DOI: 10.1534/genetics.106.065961

The Carnegie Protein Trap Library: A Versatile Tool for Drosophila Developmental Studies

Michael Buszczak,* Shelley Paterno,* Daniel Lighthouse,* Julia Bachman,* Jamie Planck,* Stephenie Owen,* Andrew D. Skora,* Todd G. Nystul,* Benjamin Ohlstein,* Anna Allen,* James E. Wilhelm,* Terence D. Murphy,* Robert W. Levis,* Erika Matunis,† Nahathai Srivali,* Roger A. Hoskins‡ and Allan C. Spradling*,[1]

**Howard Hughes Medical Institute Research Laboratories, Department of Embryology, Carnegie Institution of Washington, Baltimore, Maryland 21218, †Department of Cell Biology, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205 and ‡Department of Genome Biology, Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720*

Manuscript received September 18, 2006
Accepted for publication December 18, 2006

ABSTRACT

Metazoan physiology depends on intricate patterns of gene expression that remain poorly known. Using transposon mutagenesis in Drosophila, we constructed a library of 7404 protein trap and enhancer trap lines, the Carnegie collection, to facilitate gene expression mapping at single-cell resolution. By sequencing the genomic insertion sites, determining splicing patterns downstream of the enhanced green fluorescent protein (EGFP) exon, and analyzing expression patterns in the ovary and salivary gland, we found that 600–900 different genes are trapped in our collection. A core set of 244 lines trapped different identifiable protein isoforms, while insertions likely to act as GFP-enhancer traps were found in 256 additional genes. At least 8 novel genes were also identified. Our results demonstrate that the Carnegie collection will be useful as a discovery tool in diverse areas of cell and developmental biology and suggest new strategies for greatly increasing the coverage of the Drosophila proteome with protein trap insertions.

THE central challenge of postsequence genomics is to learn how an enhanced knowledge of genes, transcripts, and proteins can be applied to better understand the biology of multicellular organisms. Gaining an accurate picture of where and when metazoan genes are expressed remains a prerequisite for many such advances (Stathopoulos and Levine 2005). The discovery of distinctive, regulated programs of gene expression at a fine scale has the potential to reveal new cell types and substructures that make up tissues and the biological processes that govern their function. However, sensitive and widely applicable methods will be required to detect and distinguish developmentally programmed gene expression changes from those caused simply by cell cycling or environmental perturbation.

Several methods for analyzing gene expression within tissues are currently available. Particular cell types can sometimes be cultured *in vitro* into populations of useful size. However, isolated cells in artificial media frequently behave differently from cells *in vivo* interacting with precisely positioned neighbors in three-dimensional microenvironments. Another approach is to isolate tissue cells by flow sorting, microdissection, or laser capture and then determine their expression profiles in depth (reviewed in Espina *et al.* 2006). Visualizing patterns of gene expression within the intact tissues of transgenic organisms containing gene expression reporters may be the most general method (Tomancak *et al.* 2002). Epitope tagging, enhancer trapping, and gene trapping all have the added advantage that gene expression can subsequently be observed in living tissues, revealing dynamic processes that are largely beyond the reach of methods based on fixed material (reviewed in Herschman 2003; Dirks and Tanke 2006).

Protein trapping is a variation of gene trapping in which endogenous genes are engineered to produce under normal controls protein segments fused to a reporter such as GFP. The great potential of this technology has been extensively documented in yeast, where large collections of strains that each trap a different gene have been generated using transposable elements (Ross-Macdonald *et al.* 1999) or by homologous recombination (Huh *et al.* 2003). Extensive gene and protein trapping has also been carried out in cultured embryonic stem (ES) cells (Gossler *et al.* 1989; Friedrich and Soriano 1991), where fusions with more than half of annotated mouse genes have been recovered (see Skarnes *et al.* 2004). However, relatively few of these ES cell lines have so far been used to generate corresponding mouse strains where the versatility and sensitivity of the method for analyzing tissue structure can be tested.

[1]*Corresponding author:* Carnegie Institution, 3520 San Martin Dr., Baltimore, MD 21218. E-mail: spradling@ciwemb.edu

Large-scale protein trap screens may also reveal new information about genome structure and function. Identifying in an unbiased manner locations throughout a genome where a coding exon can be expressed tests the accuracy and completeness of its annotation. Characterizing the splicing patterns that lead to normal or aberrant GFP expression tests the current catalog of transcript isoforms generated by alternative splicing. Moreover, by recovering insertions in the same gene that splice differently and produce GFP with varying efficiency, such a project might generate a data set useful for studying the determinants of splice site selection and transcript stability.

Drosophila provides a favorable system for applying gene traps to diverse developmental and genomic studies. The genome sequence has been extensively annotated on the basis of experimental data (Misra *et al.* 2002). Thousands of enhancer trap lines have been generated in large-scale transposon screens and culled of redundant strains by the gene disruption project (see Bellen *et al.* 2004). In contrast, producing Drosophila protein trap lines has remained difficult. Several hundred such lines were generated using a mobile GFP-containing exon flanked by both splice acceptor and donor sites (Morin *et al.* 2001; Clyne *et al.* 2003). However, the process was highly inefficient, with as few as 1 in 1500 progeny flies expressing GFP. Positive lines often contained more than one insertion, preferentially tagged a small number of hotspot loci, and tagged many sites not predicated to fuse the GFP exon in frame to any known coding region (Morin *et al.* 2001). Kelso *et al.* (2004) found that the recovery of lines could be increased by using an automated embryo sorter to select GFP-positive embryos and established a website, FlyTrap, to gather information on Drosophila protein trap lines. Consequently, we initiated a large-scale protein trap screen to increase gene coverage, test the genome annotation, and address some of the remaining technical difficulties in efficient line production.

Here we report the production of lines that trap 600–900 Drosophila genes, including 244 where one or more trapped proteins can currently be identified. Using the Drosophila ovary as a test system we confirm that protein trap lines reveal fine-scale details of developmentally regulated protein expression, making them exceptionally valuable discovery tools for a wide range of studies. Finally, mapping RNA splicing patterns downstream from >1200 insertions provides insight into how an added exon affects splicing and suggests how the production of protein trap lines can be expanded to cover a larger fraction of the Drosophila proteome.

MATERIALS AND METHODS

Generation of *P*-element lines for protein trap screening: The *P*-element-based protein trap screens presented here utilized the pPGA, pPGB, and pPGC vectors described in Morin *et al.* (2001). These elements carry a mini-white transgene in the opposite orientation to an enhanced green gluorescent protein (EGFP) exon, which is composed of EGFP sequence, without start or stop codons, flanked by splice acceptor and donor sites from the Drosophila MHC locus. A, B, and C refer to the position of the splice sites within the first and last codons of the EGFP exon sequence. Previously used pPGA, pPGB, and pPGC third chromosome insertions (Morin *et al.* 2001) were remobilized in the presence of balancer chromosomes. New insertions that mapped to the *CyO* balancer chromosome, did not express EGFP, and exhibited remobilization rates off of the *CyO* balancer of at least 60% in single-pair mating assays were recovered and used as starting stocks in the screen (see below).

***piggyBac* protein trap vectors:** To make a shuttle vector for subcloning the EGFP exon into different transposable elements, the entire EGFP exons from the pPGA, pPGB, or pPGC plasmids were excised from the original *P*-element plasmids (kind gift of W. Chia) using *Eco*RI and *Pst*I and subcloned into pBluescript (Stratagene, La Jolla, CA). These plasmids were then cut with *Eco*RV and *Kpn*I, end filled using Klenow, and religated to themselves to create pBS-GFPA, pBS-GFPB, and pBS-GFPC. The resulting plasmids carry the EGFP exon sequence between a unique *Eco*RI site at the 5′ end and unique *Pst*I, *Sma*I, *Bam*HI, and *Xba*I sites at the 3′ end. New tagging sequences can be inserted between the splice acceptor and donor sites of the exon using unique *Nco*I and *Xho*I sites.

Two different *piggyBac* protein trap vectors (Figure 1) were constructed using pBac{D. m. w^+} (Handler and Harrell 1999) (kind gift of A. Handler). The pBac{D. m. w^+} plasmid was digested with *Cla*I to remove most of the *mini-white* sequence and a linker containing *Hpa*I, *Xho*I, and *Spe*I sites was inserted in its place to form pBAC{Δ*Cla*I}. To create pBAC{*Bgl*II-GFP}, the EGFP exons from pBS-GFPA, pBS-GFPB, and pBS-GFPC were subcloned into the *Bgl*II and *Mfe*I sites of pBAC{Δ*Cla*I}. To create pBAC{*Hpa*I-GFP}, EGFP exon sequences were inserted between the *Mfe*I and *Hpa*I sites of pBAC{Δ*Cla*I}. An intronless *yellow* transgene from the *yellow*-BSX plasmid (Bellen *et al.* 2004) was then subcloned into the unique *Spe*I site of both pBAC{*Bgl*II-GFP} and pBAC{*Hpa*I-GFP} to form either pBAC{*Bgl*II-GFP; y^+}, which has the EGFP exon and *yellow* transgene oriented away from each other, or pBAC{*Hpa*I-GFP; y^+}, which has the EGFP exon and *yellow* transgene pointing toward each other (Figure 1A). Both pBAC{*Bgl*II-GFP; y^+} and pBAC{*Hpa*I-GFP; y^+} vectors carrying the EGFP exon in the A frame were transformed into *y w* flies using the phspBac helper plasmid (Handler and Harrell 1999) (kind gift from A. Handler).

We created stable genomic sources of the *piggyBac* tranposase using *P*-element transformation vectors. To place the *piggyBac* transposase under control of the ubiquitin promoter, the *piggyBac* transposase ORF was excised from phspBac using *Bam*HI and *Dra*I and ligated into the *Bam*HI and *Sma*I sites of the pCasper3-Up2-RX poly(A) *P*-element vector (Ward *et al.* 1998) (kind gift of R. Fehon), which carried a modified multiple cloning site (kind gift of A. Hudson), to form pP{UbpBACtrans}. To make an inducible *piggyBac* transposase source, phspBac was digested with *Eco*RI and *Dra*I and the fragment containing both the Drosophila *hsp70* promoter and *piggyBac* transposase ORF was ligated into the *Eco*RI and *Stu*I sites of pCasper4 to form pP{*hsp70*-pBACtrans}. These vectors were used to transform *y w* flies, using standard *P*-element transformation techniques.

