Ohler et al. (119) identified transcription start site (TSS) candidates for close to 2,000 genes by aligning 5′ ESTs from cap-trapped cDNA libraries to the genome. Stringent criteria were used to select the 5′ EST clusters most likely to identify TSSs; a minimum of three ESTs and more than 30% of all 5′ ESTs for a gene were required to end within an 11-bp window. Examining the sequences flanking these TSSs revealed the presence of well-known core promoter motifs such as the TATA box, the initiator, and the downstream promoter element (DPE) [reviewed in (5, 6)]. Ohler et al. also defined and assessed the distribution of several new motifs prevalent in core promoters, including a variant DPE motif. Among the prevalent motifs is the DNA-replication-related element (DRE), part of the recognition site for the TATA binding protein (TBP)-related factor (TRF2).

Other *cis*-regulatory elements that regulate transcription can be located up to 100-kb 5′ or 3′ of the transcription unit or within introns; regardless of their position, the transcription factors that bind to them interact with the basal promoter to regulate gene expression. Such enhancer, silencer, and insulator elements are functionally assayed by creating transgenic animals containing constructs carrying a regulatory region and a reporter gene. Over 5700 such constructs were made to test promoters of 167 genes and are described in 973 references (45). The earliest studies include those of *hsp70* (98), chorion genes (34, 71), *white* (90), and *fushi tarazu* (62). The *cis*-regulatory regions of the homeotic complex genes are the most extensively studied. This large set of experimental data will be invaluable in developing genome-wide computational approaches to identifying

DNA sequences controlling gene expression. One promising approach takes advantage of the observation that enhancers often contain multiple transcription factor binding sites. Several recent computational approaches to identify enhancers based on searching for clusters of transcription factor binding sites have been reported (10, 59, 104, 122, 131, 132).

Evolutionary sequence comparison has long been used to facilitate enhancer dissection experiments of individual genes in Drosophila, using species such as *D. virilis* that are so sufficiently diverged that unselected sequences are generally not conserved. In 1986, the *cis*-acting elements required for proper regulation of *dopa decarboxylase* (*Ddc*) were found to be conserved between *D. virilis* and *D. melanogaster* (18). More extensive analysis of the *Rh3* and *Rh4* opsin genes revealed a near-perfect correlation between short sequences that are conserved between these two species and those sequences that produce a phenotype when mutated by site-directed mutagenesis, demonstrating the predictive power of this approach (46).

On a genome-wide scale, comparative sequence analysis will reduce the amount of noncoding sequence to be screened for *cis*-regulatory function. Draft sequence (6.5X) of the *D. pseudoobscura* genome is available (68), and a pilot study of the *cis*-regulatory regions of eight loci in five Drosophila species reveals that *D. pseudoobscura* is a suitable evolutionary distance for use in this type of analysis (9). Bergman et al.'s analysis (9) also showed that conserved noncoding sequences (CNCSs) are frequently found clustered with conserved spacing, and such clusters of CNCSs can predict enhancer sequences without prior knowledge of transcription factor binding sites. Moreover, conservation of microsynteny between genes and flanking intergenic regions throughout the frequent genome rearrangements that occurred since the separation of Drosophila species (55) can reveal the association of functional noncoding sequences with the appropriate flanking gene [for example, see (9)].

Comparative sequence analyses can be readily combined with approaches based on gene expression patterns to detect putative regulatory regions. Noncoding sequences conserved in the evolution of orthologous gene pairs in species such as *D. melanogaster* and *D. pseudoobscura*, which are also shared by groups of coexpressed genes in *D. melanogaster*, are likely to correspond to *cis*-regulatory modules. Such a combined approach, using human-mouse genome comparisons, successfully identified binding sites for three major muscle-specific transcription factors (166). Similar approaches carried out on a genome-wide basis are now possible in Drosophila, making use of the *D. pseudoobscura* draft genome sequence and the extensive information available on gene expression patterns.

