whole organs, and to systemic processes. For example, some human renal disorders are associated with defects in genes involved in fluid and electrolyte transport. *Drosophila* orthologs of these genes found in the genomic sequence should spur studies of the physiology and function of Malphigian tubules, which serve as the *Drosophila* "kidney." We anticipate that new collaborations between vertebrate and fly researchers will come about to study behavior, neurodegeneration, aging, and drugs and that important new biological principles and pathways will emerge.

Research is subject to strong selective pressures. Experimental systems that offer the most efficient and direct access to important questions attract the most attention and effort, and as techniques evolve and interests mature, the landscape will change rapidly. The *Drosophila* sequence is a critical resource that ensures that this tiny dew-lover will continue to lead the way to new biological pathways and principles. If *Drosophila* has been difficult for workers in other fields because of an arcane nomenclature and idiosyncratic husbandry, the sequence now provides access through a universal language—the DNA sequence.

**References**

1. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
2. M. D. Adams *et al.*, *Science* **287**, 2185 (2000); R. A. Hoskins *et al.*, *Science* **287**, 2271 (2000).
3. E. W. Myers, *Science* **287**, 2196 (2000).
4. G. M. Rubin and E. B. Lewis, *Science* **287**, 2216 (2000).
5. T. Xu and G. M. Rubin, *Development* **117**, 1223 (1993).
6. H. J. Bellen *et al.*, *Genes Dev.* **3**, 1288 (1989).
7. G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
8. R. D. Riddle and C. Tabin, *Sci. Am.* **280**, 74 (February 1999); J. C. Hendricks *et al.*, *Neuron* **25**, 1299 (2000); M. S. Moore *et al.*, *Cell* **93**, 997 (1998).
9. G. R. Jackson *et al.*, *Neuron* **21**, 633 (1998); J. M. Warrick *et al.*, *Cell* **93**, 939 (1998); K.-T. Min and S. Benzer, *Science* **284**, 1985 (1999).
10. M. Tessier-Lavigne and C. S. Goodman, *Science* **274**, 1123 (1996); J. A. Hoffmann *et al.*, *Science* **284**, 1313 (1999); R. Bodmer and M. Frasch, in *Heart Development*, R. P. Harvey and N. Rosenthal, Eds. (Academic Press, San Diego, CA 1999), pp. 65–90; W. J. Gehring and K. Ikeo, *Trends Genet.* **15**, 371 (1999); J. C. Dunlap, *Cell*, **96**, 271 (1999); A. P. McMahon, *Cell*, **100**, 185 (2000).
11. A. C. Spradling *et al.*, *Genetics* **153**, 135 (1999).
12. G. M. Rubin *et al.*, *Science* **287**, 2222 (2000); K. P. White *et al.*, *Science* **286**, 2179 (1999).
13. G. A. Kerkut and L. I. Gilbert, *Comprehensive Insect Physiology, Biochemistry, and Pharmacology* (Pergamon, New York, 1985); V. B. Wigglesworth, *The Principles of Insect Physiology* (Chapman & Hall, London, 1939).

VIEWPOINT

# From Sequence to Chromosome: The Tip of the *X* Chromosome of *D. melanogaster*

**Panayiotis V. Benos,[1] Melanie K. Gatt,[2,11] Michael Ashburner,[1,2] Lee Murphy,[3] David Harris,[3] Bart Barrell,[3] Concepcion Ferraz,[4] Sophie Vidal,[4] Christine Brun,[4] Jacques Demailles,[4] Edouard Cadieu,[5] Stephane Dreano,[5] Stéphanie Gloux,[5] Valerie Lelaure,[5] Stephanie Mottier,[5] Francis Galibert,[5] Dana Borkova,[6] Belen Minana,[6] Fotis C. Kafatos,[6] Christos Louis,[7,8] Inga Sidén-Kiamos,[7] Slava Bolshakov,[6,7] George Papagiannakis,[7] Lefteris Spanos,[7] Sarah Cox,[7] Encarnación Madueño,[9] Beatriz de Pablos,[9] Juan Modolell,[9] Annette Peter,[10] Petra Schöttler,[10] Meike Werner,[10] Foteini Mourkioti,[10] Nicole Beinert,[10] Gordon Dowe,[10] Ulrich Schäfer,[10] Herbert Jäckle,[10] Alain Bucheton,[4] Deborah M. Callister,[11] Lorna A. Campbell,[11] Areti Darlamitsou,[11] Nadine S. Henderson,[11] Paul J. McMillan,[11] Cathy Salles,[11] Evelyn A. Tait,[11] Phillipe Valenti,[11] Robert D. C. Saunders,[11,12] David M. Glover[2,11]**