To test the activity of the *piggyBac* transposase transgenes, single-pair matings were set up using the pBAC{*Hpa*I; y^+}24.3 insertion, which mapped to the X chromosome, and pP{UbpBACtrans} or pP{*hsp70*-pBACtrans} stocks. The pBAC{*Hpa*I; y^+}24.3 insertion was mobilized in males that were then outcrossed to *y w* females. Phenotypically *yellow*+ males in

the next generation were scored as new insertions. The pP{*hsp70*-pBACtrans} was able to remobilize the pBAC{*Hpa*I; y^+}24.3 insert in 43% ($n = 30$) of single-pair matings tested whereas the pP{Ub-pBACtrans} was able to remobilize the pBAC{*Hpa*I; y^+}24.3 insert in 48% ($n = 21$) of single-pair matings tested. The pBAC{*Hpa*I; y^+}24.3 insertion was remobilized in the presence of a *CyO* balancer chromosome. New insertions that did not express EGFP and mapped to the *CyO* chromosome were used in the pBAC-based protein trap screen.

Generation of embryos with novel transpositions: We isolated new EGFP-expressing *P*-element insertions using the following genetic scheme:

$$F_1: \frac{yw}{Y};\frac{Sco}{CyO\ pPG\{w+;\ GFP\ exon\}};\frac{+}{+}\times\frac{yw}{yw};\frac{+}{+};\frac{Ki, P\{ry^{+t7.2}, \Delta 2\text{-}3\}99B}{Ki, P\{ry^{+t7.2}, \Delta 2\text{-}3\}99B}$$

$$F_2: \frac{yw}{Y};\frac{+}{CyO\ pPG\{w+;\ GFP\ exon\}};\frac{+}{Ki, P\{ry^{+t7.2}, \Delta 2\text{-}3\}99B}\times\frac{yw}{yw};\frac{+}{+};\frac{+}{+}$$

Sort embryos.

New pBAC insertions were generated through a similar genetic scheme:

$$F_1: \frac{yw}{Y};\frac{Sco}{CyO\ pBAC\{y+;\ GFP\ exon\}};\frac{+}{+}\times\frac{yw}{yw};\frac{+}{+};\frac{P\{w+, pBACtrans\}}{P\{w+, pBACtrans\}}$$

$$F_2: \frac{yw}{Y};\frac{+}{CyO\ pBAC\{y+;\ GFP\ exon\}};\frac{+}{P\{w+, pBACtrans\}}\times\frac{yw}{yw};\frac{+}{+};\frac{+}{+}$$

Sort embryos.

Hereafter, *P*-element and *piggyBac* element-based protein trap lines were treated the same. For the F_1 cross several hundred males and females of the appropriate genotypes were mated in bottles to produce several thousand males in which the elements were mobilized for the F_2 cross. These males were crossed to 8000–10,000 virgin *y w* females in a population cage. These virgin females were obtained using a virgining stock that carried a heat-shock-inducible *hid* transgene on the *Y* chromosome (kind gift of R. Lehmann). Two separate overnight embryo collections from each population cage were screened for EGFP expression. We limited the number of times we screened embryos from a particular cage to try to minimize the number of identical insertions recovered due to premeoitic insertion events.

Embryo sorting and line establishment: We screened for EGFP expression in embryos using a COPAS Drosophila embryo sorter (Union Biometrica). Embryos were dechorionated in 50% bleach for 2.5 min and washed extensively with water. Dechorionated embryos were then washed into sorting solution (0.5× PBS, 2% Tween-20). With the exception of the sorting solution, the COPAS sorter was used according to the manufacturer's protocol, using the manufacturer's solutions. The sorter and sample pressures of the COPAS machine and embryo density were maintained so that the COPAS sorter screened 15–20 embryos/sec. The sorter used a 488, 514 nm multiline argon laser. EGFP fluorescence was detected using PMT1 set to 510 nm. Red fluorescence, used as a measure of embryo autofluorescence, was detected using a second PMT set to 580 nm. Baseline values for each fluorescent axis were set empirically using previously isolated fly strains that express low levels of EGFP and *y w* non-EGFP expressing embryos (Figure 1). Approximately 250,000 embryos were sorted in five 50,000-embryo batches per day. Sorted embryos were collected and washed in dH_2O. All the embryos from a single batch were placed together in standard food vials. We estimate that ~80% of the sorted embryos survived to adulthood. Sorted flies that survived to adulthood and did not carry the *Ki, P{ry*$^+$*t7.2; Δ2-3}99B* or *P{w*$^+$*; pBACtrans}* chromosomes were individually outcrossed to a *y w* stock. New lines that carried EGFP-expressing insertions that did not map to the starting *CyO* chromosome were maintained as stocks.

DNA sequencing, RT–thermal asymmetric interlaced PCR analyses, and prediction of fusion potential: Genomic sequences flanking either *P*-element or *piggyBac* protein trap insertions were determined by members of the Lawrence Berkeley Lab group using an established protocol for sequencing inverse PCR products from genomic DNA (Bellen *et al.* 2004). Database software developed for the annotation of the Drosophila gene disruption project (Bellen *et al.* 2004) was used to manage the sequence data. Once the insertion site of a given protein trap line was determined, a FileMaker Pro database that contained information [version 3.2 of the Drosophila genome annotation (Misra *et al.* 2002)] for all Drosophila transcripts, exons and introns, and their reading frames was used to predict which gene(s) and transcript(s) were being trapped by a given protein trap insertion.

We developed a reverse transcriptase coupled thermal asymmetric interlaced PCR (RT–TAIL) protocol largely on the basis of methods used to determine T-DNA insertion sites in Arabidopsis (Singer and Burke 2003). This method allowed us to determine the mRNA sequence adjacent to the EGFP exon without using gene-specific primers. Total RNA was isolated from 15 adult flies using an RNAqueous-96 automated kit (catalog no. 1812; Ambion, Austin, TX). The samples were ground in 200 μl of sample buffer and spun for 5 min at 14,000 rpm. The supernatant was placed in a 96-well plate and 100 μl of 100% EtOH was mixed with each sample. The sample was transferred to the filter plate, washed, and then treated with Dnase I (Ambion) for 15 min. Rebinding buffer was added to each well of the filter plate, and the plate was washed extensively. The RNA was eluted off the filter and precipitated with 7.5 M LiCl solution (Ambion). The resulting RNA pellet was washed with 75% EtOH and then retreated with Dnase I for 30 min at 37°. Dnase inactivation reagent (Ambion) was added to the samples. The RNA samples were spun and the supernatant was transferred to a new plate. A detailed protocol is available upon request.

The following GFP-specific primers were used for RT–TAIL PCR:

GFP-For1, 5′-GGAGGACGGCAACATCCTGG-3′;
GFP-For2, 5′-CAACGTCTATATCATGGCCG-3′;
GFP-For3, 5′-AGACCCCAACGAGAAGCGCG-3′;
GFP-Rev1, 5′-GTCGTGCTGCTTCATGTGGTCG-3′;
GFP-Rev2, 5′-GACACGCTGAACTTGTGGCCG-3′;
GFP-Rev3, 5′-AGCTCCTCGCCCTTGCTCACC-3′ .

The arbitrary degenerate (AD) primers used in this study were originally described by Singer and Burke (2003) but are listed here for convenience:

AD3, 5′-AGWGNAGWANCAWAGG-3′;
AD4, 5′-STTGNTASTNCTNTGC-3′;
AD5, 5′-NTCGASTWTSGWGTT-3′;
AD6, 5′-WGTGNAGWANCANAGA-3′.

A pool of the AD primers was mixed according to Singer and Burke (2003).

The first round of RT–TAIL PCR was set up in 96-well format using a one-step RT–PCR kit (QIAGEN, Valencia, CA). For every reaction 5 μl of total RNA was mixed with 10 μl 5× buffer, 2 μl 10 mm dNTP solution, 1 μl GFP-For1 or -Rev1 primer, 12.5 μl AD primer mix, 2 μl enzyme mix, and 17.5 μl of dH_2O. The reverse transcription reaction was carried out at 50° for 30 min. The sample was then heated to 95° for 15 min and then cycled for primary TAIL–PCR according to Singer and Burke (2003), using a MJ Research (Watertown, MA) thermal cycler. The secondary and tertiary TAIL–PCR reactions were carried out according to Singer and Burke (2003), using GFP-For2 or -Rev2 primers and GFP-For3 or -Rev3 primers, respectively, and regular TAQ DNA polyermase (Roche, Indianapolis). The PCR products of the tertiary reaction were treated with exoSAP (United States Biochemical, Cleveland) and sequenced using GFP-For3 or -Rev3 primers.

The RT–TAIL PCR protocol using the three GFP-For primers, which amplified off of the 3′ end of EGFP, consistently yielded better results than the same reaction using the Rev primers. Therefore most of the RT–TAIL PCR data define splicing products at the 3′ end of the EGFP sequence. To identify sequence fusing to the 5′ end of EGFP, we employed a 5′ RACE kit according to the manufacturer's protocol (Ambion), using the EGFP reverse primers listed above.

Analysis of protein expression in tissues: Samples were dissected in Grace's medium, placed in 48-well plates outfitted with a nylon mesh bottom, and fixed in 4% paraformaldehyde buffered in 1× PBS for 10 min at room temperature. The plate was washed extensively with PBT (1× PBS, 0.5% Triton X-100, 0.3% BSA) and incubated overnight at 4° with rabbit anti-GFP antibody (Torrey Pines) (1:2000) in PBT. The samples were then washed extensively with PBT and incubated with goat anti-rabbit Alexa488 (Molecular Probes, Eugene, OR) (1:400) for 4 hr at room temperature. The samples were then washed with PBT, stained with 2 μg/ml of DAPI, and mounted in Vectashield (Vector Laboratories, Burlingame, CA). Images were collected using a Leica SP2 confocal microscope.

RESULTS

Generating a large initial collection of tagged strains expressing EGFP: Our initial strategy was to generate a much larger number of lines containing new protein trap vector insertions than in previous screens and to institute additional technical improvements. Because of their proven utility, we used the same *P*-element-based protein trap vectors employed by Morin *et al.* (2001), but we also constructed a similar set of vectors with *piggyBac* (Figure 1A). As described in materials and methods, we set up crosses in small population cages to limit the recovery of clusters, utilized dominant markers to remove the transposase source from all new lines, and identified rare GFP-expressing embryos rapidly and sensitively using an automated embryo sorter (Figure 1B). This protocol allowed us to screen >60 million embryos over a period of 2.5 years, to identify >7500 "green" embryos, and to use each one to start an individual culture (see Table 1). Ultimately, 7404 strains were successfully established, maintained by selection for *white*+ eye color, and analyzed further as diagrammed in Figure 1C.