ACKNOWLEDGMENTS

We thank Catherine Nelson, Audrey Huang, and Roger Hoskins for critical comments on the manuscript. We thank Joseph Carlson for generating Figure 2. We thank Madeline Crosby, Beverly Matthews, and David Emmert for help with

FlyBase queries. We thank all members of the BDGP, especially Casey Bergman, Sima Misra, Chris Smith, and Chris Mungall for discussions and Gadfly queries. This work was supported by the Howard Hughes Medical Institute and NIH Grant P50-HG00750 (GMR) and performed under Department of Energy Contract DE-AC0376SF00098 to the University of California.

The *Annual Review of Genomics and Human Genetics* is online at http://genom.annualreviews.org

LITERATURE CITED

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, et al. 2000. The genome sequence of Drosophila melanogaster. *Science* 287:2185–95
2. Andrews J, Bouffard GG, Cheadle C, Lu J, Becker KG, Oliver B. 2000. Gene discovery using computational and microarray analysis of transcription in the Drosophila melanogaster testis. *Genome Res.* 10:2030–43
3. Andrews J, Smith M, Merakovsky J, Coulson M, Hannan F, Kelly LE. 1996. The stoned locus of Drosophila melanogaster produces a dicistronic transcript and encodes two distinct polypeptides. *Genetics* 143:1699–711
4. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, et al. 2002. Gene expression during the life cycle of Drosophila melanogaster. *Science* 297: 2270–75
5. Arnosti DN. 2002. Design and function of transcriptional switches in Drosophila. *Insect Biochem. Mol. Biol.* 32:1257–73
6. Arnosti DN. 2003. Analysis and function of transcriptional regulatory elements: insights from Drosophila. *Annu. Rev. Entomol.* 48:579–602
7. Bass BL. 2002. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* 71:817–46
8. Bell AC, West AG, Felsenfeld G. 2001. Insulators and boundaries: versatile regulatory elements in the eukaryotic. *Science* 291:447–50
9. Bergman CM, Pfeiffer BD, Rincón-Limas DE, Hoskins RA, Gnirke A, et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol.* 3:research0086.1–20
10. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. USA* 99:757–62
11. Bibikova M, Golic M, Golic KG, Carroll D. 2002. Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* 161:1169–75
12. Biggs WH 3rd, Zavitz KH, Dickson B, van der Straten A, Brunner D, et al. 1994. The Drosophila rolled locus encodes a MAP kinase required in the sevenless signal transduction pathway. *EMBO J.* 13:1628–35
13. Bolshakov VN, Topalis P, Blass C, Kokoza E, della Torre A, et al. 2002. A comparative genomic analysis of two distant diptera, the fruit fly, Drosophila melanogaster, and the malaria mosquito, Anopheles gambiae. *Genome Res.* 12:57–66
14. Bonaccorsi S, Gatti M, Pisano C, Lohe A. 1990. Transcription of a satellite DNA on two Y chromosome loops of Drosophila melanogaster. *Chromosoma* 99:260–66
15. Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI. 2002. Large