**One of the rewards of having a *Drosophila melanogaster* whole-genome sequence will be the potential to understand the molecular bases for structural features of chromosomes that have been a long-standing puzzle. Analysis of 2.6 megabases of sequence from the tip of the *X* chromosome of *Drosophila* identifies 273 genes. Cloned DNAs from the characteristic bulbous structure at the tip of the *X* chromosome in the region of the *broad* complex display an unusual pattern of in situ hybridization. Sequence analysis revealed that this region comprises 154 kilobases of DNA flanked by 1.2-kilobases of inverted repeats, each composed of a 350–base pair satellite related element. Thus, some aspects of chromosome structure appear to be revealed directly within the DNA sequence itself.**

Fewer than 90 years have elapsed since Alfred H. Sturtevant presented the world with the first-ever genetic map of six visible markers on the *X* chromosome of *Drosophila* (*1*). Now that the sequence of almost the entire euchromatic genome of *Drosophila* has been determined (*2*), we have the opportunity to study the function of each gene. In addition, we can study the relation between DNA sequence and chromosome structure.

The European *Drosophila* Genome Project (EDGP) (*3*) has determined the sequence of a contiguous segment of DNA extending some 2.6 Mb from the tip of the *X* chromosome, that is, from subdivision 1A to subdivision 3C on the standard polytene chromosome map (*4*). We predict the existence of 273 protein-coding genes in this region (one gene every 9.6 kb), which is of some sentimental, as well as much scientific, interest to geneticists. It extends from a position 120-kb distal to the *yellow* locus to 150-kb proximal to the *white* locus, whose mutation was the first clearly visible mutation found in *Drosophila* and whose study led to the discovery of sex-linked inheritance and, hence, to the proof of the chromosome theory of heredity (*5*).

In the region sequenced, we have identified 17 transposon insertions. The most common element was *roo* (six copies), but six other retroviral-like elements, two LINE-like elements, an element with inverted repeat ends (*S*-element), and a foldback (FB) element were also found. The overall density of insertions (one insertion every 155 kb) is similar to that in the *Adh* region [one every 170 kb (*6*)]. This is of some interest because the tip of the *X* is a region of low genetic recombination and, on theoretical grounds, might have been expected to accumulate transposable elements. However, it has been suspected that transposon insertion might reduce fitness. Because the *X* chromosome is hemizygous in the male fly, it is subject to stronger selection pressures. This would lead to the prediction of a lower frequency of insertions on the *X* (*7*). Our finding that the overall transposable element densities in the *X* tip and region 35 to 36 are comparable argues against the maintenance of element copy number by negative selection.

The first physical map of any genome was the description of the polytene chromosomes of *Drosophila* (*8*). These chromosomes arise from endoreduplication resulting in a large number of parallel fibers with each fiber rep-

resenting a single haploid chromosome. They are characterized by an aperiodic pattern of dark-stained bands and light-stained interbands, reflecting differences in the extent of DNA packing. These patterns are colinear with the genetic map, as proven by Bridges (*9*) and must be remarkably stable because they are recognizable in species that have diverged many millions of years ago (*10*). Our contiguous sequence covers 102 of the 5072 polytene chromosome bands. The average DNA content per band is 26.2 kb, similar to that estimated by Sorsa (*11*) for the genome as a whole but less than that estimated (49 kb/band) for the *Adh* region (*6*). Little is known about the mechanism that determines the banding pattern of polytene chromosomes. Although the answer must be based on the DNA sequence, any satisfactory solution to this problem will require further understanding of the DNA-protein interactions inherent to chromosome structure. We can now suggest an explanation for a long-standing observation concerning the morphology of the polytene *X* chromosome tip.