The same scheme was used with both transposons; however, in practice the *piggyBac* vectors were not nearly as efficient at generating EGFP-positive candidate lines as the *P*-element vectors (Table 1). *P*-element vectors typically exhibited 70% mobilization and generated ~1 EGFP-expressing embryo per 1000 sorted. In comparison, the *piggyBac* vectors displayed nearly 50% mobilization, but they yielded only 1 EGFP-expressing embryo per 50,000 sorted. Thus, the *piggyBac* vectors were slightly less efficient at mobilization, but drastically less efficient at generating EGFP-positive lines upon insertion. Consequently, we soon abandoned attempts to generate large numbers of *piggyBac* protein trap insertions (Table 1), but continued to characterize the lines we did recover to learn if they would shed any light on the lower frequency of trapping observed.

Localizing insertions on the annotated genome: To identify candidate proteins that may have been fused within individual lines, we determined the genomic DNA sequence flanking the insertion(s) in collaboration with the Berkeley Drosophila Genome Project (BDGP) gene disruption project (materials and methods). In most cases, the sequences from both the 5′ and 3′ vector end junctions mapped by BLAST analysis to a unique insertion site within the Drosophila genome sequence. Lines for which the sequencing reaction failed, the sequence matched repetitive DNA, or the 5′ and 3′ sequences differed (indicating that two or more insertions were present in the stock) were recycled back into the starting pool, and frequently a unique single insert was eventually identified. Altogether the insertions in 1375 C frame, 3172 B frame, and 1009 A frame *P*-element and 164 *piggyBac* A frame lines were localized to unique genomic sites.

Knowing the genomic location of an insertion allowed us to predict which transcripts would incorporate the EGFP exon and whether they would undergo splicing and translation into a functional fusion protein. First, we removed ~1550 duplicate lines derived from premeiotic clusters that were identified because they bore insertions identical in position and orientation to those in one or more sibling lines. Of the 4170 independent lines remaining, 2149 (52%) were associated with a gene correctly oriented for possible fusion (*i.e.*, located between −500 and the 3′ end). We also classified the ways an insert can be located relative to its closest annotated transcript into general categories as diagrammed in Figure 1D and classified all the lines (Table 2).

Expression of a fusion protein is expected when the GFP exon resides between two coding exons within an intron of matching reading frame (class 1A). Such insertions made up only 23% of the total localized insertions and defined 192 different genes (Table 4). Forty percent of insertions were close to an annotated gene but were not predicted to express the EGFP exon (classes 2–4), while 37% of the lines were not located within 0.5 kb of a correctly oriented gene. EGFP production from these lines might be explained by the use of unannotated

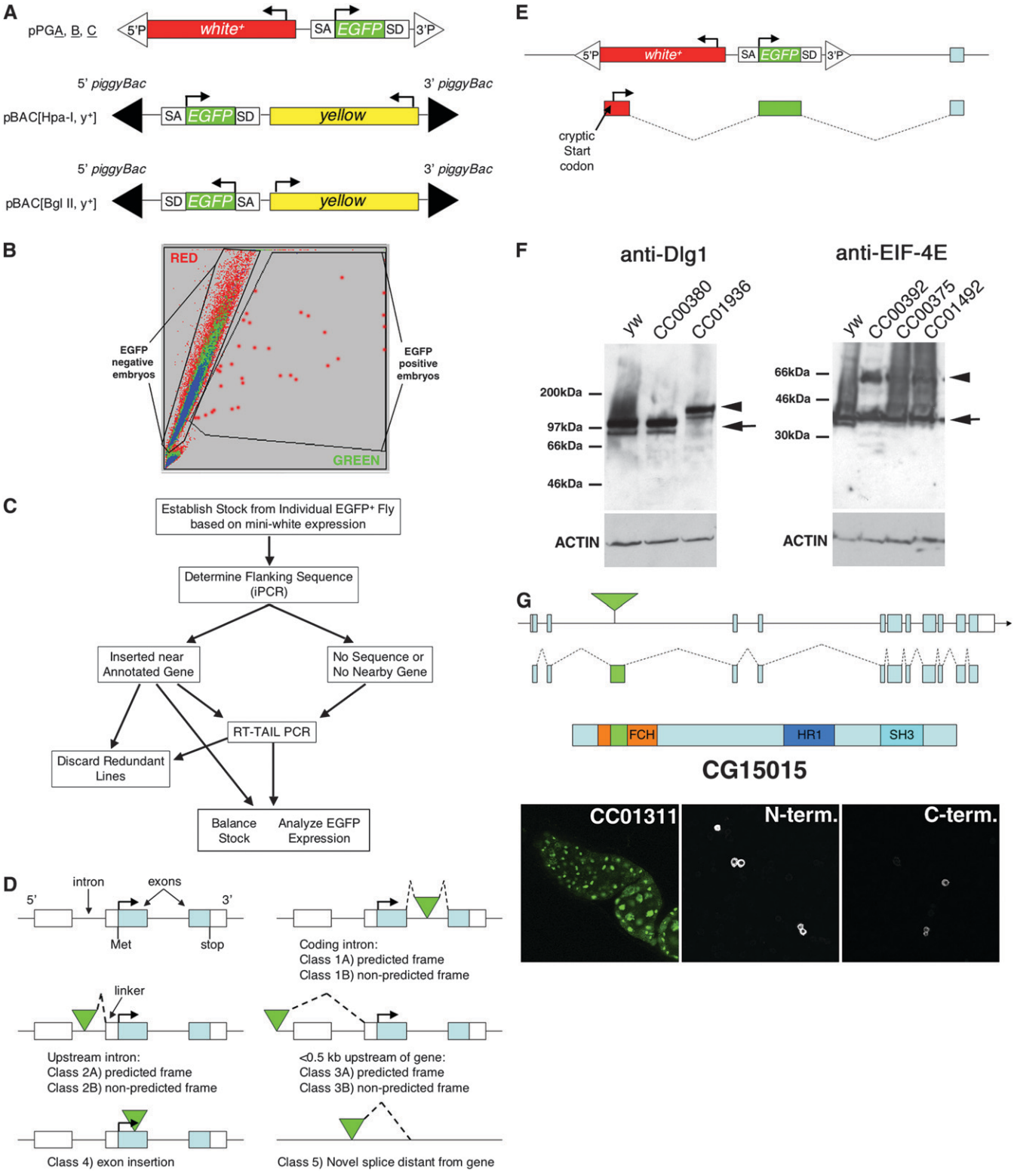

FIGURE 1.—Generation and classification of protein trap vector insertions. (A) Schematic of protein trap vectors (after MORIN *et al.* 2001). (B) Sample output from automated sorting of Drosophila embryos mobilized from site not expressing GFP. Rare GFP+ embryos (red circles) registering above a threshold value are diverted by the machine and later used to start individual cultures. (C) Scheme for characterization of putative protein trap lines (see text). (D) Classification of the general types of relationships between transposon inserts and the local genome annotation. Classes 1–4 consist of insertions in the appropriate orientation located within a codon intron (class 1), a noncoding transcribed region (class 2), an upstream genomic region (class 3), or an exon (class 4). For each class, the insert was either of the appropriate frame (subclass A) or of nonappropriate frame

(*Continued*)

TABLE 1

Project summary

Type	No.	%
Total setup	7404	100
CA	1344	18
CB	4000	54
CC	1870	25
piggyBac A	190	3
Aligned sequence	5720	77
Clusters	1550	21
Independent aligned	4170	56
Gene hits	2149	29
Spacing hits	2093	28
Different genes	1154	16
Protein traps	244	
Enhancer traps	256	
Novel gene/exon	50	
Unclassified	300	
Balanced stocks	878	12

TABLE 2

Line types

	Lines		
Type	Code	*N*	%
Coding intron, in frame	1A	1337	23
Coding intron, out of frame	1B	173	3
Noncoding intron, in frame	2A	29	1
Noncoding intron, out of frame	2B	429	8
1–500 bp upstream, in frame	3A	56	1
1–500 bp upstream, out of frame	3B	756	13
Exon	4	804	14
>500 bp to next oriented gene	5	2093	37
Total		5677	100

See class definitions in Figure 1D.

genic elements, by noncanonical splicing events, or by the presence of a second insertion at a canonical site.

The Drosophila genome annotation is highly supported by experimental evidence (Misra *et al.* 2002); hence the frequency of these discrepancies was surprising, but a similar outcome has been reported in previous protein trap screens using both yeast (Ross-Macdonald *et al.* 1999) and Drosophila (Morin *et al.* 2001). Some protein-coding genes may have been missed within cDNA libraries (Hild *et al.* 2003), and a substantial number of genes may contain unannotated far-upstream promoters and alternative translation start sites. There might be a large class of RNA genes that have escaped detection but that can drive expression of EGFP using cryptic start sites. Alternatively, the high selective pressure used to isolate EGFP+ strains may have led to the recovery of rare events in which the normal gene or transcript structure has changed.

Analysis of fusion transcripts: We sequenced portions of the fusion transcripts to address how EGFP expression arises in lines of various classes. A limited number of 5′ RNA sequences were obtained by 5′ RACE analysis or by TAIL–PCR (see materials and methods). As expected, several lines in class 1A were found to initiate in normal exons upstream from the insertion site. However, we discovered several CB lines that spliced into the EGFP exon from the 5′ *P*-element sequences of the vector. The *P*-element promoter is highly efficient at enhancer trapping, and the entire first exon of the transposase gene is present in the vector along with the start of intron 1, which is in the frame compatible with CB lines. These lines were associated with nuclear localized EGFP, possibly due to fusion of the first exon of *P* transposase with EGFP. Further evidence of enhancer trapping was observed in the analysis of line CA07138. The EGFP RNA was fused to sequences, including an in-frame ATG start codon, derived by transcription and splicing from the noncoding strand of the mini-white transgene carried in the *P*-element vector (Figure 1E). These observations suggest that EGFP expression in a significant number of the lines depends on transcripts initiated from within the transposon itself by enhancer trapping rather than on EGFP exon addition by splicing into exogenous transcripts.