clusters of co-expressed genes in the Drosophila genome. *Nature* 420:666–69
16. Boutros M, Agaisse H, Perrimon N. 2002. Sequential activation of signaling pathways during innate immune responses in Drosophila. *Dev. Cell* 3:711–22
17. Brand AH, Perrimon N. 1993. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development* 118:401–15
18. Bray SJ, Hirsh J. 1986. The Drosophila virilis dopa decarboxylase gene is developmentally regulated when integrated into Drosophila melanogaster. *EMBO J.* 5:2305–11
19. Bridges CB. 1916. Non-disjunction as a proof of the chromosome theory of heredity. *Genetics* 1:1–52, 107–63
20. Bridges CB. 1935. Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster. J. Hered.* 26:60–64
21. Brogna S, Ashburner M. 1997. The Adh-related gene of Drosophila melanogaster is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. *EMBO J.* 16:2023–31
22. Brosseau GE. 1960. Genetic analysis of the male fertility factors on the *Y* chromosome of *Drosophila melanogaster*. *Genetics* 45:257–74
23. Brown PO, Botstein D. 1999. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* 21:33–37
24. Carvalho AB. 2002. Origin and evolution of the Drosophila Y chromosome. *Curr. Opin. Genet. Dev.* 12:664–68
25. Carvalho AB, Lazzaro BP, Clark AG. 2000. Y chromosomal fertility factors kl-2 and kl-3 of Drosophila melanogaster encode dynein heavy chain polypeptides. *Proc. Natl. Acad. Sci. USA* 97:13239–44
26. Celniker SE, Wheeler DL, Kronmiller B, Carlson J, Halpern A, et al. 2002. Finishing a whole genome shotgun: Release 3 of the Drosophila melanogaster euchromatic genome sequence. *Genome Biol.* 3:research0079.1–14
27. Celotto AM, Graveley BR. 2001. Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. *Genetics* 159:599–608
28. Celotto AM, Graveley BR. 2002. Exon-specific RNAi: a tool for dissecting the functional relevance of alternative splicing. *RNA* 8:718–24
29. Chudin E, Walker R, Kosaka A, Wu SX, Rabert D, et al. 2002. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays. *Genome Biol.* 3:research0005.1–10
30. Clemens JC, Worby CA, Simonson-Leff N, Muda M, Maehama T, et al. 2000. Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways. *Proc. Natl. Acad. Sci. USA* 97:6499–503
31. Cooley L. 2002. GFP Protein Trap Database. http://info.med.yale.edu/cooley
32. Cooley L, Kelley R, Spradling A. 1988. Insertional mutagenesis of the Drosophila genome with single P elements. *Science* 239:1121–28
33. Cumberledge S, Zaratzian A, Sakonju S. 1990. Characterization of 2 RNAs transcribed from the *cis*-regulatory region of the Abd-A domain within the *Drosophila* Bithorax complex. *Proc. Natl. Acad. Sci. USA* 87:3259–63
34. de Cicco DV, Spradling AC. 1984. Localization of a cis-acting element responsible for the developmentally regulated amplification of Drosophila chorion genes. *Cell* 38:45–54
35. De Gregorio E, Spellman PT, Rubin GM, Lemaitre B. 2001. Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. *Proc. Natl. Acad. Sci. USA* 98:12590–95
36. De Gregorio E, Spellman PT, Tzou P, Rubin GM, Lemaitre B. 2002. The Toll and Imd pathways are the major regulators