The tip of the polytene *X* chromosome in *Drosophila* and related species is characterized by a bulbous structure that often exhibits an unusual arrangement of chromosome bands. This arrangement was noted by those who pioneered the study of polytene chromosomes, including Bridges (*12*) and Offermann (*13*). Normally, in polytene chromosomes of *Drosophila*, each band crosses the entire width of the chromosome and lies perpendicular to its long axis. In the bulbous region in division 2 of the *X* chromosome, the bands within the cytological interval 2B3-8 are often roughly parallel to the chromosome's long axis (Fig. 1). This pattern is unique to the wild-type chromosomes of *Drosophila*. Both Bridges and Offermann concluded that it resulted from local duplications of polytene chromosome bands, in particular from reverse duplications. These could, in molecular terms, have represented large-scale duplications up to 25 kb, although Bridges and Offerman did not have the tools to resolve this. Banding patterns not dissimilar to this are seen in chromosome aberrations that have been interpreted as reverse duplications (*14*) and are common in the genomes of some other *Drosophila* species (*15*).

To determine whether this unusual polytene structure might reflect the organization of the DNA in the 2B3-8 interval, we carried out in situ hybridization studies with more than 50 independent cosmid clones previously localized to this region by overlapping restriction endonuclease digestion patterns (*3*). Independent overlapping clones from this region give the striking pattern of in situ hybridization seen in Fig. 1. Rather than lying transverse to the chromosome as it would normally, the hybridization appears to be restricted to the lateral parts of the 2B bulb. Thus, it seems that the exceptional banding pattern in region 2B of the polytene chromosomes has its counterpart in the patterns of in situ hybridization.

We know from genetic studies that region 2B includes the *broad* (*br*) gene. This gene encodes a family of zinc finger proteins that possess common and unique exons (*16*). A consequence of its molecular organization is that the *br* region has a complex pattern of complementation between mutant alleles: It was originally defined as four mutually complementing, lethally mutable loci. The region is expressed as an "early" ecdysone puff in salivary gland polytene chromosomes (*17*). We had no difficulty in assembling contiguous DNA sequence from the DNA sequence of 12 cosmid clones spanning the 2B3-8 region and including the *br* locus. Thus, we could not account for the unusual polytene chromosome structure or for in situ hybridization patterns of region 2B cosmids by the presence of any large-scale sequence repeat.

To determine whether there might be any other aspects of DNA sequence responsible for both of these features, we used computational methods (*18*) to look for both direct (i.e., *abab* ) and reverse (i.e., *abba*) repeat motifs in a 1.4-Mb region that spans from polytene chromosome region 1D (690 kb distal to *br*) to 3A1 (720 kb proximal to *br*). We find at a position some 82 kb distal to *broad* a 1.2 kb sequence composed of three and a half tandem repeats of a 350 base pair (bp)–element. A corresponding 3.5 repeat of the same element is found in inverted orientation