Analysis of downstream transcript sequences: To gain additional information, we analyzed the sequences downstream from the EGFP exon from many lines in the collection by carrying out RT–TAIL–PCR in a 96-well format (materials and methods). 3′ sequences up to 700 bp in length and defining the location of one to six downstream exons were obtained for >1200 lines (Table 3). The pattern of downstream splicing allowed productive fusions to be identified and indicated lines that splice out-of-frame and likely become subject to nonsense-mediated decay (NMD) (Vasudevan and Peltz 2003). Fusions within lines of classes 2–4 often

(subclass B) to fuse to the protein if splicing continued to the next annotated exon splice acceptor site. Class 5 consists of transposons inserted >0.5 kb from a correctly oriented annotated gene. (E) The structure of cryptic transcripts initiated within the Drosophila mini-white marker gene that contain an ATG codon and splice in frame to EGFP, thereby allowing expression independent of an endogenous transcript in some lines. (F) Western blot analysis of Dlg1 and eIF-4E protein production in control animals (*y w*, CC00380) and insertion lines predicted to trap Dlg1 (CC01936) or eIF-4E (CC00392, CC00375, and CC01492). (G) Abnormal nuclear accumulation of CG15015-EGFP in line CC01311 whose insertion lies within the FHC domain (left). Tissue culture cells expressing N-terminal or C-terminal fusions are found in the cytoplasm (center and right).

TABLE 3

RNA analysis

Type	Successful	Confirmed DNA	Second insert	% confirmed
CA	316	224	50	82
CB	328	166	75	69
CC	572	165	72	70
Piggy A	13			
Totals	1229	555	197	74

were predicted to encode a "linker peptide," which might or might not include a stop codon, derived from the translation of a small segment of upstream nucleotides. The RNA analysis also revealed the presence of a second insertion in 14% of type 1A lines, but between 24 and 50% of the other classes. The second insertions found within class 2–5 lines were often valid protein trap alleles (class 1A) and were frequently the true source of the lines' EGFP production. This information allowed us to identify additional candidate fusions (Table 4), to correct many initial line classifications, and to more accurately estimate the total number of trapped genes (Table 1: 600–900). By the time lines were selected and balanced, secondary insertions or damage did not contribute substantially to the phenotypes reported in Table 4. Tests estimated the frequency of background lethal mutations among balanced, saved lines at 7–21%, similar to the best transposon screens (Spradling *et al.* 1999).

Novel splicing suggests new genes and exons: We compared the splicing observed downstream from the inserted exons with that of the genome annotation (Misra *et al.* 2002) to identify new Drosophila gene and transcript isoform candidates. To identify new candidate genes, we focused on lines inserted >0.5 kb from an appropriately oriented known gene and for which RNA sequence data were also available. In 114 of these 205 lines, the RNA sequence coincided with the position and orientation of the insertion and therefore indicated the splicing pattern downstream of the single EGFP exon. Most of the lines spliced to one or more novel exons. At least 8 probably correspond to unannotated genes because they match previous gene predictions (Hild *et al.* 2003) or are supported by EST data (see Table 4). Most of the remaining exons do not predict proteins with homologs in other species and represent either aberrant splicing events or novel or untranslated exons.

Similar analysis of 297 lines with intron insertions allowed us to test for novel exons and transcript isoforms. We examined 443 splicing events and identified a total of 35 (7.9%) that did not correspond to current gene models (Misra *et al.* 2002). Since at least some of these differences probably resulted from aberrant splicing induced by the insertions, this represents a maximum estimate of the fraction of unannotated genomic exons and emphasizes the high accuracy and completeness of current Drosophila gene models, at least for abundant transcript forms. Often, the RNA data indicated which isoform among several predicted to fuse in frame is likely to predominate in ovarian tissue. For example, we could determine that line CA06613 in ovarian tissue predominantly fuses the *Su(Tpl)* gene rather than *Mi-2*, in whose transcription unit it also lies in frame.

The nature of the noncanonical splices observed was interesting. The most common events (21/35) were for insertions in large introns to splice to a novel exon(s) prior to joining the predicted downstream exon. Some simply appear to define alternative isoforms that skip exons or utilize different exon combinations not previously documented. Some of these events may have been induced by the abnormal position of the EGFP exon within the primary transcript. However, several lines appear to define alternative isoforms because they utilize different combinations of known exons in no previously documented transcript isoforms. Three lines utilized 5′ start sites for exons that differed by 6, 21, or 27 bp from the annotated exon. The CC01473 transcript reads through an annotated exon into the adjacent intron and probably defines a novel alternate transcript 3′ end. Although we consider it likely that many of these differences reflect endogenous Drosophila gene expression, all of the candidate novel genes and transcript isoforms require independent confirmation in strains lacking protein trap insertions. Such tests were beyond the scope of our project.

Protein trap insertions likely vary in the fraction of the endogenous protein that is tagged with EGFP for a variety of reasons. First, in many lines only some of the multiple-transcript isoforms contributing to protein production are tagged by the insertion and fused in frame. Second, the splicing efficiency of the EGFP exon might vary due to its surrounding genomic context. To investigate this issue, we analyzed the protein products of tagged genes by Western blotting. The tagged proteins were easily distinguished from their wild-type counterparts on the basis of size and by probing with protein-specific and anti-EGFP antibodies (Figure 1F). In line CC01936 all three isoforms are predicted to incorporate the EGFP exon in frame, and nearly all of the ovarian Dlg1 protein incorporated EGFP as indicated by its mobility. A similar result was reported previously in the case of line *CB02119* (Buszczak and Spradling 2006), where the precursors of five of six annotated transcripts are predicted to contain the insertion, although only two fuse in frame. In contrast, only ~50% of the ovarian wild-type eIF-4E protein is tagged with EGFP (Figure 1F) despite the fact that six of seven annotated eIF-4E transcripts initiate upstream

TABLE 4

Identified trapped proteins

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
14-3-3ε	CA06506	14068962	3R	+	Lethal	RA	1.5	1	4	RB	1.5	1	4	RC	1.5	1	4	RD	1.5	1	4	1A
26-29-p	CA06735	13987313	3L	+	hv	RA	0.5	1	3													3A
Acon	CC00758	21169155	2L	+	Lethal	RB	1.5	1	4													1A
Actn	CC01961	1927332	X	−	hv	RB	2.5	2	10	RA	2.5	2	10	RC	2.5	2	10					1A
AGO1	CA06914	9841633	2R	−	hv	RC	3.5	3	8	RA	3.5	2	7	RB	0.5	2	7					1A
Alh	CC01367	2935653	3R	−	hv	RA	5.5	2	10	RD	5.5	2	10									1A
apt	CC01392	19468319	2R	+	hv	RB	1.5	1	5	RD	1.5	2	6	RE	1.5	3	6	RC	1.5	1	5	1A
apt	CC01186	19473808	2R	+	hv	RB	1.5	1	5	RD	2.5	2	6	RE	2.5	3	6	RC	1.5	1	5	1A
arg	CB03579	417662	X	+	hv	RA	3.5	1	5													1A
Argk	CB05492	9051633	3L	−	Lethal	RA	3.5	2	4	RB	2.5	2	3									1A
Argk	CB03789	9056078	3L	−	Lethal	RA	2.5	2	4	RB	0.5	2	3									1A
Atpα	CC00319	16783418	3R	+	Lethal	RC	1.5	3	10	RA	1.5	1	10	RB	0.5	2	9	RD	0.5	3	6	1A
baz	CC01941	17072582	X	+	hv	RA	1.5	1	7													1A
bel	CC00869	4485350	3R	−	hv	RA	1.5	1	4													1A
Best1	CB02354	5996196	3R	−	hv	RA	1	2	7													4
βTub56D	CC02069	15338563	2R	−	Lethal	RB	1.5	1	2	RC	1.5	2	2	RD	1.5	2	2					1A
bon	CB02667	16418866	3R	+	Semilethal	RA	0.5	1	10													3B
Bsg	CA06978	8104393	2L	+	hv	RB	2.5	2	7	RA	2.5	2	7	RD	2.5	2	7	RC	2.5	2	7	1A
bun	CB03431	12482943	2L	−	hv	RA	3.5	1	5	RB	1	1	3	RD	1.5	1	3	RE	2.5	2	4	1A
Cam	CC00814	8149270	2R	+	Lethal	RA	2.5	2	5													1A
CAP	CA06924	6190378	2R	+	hv	RI	9.5	1	14	RH	7.5	3	12	RJ	8.5	2	13	RG	7.5	3	12	1A
CAP	CA07185	Transposon	2R	+	hv	RI/RF	3.5	1	14													1A
Cat	CC00907	18815951	3L	+	Lethal	RA	1.5	1	3													1A
Cg	CC01469	10063184	2R	+	hv	RD	1.5	1	10	RB	1.5	1	11	RC	2.5	3	11	RA	2.5	2	11	1A
CG10724	CA07499	13406231	3L	+	hv	RA	1.5	1	7	RB	1.5	1	7									1A
CG11138	CA06844	12472570	X	−	hv	RC	1.5	1	4													1A
CG11255	CB04917	13015302	3L	−	hv	RA	1.5	1	3													1A
CG11266	CC01391	7033157	2L	+	hv	RA	2.5	2	6	RD	2.5	8	9	RG	2.5	7	8	RF	1.5	6	7	1A
CG11963	CC06238	4764704	3R	+	Semilethal	RA	2.5	2	9													1A
CG12163	CC00625	1076935	3R	−	hv	RA	1.5	1	6	RB	1.5	1	5									1A
CG12785	CC06135	12098103	3R	−	Lethal	RA	1	1	6													4
CG13920	CC01646	1650510	3L	+	Lethal	RA	1.5	1	3													1A
CG14207	CB02069	19502492	X	+	Lethal	RA	1.5	1	4	RB	1.5	1	5									1A
CG1440	CA07287	8331699	X	+	hv	RA	1.5	1	5													1A
CG14648	CA06610	229231	3R	+	hv	RA	1.5	1	6	RB	1.5	3	6									1A
CG14656	CA06996	624015	3R	−	Lethal	RA	2.5	1	3													1A
CG15926	CB04063	12365107	X	+	hv	RA	0.5	2	5													3A
CG1600	CB03410	3416244	2R	−	hv	RA	1.5	2	3	RC	1.5	1	3	RB	1.5	2	3					1A
CG17273	CC01294	16654986	3R	−	Lethal	RA	1.5	1	5													1A
CG17646	CB02833	1737431	2L	+	hv	RB	1.5	2	12	RA	0.5	2	12									2B
CG1888	CB02075	5434043	2R	−	hv	RA	1.5	1	2													1B

(*continued*)