of the immune response in Drosophila. *EMBO J.* 21:2568–79
37. Devlin RH, Bingham B, Wakimoto BT. 1990. The organization and expression of the light gene, a heterochromatic gene of Drosophila melanogaster. *Genetics* 125: 129–40
38. Dorn R, Reuter G, Loewendorf A. 2001. Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in Drosophila. *Proc. Natl. Acad. Sci. USA* 98:9724–29
39. Echols N, Harrison P, Balasubramanian S, Luscombe NM, Bertone P, et al. 2002. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res.* 30:2515–23
40. Eddy SR. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* 2:919–29
41. Egger B, Leemans R, Loop T, Kammermeier L, Fan Y, et al. 2002. Gliogenesis in Drosophila: genome-wide analysis of downstream genes of glial cells missing in the embryonic nervous system. *Development* 129:3295–309
42. Erdmann VA, Szymanski M, Hochberg A, de Groot N, Barciszewski J. 1999. Collection of mRNA-like non-coding RNAs. *Nucleic Acids Res.* 27:192–95
43. Erdmann VA, Szymanski M, Hochberg A, Groot N, Barciszewski J. 2000. Non-coding, mRNA-like RNAs database Y2K. *Nucleic Acids Res.* 28:197–200
44. Filipowicz W, Pogacic V. 2002. Biogenesis of small nucleolar ribonucleoproteins. *Curr. Opin. Cell Biol.* 14:319–27
45. FlyBase, Project BDG. 2003. FlyBase: A Database of the Drosophila Genome. http://www.flybase.org
46. Fortini ME, Rubin GM. 1990. Analysis of cis-acting requirements of the Rh3 and Rh4 genes reveals a bipartite organization to rhodopsin promoters in Drosophila melanogaster. *Genes Dev.* 4:444–63
47. Fujita SC, Zipursky SL, Benzer S, Ferrus A, Shotwell SL. 1982. Monoclonal antibodies against the Drosophila nervous system. *Proc. Natl. Acad. Sci. USA* 79: 7929–33
48. Furlong EE, Andersen EC, Null B, White KP, Scott MP. 2001. Patterns of gene expression during Drosophila mesoderm development. *Science* 293:1629–33
49. Gatti M, Bonaccorsi S, Pimpinelli S. 1994. Looking at Drosophila mitotic chromosomes. *Methods Cell Biol.* 44:371–91
50. Gatti M, Pimpinelli S, Santini G. 1976. Characterization of Drosophila heterochromatin. I. Staining and decondensation with Hoechst 33258 and quinacrine. *Chromosoma* 57:351–75
51. Gepner J, Hays TS. 1993. A fertility region on the Y chromosome of Drosophila melanogaster encodes a dynein microtubule motor. *Proc. Natl. Acad. Sci. USA* 90:11132–36
52. Gerstein M. 2003. Pseudogenes and intergenic analysis. http://bioinfo.mbb.yale.edu/genome/pseudogene/fly
53. Giordano E, Rendina R, Peluso I, Furia M. 2002. RNAi triggered by symmetrically transcribed transgenes in Drosophila melanogaster. *Genetics* 160:637–48
54. Golic KG, Lindquist S. 1989. The FLP recombinase of yeast catalyzes site-specific recombination in the Drosophila genome. *Cell* 59:499–509
55. Gonzalez J, Ranz JM, Ruiz A. 2002. Chromosomal elements evolve at different rates in the Drosophila genome. *Genetics* 161:1137–54
56. Gopal S, Schroeder M, Pieper U, Sczyrba A, Aytekin-Kurban G, et al. 2001. Homology-based annotation yields 1,042 new candidate genes in the Drosophila melanogaster genome. *Nat. Genet.* 27: 337–40
57. Gorski SM, Chittaranjan S, Pleasance ED, Freeman JD, Anderson CL, et al. 2003. A SAGE approach to discovery of genes involved in autophagic cell death. *Curr. Biol.* 13:358–63
58. Gray TA, Nicholls RD. 2000. Diverse

splicing mechanisms fuse the evolutionarily conserved bicistronic MOCS1A and MOCS1B open reading frames. *RNA* 6: 928–36

59. Halfon MS, Grad Y, Church GM, Michelson AM. 2002. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.* 12:1019–28
60. Harrison PR, Conkie D, Paul J, Jones K. 1973. Localisation of cellular globin messenger RNA by in situ hybridisation to complementary DNA. *FEBS Lett.* 32:109–12
61. Harvey AJ, Bidwai AP, Miller LK. 1997. Doom, a product of the Drosophila mod(mdg4) gene, induces apoptosis and binds to baculovirus inhibitor-of-apoptosis proteins. *Mol. Cell. Biol.* 17: 2835–43
62. Hiromi Y, Kuroiwa A, Gehring WJ. 1985. Control elements of the Drosophila segmentation gene fushi tarazu. *Cell* 43:603–13
63. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. 2002. The genome sequence of the malaria mosquito Anopheles gambiae. *Science* 298:129–49
64. Hoskins RA, Smith CD, Carlson J, Carvalho BA, Halpern A, et al. 2002. Heterochromatic sequences in a Drosophila whole genome shotgun assembly. *Genome Biol.* 3:research0085.1–.16
65. Hovemann B, Walldorf U, Ryseck RP. 1986. Heat-shock locus 93D of Drosophila melanogaster: an RNA with limited coding capacity accumulates precursor transcripts after heat shock. *Mol. Gen. Genet.* 204:334–40
66. Howe M, Dimitri P, Berloco M, Wakimoto BT. 1995. Cis-effects of heterochromatin on heterochromatic and euchromatic gene activity in Drosophila melanogaster. *Genetics* 140:1033–45
67. Huet F, Lu JT, Myrick KV, Baugh LR, Crosby MA, Gelbart WM. 2002. A deletion-generator compound element allows deletion saturation analysis for genomewide phenotypic annotation. *Proc. Natl. Acad. Sci. USA* 99:9948–53
68. Human Genome Sequencing Center Baylor College of Medicine. 2003. *D. pseudoobscura* Genome Project. http://www.hgsc.bcm.tmc.edu/projects/drosophila/
69. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. 2001. The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. *Nat. Genet.* 29: 389–95
70. Jones KW, Robertson FW. 1970. Localisation of reiterated nucleotide sequences in Drosophila and mouse by in situ hybridisation of complementary RNA. *Chromosoma* 31:331–45
71. Kalfayan L, Wakimoto B, Spradling A. 1984. Analysis of transcriptional regulation of the s38 chorion gene of Drosophila by P element-mediated transformation. *J. Embryol. Exp. Morphol.* 83 (Suppl.):137–46
72. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, et al. 2002. The transposable elements of the Drosophila melanogaster euchromatin—a genomics perspective. *Genome Biol.* 3:research0084.1–20
73. Karlin S, Bergman A, Gentles AJ. 2001. Genomics. Annotation of the Drosophila genome. *Nature* 411:259–60
74. Kennerdell JR, Carthew RW. 1998. Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway. *Cell* 95: 1017–26
75. Kennison JA. 1981. The genetic and cytological organization of the Y-chromsome of Drosophila melanogaster. *Genetics* 98:529–48
76. Klebes A, Biehs B, Cifuentes F, Kornberg TB. 2002. Expression profiling of Drosophila imaginal discs. *Genome Biol* 3:research0038.1–16
77. Kopczynski CC, Noordermeer JN, Serano TL, Chen WY, Pendleton JD, et al.