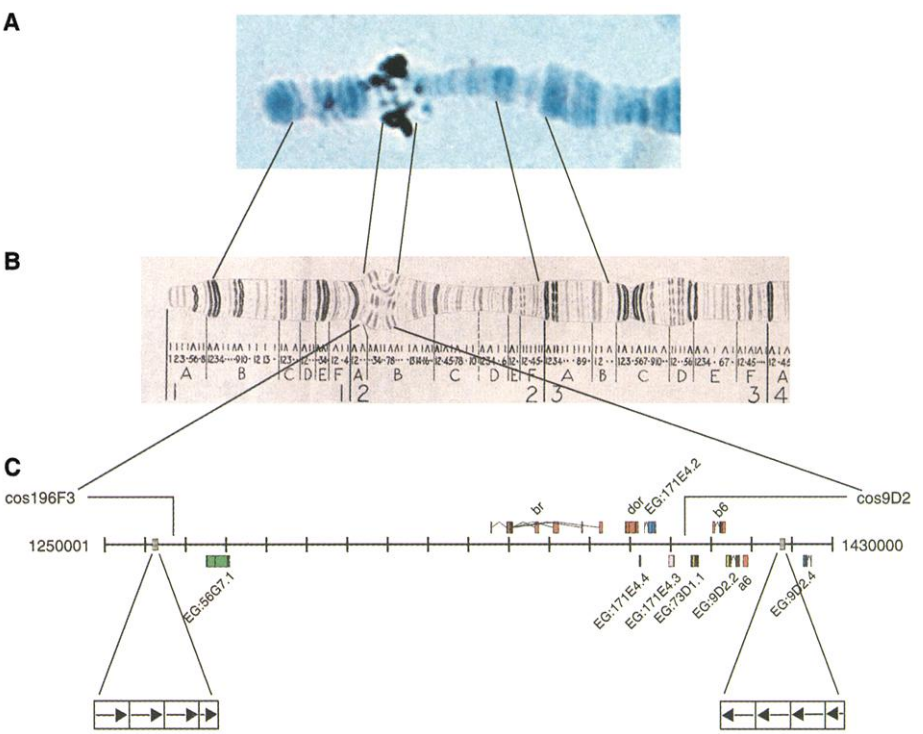


**Fig. 1.** The 2B region of the *X* chromosome. (**A**) A polytene chromosome to which the cosmid clone 9E2 has been hybridized in situ. This ~40-kb clone apparently covers six polytene chromosome bands and the hybridization signals are concentrated laterally. This is seen with many independent cosmid clones hybridizing to this region, including 196F3 and 9D2. (**B**) A reproduction of the drawing of this region by Bridges (*12*). He depicted the bands 2B3-4 and 2B7-8 as being connected. (**C**) A molecular map of this region (180 kb) showing the inverted 1.688 satellite DNA-like repeats distal to *br* (within cosmid 196F3) and proximal to a6 (within cosmid 9D2). Four genes in this region were known previously (*br*, *dor*, *a6*, *b6*); the others have been computationally predicted.

[1]The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Hall, Cambridge CB10 1SD, UK. [2]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK. [3]Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [4]Montpellier University Medical School, Institut de Génétique Humaine (IGH), CNRS, 114 Rue de la Cardonille, 34396 Montpellier Cedex 5, France. [5]Unité Propre de Recherche 41, CNRS, Recombinaisons Genetiques, Faculte de Medecine, 2 Avenue du Pr Leon Bernard, 35043 Rennes Cedex, France. [6]European Molecular Biology Laboratory, Heidelberg, Germany. [7]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology–Hellas, Heraklion, Greece. [8]Department of Biology, University of Crete, Heraklion, Greece. [9]Centro de Biología Molecular Severo Ochoa, Consejo Superior de Investigaciones Cientificas and Universidad Autónoma de Madrid, 28049 Madrid, Spain. [10]Max-Planck-Institut für Biophysikalische Chemie, Department of Molecular Developmental Biology, D-37070 Göttingen, Germany. [11]Department of Anatomy and Physiology, Cancer Research Campaign Cell Cycle Genetics Group, University of Dundee, Dundee DD1 4HN, UK. [12]Department of Biological Sciences, The Open University, Milton Keynes MK7 6AA, UK.

some 44 kb proximal to *broad*. This 350-bp element is related to the 1.688 satellite repeat, of which three dispersed subfamilies have been described previously (*19*). They are scattered in units of 1 to 4 and are at sites largely restricted to the *X* chromosome, where it has been speculated that they have a role in dosage compensation. The two internally repetitive 1.2-kb inverted repeats in region 2B are located precisely where they could define the ends of the inverted repeat band region suggested by Bridges and Offermann some 60 years ago.

Our findings suggest that inverted repeats of DNA can influence the architecture of chromosomes even when they are widely separated, in this case by 154 kb of sequences that have no obvious repetitive structure within them. Factors that influence the three-dimensional organization of chromosomes within the nucleus are poorly understood. We suggest that some of them might be recognized within the long-range sequence of the DNA itself. Because sequence repeats of this type are not uncommon within eukaroytic genomes, this could be one general means influencing the organization of chromosomal domains.