TABLE 4

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
CG1910	CC01491	27573862	3R	+	hv	RB	1.5	2	4	RA	1.5	1	4	RD	0.5	2	4					1A
CG3036	CA06801	4894157	2L	+	hv	RA	2.5	2	7													1A
CG31012	CA06686	26640622	3R	+	hv	RC	1.5	1	6													1A
CG31012	CA06810	26646037	3R	+	hv	RC	4.5	1	6	RD	1.5	1	3									1A
CG31694	CA07748	2873276	2L	−	hv	RA	1.5	1	5													1A
CG32062	CC00511	10515608	3L	+	hv	RB	2.5	2	13	RD	2.5	2	14									1A
CG32423	CC00236	5177892	3L	−	hv	RA	3.5	3	10	RD	2.5	2	9	RB	3.5	3	10	RC	1	2	8	1A
CG32479	CA06614	882041	3L	+	Lethal	RA	4.5	2	10													1A
CG32486	CC00904	3060297	3L	−	hv	RD	1.5	1	8													1A
CG32560	CA06772	17632469	X	+	hv	RA	3.5	1	8													1A
CG3287	CB04483	2699967	2R	+	hv	RB	2.5	2	5	RC	1.5	1	4									1A
CG33936	CC01586	5176796	3R	+	Lethal	RA	2.5	3	6	RB	1.5	2	5									2B
CG3810	CA07694	1802294	X	−	hv	RC	1.5	2	4	RB	1	2	4	RA	0.5	1	3					2B
CG3939	CA07562	3069871	X	+	hv	RA	1.5	1	2													1A
CG5059	CA06926	20509742	3L	−	hv	RA	1.5	1	5	RC	1.5	2	5	RB	1.5	1	5	RD	1.5	2	5	1A
CG5060	CC00526	16079041	3R	+	Lethal	RA	1.5	1	7													1A
CG5130	CB03619	20486552	3L	−	hv	RA	1.5	2	4	RB	1	2	4									2A
CG5174	CA07176	14309292	2R	+	hv	RJ	1.5	1	6	RI	1.5	1	6	RA	1.5	1	6	RH	1.5	1	5	1A
CG5392	BA00207	15001227	3L	−		RA	2.5	2	8													1A
CG6151	CA07529	15808775	3L	−	Lethal	RA	2.5	1	5	RC	2.5	1	5	RB	2.5	1	5					1A
CG6330	CB03223	22779020	3R	−	Lethal	RB	2.5	2	6	RA	0.5	1	5									1A
CG6416	CC00858	8627040	3L	+	Lethal	RE	3.5	1	9	RF	3.5	1	9	RA	2.5	2	8	RG	2.5	2	5	1A
CG6424	CC00677	13604099	2R	−	hv	RA	2	3	4	RB	0.5	2	3									4
CG6783	CA06960	7392313	3R	−	hv	RB	1.5	1	3	RC	1.5	1	3	RA	1.5	2	4					1A
CG6854	CA07332	15099355	3L	+	Lethal	RC	1.5	1	4	RA	1.5	2	4	RB	1.5	2	5					1A
CG6930	CA06556	7597410	3R	−	hv	RA	0.5	1	5	RB	2.5	2	6	RC	1.5	1	5					1A
CG6945	CC00864	15088263	3L	−	Lethal	RA	0.5	3	3													3A
CG7185	CC00645	8308089	3L	−	hv	RA	1.5	1	7													1A
CG7484	CB04101	17647286	3L	+	hv	RB	0.5	1	3													3A
CG8209	CB02086	7969693	3L	+	hv	RA	1.5	1	3													1A
CG8213	BA00169	4863309	2R	−		RA	1.5	1	8													1A
CG8351	CA07228	4631299	3R	+	Lethal	RA	4.5	1	5													1A
CG8443	CA06604	12075198	2R	+	hv	RA	1.5	1	7													1A
CG8552	CA07352	8159580	2L	−	hv	RA	0.5	1	7													3A
CG8583	CA06603	7354388	3L	+	hv	RA	1.5	1	4													1A
CG8920	CC00825	16208199	2R	+	hv	RC	2.5	2	8	RB	2.5	2	8	RA	1.5	2	4					1A
CG9331	CB04962	20824559	2L	+	hv	RB	1.5	2	5	RD	1.5	2	5	RC	1.5	2	6	RE	1.5	1	5	1A
CG9772	CB02188	163496	3R	−	hv	RB	1.5	1	5	RA	0.5	1	5	RC	0.5	1	2					1A
CG9796	CC00817	9226998	3R	−	hv	RA	1.5	1	4													1A
CG9894	CC00719	2755189	2L	+	hv	RB	2.5	2	4	RA	1.5	1	3									1A
Cp1	CC01377	9852057	2R	+	hv	RB	2.5	2	4	RA	2.5	2	4	RC	2.5	1	4					1A

(*continued*)

TABLE 4

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
Crc	CA06507	5456158	3R	−	hv	RA	2.5	1	4													1A
Crp	CB03073	16280297	2L	−	hv	RA	1.5	1	3													1A
CycB	CC01846	18694008	2R	−	hv	RA	1.5	1	5	RD	1.5	1	5	RB	1	1	5	RC	0.5	1	4	1A
Dek	CC00921	12743940	2R	+	hv	RA	1.5	1	11	RC	1.5	1	11	RB	0.5	1	10	RD	1.5	1	9	1A
Dek	CA06616	12744777	2R	+	hv	RA	2.5	1	11	RC	2.5	1	11	RB	1.5	1	10	RD	2.5	1	9	1A
desat1	CC01694	8269738	3R	+	hv	RA	1.5	2	5	RC	1.5	2	5	RE	1.5	2	5	RB	1.5	2	5	2B
Df31	CB02104	21629393	2L	−	hv	RB	2.5	2	4	RA	1.5	1	3	RF	1.5	1	3					1A
dlg1	CC01936	11286274	X	+	hv	RF	5.5	1	9	RB	6.5	2	17	RH	5.5	2	16	RE	2.5	2	13	1A
DLP	CA06573	6480860	2L	+	hv	RA	0.5	1	4													3A
Doa	CB03889	24717483	3R	+	hv	RA	3.5	1	13	RB	3.5	2	13	RE	0.5	2	12	RF	3.5	1	13	1A
dom	BA00164	17211471	2R	+		RD	1.5	2	15	RA	1.5	2	14	RE	1.5	2	11					2A
Dp	CA06594	9111298	2R	+	hv	RA	1.5	1	9													1A
drl	CC00251	19190343	2L	+	hv	RA	0.5	1	4													3B
Ef2b	CC01924	21681694	2L	−	Lethal	RA	1.5	1	5	RC	1.5	2	5	RB	1	2	5					1A
eff	CC01915	10565092	3R	−	Lethal	RA	2.5	2	6													1A
eIF-2β	CC06208	12519527	3L	−	Lethal	RA	1.5	1	3													1A
eIF3-S9	CB04769	13423919	2R	+	hv	RB	1.5	2	5	RA	1	1	4									2A
eIF-4a	CB03721	5982474	2L	+	Lethal	RA	1.5	1	5	RC	1.5	1	5	RB	0.5	1	5	RD	0.5	1	5	1A
eIF-4E	CC00392	9395078	3L	−	hv	RD	1.5	2	6	RB	1.5	2	6	RC	1.5	1	5	RA	1.5	2	6	1A
eIF-5A	BA00155	19945688	2R	+		RB	1.5	2	4	RA	1.5	2	4									2A
eIF-5C	BA00280	1425192	3R	−		RA	1.5	2	8	RC	2.5	3	9	RF	1.5	2	8	RD	1.5	2	8	2A
Eip63E	CA06742	3569399	3L	+	hv	RD	2.5	1	11	RE	3.5	2	12	RA	4.5	3	12	RB	3.5	2	11	1A
Elf	CA06515	12435789	2L	+	hv	RA	1.5	1	7													1A
eRF1	CB03931	20342600	3L	+	Lethal	RC	2.5	2	8	RB	2.5	2	8	RE	2.5	2	8	RG	2.5	2	8	1A
Fas2	CB03613	4029454	X	−	hv	RA	9	2	10													4A
fax	CC01359	16403817	3L	−	Lethal	RA	1.5	1	5	RC	1.5	1	5									1A
Fer1HCH	CA06503	26212791	3R	−	Lethal	RA	1.5	1	3	RB	2.5	2	4	RC	2.5	2	4	RD	2.5	2	4	1A
Fer2LCH	CA07607	26215006	3R	+	Lethal	RA	3	3	4	RB	3	3	4	RC	1	1	2					4
Fim	CC01493	17185787	X	−	hv	RA	1.5	1	5	RC	2.5	2	6	RD	0.5	1	5					1A
Fkbp13	CA07340	17385064	2R	−	hv	RA	0.5	2	5	RB	1.5	1	5									1A
Fpps	CB04937	7194698	2R	−	hv	RA	1.5	1	6													1A
Fs(2)Ket	CA07301	20735659	2L	+	hv	RA	2.5	2	6													1A
Gdi	CA07108	9493967	2L	−	Lethal	RA	1.5	1	3													1A
gish	CB02804	12106314	3R	+	hv	RD	2.5	2	12	RB	2.5	2	12	RE	2.5	2	12	RA	0.5	3	12	1A
GlcAT-S	CA07168	9616795	2L	+	hv	RA	1.5	1	5	RB	0.5	1	5									1A
Gli	CB02989	15762784	2L	−	hv	RA	0.5	2	7	RB	0.5	2	5	RC	0.5	3	8	RD	0.5	2	7	3A
G-oα47A	CA06658	6331783	2R	+	Semilethal	RB	2.5	2	8	RC	2.5	3	9	RD	2.5	2	8	RE	2.5	2	8	1A
gp210	CC00195	1647884	2R	+	hv	RA	0.5	1	19													5
HDAC4	CA07134	13172850	X	−	hv	RA	2.5	2	14	RC	1.5	1	12									1A
heph	CC00664	27763272	3R	−	Lethal	RB	4.5	4	14	RA	4.5	4	14	RK	3.5	3	13	RH	5.5	5	16	1A
His2Av	CC00358	22693293	3R	+	Lethal	RA	2.5	1	4													1A

(*continued*)