1998. A high throughput screen to identify secreted and transmembrane proteins involved in Drosophila embryogenesis. *Proc. Natl. Acad. Sci. USA* 95:9973–78

78. Koryakov DE, Zhimulev IF, Dimitri P. 2002. Cytogenetic analysis of the third chromosome heterochromatin of Drosophila melanogaster. *Genetics* 160:509–17
79. Kumar M, Carmichael GG. 1998. Antisense RNA: function and fate of duplex RNA in cells of higher eukaryotes. *Microbiol. Mol. Biol. Rev.* 62:1415–34
80. Labrador M, Mongelard F, Plata-Rengifo P, Baxter EM, Corces VG, Gerasimova TI. 2001. Protein encoding by both DNA strands. *Nature* 409:1000
81. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* 294:853–58
82. Lai EC. 2002. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* 30:363–64
83. Lakhotia SC, Ray P, Rajendra TK, Prasanth KV. 1999. The non-coding transcripts of hsr-omega gene in Drosophila: do they regulate trafficking and availability of nuclear RNA-processing factors? *Curr. Sci.* 77:553–63
84. Lam G, Thummel CS. 2000. Inducible expression of double-stranded RNA directs specific genetic interference in Drosophila. *Curr. Biol.* 10:957–63
85. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921
86. Lee RC, Ambros V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science* 294:862–64
87. Lee RC, Feinbaum RL, Ambros V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75:843–54
88. Lee T, Luo L. 2001. Mosaic analysis with a repressible cell marker (MARCM) for Drosophila neural development. *Trends Neurosci.* 24:251–54
89. Levine F, Yee JK, Friedmann T. 1991. Efficient gene expression in mammalian cells from a dicistronic transcriptional unit in an improved retroviral vector. *Gene* 108:167–74
90. Levis R, Hazelrigg T, Rubin GM. 1985. Separable cis-acting control elements for expression of the white gene of Drosophila. *EMBO J.* 4:3489–99
91. Lewis EB. 1978. A gene complex controlling segmentation in *Drosophila*. *Nature* 276:565–70
92. Lewis EB, Bacher F. 1968. Method of feeding ethyl methane sulfonate (EMS) to Drosophila males. *Drosophila Info. Serv.* 43:193
93. Lewis SE, Searle SMJ, Harris NL, Gibson M, Iyer VR, et al. 2002. Apollo: A sequence annotation editor. *Genome Biol.* 3:research0082.1–.14
94. Lindsley DL, Sandler L, Baker BS, Carpenter AT, Denell RE, et al. 1972. Segmental aneuploidy and the genetic gross structure of the Drosophila genome. *Genetics* 71:157–84
95. Lipman DJ. 1997. Making (anti)sense of non-coding sequence conservation. *Nucleic Acids Res.* 25:3580–83
96. Lipshitz HD, Peattie DA, Hogness DS. 1987. Novel transcripts from the *Ultrabithorax* domain of the Bithorax Complex. *Gene Dev.* 1:307–22
97. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. 1999. High density synthetic oligonucleotide arrays. *Nat. Genet.* 21:20–24
98. Lis JT, Simon JA, Sutton CA. 1983. New heat shock puffs and beta-galactosidase activity resulting from transformation of Drosophila with an hsp70-lacZ hybrid gene. *Cell* 35:403–10
99. Liu H, Jang JK, Graham J, Nycz K, McKim KS. 2000. Two genes required for meiotic recombination in Drosophila are expressed from a dicistronic message. *Genetics* 154:1735–46

100. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14:1675–80
101. Lohe AR, Brutlag DL. 1986. Multiplicity of satellite DNA sequences in Drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* 83:696–700
102. Lohe AR, Hilliker AJ, Roberts PA. 1993. Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. *Genetics* 134:1149–74
103. Ma E, Tucker MC, Chen Q, Haddad GG. 2002. Developmental expression and enzymatic activity of pre-mRNA deaminase in Drosophila melanogaster. *Brain Res. Mol. Brain. Res.* 102:100–4
104. Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc. Natl. Acad. Sci. USA* 99:763–68
105. Mattick JS. 2001. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO J.* 2:986–91
106. Meller VH, Gordadze PR, Park Y, Chu X, Stuckenholz C, et al. 2000. Ordered assembly of roX RNAs into MSL complexes on the dosage-compensated X chromosome in Drosophila. *Curr. Biol.* 10:136–43
107. Miklos GLG, Rubin GM. 1996. The role of the genome project in determining gene function: insights from model organisms. *Cell* 86:521–29
108. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell K, et al. 2002. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Bio.* 3:research 0083.1–22
109. Montalta-He H, Leemans R, Loop T, Strahm M, Certa U, et al. 2002. Evolutionary conservation of otd/Otx2 transcription factor action: a genome-wide microarray analysis in Drosophila. *Genome Biol.* 3:research0015.1–15
110. Morgan TH. 1910. Sex limited inheritance in Drosophila. *Science* 32:120–22
111. Morin X, Daneman R, Zavortink M, Chia W. 2001. A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in Drosophila. *Proc. Natl. Acad. Sci. USA* 98:15050–55
112. Mounsey A, Bauer P, Hope IA. 2002. Evidence suggesting that a fifth of annotated Caenorhabditis elegans genes may be pseudogenes. *Genome Res.* 12:770–75
113. Muller HJ. 1918. Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics* 3:422–99
114. Muller HJ. 1927. Artificial transmutation of the gene. *Science* 66:84–87
115. Mungall CJ, Misra S, Berman BP, Carlson J, Frise E, et al. 2002. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol.* 3:research0081.1–10
116. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. 2000. A whole-genome assembly of Drosophila. *Science* 287:2196–204
117. Niimi T, Yokoyama H, Goto A, Beck K, Kitagawa Y. 1999. A Drosophila gene encoding multiple splice variants of Kazal-type serine protease inhibitor-like proteins with potential destinations of mitochondria, cytosol and the secretory pathway. *Eur. J. Biochem.* 266:282–92
118. Nusslein-Volhard C, Wieschaus E. 1980. Mutations affecting segment number and polarity in Drosophila. *Nature* 287:795–801
119. Ohler U, Liao G, Niemann H, Rubin GM. 2002. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.* 3:research0087.1–12
120. Painter TS. 1933. A new method for the study of chromosomal rearrangements and the plotting of chromosomal maps. *Science* 78:585–86

