The *Drosophila* Gene Collection: Identification of Putative Full-Length cDNAs for 70% of *D. melanogaster* Genes

Mark Stapleton,[1,2,6] Guochun Liao,[1,3] Peter Brokstein,[1,3] Ling Hong,[1,3] Piero Carninci,[4] Toshiyuki Shiraki,[4] Yoshihide Hayashizaki,[4] Mark Champe,[1,2] Joanne Pacleb,[1,2] Ken Wan,[1,2] Charles Yu,[1,2] Joe Carlson,[1,2] Reed George,[1,2] Susan Celniker,[1,2] and Gerald M. Rubin[1,3,5]

[1]*Berkeley* Drosophila *Genome Project,* [2]*Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA;* [3]*Department of Molecular and Cell Biology, University of California, Berkeley, California 94720-3200, USA;* [4] *Genome Exploration Research Group, RIKEN Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan;* [5]*Howard Hughes Medical Institute, University of California, Berkeley, California 94720, USA*

Collections of full-length nonredundant cDNA clones are critical reagents for functional genomics. The first step toward these resources is the generation and single-pass sequencing of cDNA libraries that contain a high proportion of full-length clones. The first release of the *Drosophila* Gene Collection Release 1 (DGCr1) was produced from six libraries representing various tissues, developmental stages, and the cultured S2 cell line. Nearly 80,000 random 5′ expressed sequence tags (5′ expressed sequence tags [ESTs]from these libraries were collapsed into a nonredundant set of 5849 cDNAs, corresponding to ~40% of the 13,474 predicted genes in *Drosophila*. To obtain cDNA clones representing the remaining genes, we have generated an additional 157,835 5′ ESTs from two previously existing and three new libraries. One new library is derived from adult testis, a tissue we previously did not exploit for gene discovery; two new cap-trapped normalized libraries are derived from 0–22-h embryos and adult heads. Taking advantage of the annotated *D. melanogaster* genome sequence, we clustered the ESTs by aligning them to the genome. Clusters that overlap genes not already represented by cDNA clones in the DGCr1 were analyzed further, and putative full-length clones were selected for inclusion in the new DGC. This second release of the DGC (DGCr2) contains 5061 additional clones, extending the collection to 10,910 cDNAs representing >70% of the predicted genes in *Drosophila*.

[The sequence data described in this paper have been submitted to the GenBank data library under accession nos. BF485518-BF503517, BF503521-BF506780, BG631888-BG631996, BG633696-BG637540, BG640063-BG641469, BI141709-BI142246, BI161485-BI173971, BI212109-BI216987, BI227448-BI233322, BI234009-BI243989, BI351612-BI354228, BI354231-BI355901, BI355935-BI358751, BI361285-BI376197, BI481532-BI487261, BI563331-BI593695, BI604243-BI620155, BI620158-BI635012, BI635064-BI638027, and BI638030-BI642053. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: J. Pringle and M. Fuller.]

The identification of all expressed genes and the structure(s) of their transcripts are prerequisites for many structural and functional genomic studies. Gene-finding programs are valuable tools for identifying gene structure, but they are error-prone and suffer from the inability to predict untranslated regions (UTRs) (Ashburner 2000; Reese et al. 2000). Direct analysis of gene transcripts is the only proven way to establish gene structures with confidence. Generating a collection of expressed sequence tags (ESTs) from high quality cDNA libraries is a widely used approach for acquiring this information (Adams et al. 1991). The sequences of ESTs and full-length nonredundant cDNA collections provide ideal tools for genome annotation and for the further training of gene prediction algorithms. Our first *D.melanogaster* EST project yielded putative full-length clones corresponding to >5000 different genes (Rubin et al. 2000). This was accomplished by generating 79,636 5′ ESTs from libraries, derived from four different tissues and the Schneider-2 cultured cell line, that contained a high proportion of full-length clones. These 5′ ESTs were clustered by inter se comparison, and the clone that extended the farthest 5′ in each cluster was selected for further analysis. From these clones, 3′ ESTs were then generated, and any clone not containing a polyA tail or that was redundant with another selected clone was eliminated. This collection, the *Drosophila* Gene Collection Release 1 (DGCr1) comprises full-length clones from ~40% of the 13,474 genes predicted in *D. melanogaster*

To obtain cDNA clones for the remaining genes, we generated 5′ ESTs from another 157,835 clones. Because our goal is a collection of full-length cDNA clones, we require that the libraries from which the ESTs are generated have a high per-

[6]**Corresponding author.**
E-MAIL staple@fruitfly.org; FAX (510) 486-6798.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.269102.