TABLE 4

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
homer	CB02121	6722933	2L	−	hv	RA	1.5	1	6	RB	0.5	1	6	RC	1.5	1	7					1A
how	CA07414	17881934	3R	+	Lethal	RA	2.5	2	8	RB	2.5	2	7	RC	2.5	2	8					1A
Hrb87F	CC00189	9485980	3R	−	hv	RA	1.5	1	3	RB	1.5	1	3									1A
Hrb98DE	CC01563	24425969	3R	+	hv	RA	1.5	1	5	RE	1.5	1	5	RB	0.5	1	5	RC	0.5	1	5	1A
Hrb98DE	CA06921	24427038	3R	+	Lethal	RA	2.5	1	5	RE	2.5	1	5	RB	2.5	1	5	RC	2.5	1	5	1A
Hsc70Cb	CB02656	14031299	3L	+	Lethal	RA	2.5	2	6	RB	2.5	2	6	RC	1.5	1	5					1A
Imp	CB04573	10702676	X	−	hv	RF	2.5	2	6	RH	3.5	3	7	RG	2.5	2	6	RD	2.5	2	6	1A
Indy	CC00377	18833638	3L	−	hv	RB	1.5	1	9	RC	1.5	2	9	RA	0.5	2	9					1A
inx7	CB04539	6892590	X	−	hv	RB	4.5	3	5													1B
jumu	CC00294	6182245	3R	+	hv	RA	1.5	1	3													1A
Jupiter	CB05190	7430651	3R	−	hv	RD	1.5	1	4	RH	1.5	1	5	RA	1.5	1	5	RE	0.5	2	6	1A
kay	CC01156	25608214	3R	+	hv	RA	1.5	1	3	RB	0.5	1	3									1A
kek1	CB02190	12822834	2L	−	hv	RA	0.5	1	2													3A
kis	CC01466	220632	2L	−	hv	RA	12.5	2	18	RB	1.5	1	7									1A
kis	CC00801	221924	2L	−	hv	RA	12.5	2	18	RB	1	1	7									1A
l(1)G0084	CC01368	19525602	X	−	hv	RA	4.5	2	12													1A
l(1)G0168	CC00492	15393000	X	+	hv	RA	2.5	1	7	RB	0.5	2	6									1A
l(1)G0320	CA06684	9447368	X	−	hv	RA	1.5	1	2													1A
l(2)08717	CA06962	14688632	2R	−	hv	RB	2.5	2	5	RA	1.5	1	4									1A
l(3)02640	CA07460	1336285	3L	+	Lethal	RA	2.5	1	4													1A
l(3)82Fd	CA07520	1123169	3R	−	Lethal	RL	7.5	3	19	RB	7.5	3	19	RJ	7.5	3	19	RF	1	1	13	1A
Lam	CB03749	5543070	2L	+	Lethal	RA	1.5	2	4													2A
LamC	CB04957	10462132	2R	−	Lethal	RA	1.5	1	4													1A
larp	CC06230	24152038	3R	−	hv	RB	1.5	4	6	RC	1.5	2	6	RA	1.5	4	6	RD	1.5	1	6	2A
Lsd-2	CA07051	14969607	X	−	hv	RA	1.5	1	4													1A
M6	CA06602	21502447	3L	+	hv	RA	1.5	2	5													2B
Map205	CC00109	27891553	3R	−	hv	RA	1.5	1	2													1A
Mapmodulin	CC01398	13756894	2R	−	hv	RB	3.5	3	7	RA	2.5	2	6									1A
mask	CC00924	20060119	3R	+	hv	RA	1.5	1	16	RB	1.5	1	16									1A
Mdh	CB04968	22978192	3R	+	hv	RA	1.5	1	7													1A
me31B	CB05282	10240237	2L	+	hv	RA	1.5	1	5	RB	1.5	2	5									1A
Men	CC06325	8545125	3R	−	hv	RB	2.5	1	4	RA	2.5	1	4									1A
Mi-2	CA06598	19901556	3L	−	Lethal	RA	1.5	1	5	RA	1.5	5	5	RB	1.5	5	5					1A
Mob1	CB04396	11546502	3L	−	hv	RC	1.5	1	4	RD	1.5	1	4									1A
mod(mdg4)	CA07012	17200382	3R	−	hv	RR	4.5	2	5	RA	4.5	2	5	RF	4.5	2	5	RD	4.5	2	5	1A
mub	CC01995	21909824	3L	+	hv	RA	7.5	2	9	RB	7.5	2	9									1A
NetB	BA00253	14596460	X	−	hv	RA	3.5	2	9													1A
NFAT	CA07788	13534781	X	+	hv	RA	1.5	1	10													1A
Nlp	CC01224	25831370	3R	+	hv	RA	2.5	1	3													1A
Nmdmc	CB02647	4873684	3R	−	hv	RA	1.5	2	3	RB	1.5	1	3	RA	1	1	2					1A
Nrx-IV	CA06597	12141797	3L	+	hv	RA	1.5	1	12	RB	1.5	1	14									1A

(*continued*)

TABLE 4

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
Oda	CC01311	Transposon	3L	–	Lethal	RA	2.5	1	11													1A
Oda	CB03751	8060372	2R	–	hv	RA	1.5	1	2													1A
Orc2	CB04400	9789604	3R	+	hv	RA	0.5	1	5	RA	0.5	2	4									3A
osa	CC00445	13529472	3R	–	Lethal	RB	4.5	2	16	RA	4.5	2	16									1A
Pabp2	CC00380	4019736	2R	+	hv	RA	2.5	2	5	RB	2.5	2	5									1A
Past1	CB02132	8523601	3R	+	hv	RA	1.5	1	4	RB	1	1	3									1A
Pde8	CA07101	19554469	2R	+	hv	RA	3.5	2	19	RE	4.5	3	20	RB	0.5	2	18					1A
Pdi	CA06526	15134669	3L	–	Lethal	RA	1.5	1	2	RB	1.5	1	2	RD	1.5	2	2	RC	0.5	1	2	1A
Pdp1	CB02246	7847944	3L	–	hv	RF	1.5	1	5	RB	1.5	2	6	RG	1.5	2	7					1B
Picot	CA07474	12548787	2R	–	Lethal	RA	2.5	2	6													1A
Pkn	CC01654	5157995	2R	–	hv	RB	2.5	2	10	RC	2.5	2	11	RF	2.5	2	10	RD	2.5	2	10	1A
Pli	CB03040	19712573	3R	–	hv	RA	2.5	2	10													1A
Pmm45A	CB02099	4996761	2R	–	hv	RB	1	1	4	RA	1	1	3									4A
polo	CC01326	20303643	3L	+	Lethal	RA	1.5	1	5													1A
psq	CC01645	Transposon	2R	+	Lethal	RB	3.5?	2	10													1A
Ptp10D	CC06344	11538719	X	+	hv	RB	2.5	2	14	RC	1.5	1	10									1A
pUf68	CA06961	1501291	3L	–	hv	RC	3.5	4	6	RD	3.5	6	8	RA	2.5	1	5	RB	2.5	4	6	1A
pum	CC00479	4983771	3R	–	Lethal	RA	8.5	2	13	RD	8.5	2	13	RC	8.5	2	13	RB	6.5	1	11	1A
Rab11	CA07717	16937950	3R	+	Lethal	RA	1.5	1	4	RB	2.5	2	5									1A
Rab2	CA07465	2584592	2R	+	Lethal	RA	1.5	1	4													1A
Rm62	CB02119	1833559	3R	–	Mfsterile	RE	1.5	2	6	RC	2.5	3	7	RD	2.5	2	6	RB	1.5	2	6	1A
RpL10Ab	CB02653	11815918	3L	+	Lethal	RA	0.5	1	3	RC	0.5	2	3	RB	0.5	1	2					3A
RpL13A	CC01920	1449810	3R	+	hv	RB	0.5	1	3													3A
RpL30	CB03373	19009261	2L	–	Lethal	RB	0.5	2	3													3A
Rtnl1	CA06523	5000772	2L	–	hv	RB	3.5	2	7	RE	2.5	1	6	RD	0.5	2	6	RA	1.5	1	5	1A
Rtnl1	CA06547	4997815	2L	–	hv	RB	3.5	2	7	RE	2.5	1	6	RD	2.5	2	6	RA	0.5	1	5	1A
S6k	CC01583	5802962	3L	–	hv	RA	1.5	1	10													1A
Sap-r	CA07241	26714539	3R	–	hv	RA	1.5	1	7	RB	1	2	6									1A
sar1	CA07674	18184952	3R	+	Lethal	RA	4.5	2	6													1A
scrib	CA07683	22393784	3R	+	hv	RC	10	2	14													4
sd	CA07575	15712098	X	+	hv	RB	3.5	3	12	RA	1.5	2	10									1A
Sdc	CC00871	17362946	2R	–	hv	RC	2.5	2	6	RA	2.5	2	7	RB	2.5	2	6					1A
Sec61α	CC00735	6479544	2L	–	Lethal	RA	2.5	1	4													1A
Sema-1a	CA07125	8592539	2L	+	hv	RA	1.5	1	20													1A
Sema-2a	CA06989	12411885	2R	+	hv	RC	2.5	2	15	RB	2.5	2	15	RA	2.5	2	15					1A
sgg	CA06683	2536739	X	+	hv	RB	2.5	2	10	RA	2.5	2	10	RE	2.5	2	10	RF	2.5	2	10	1A
Sh3β	CC01823	7361764	3L	–	Lethal	RB	1.5	1	2	RA	2.5	2	3									1A
Sin	CC01921	21024944	3L	+	Lethal	RA	1	1	3													4
sls	CA06744	2107593	3L	–	hv	RC	0.5	1	1	RA	2.5	2	14									1A
sm	CC00233	15504045	2R	–	hv	RA	2.5	2	10	RC	2.5	2	9									1A
smi21F	CA07211	1119094	2L	–	hv	RB	2.5	2	5	RA	1	2	4									1A

(*continued*)