**References and Notes**


1. A. H. Sturtevant, *J. Exp. Zool.* **14**, 43 (1913).
2. M. Adams *et al.*, *Science* **287**, 2185 (2000); G. M. Rubin *et al.*, *Science* **287**, 2204 (2000).
3. R. D. C. Saunders *et al.*, *Nucleic Acids Res.* **17**, 9027 (1989); I. Sidén-Kiamos *et al.*, *Nucleic Acids Res.* **18**, 6261 (1990); E. Madueno *et al.*, *Genetics* **139**, 1631 (1995).
4. P. V. Benos *et al.*, in preparation; A. Peter *et al.*, in preparation. The sequence was determined from a minimum tiling path of the cosmid and bacterial artificial chromosome (BAC) clones that had been used to construct a physical map of the *X* chromosome (*6*). Two BAC libraries were made by Alain Billaud at the Centre d'Etude du Polymorphisme Humaine in a collaboration with the EDGP and one was obtained from the Berkeley *Drosophila* Genome Project [R. A. Hoskins *et al.*, *Science* **287**, 2271 (2000)]. A BAC, NdeI (BACN) library was prepared with Nde II inserts and a BAC, Hind II (BACH) library with Hind III inserts in the vector pBeloBACII. These libraries were made with pools of size-fractionated DNA that gives mean insert sizes of up to 90 kb. The 23,400 clones give about a 10-fold coverage of the genome. These libraries are available at www.hgmp.mrc.ac.uk/Biology/Bio.html. Sequence-tagged sites of both terminal insert sequences were determined for a total of about 6350 BACs from these libraries (www.genoscope.cns.fr/externe/English/Projets/Resultats/rapport.html). The assembled nonredundant 2.6 Mb sequence is at edgp.ebi.ac.uk/cgi-bin/progress.pl, which links to the following European Molecular Biology Laboratory database accession numbers: AL009146, AL009147, L009171, AL009188-AL009196, AL021067, AL021086, AL021106 to AL021108, AL021726, AL021728, AL022017, AL022018, AL022139, AL023873, AL023874, AL023893, AL024453-AL024457, AL024484, AL024485, AL030993, AL030994, AL031024-AL031131, AL031173, AL031227, AL031366, AL031367, AL031581-AL031583, AL031640, AL031765, AL031766, AL031863, AL031883, AL031884, AL033125, AL034388, AL034544, AL035104, AL035105, AL035207, AL035245, AL035311, AL035312, AL035331, AL035395, AL035436, AL035631, AL035632, AL049535, AL050231, AL050232, AL109630, AL121800, AL121803-AL121806, AL132651, AL132792, AL132797, AL133503 to AL133506, AL138678, AL138971, AL138972, Z98254, and Z98269. Sequences were analyzed by methods similar to those described (*6*). Data were managed with AceDB.
5. T. H. Morgan, *Science* **32**, 120 (1910); C. B. Bridges, *Genetics* **1**, 1, 107 (1916).
6. M. Ashburner *et al.*, *Genetics* **153**, 179 (1999).
7. C. H. Langley, E. A. Montgomery, R. Hudson, N. L. Kaplan, B. Charlesworth, *Genet. Res.* **52**, 223 (1988); B. Charlesworth and A. Lapid, *Genet. Res.* **54**, 113 (1989); P. D. Sniegowski and B. Charlesworth, *Genetics* **137**, 815 (1994); C. Hoogland and C. Biemont, *Genetics* **144**, 197 (1996).
8. C. B. Bridges, *J. Hered.* **26**, 60 (1935).
9. _____, *Cytologia* (Fujii Jubilee Vol.) 745 (1937).
10. There have been many studies of polytene chromosome banding patterns in related species. Perhaps the most complete is that of the Hawaiian fauna; see H. L. Carson J. Tonzetich, L. T. Doescher, in *Drosophila Inversion Polymorphism*, C. B. Krimbas and J. R. Powell, Eds. (CRC, Boca Raton, FL, 1992), pp. 441–453. For the persistence of banding patterns over long evolutionary time, see H. D. Stalker, *Genetics* **70**, 457 (1972).
11. V. Sorsa, *Chromosome Maps of Drosophila*: vols. 1 and 2 (CRC, Boca Raton, FL, 1988).
12. C. B. Bridges, *J. Hered.* **29**, 11 (1938).
13. C. A. Offermann, *J. Genet.* **32**, 102 (1936).
14. See figure 17.2 of M. Ashburner, *Drosophila: A Laboratory Handbook* (Cold Spring Harbor Press, Cold Spring Harbor, NY, 1989); D. Gubb, M. Ashburner, J. Roote, T. Davis, *Genetics* **126**, 167 (1990).
15. C. D. Kastritsis, Z. G. Scouras, M. Ashburner, *Chromosoma* **93**, 381 (1986).
16. I. F. Zhimulev, E. S. Belyaeva, O. M. Mazina, M. C. Balasov, *Eur. J. Entomol.* **92**, 263 (1992); G. Tzolovsky, W.-M. Deng, T. Schlitt, M. Bownes, *Genetics* **153**, 1371 (1999).
17. H. J. Becker, *Chromosoma* **10**, 654 (1959).
18. Methods for repeat analysis used were DOTTER [E. L. L. Sonnhammer and R. Durbin, *Gene* **167**, GC1 (1995)] and MIROREPEATS [J. D. Parsons, *Comput. Appl. Biosci.* **11**, 615 (1995)]. Sequence alignments were made with CLUSTALW [J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4573 (1994)].
19. G. L. Waring and J. C. Pollack, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2843 (1987); S. M. DiBartolomeis, K. D. Tartof, F. R. Jackson, *Nucleic Acids Res.* **20**, 1113 (1992).