12:1294–1300 ©2002 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

centage of full-length clones. Improvements in the methodology for cDNA library construction, such as the use of the reverse transcriptase-stabilizing additive trehalose and 5′ cap-trapping methods, have greatly increased the efficiency of generating full-length clones (Carninci et al. 1998; Sugahara et al. 2001). Normalization of cDNA libraries by decreasing the prevalence of clones representing abundant transcripts before sequencing can be used to increase gene discovery rates (Bonaldo et al. 1996; Carninci et al. 2000; Clark et al. 2001). Cap-trapped and normalized embryonic (RE) and head (RH) libraries were constructed and used to generate >115,000 new ESTs. Another requirement for generating a well-represented cDNA collection is to sample from many different tissue types and developmental stages. In addition to sequencing more clones from the S2 and ovary libraries, both of which appear to be good candidates for gene discovery and were not heavily targeted in creating the DGCr1, we also generated 23,215 ESTs from a non-normalized library derived from the adult testis (the AT library). Previous studies (Andrews et al. 2000) indicated that *Drosophila* testes are a rich source of novel ESTs.

All ESTs in our collection were aligned to Release 2 of the *D. melanogaster* genomic sequence (Adams et al. 2000) using the cDNA alignment tool `Sim4` (Florea et al. 1998). A stringent clustering algorithm using the coordinates from the `Sim4` alignments identified 5061 additional putative full-length clones. The DGC now consists of 10,910 cDNA clones and is estimated to contain cDNAs for >70% of the predicted genes in the Release 2 annotation of the *D. melanogaster* genome.

RESULTS AND DISCUSSION

Libraries and Sequencing

To increase the gene discovery rate for the project, we sequenced clones from three new libraries and two of our original libraries. One of the new libraries is the AT library, which was made from dissected adult male testes and seminal vesicles. The other two new libraries are the RE (0–22-h embryos) and RH (adult head) libraries, which were constructed using 5′ cap-trapping and normalization methodologies. We also sequenced additional clones from each of the original GM (ovary) and S2 libraries because analysis of the limited number of ESTs previously generated from these libraries indicated that they were likely to yield cDNAs for genes not already represented in the DGCr1. The single-pass sequencing for these directionally cloned libraries was from the 5′ end of the cDNA, except for 1282 randomly selected 3′-end sequences that were used to confirm that nearly all clones in each library extended to the polyA tail (Table 1).

A total of 23,215 high quality 5′ reads from the AT library were generated and deposited in GenBank. Analysis of the initial 400 clones indicated a high potential gene discovery rate because 23% of these ESTs did not overlap with the 80,000 ESTs previously sequenced, although some of these 5′ ESTs represent alternative germ-line-specific transcriptional start sites used for genes already represented by other ESTs. In contrast, the RE and RH libraries had only ~10% nonoverlap with previous ESTs. A total of 60,254 high quality ESTs from the RE library and 54,915 from the RH library were submitted to GenBank. The AT, RH, and RE libraries were all judged to be of high quality by a number of criteria described in Table 1 and have clones with average insert sizes of 2.0, 1.6, and 2.1 kb, respectively. An additional 5281 ESTs were generated from the GM library and 14,170 from the SD library. In total, we generated 157,835 ESTs to add to the Berkeley *Drosophila* Genome Project (BDGP) EST collection.

EST Alignments and Clustering

Aligning ESTs by inter se sequence comparison allows the grouping of clones derived from the same transcript and is the most common method used to estimate the number of unique transcripts represented in an EST collection. We used this approach in our first EST project to produce the DGCr1; Rubin et al. 2000). 5′ ESTs were grouped using `BLAST` (Altschul et al. 1997) and then assembled using `Phrap` (Ewing and Green 1998; Ewing et al. 1998). The 3′-end sequence was then determined from the clone that extended the most 5′ in each cluster. Comparison of the 3′ ESTs was used as a final

Table 1. cDNA Library and EST Characterization

	Riken embryo	Riken head	Adult testis
Libraries			
cDNA cloning vector	pFLC1	pFLC1	pOTB7
PolyA presence	100% (n = 485)	92% (n = 352)	99% (n = 445)
Inverted inserts[a]	0% (n = 454)	0% (n = 355)	0.7% (n = 706)
Average insert length[b]	2.1 kb (n = 96)	1.6 kb (n = 96)	2.0 kb (n = 96)
Chimeric insert[c]	<1% (n = 488)	1.6% (n = 668)	2.8% (n = 313)
Initial gene discovery rate[d]	10%	9%	23%
ESTs			
Attempts	71807	67870	29664
Failed quality[e]	10181	11731	6146
Contaminant[f]	1372	1224	303
Total high quality[g]	60254	54915	23215
Average high quality read length	484	472	528

[a]Determined by the presence of a polyA tract in the 5′-end sequence.
[b]Determined by PCR amplification using primers in the cloning vector.
[c]Clones whose 5′ and 3′ reads aligned to different chromosomal arms or >300 kb apart using `Sim4`.
[d]Originally determined by pairwise `Blast` using all previous ESTs.
[e]Reads of <150 bp after vector and quality trimming.
[f]Reads that were discarded because of significant hits to the Genbank GB.vector dataset.
[g]See Methods for details.
EST, expressed sequence tag.

check for redundancy because 5′ ESTs from incomplete cDNA clones often do not overlap with ESTs from full-length clones derived from the same transcript.