TABLE 4

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
sno	CC01032	13104608	X	–	hv	RB	1.5	1	12	RA	1.5	1	3									1A
snRNP69D	CB02932	12727708	3L	–	Lethal	RA	0.5	1	2													3A
sop	CB02294	9897503	2L	–	Lethal	RA	1	2	2													4A
SPoCk	CA06644	22766451	3L	+	hv	RA	0.5	2	4	RC	0.5	2	5									3A
Spt6	CA07692	6162464	X	+	hv	RA	2.5	1	6													1A
sqd	CB02655	9470746	3R	–	hv	RB	1.5	1	7	RC	1.5	1	6	RA	1.5	1	5					1A
stwl	CA07249	14402679	3L	–	Lethal	RA	1.5	1	2													1A
Su(var)2-10	CC02013	5004591	2R	+	hv	RI	1.5	1	8	RH	1.5	1	7	RD	1.5	1	9	RC	1	1	8	1A
Surf4	CC01684	11171291	3R	–	Lethal	RA	1.5	1	4	RB	2.5	2	5									1A
sws	CC01711	7861288	X	–	hv	RA	1.5	1	11													1A
Sxl	CB05562	6986034	X	–	Semilethal	RB	2	2	3	RC	1.5	1	6	RN	1.5	1	6	RO	1.5	1	8	1A
TER94	CB04973	5877016	2R	+	hv	RA	1.5	1	5	RB	1.5	2	4									1A
Tm1	CC01710	11116123	3R	+	Lethal	RB	3.5	2	10	RJ	3.5	2	10	RD	3.5	2	10	RG	3.5	2	10	1A
Tm1	CC00578	11117364	3R	+	Lethal	RB	3.5	2	10	RJ	3.5	2	10	RD	3.5	2	10	RG	3.5	2	10	1A
tmod	CC00416	26389239	3R	+	hv	RE	2.5	2	7	RD	1.5	1	6	RF	2.5	2	7	RC	2.5	2	7	1A
tmod	CA07346	26400579	3R	+	Lethal	RE	6.5	2	7	RD	5.5	1	6	RF	6.5	2	7	RC	6.5	2	7	1A
Top1	CC01414	15214506	X	+	hv	RA	1.5	1	8													1A
tra2	CC01925	10491357	2R	–	hv	RC	3	3	7	RB	2.5	2	5	RA	2.5	2	6	RE	1	1	4	4
tral	CA06517	12508905	3L	+	Lethal	RA	1.5	1	7													1A
Trxr-1	CA06750	8137659	X	+	hv	RA	1.5	1	4	RB	1	1	4									1A
Tsp42Ee	CC01420	2903327	2R	+	hv	RA	2.5	2	5													1A
Tsp96F	CC01830	21707093	3R	–	Lethal	RA		1	3	RA	1	1	5									1B
tsr	CC01393	19932991	2R	–	Lethal	RA	1.5	1	4													1A
Tudor-SN	CC00737	262842	3L	–	hv	RA	1.5	1	4													1A
tun	CC00482	11681495	2R	–	hv	RG	2.5	2	16	RA	2.5	2	16	RC	2.5	2	16	RE	2.5 2	15		1A
twin	CA06641	20044629	3R	–	hv	RB	4.5	2	7	RE	5.5	2	8	RC	5.5	1	8	RD	5.5	3	4	1A
Uev1A	CA07496	5358821	3L	–	Lethal	RA	1.5	1	4													1A
VAChT	CA06666	14538229	3R	+	Lethal	RA	2	2	2	RA	1.5	1	8									1A
Vha13	CA07644	15469816	3R	–	Lethal	RA	1.5	1	3													1A
Vha16	CA06708	2518996	2R	–	Lethal	RA	2.5	2	4	RB	2.5	2	4	RC	1.5	1	3	RD	2.5	2	4	1A
Vha26	CC01380	1417892	3R	+	Lethal	RB	2.5	2	5	RA	1.5	1	4									1A
Vha55	CA07634	8452738	3R	–	Lethal	RB	2.5	2	4	RA	1.5	1	3									1A
vib	CB05330	15045962	3R	–	Lethal	RA	2.5	2	9													1A
vkg	CC00791	5019005	2L	–	hv	RA	2.5	2	9													1A
vsg	CA07004	9707771	3L	+	hv	RA	1.5	1	2	RD	1.5	1	2	RB	2.5	2	3	RC	2.5	2	3	1A
xl6	CB03248	6918786	2L	–	hv	RA	1.5	1	2	RA	0.5	1	2									1A
yps	CA06791	12117267	3L	–	Lethal	RA	1.5	1	4													1A
zip	CC01626	20896845	2R	–	Lethal	RB	2.5	2	14	RA	2.5	1	14									1A
Zn72D	CA07703	16102655	3L	–	hv	RA	4.5	3	5	RB	4.5	3	6									1A
	CB02318	4638102	3R	+	hv																	5B
	CC01309	8022556	3L	+	hv																	5B

(*continued*)

TABLE 4

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
	CB02658	12522067	2R	−	hv																	5B
	CC01670	12944837	3R	+	hv																	5B
	CC01932	21856862	3R	−	hv																	5B
	CA07436	5644358	2R	−	sl																	5B
	CB03064	8536402	2L	+	hv																	5B
	CC00523	22806145	3R	−	hv																	5B

[a] Annotation release 5.0.

from the insertion site. These examples suggest that the EGFP incorporation level varies between genes in large part due to the tagging of a subset of gene isoforms that themselves display differing expression levels.

In contrast, we found little evidence of short-range context effects. Less than twofold variation in protein expression as measured by Western blotting with anti-EGFP antibodies was observed between lines with insertions at sites within the same intron (N. Srivali and A. Spradling, unpublished data). However, these experiments did reveal that insertions of the *piggyBac*-based vector consistently produced less EGFP protein than lines with the corresponding *P*-element-based vector that were inserted in the same intron (N. Srivali and A. Spradling, unpublished data). This suggests that some aspect of the structure of the *piggyBac* vector used compromised splicing efficiency.

Identification of protein trap and enhancer trap core collections: To help identify a core set of valid gene trap lines we examined the EGFP expression patterns of many nonredundant lines in both the adult ovary and larval salivary gland. There was a strong correlation between insert location and the nature of the staining patterns observed. More than 95% of lines in class 1A, the in-frame fusions, produced patterned EGFP expression above background in at least some ovarian cells or in the salivary gland. In contrast, a much smaller, but still significant, fraction of lines in classes 2–5 also expressed EGFP in a regulated manner. Combining information on insert location, genome annotation, RNA transcript sequence, and EGFP pattern, we identified a set of 244 lines predicted to produce fusions between EGFP and 431 protein isoforms of 233 distinct genes (Table 4). These new protein trap lines express EGFP in a wide variety of cellular compartments under diverse developmental controls (see Figures 2 and 3).

A second major class of lines in the collection showed the properties expected of EGFP enhancer traps (Table 5). These CB lines were susceptible to enhancer trapping from the *P*-element promoter, were located mostly upstream of the annotated start site, expressed EGFP in nuclei, and the RNA analysis, if available, did not indicate fusion in frame downstream. The expression patterns of such insertions in well-characterized genes supported this interpretation. For example, line CB02030 in *ptc* showed strong expression in the inner sheath cells of the germarium (Forbes *et al.* 1996), while line CB04353 in *Dad* strongly expressed in the germline stem cells and immediately downstream germ cells (Kai and Spradling 2003; Casanueva and Ferguson 2004).

The characterization of a significant number of lines in the collection remains incomplete (Table 1). Many of these contain insertions located >0.5 kb from an appropriately oriented annotated gene but where RNA sequence data were not obtained. Others are inserted within genes at locations not predicted to generate protein fusions or gene traps. Some of these lines

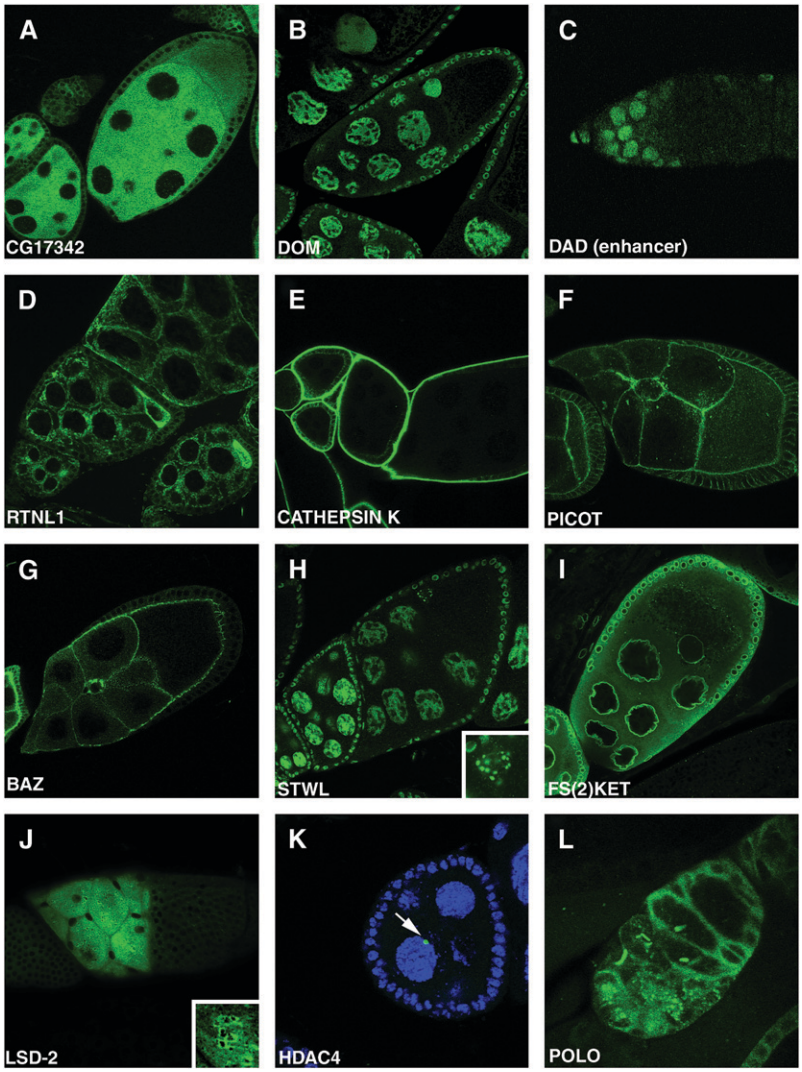

FIGURE 2.—Protein traps for the study of protein subcellular localization. Patterns of subcellular localization of EGFP expressed from the following lines that trap the indicated genes were observed: (A) cytoplasmic, CA06607 (CG17342); (B) nuclear, BA00164 (dom); (C) enhancer trap nuclear, CB04353 (Dad) in stem cells and early cystocytes; (D) endoplasmic reticulum, CA06523 (Rtnl1); (E) extracellular, CA06735 (cathepsin K); (F) membrane, CA07474 (Picot); (G) apical, CC01941 (Baz); (H) chromatin, CA07249 (stwl); (I) nuclear membrane, CA07301 (Fs(2)Ket); (J) lipid droplets, CA07051 (Lsd-2); (K) novel structure, CA07332 (CG6854); (L) novel structure, CC01326 (polo).

express EGFP in the ovary, and we cannot rule out that others express transcripts in other tissues. On the basis of the processing of previous lines in these same classes this suggests that a significant number of new enhancer traps and a handful of new protein traps could be sorted out from a larger number of lines with secondary insertions in already trapped genes. Consequently, the number of different genes trapped in the collection is likely to increase beyond the 600 or so currently characterized.

