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-

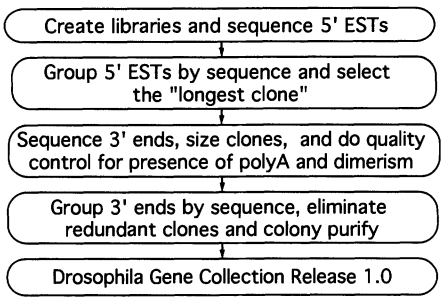

Fig. 1. Diagram of the process used to generate the DGC. See text for details.

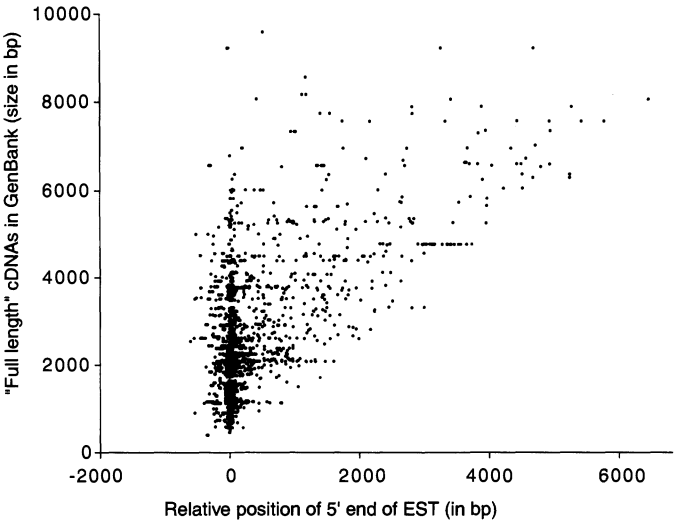

Fig. 2. Estimating the quality of the LD cDNA library. 5′ ESTs derived from LD library clones were compared with the 1213 sequenced *Drosophila* cDNAs in GenBank that are reported to extend farther 5′ than the start of the ORF. When an EST corresponded to one of these 1213 genes, it was aligned to its GenBank counterpart with LALIGN (*13*). Each dot represents the result of one such alignment. A position of 0 on the *x* axis indicates that the GenBank clone and the EST are the same length; a negative number indicates the EST extends farther 5′. As reported in Table 1, 33.6% of LD clones are as long as or longer than the corresponding GenBank clone. This percentage is higher for clones under 4 kb and drops off markedly with increasing clone size, although apparently full-length clones are seen up to 6.5 kb.

pressed sequence tags (ESTs) (*4*) from the 5′ ends of 80,000 cDNAs (*5*). A comparison with the 13,600 genes predicted from the genomic sequence indicates that these ESTs represent 8900 different genes, 65% of all *Drosophila* genes (*6*).

The use of 5′ ESTs allows us to evaluate the quality of each library rapidly and to identify the clone that extends farthest toward the 5′ end of each gene. We assessed the fraction of clones in each library likely to be full length by aligning the 5′ EST sequences derived from that library to the sequences of a test set of clones. This test set consists of the 9% of all *Drosophila* genes for which a cDNA clone having the full-length ORF exists in GenBank (see Fig. 2). About 80% of the clones that matched the test set contain the full-length ORF; for 33%, the 5′ extent is equal to or greater than the GenBank sequence (Table 1).

We then clustered the 5′ ESTs by sequence (*7*) and selected the one clone representing each gene that extends farthest 5′. We estimate that by selecting the longest clone for each gene, we increased the percentage of clones containing the full ORF to greater than 95%. We next obtained the sequence of the 3′ ends of 9080 of the selected clones (*8*). We performed two quality-control tests at this point: First, we discarded clones for which a polyadenylate [poly(A)] tail was not apparent. Second, we aligned the 5′ and 3′ sequences of each clone to the genomic DNA sequence (*9*) and discarded clones for which the two sequences were not in proximity. This eliminated clones that contain two unrelated cDNAs coligated into the same cloning vector, which occurred in about 5% of clones, as well as clones for which a data tracking error occurred such that the 5′ and 3′ reads of that clone were not appropriately associated in our database. We also determined the insert size of each clone (*10*).

We clustered the remaining clones on the basis of their 3′ end sequences. This allowed us to eliminate remaining duplicate clones; such clones might escape detection in the 5′ end clustering if they so differ in length that their 5′ ESTs do not overlap. In these cases, the longer clone was retained.

These steps resulted in the generation of a validated set of 5849 clones, estimated to represent 42% of all predicted *Drosophila* genes. The average size of the cDNAs in this set is 2.2 kb (Table 1). These clones are now being colony purified and arrayed to generate what we call the *Drosophila* Gene Collection (DGC), Release 1.0 (*11*). We are also selecting replacements, if they exist in our EST collection, for clones that failed to pass quality-control tests. We anticipate that this selection of replacements will increase our representation from 42% to over 50% of all genes.