Sim4 is a software tool designed for aligning ESTs or cDNAs with genomic sequence allowing for gaps at the positions of introns. Using Sim4 to scan the entire genome for sequences matching each EST proved impractical because of the long compute time required. We first used BLAST to align the ESTs with 100-kb genomic segments and then further characterized the transcript structure of each EST using Sim4. We successfully aligned 213,238 ESTs to the genome using a high stringency cutoff that required >90% of the EST sequence to be aligned.

We used an iterative approach to cluster ESTs that was based on their alignments with the genomic sequence. If any exons in an EST alignment overlapped exons within an existing cluster, the EST was incorporated into that cluster and the intron–exon structure(s) of the cluster updated. Because one gene could nest in the intron of another gene, EST alignments were grouped on the basis of overlap between exons and not introns. Figure 1 shows a user interface we developed to view the structures of EST clusters. We subdivided EST clusters such that each subcluster contained only ESTs whose alignments had the same intron–exon structure. For 3′ ESTs, we placed that EST in the same subcluster as the 5′ EST from the same cDNA clone. For the EST cluster shown in Figure 1A, there are six subclusters (numbered SC2 to SC7) that have been color coded to indicate the number of their constituent ESTs. These subclusters were further combined into merged subclusters that most likely reflect the structure(s) of alternatively spliced transcripts (Fig. 1B). However, the data are insufficient to establish whether all the ESTs that make up a merged subcluster are in fact derived from a single splice form. Thus the gene models indicated by the merged subclusters, whereas consistent with the available data, are not proven and do not represent all possible models. Because our initial goal was to identify a single full-length cDNA clone for each gene, we used a very conservative process in selecting clones for the DGCr2, which is described in detail below. We selected only a single clone from each cluster for inclusion, irrespective of the number of subclusters. In the case illustrated in Figure 1B, we selected the clone corresponding to the EST extending the farthest 5′ in subcluster 2 (SC2). Our analysis identified cDNAs derived from apparent alternatively spliced transcripts for >20% of the genes with ESTs. These data and clones will be used in the future to isolate and further characterize cDNAs representing alternative splice forms.

This clustering approach identified 16,744 clusters from the 213,238 successful EST alignments. Annotations of the Release 2.0 *Drosophila* genome sequence are composed of gene predictions and previously characterized genes, designated as Curated Genes (CGs). Of the 16,744 clusters, 12,154 overlap 9664 different CGs. The remaining 4590 clusters do not directly overlap a CG. We expect these nonoverlapping clusters for two reasons. First, many 5′ ESTs are composed entirely of 5′-UTR sequences, whereas the majority of CGs result from gene-prediction algorithms that only predict open reading frames (ORFs) and thus do not include the sequences of the UTRs. For example, 13% of full-length sequences from our gene collection have a 5′ UTR >500 bp in length. The average sequence trace from the ABI-PRISM DNA sequencer (Applied Biosystems) is ~500 bp in length after quality trimming, so we would expect that some of these clusters would not overlap with a gene prediction. Second, 36% of the nonoverlapping clusters are >20 kb upstream of a CG and likely represent new genes that were not detected by gene-finding programs. Verifying that these nonoverlapping clusters represent authentic genes will require additional sequence and gene-expression information.

Genomic contamination or reverse transcription of unspliced mRNA precursors has the potential to corrupt gene collections. To evaluate our EST clusters for such artifacts, we examined separately the position, relative to CGs, of EST clusters whose alignments indicate that they contain spliced ESTs and those containing only unspliced ESTs. Of the 12,154 clusters that overlap a CG, 72% contain spliced ESTs. Those 1455 clusters that do not overlap a CG but lie within 5 kb upstream of a CG contained spliced ESTs 28% of the time, whereas only 18% of the 3135 clusters that lie >5 kb upstream of a CG contained spliced ESTs. To minimize the number of cDNAs derived from unspliced transcripts that were selected, we made the conservative decision to select ESTs only from clusters that contained spliced ESTs, unless the cluster directly overlapped a CG, in which case, we also accepted clusters comprised only of unspliced ESTs.