121. Palladino MJ, Keegan LP, O'Connell MA, Reenan RA. 2000. A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. *Cell* 102:437–49
122. Papatsenko DA, Makeev VJ, Lifanov AP, Regnier M, Nazina AG, Desplan C. 2002. Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers. *Genome Res.* 12:470–81
123. Pauli D, Tonka CH, Ayme-Southgate A. 1988. An unusual split Drosophila heat shock gene expressed during embryogenesis, pupation and in testis. *J. Mol. Biol.* 200:47–53
124. Peabody DS, Berg P. 1986. Termination-reinitiation occurs in the translation of mammalian cell mRNAs. *Mol. Cell Biol.* 6:2695–703
125. Pimpinelli S, Berloco M, Fanti L, Dimitri P, Bonaccorsi S, et al. 1995. Transposable elements are stable structural components of Drosophila melanogaster heterochromatin. *Proc. Natl. Acad. Sci. USA* 92:3804–8
126. Pimpinelli S, Santini G, Gatti M. 1976. Characterization of Drosophila heterochromatin. II. C- and N-banding. *Chromosoma* 57:377–86
127. Pollet N, Niehrs C. 2001. Expression profiling by systematic high-throughput in situ hybridization to whole-mount embryos. *Methods Mol. Biol.* 175:309–21
128. Powell JR. 1997. *Progress and Prospects in Evolutionary Biology The Drosophila Model*. New York, Oxford: Oxford University Press 562 pp.
129. Project BDG. 2003. Patterns of Gene Expression in *Drosophila* Embryogenesis. http://www.fruitfly.org/cgi-bin/ex/insitu.pl
130. Project BDG. 2003. Transposon Insertions. http://www.fruitfly.org/p_disrupt/TE.html
131. Rajewsky N, Vergassola M, Gaul U, Siggia ED. 2002. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* 3:30
132. Rebeiz M, Reeves NL, Posakony JW. 2002. SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl. Acad. Sci. USA* 99:9888–93
133. Reenan RA, Hanrahan CJ, Barry G. 2000. The mle(napts) RNA helicase mutation in drosophila results in a splicing catastrophe of the para Na+ channel transcript in a region of RNA editing. *Neuron* 25:139–49
134. Reinke V, White KP. 2002. Developmental genomic approaches in model organisms. *Annu. Rev. Genomics Hum. Genet.* 3:153–78
135. Reugels AM, Kurek R, Lammermann U, Bunemann H. 2000. Mega-introns in the dynein gene DhDhc7(Y) on the heterochromatic Y chromosome give rise to the giant threads loops in primary spermatocytes of Drosophila hydei. *Genetics* 154:759–69
136. Rong YS, Golic KG. 2000. Gene targeting by homologous recombination in Drosophila. *Science* 288:2013–18
137. Rong YS, Titen SW, Xie HB, Golic MM, Bastiani M, et al. 2002. Targeted mutagenesis by homologous recombination in D. melanogaster. *Genes Dev* 16:1568–81
138. Rubin GM, Hong L, Brokstein P, Evans-Holm M, Frise E, et al. 2000. A Drosophila complementary DNA resource. *Science* 287:2222–24
139. Rubin GM, Spradling AC. 1982. Genetic transformation of Drosophila with transposable element vectors. *Science* 218:348–53
140. Rubin GM, Yandell MD, Wortman JR, Miklos GLG, Nelson CR, et al. 2000. Comparative genomics of the eukaryotes. *Science* 287:2204–15
141. Sanchez-Herrero E, Akam M. 1989. Spatially ordered transcription of regulatory DNA in the bithorax complex