20. We thank the European Commission for financial support under the Biotechnology Programme of Framework 4. This work was also supported by a Medical Research Council (UK) project grant to D.M.G. and M.A. and by a Dirección General de Investigación Científica y Técnica grant to J.M. Work in Göttingen was supported by the Deutsche Humangenomprojekt. R.D.C.S. held a Wellcome Trust Senior Fellowship. We thank G. M. Rubin and his colleagues in the Berkeley *Drosophila* Genome Project for their long-standing collaboration and, in particular, for their BAC, Eco RI (BACR) clone library and their expressed sequence tag and EP-element sequences. We thank R. Durbin and his colleagues in the Sanger Centre for help with ACeDB.


VIEWPOINT

# A *Drosophila* Complementary DNA Resource


**Gerald M. Rubin,[1,2,3] Ling Hong,[1,3] Peter Brokstein,[1,3] Martha Evans-Holm,[1,3] Erwin Frise,[1,3] Mark Stapleton,[4] Damon A. Harvey[1,2,3]**



**Collections of nonredundant, full-length complementary DNA (cDNA) clones for each of the model organisms and humans will be important resources for studies of gene structure and function. We describe a general strategy for producing such collections and its implementation, which so far has generated a set of cDNAs corresponding to over 40% of the genes in the fruit fly *Drosophila melanogaster*.**


Collections of full-length sequenced cDNAs corresponding to each gene in an organism are widely recognized to be of great utility (*1*). They allow expression of the encoded proteins in a variety of contexts, which facilitates comprehensive structural and functional studies. In addition, they allow the accurate prediction of gene structures, particularly of 5′ and 3′ untranslated regions (UTRs) that are refractory to computational prediction based on genomic DNA sequence alone. The first steps in producing such a collection are the generation of high-quality cDNA libraries and the identification of a full-length clone, or minimally a clone containing the full-length open reading frame (ORF), for each gene. Here we present a strategy that has so far allowed us to obtain such clones for over 40% of all *Drosophila* genes. We also discuss how clones corresponding to less highly expressed genes might be obtained.

Our approach is outlined in Fig. 1. We first constructed oligo(dT)-primed cDNA libraries from high-quality RNA isolated from a variety of developmental stages and tissues using well-established methods (*2*) (Table 1). We did not attempt to decrease the contribution of abundant mRNAs to these libraries by normalization because such protocols are difficult to perform without compromising cDNA length (*3*). We then generated ex-