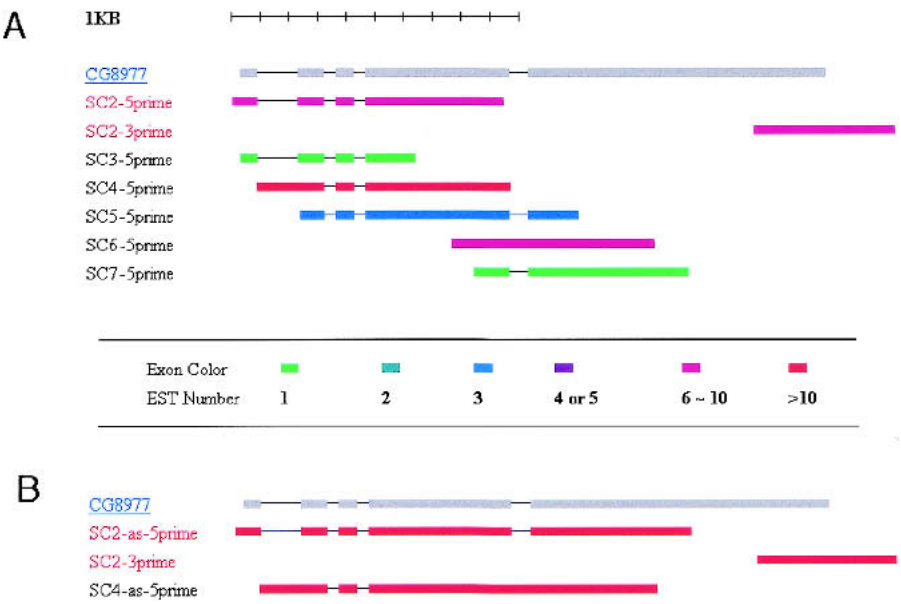

Figure 1 Graphical display of expressed sequence tag (EST) clusters. (*A*) is an example of six different subclusters aligning to the Curated Gene CG8977 and numbered SC2-SC7. The subclusters are color coded with respect to the number of EST members as shown. (*B*) is a gene model based on the merged subclusters illustrating two possible splice variants and numbered SC2-as and SC4-as. SC2-as is a merge of SC2, -3, -5, and -7 and SC4-as is a merge of SC4 and SC6.

Selecting DGC2

In the absence of direct overlap be-

tween an EST cluster and the closest downstream CG, it proved difficult to determine unambiguously if they represent the same gene. We developed the following rule-based approach to select clones for the DGCr2 in such cases. We defined a "CG region" as the genomic region extending from the 3′ end of a CG to the 3′ end of the adjacent upstream CG (Fig. 2A). In cases in which the intergenic distance was >5 kb, we arbitrarily limited the CG region to the span extending from the 3′ end of the CG to 5 kb upstream of its 5′ end and considered the remainder of the intergenic space as a distinct CG region (Fig. 2B). We made the further assumption that a CG region contains no more than one gene. The types of alignments observed for DGCr2 clusters and their frequencies are depicted in Figure 2, panels C through G. In 310 cases, clusters overlapped more than one CG as in Figure 2F. Such cases were evidence that original gene predictions erroneously split a single gene into two CGs.

After defining the CG regions, the rules for choosing clusters and the clone in each cluster for inclusion in DGCr2 are straightforward. CG regions that already had a DGCr1 clone aligned within them were omitted from the analysis. For the CG regions that had no DGCr1 clone aligned, we examined each cluster within the region progressing from 5′ to 3′. For regions that contained both a cluster directly overlapping the CG and a nonoverlapping cluster more 5′, we chose the most 5′ cluster if it contained a spliced EST. In cases in which the more 5′ cluster contained only unspliced ESTs, we selected the EST from the cluster that overlapped the CG. After a cluster was identified, we selected the EST alignment that extended the most 5′ within the cluster. Using these criteria, we identified 5061 new clones for the DGC. Preferentially picking the DGCr2 clones from clusters that directly overlap CGs might result in our choosing a significant percentage of clones that were not full length; however, this does not appear to be the case. We compiled a test set of 328 CGs represented by a DGCr2 clone and for which full-length cDNA sequences previously existed in GenBank. We asked whether the DGCr2 clone extended as far 5′ as the GenBank sequence and whether the DGCr2 clone contains the complete ORF. For the 53 cDNAs in our test set <1 kb in length, the corresponding DGC cDNA was as long or longer than the test set cDNA 77% of the time and contained the complete ORF 100% of the time. For the 101 cDNAs in the test set between 1 kb and 2 kb, these numbers were 71% and 99%; for those 74 cDNAs between 2 and 3 kb, these numbers were 66% and 88%; and for those 100 cDNAs >3 kb, these numbers were 59% and 91%. This analysis shows that larger genes are more likely to be represented as an incomplete cDNA. However, we conclude that our selection criteria did not significantly bias the DGC toward incomplete clones.