We envision two complementary strategies to isolate cDNAs representing the remaining genes. Because we want to determine the sequence of the 5′ and 3′ UTRs, we do not intend to simply amplify predicted ORFs using reverse transcription polymerase chain reaction (RT-PCR). Given funding, we would propose to generate an additional 200,000 5′ ESTs from both existing as well as newly constructed libraries. Given the availability of a highly annotated genome sequence, we need only sequence 50 to 100 base pairs (bp) to obtain an unambiguous alignment with the genome. We can then computationally determine whether a particular EST is likely to derive from a clone containing a complete ORF not represented in our current DGC set. Promising clones would then be sequenced from the 3′ end and subjected to our other quality-control criteria. We anticipate that 100-bp ESTs can be generated for a fraction of the cost of the 500-bp ESTs used in our initial work and that

Table 1. Summary of construction of the *Drosophila* Gene Collection. RNA for the various libraries was obtained from the following sources: LD, 0- to 22-hour embryos; GM, ovaries, stage 1 to 6 of oogenesis; HL and GH, adult head; LP, mixed larval and early pupal stages; and SD, Schneider L2 cell line. Sequence reads were quality trimmed before submission to GenBank essentially as described in (*14*); we estimate the accuracy of the high-quality region to be better than 99% and that of the additional bases included in the total submission to be 97%. A list of the clones that make up the current DGC can be found at www.fruitfly.org/DGC.

Library name	LD	GM	HL	GH	LP	SD	Totals
Number of 5′ ESTs sequenced	32,714	6,201	2,981	21,212	8,935	6,962	79,636
Average submitted length inbp	554	505	508	577	584	544	546
Average high-quality length inbp	457	408	405	478	488	448	447
Estimated percentage of clones extending 5′ of AUG	83.4	78.6	46.0	78.9	79.4	81.0	80.0
Percentage of clones longer than the corresponding clone in the GenBank test set	33.6	36.9	21.7	35.7	33.4	40.5	34.5
Number of 3′ ESTs sequenced	3,813	714	204	3,111	381	857	9,080
Number of clones selected for DGC	2,594	467	137	1,867	209	575	5,849
Average size in kb of cDNAs in DGC	2.3	1.7	2.0	2.0	2.2	2.7	2.2

[1]Berkeley *Drosophila* Genome Project, [2]Howard Hughes Medical Institute, and [3]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720–3200, USA. [4]Berkeley *Drosophila* Genome Project, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

200,000 additional ESTs will be sufficient to bring our DGC set to 80% of all genes. The remaining 20% of clones can be isolated by library screening with PCR-based methods (*12*); our existing libraries have an estimated total complexity in excess of 5 million clones.

The approach we have demonstrated, as well as the extensions outlined above, will serve as a useful model for the generation of similar clone sets in other organisms. Indeed, some aspects of these ideas have already been adopted by the Mammalian Gene Collection project (*1*).

References and Notes

1. R. L. Strausberg, E. A. Feingold, R. D. Klausner, F. S. Collins, *Science* **286**, 455 (1999).
2. Total RNA was prepared by hot phenol extraction. mRNA was purified with the Stratagene Poly(A) Quik mRNA isolation kit. mRNA quality was assessed by RNA blots with Delta and Notch clones as probes. First strand cDNA synthesis was carried out with the Stratagene λZAPII-cDNA synthesis kit with the substitution of SuperscriptII reverse transcriptase from GIBCO-BRL. The cDNAs were directionally cloned into the λZAPII and pOT2a vectors. Phage and plasmid libraries were amplified once. The plasmid libraries were size fractionated with GIBCO-BRL S-500 cDNA size fractionation columns, and clones with inserts larger than 1 kb were pooled and transformed.
3. M. F. Bonaldo, G. Lennon, M. B. Soares, *Genome Res.* **6**, 791 (1996).
4. M. D. Adams *et al.*, *Science* **252**, 1651 (1991).
5. 5′ ESTs were generated with either dye primer or dye terminator chemistries on ABI373 and ABI377 sequencers.
6. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
7. Clustering was done by first collecting similar sequences with BLAST and then aligning these sequences with PHRAP. Details of the computational methods used in this work will be described elsewhere.
8. 3′ ESTs were generated on an ABI377 sequencer with rhodamine dye terminators or by dye primer sequencing on a Licor sequencer.
9. Alignments were performed with SIM4.
10. The length of the cDNA insert was determined by PCR amplification with vector primers that flank the cloning sites, and the products were sized by agarose gel electrophoresis.
11. The individual clones corresponding to all 80,000 of our ESTs are available through Research Genetics, including all clones that are in the DGC set. We anticipate that the DGC as a separate collection of cDNAs will be available for distribution in May 2000.
12. D. J. Monroe *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 2209 (1995).
13. X. Huang and W. Miller, *Adv. Appl. Math.* **12**, 373 (1991).
14. L. D. Hillier *et al.*, *Genome Res.* **6**, 807 (1996).

15. The following individuals helped generate the sequence data reported here: E. Baxter, R. Blazej, M. Chew, C. Doyle, R. Galle, R. George, R. Hoskins, D. Kruse, J. Landau, H. Meagher, A. Pinder, S. Richards, C. Suh, G. Tsang, and C. Yu. We are especially grateful to S. Lewis for her help in developing the informatics tools used to support this work, K. Wan and M. Champe for their technical contributions to the sequencing, and S. Celniker for her management of the Lawrence Berkeley National Laboratory sequencing facility in which some of this work was done. A. Spradling provided the RNA used to construct the GM library. C. Nelson and A. Huang helped improve the writing. The generation of the 5′ ESTs was supported by the Howard Hughes Medical Institute. Other aspects of this work were supported by grant DE-FG03-98ER62625 from the Department of Energy.