of *Drosophila*. *Development* 107:321–29

142. Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467–70
143. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, et al. 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101:671–84
144. Schulz RA, Miksch JL, Xie XL, Cornish JA, Galewsky S. 1990. Expression of the Drosophila gonadal gene: alternative promoters control the germ-line expression of monocistronic and bicistronic gene transcripts. *Development* 108:613–22
145. Semenov EP, Pak WL. 1999. Diversification of Drosophila chloride channel gene by multiple posttranscriptional mRNA modifications. *J. Neurochem.* 72: 66–72
146. Shendure J, Church GM. 2002. Computational discovery of sense-antisense transcription in the human and mouse genomes. *Genome Biol.* 3:research 0044.1–14
147. Simin K, Scuderi A, Reamey J, Dunn D, Weiss R, et al. 2002. Profiling patterned transcripts in Drosophila embryos. *Genome Res.* 12:1040–47
148. Smith LA, Peixoto AA, Hall JC. 1998. RNA editing in the Drosophila DMCA1A calcium-channel alpha 1 subunit transcript. *J. Neurogenet.* 12:227–40
149. Spellman PT, Rubin GM. 2002. Evidence for large domains of similarly expressed genes in the Drosophila genome. *J. Biol* 1:5
150. Spencer CA, Gietz RD, Hodgetts RB. 1986. Overlapping transcription units in the dopa decarboxylase region of Drosophila. *Nature* 322:279–81
151. Spradling AC, Rubin GM. 1982. Transposition of cloned P elements into Drosophila germ line chromosomes. *Science* 218:341–47
152. Spradling AC, Stern D, Beaton A, Rhem EJ, Laverty T, et al. 1999. The Berkeley Drosophila Genome Project gene disruption project: Single P-element insertions mutating 25% of vital Drosophila genes. *Genetics* 153:135–77
153. Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, Rubin GM. 1995. Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. *Proc. Natl. Acad. Sci. USA* 92:10824–30
154. Stapleton M, Carlson J, Brokstein P, Yu C, Champe M, et al. 2002. A Drosophila full-length cDNA resource. *Genome Biol.* 3:research0080.1–8
155. Stapleton M, Liao G, Brokstein P, Hong L, Carninci P, et al. 2002. The Drosophila gene collection: identification of putative full-length cDNAs for 70% of D. melanogaster genes. *Genome Res.* 12:1294–300
156. Stathopoulos A, Levine M. 2002. Dorsal gradient networks in the Drosophila embryo. *Dev. Biol.* 246:57–67
157. Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J. 2001. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* 29:82–86
158. Struhl G, Basler K. 1993. Organizing activity of wingless protein in Drosophila. *Cell* 72:527–40
159. Sturtevant AH. 1965. *A History of Genetics*. New York: Harper & Row 167 pp.
160. Sun X, Le H, Wahlstrom J, Karpen GH. 2002. Sequence analysis of a functional Drosophila centromere. *Genome Res.* 13:182–94
161. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, et al. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.* 3:research0088.1–14
162. Tulin A, Stewart D, Spradling AC. 2002. The Drosophila heterochromatic gene encoding poly(ADP-ribose) polymerase (PARP) is required to modulate chromatin

structure during development. *Genes Dev.* 16:2108–19
163. Vanhee-Brossollet C, Vaquero C. 1998. Do natural antisense transcripts make sense in eukaryotes? *Gene* 211:1–9
164. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. 2001. The sequence of the human genome. *Science* 291:1304–51
165. Walker DL, Wang D, Jin Y, Rath U, Wang Y, et al. 2000. Skeletor, a novel chromosomal protein that redistributes during mitosis provides evidence for the formation of a spindle matrix. *J. Cell Biol.* 151:1401–12
166. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* 26: 225–28
167. White KP, Rifkin SA, Hurban P, Hogness DS. 2000. Microarray analysis of *Drosophila* development during metamorphosis. *Science* 286:2179–84
168. WormBase. 2003. The Genome and Biology of *C. elegans*, Release WS88. http://www.wormbase.org
169. Xu T, Rubin GM. 1993. Analysis of genetic mosaics in developing and adult Drosophila tissues. *Development* 117: 1223–37
170. Yamamoto MT, Mitchelson A, Tudor M, O'Hare K, Davies JA, Miklos GL. 1990. Molecular and cytogenetic analysis of the heterochromatin-euchromatin junction region of the Drosophila melanogaster X chromosome using cloned DNA sequences. *Genetics* 125:821–32
171. Zdobnov EM, Von Mering C, Letunic I, Torrents D, Suyama M, et al. 2002. Comparative genome and proteome analysis of Anopheles gambiae and Drosophila melanogaster. *Science* 298:149–59
172. Zinke I, Schutz CS, Katzenberger JD, Bauer M, Pankratz MJ. 2002. Nutrient control of gene expression in Drosophila: microarray analysis of starvation and sugar-dependent response. *EMBO J.* 21:6162–73

Annual Review of Genomics and Human Genetics
Volume 4, 2003

CONTENTS