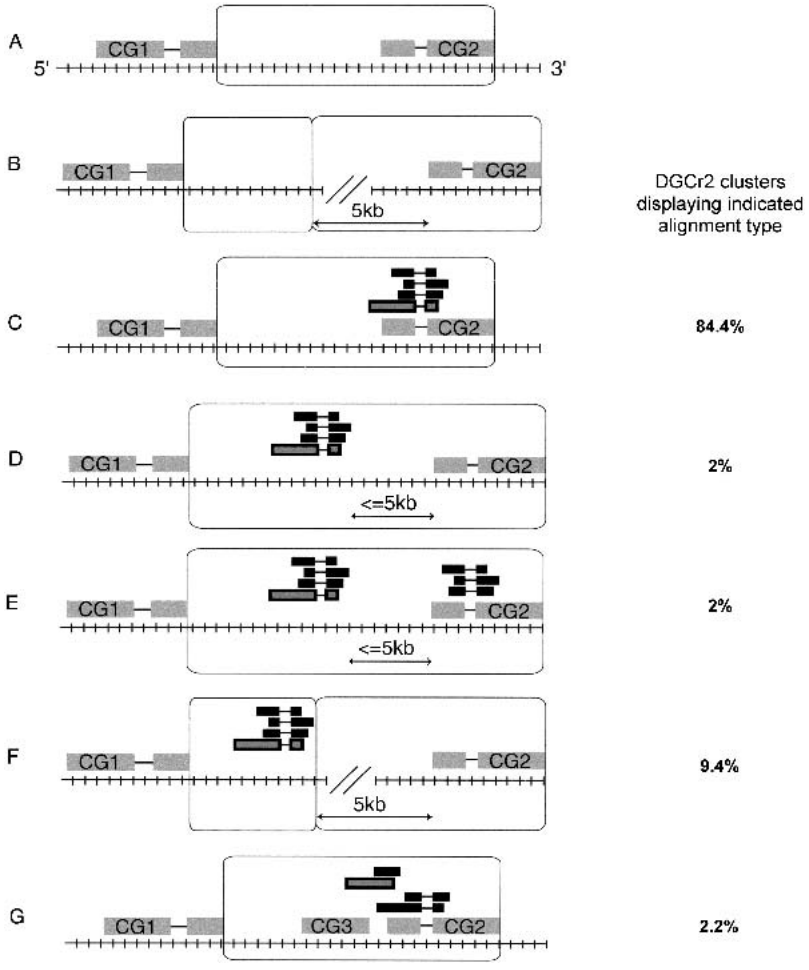

Figure 2 CG regions. The observed percentage of each alignment type for clusters representing DGCr2 is indicated in the right column. CG regions are represented as open boxes. CGs are shown numbered in gray boxes (exons) connected by lines (introns). The genomic sequence is shown as a hatched line. Aligned EST clusters are shown in black, and the ESTs chosen for DGC2 are gray with a black border. All CGs and alignments are shown representing one strand of the genome proceeding 5′ to 3′ as indicated in (*A*). See text for a full description of CG regions.

We were also concerned that the high initial gene discovery rate of the AT library, which had the consequence that 1000 (20%) of the new clones chosen for the DGCr2 are AT clones, was caused by testis-specific transcription start sites of genes also expressed in other tissues. However, we found that 768 of the 1000 AT clones in DGCr2 represented clusters consisting of only AT ESTs. Furthermore, 663 (86%) of these AT-only clusters did not contain ESTs from other tissues in the same CG region. Moreover, preliminary evidence from `GeneChip` Array (Affymetrix) experiments comparing RNA from male and female adults indicates that transcripts corresponding to 30% of these AT-only clusters are present at >25-fold higher levels in adult males (P. Spellman and G.M. Rubin, pers. comm.). Although a more rigorous analysis of transcript structures and expression

patterns will be necessary to address this question more directly, our initial analysis indicates that many of the testis clones in the DGC are abundantly expressed in testis and may be testis-specific genes.

Table 2 shows the number of clones from each library included in DGCr1 and DGCr2. Clones from the AT, RE, and RH libraries account for 76% of DGCr2, indicating the value of these new libraries in identifying clones for the DGC. Sequencing more clones from older libraries as well as realigning all ESTs from the original project also proved valuable. A total of 1205 (24%) DGCr2 clones are from the original EST collection; of those 1205 clones, 328 (27%) are new ESTs from the SD and GM libraries, and the other 882 (73%) are ESTs from the original sequencing effort. Thus, 17% of the new DGCr2 clones already existed but were not included in DGCr1. Many of these clones represent replacements for clones that were originally selected for the DGCr1 but which subsequently failed one of the quality control tests used to validate this collection (Rubin et al. 2000). Moreover, in selecting clones for the DGCr1 we generally did not include CGs represented by a single EST. Genome alignments allowed us to identify which of these singleton clusters contained a spliced EST and gave us confidence that the EST represented a legitimate transcript. In total, DGCr2 represents 4274 ESTs that overlap a single CG, 113 that overlap two adjacent CGs, 198 that are within 5 kb upstream of a CG, and 476 that are spliced but align >5 kb upstream of a CG. If the 10,910 clones in our combined DGCr1 plus DGCr2 collections do indeed represent unique genes, the DGC would consist of ~81% of genes predicted to exist in *D. melanogaster*.

Concluding Remarks

Large-scale EST projects are a vital first step toward cataloging the expressed portion of a genome. The BDGP EST resource has greatly accelerated progress toward our long-term goal of generating a comprehensive transcript map of all *Drosophila* genes by providing information on intron–exon structure, alternative splicing, and transcriptional start and stop sites. The FlyBase consortium (www.flybase.org) and the BDGP have begun a comprehensive reannotation of the *D. melanogaster* genome sequence and are making extensive use of the ESTs described here, as well as >5500 full-length cDNA sequences that the BDGP has so far produced from DGC clones. This information is used to verify predicted genes, identify previously unannotated genes, and refine existing gene models. EST and full-length sequences also provide important linking information that can be used to order and orient genomic sequence contigs from heterochromatic regions of the genome that cannot otherwise be assembled into contiguous sequences (Carvalho et al. 2000; Carvalho et al. 2001). It is most important that the resource is an entry point for functional genomic studies.

The methods we applied to cluster ESTs and select the DGCr2 are based on analysis of the genomic alignments of the EST sequences rather than alignments of the EST sequences to themselves. This approach has several key advantages over methods that rely solely on comparing ESTs to one another and takes advantage of the information provided by annotated genomic sequence. First, sequencing of 3′ ends was not necessary for detecting redundancy because EST clusters derived from the same gene could be identified even if they did not directly overlap in sequence. Second, EST alignments to the genome delineated intron–exon structures and allowed us to select cDNA clones that represented spliced transcripts, which helped to minimize contamination of the DGC by clones derived from genomic DNA or unspliced precursors. Third, these alignments revealed alternative splicing events.

We plan to extend the DGC in several ways in the future. In the course of reannotating the *Drosophila* genome sequence, human curators are examining the genomic alignment of ESTs and full-length cDNA sequences in relation to the output of gene-prediction algorithms and sequence-similarity searches. These individuals are identifying additional clones for inclusion in the DGC. Our selection rules for the DGCr2 excluded all genes for which a DGCr1 clone had been selected. The analysis of the full-length sequences of several thousand of these clones show that ~10% do not contain the complete ORF. Replacement clones for these ~500 members of the DGCr1 will be selected. Likewise, the curators will identify clones that correspond to alternative splice products that affect the coding potential of the transcript; our selection criteria for the DGCr2 were to pick only one cDNA per gene regardless of the evidence for alternative transcripts. We also excluded clones on the basis of EST clusters that did not either directly overlap a CG or contain a spliced EST. We expect that human curators will identify meaningful cDNAs for inclusion in the DGC from among these clusters. Finally, we plan to use high-throughput screening methods to identify clones in our cDNA libraries that correspond to predicted genes that are still not represented in the DGC.

METHODS

Library Creation

The AT (adult testis) library was made from *Drosophila* adult male testes RNA kindly provided by J. Pringle and M. Fuller. Testes and seminal vesicles were hand dissected from 0–3-d-old Ore-R males, and RNA was prepared by hot phenol extraction and selected twice for polyA+. cDNA was made using a `Stratagene` ZAP-cDNA synthesis kit substituted with SuperscriptII reverse transcriptase (Lifetech) using an oligo(dT)

Table 2. The *Drosophila* Gene Collection

Library	AT	GH	GM	HL	LD	LP	RE	RH	SD	Total
DGCr1	0	1867	467	137	2594	209	0	0	575	5849
DGCr2	1000	268	115	30	189	315	1947	909	288	5061
Total	1000	2135	582	167	2783	524	1947	909	863	10910

RNA for the libraries was obtained from the following sources: AT, adult male testes and seminal vesicles; GH, HL, RH, adult heads; GM, ovaries from stages 1–6; LD, RE, 0–22-h embryos; LP, mixed larval and early pupal stages; SD, Schneider S2 cell line.

primer with an XhoI site at the end for first-strand synthesis plus an EcoRI site adapter ligated onto the 5′ ends of clones. cDNAs were size fractionated on Sephacryl S-500 and inserts of 1–6 kb were directionally cloned into EcoRI/XhoI-digested pOTB7 plasmid.

The methods for making the RE and RH libraries have been described elsewhere (Carninci et al. 2000; Carninci et al. 2001). The RE (Riken embryo) library was made from RNA extracted from *Drosophila* 0–22-h mixed-stage embryos of the isogenic *y; cn bw sp* strain and polyA+ were selected twice. The RH (Riken head) library was made from RNA extracted from *Drosophila* adult heads, from the isogenic *y; cn bw sp* strain and polyA+ were selected once. cDNA for both RE and RH libraries was synthesized by priming with the oligo(dT) primed adapter (5′-GAGAGAGAGAGGATCCAATACTGGAG AGTTTTTTTTTTTTTTT-VN-3′). The first strand was synthesized in the presence of trehalose. Subsequently, full-length cDNA was selected with a biotinylated cap-trapper. A linker was then ligated to the single-strand cDNA (Shibata et al. 2001). The cDNA was normalized by using RoT = 1.0. Second-strand cDNA synthesis was primed with a (5′-AGAGAGAGAGCTCGAGCTCTAATAAGGTGACACTA TAGAACCA-3′) primer. After restriction digestion of the hemi-methylated cDNA with BamHI and XhoI, the cDNA was cloned into the lambda FLC-I vector. Subsequently, the library was bulk-excised into pFLC-I plasmid. Plasmid sequence and information concerning these vectors can be found at http://www.fruitfly.org/about/methods/index.html.

End Sequencing and Analysis

Random cDNA clones were end sequenced from the 5′ end using ABI Big Dye II Dye terminator chemistry. End sequencing of 3′ ends was performed using ABI d-rhodamine dye terminator chemistry on the AT, GM, and SD clones and ABI Big Dye II Dye terminator chemistry for the RE and RH clones. Primers for each of the cDNA cloning vectors are described in detail on the BDGP web page (http://www.fruitfly.org/about/methods/index.html). Reactions were run on ABI 3700 capillary sequence machines and chromatograms were evaluated using `Phred` (Ewing and Green 1998; Ewing et al. 1998). Reads were vector trimmed using `Crossmatch` (P. Green, http://bozeman.mbt.washing-ton.edu/phrap.docs/phrap.html). The ESTs were quality trimmed by evaluating chromatogram trace peaks using a *Java* program written by S. Lewis that implements an algorithm described previously (Hillier et al. 1996). The high quality cutoff was defined at the point in which either of the following conditions was true: The ratio between peak height and valley fell below a threshold or the ratio of the areas beneath the highest peak and the next highest fell below a threshold. This point was extended to the location where 4 consecutive bases with scores <q15 occurred. This marked the end of the "high quality" sequence. By further evaluating the `Phred` quality scores, the reads were then trimmed for submission at the base before a run of 6 bases with scores <q10. This marked the end of the submitted sequence. Reads were also screened for *Escherichia coli* transposons using the Genbank data set GB.vector. Reads that were >150 bp of contiguous high quality sequence were submitted to the National Center for Biotechnology Information (NCBI) and have accession numbers BF485518-BF503517, BF503521-BF506780, BG631888-BG631996, BG633696-BG637540, BG640063-BG641469, BI141709-BI142246, BI161485-BI173971, BI212109-BI216987, BI227448-BI233322, BI234009-BI243989, BI351612-BI354228, BI354231-BI355901, BI355935-BI358751, BI361285-BI376197, BI481532-BI487261, BI563331-BI593695, BI604243-BI620155, BI620158-BI635012, BI635064-BI638027, and BI638030-BI642053. Before high quality trimming, the average number of bases with phred quality ≥q20 for ESTs submitted to Genbank was 554. Clones corresponding to all ESTs, including those that make up DGCr2, are available from Research Genetics (http://www.resgen.com/products/DEST.php3). We are in the process of colony purifying and rearraying the DGCr2 clone set and expect it to be available for distribution by July of 2002.

Aligning and Clustering ESTs

Sequences of *Drosophila* chromosome arms were divided into 100-kb nonoverlapping genomic segments, and 237,471 5′ EST sequences were screened against these genomic segments using `WU-BLASTN` 2.0. The genomic segments showing ≥90% similarity to each EST were identified. 5′ EST sequences were then aligned to their corresponding genomic segments using `Sim4`. The ESTs that spanned two adjacent genomic segments were realigned to entire chromosome arm sequences using `Sim4`. Only ESTs that had ≥90% of their sequence traces aligned were analyzed further.

Successfully aligned 5′ ESTs were grouped on the basis of the `Sim4` alignments to the genomic using an iterative algorithm. Initially, all 5′ EST alignments were kept in a data set. As long as the data set was not empty, one EST alignment was randomly picked to construct an initial cluster, which was used to continually scan the entire data set. Whenever any exons in an EST alignment were found to overlap exons of a cluster, the exons of the cluster were updated and the EST alignment was assigned to the cluster. When the cluster was not updated in an entire scan, it was added to the data set as a new cluster. This process continued until there were no EST alignments left in the data set.

ACKNOWLEDGMENTS

We thank R. Hoskins, A. Huang, and S. Misra for critically reading and improving the manuscript. We thank H. Guarin for excellent technical assistance and the entire staff of the BDGP sequencing center. This work was funded by grants from the Department of Energy (grant no. DE-FG03-98 ER62625 and DE-FG03-99ER62739 and National Institutes of Health (grant no. P50 HG00750) to G.M. Rubin. This work was also supported by the Riken Genome Exploration Research grant to Yoshihide Hayashizaki.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. 1991. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252:** 1651–1656.

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped `BLAST` and `PSI-BLAST`: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Andrews, J., Bouffard, G.G., Cheadle, C., Lu, J., Becker, K.G., and Oliver, B. 2000. Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. *Genome Res.* **10:** 2030–2043.

Ashburner, M. 2000. A biologist's view of the *Drosophila* genome annotation assessment project. *Genome Res.* **10:** 391–393.

Bonaldo, M.F., Lennon, G., and Soares, M.B. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6:** 791–806.

Carninci, P., Nishiyama, Y., Westover, A., Itoh, M., Nagaoka, S., Sasaki, N., Okazaki, Y., Muramatsu, M., and Hayashizaki, Y. 1998. Thermostabilization and thermoactivation of thermolabile

enzymes by trehalose and its application for the synthesis of full length cDNA. *Proc. Natl. Acad. Sci.* **95:** 520–524.
Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K., Itoh, M., Konno, H., Okazaki, Y., Muramatsu, M., and Hayashizaki, Y. 2000. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* **10:** 1617–1630.
Carninci, P., Shibata, Y., Hayatsu, N., Itoh, M., Shiraki, T., Hirozane, T., Watahiki, A., Shibata, K., Konno, H., Muramatsu, M., et al. 2001. Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* **77:** 79–90.
Carvalho, A.B., Lazzaro, B.P., and Clark, A.G. 2000. Y chromosomal fertility factors kl-2 and kl-3 of *Drosophila melanogaster* encode dynein heavy chain polypeptides. *Proc. Natl. Acad. Sci.* **97:** 13239–13244.
Carvalho, A.B., Dobo, B.A., Vibranovski, M.D., and Clark, A.G. 2001. Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **98:** 13225–13230.
Clark, M.D., Hennig, S., Herwig, R., Clifton, S.W., Marra, M.A., Lehrach, H., Johnson, S.L., and WU-GSC EST Group. 2001. An oligonucleotide fingerprint normalized and expressed sequence tag characterized zebrafish cDNA library. *Genome Res.* **11:** 1594–1602.
Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using `Phred`. II. Error probabilities. *Genome Res.* **8:** 186–194.
Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using `Phred`. I: Accuracy assessment. *Genome Res.* **8:** 175–185.
Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8:** 967–974.
Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6:** 807–828.
Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., Abril, J.F., and Lewis, S.E. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* **10:** 483–501.
Rubin, G.M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M., and Harvey, D.A. 2000. A *Drosophila* complementary DNA resource. *Science* **287:** 2222–2224.
Shibata, Y., Carninci, P., Watahiki, A., Shiraki, T., Konno, H., Muramatsu, M., and Hayashizaki, Y. 2001. Cloning full-length, cap-trapper-selected cDNAs by using the single-strand linker ligation method. *Biotechniques* **30:** 1250–1254.
Sugahara, Y., Carninci, P., Itoh, M., Shibata, K., Konno, H., Endo, T., Muramatsu, M., and Hayashizaki, Y. 2001. Comparative evaluation of 5′-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries. *Gene* **263:** 93–102.

WEB SITE REFERENCES

http://bozeman.mbt; FlyBase consortium.
http://www.fruitfly.org/about/methods/index.html; Berkeley *Drosophila* Genome Project.
http://www.resgen.com/products/DEST.php3; Research Genetics.

Received March 13, 2002; accepted in revised form June 12, 2002.