Even when a protein is tagged in frame, the insertion of the EGFP sequence is expected to disrupt its normal structure and localization some fraction of the time. For example, line CC01311 traps CG15015, the Drosophila homolog of mammalian Cip4, a modular protein that interacts with Cdc42 and helps to regulate the actin cytoskeleton (ASPENSTROM 1997). The CC01311 *P* element is inserted between the first two coding exons and thus disrupts the FES/Cip4 domain of CG15015 (Figure 1G). The protein trap fusion product localizes to the nucleolus while transgenes of CG15015 tagged at either the very N or C termini localize to the cytoplasm when expressed in S2 cells. We observed that 3 other protein trap fusion products of 107 analyzed accumulated in the nucleus when they were expected to reside in the cytoplasm.

Diverse behavior of tagged proteins: The 244 identified protein trap lines of the core collection exhibit extremely diverse patterns of EGFP expression, suggesting that proteins occupying a wide range of cellular compartments can be tagged *in vivo*. We observed many lines with EGFP fluorescence in the cytoplasm (Figure 2A) or nucleus (Figure 2B) as expected. Localization to intracellular membranous structures was also commonly seen, as illustrated by a trap in the ER structural component Reticulon-1 (Figure 2D). Gene trapping of secretory proteins is thought to be inefficient due to retention of the fusion proteins in the ER where the activity of the fusion gene may be affected (SKARNES *et al.* 1995). The full-length protein traps we constructed could label secreted proteins, as indicated by the extracellular localization of EGFP in CA06735, a fusion

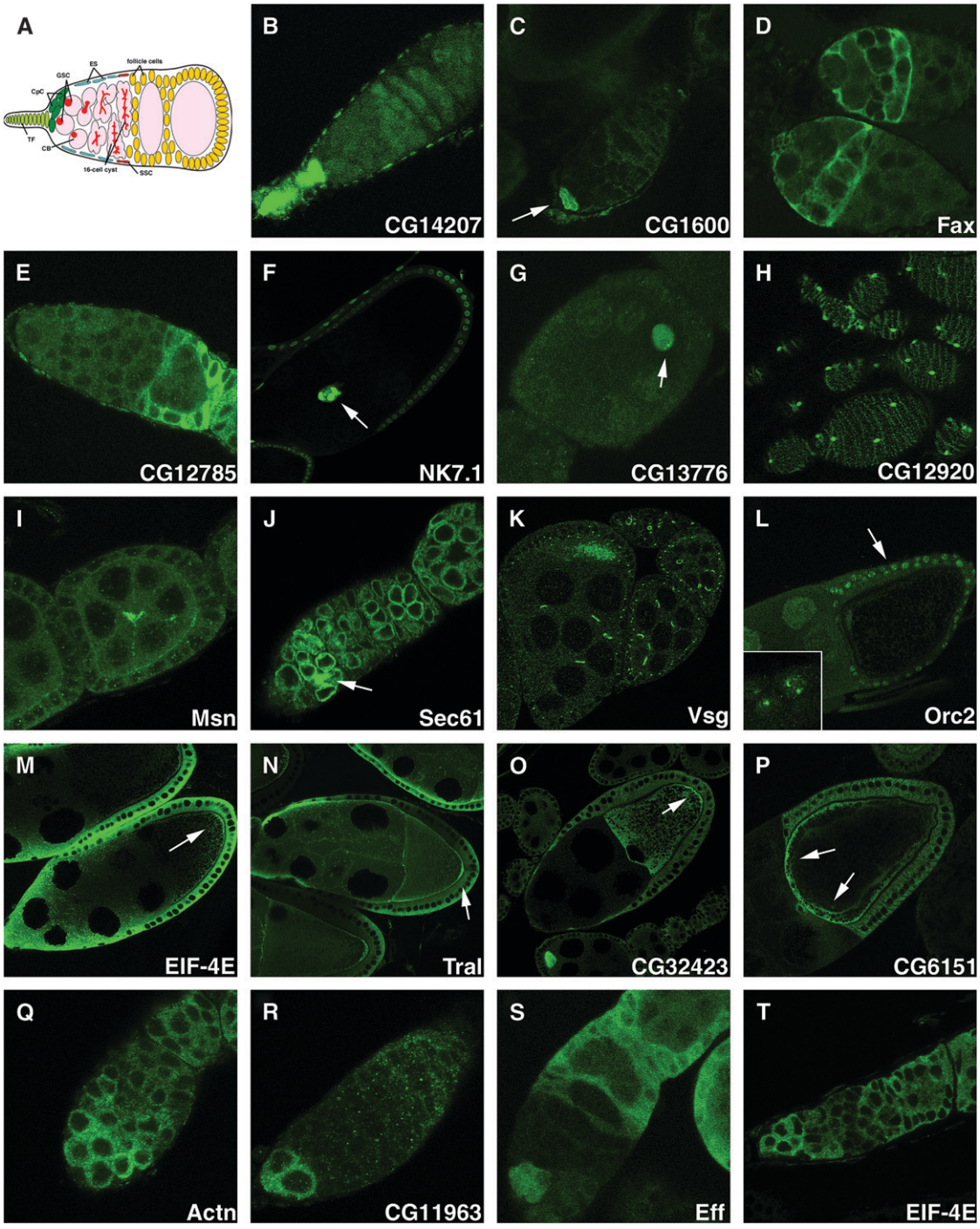

FIGURE 3.—Protein traps for the study of developmental regulation during oogenesis. The expression in the ovary of various protein trap lines is shown to illustrate how they can be used to associate genes with developmental processes. (A) Schematic of an ovariole tip. The terminal filament (TF), cap cells (CpC), germline stem cells (GSC), cystoblast (CB), and escort cells (ES) are illustrated. (B–H) Cell type identification. (B) Terminal filament, CB02069 (CG14207); (C) cap cells, CB03410 (CG1600); (D) escort cells, CC01359 (fax); (E) follicle cells, CC06135 (CG12785); (F) outer border cells and posterior follicle cells, CB02349 (NK7.1); (G) oocyte nucleus equals the germinal vesicle (arrow), CB04219 (CG13776); (H) novel sheath cell type, CC01646 (CG12920). (I–L) Analyzing developmental processes. (I) Novel structure in center of midstage follicle, CC00523 (Msn); (J) fusome, CC01436 (Sec61); (K) germline and somatic ring canals, CA07004 (Vsg); (L) chorion gene amplification, CB04400 (Orc2). (M–P) Localization of proteins in the oocyte. (M) Posterior pole, CC01442 (EIF-4E); (N) posterior pole, CA06517 (Tral); (O) posterior pole, CC00236 (CG32423); (P) anterior pole, CA07529 (CG6151). (Q–T) Developmental regulation of gene expression in early germ cells. (Q) Control with little change, CC01961 (Actn); (R) GSC/CB enriched, CC06238 (CG11963); (S) GSC and early cyst enriched, CC01915 (Eff); (T) GSC and forming cyst enriched, CC01442 (EIF-4E).

TABLE 5

Enhancer trap alleles

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
1.28	CB04492	2389425	2R	+	hv	RA	0.5	1	1													3
4EHP	CB05417	19934636	3R	−	hv	RA	1	1	4													4
abo/l(2)06225	CB03272	10975307	2L	−	hv	RA	0.5	1	2													3
Adf1	CB02276	2550455	2R	−	Lethal	RA	2.5	1	4													1
Adk3/CG4674	CB04307	6715035	3R	−	hv	RB	0.5	2	2	RA	0.5	2	2									3
aop	CB02762	2178398	2L	−	Lethal	RB	1	2	4													4
ATPCL/CG8370	CB04427	11888767	2R	−	hv	RA	0.5	2	9													3
bic	CB03855	8757727	2R	+	Lethal	RA	0.5	2	2	RB	0.5	1	1									3
bnl	CB02854	15662816	3R	−	hv	RA	0.5	2	5	RB	0.5	2	5									3
boca	CB02070	3454283	2R	−	hv	RA	0.5	1	2													3
bocksbeutel	CB03586	5416706	3R	−	hv	RA	1	1	2													4
Brd	CB02665	14965467	3L	−	hv	RA		1	1													4
btn	CB03501	18414077	3R	+	hv	RA	1	1	1													4
CBP	CB03762	7230965	X	+	hv	RA	0.5	1	4													3
Cct1	CB02171	1546105	3L	+	hv	RA	1	2	4													4
cdi	CB05129	14919950	3R	+	hv	RA	1	2	8								2					2
CG10225/Tbp-1	CB02343	19578503	3R	+	hv	RA	0.5	1	3													3
CG10272	CB03567	2233083	3R	+	Lethal	RA	1.5	3	9	RB	1	2	8	RC	1.5	3	6	RD	1.5	3	9	2
CG10399	CB04753	6960416	2L	+	hv	RA	0.5	1	2													3
CG10444	CB05265	16161990	2R	−	Lethal	RA	1	1	2													4
CG10863	CB05229	3942458	3L	+	hv	RB	1	1	6													3
CG10990	CB02351	13648912	X	+	hv	RA	1	2	4													5
CG11042	CB03720	9046478	X	+	hv	RA	0.5	1	1													3
CG1116	CB03511	1045469	3R	−	hv	RA	1	1	5	RB	1	1	4	RC	1	1	5					4
CG11382	CB02249	1103926	X	+	hv	RB	0.5	1	1													3
CG11526	CB02228	3325883	3L	+	hv	RA	1.5	2	3	RB	1.5	1	3									2
CG11537	CB03533	3143415	3L	−	hv	RC	1	1	8													4
CG11638/CG32814	CB02337	920725	X	+	hv	RA	1	2	5													4
CG11779	CB05200	14983800	3R	+	hv	RC	1.5	1	5	RA	1.5	1	4	RD	1.5	1	4					1
CG11940	CB05024	19742560	X	−	hv	RA	1.5	1	3	RB	0.5	2	3									1
CG12360	CB02135	8856387	3R	+	hv	RA	0.5	2	5	RB	0.5	2	5									3
CG12367	CB05312	8033260	2R	+	hv	RA	1	1	3													4
CG1240	CB02937	2767042	3L	+	Lethal	RA	1	2	2													4
CG12418	CB05329	5700359	3R	+	hv	RA	0.5	1	1													3
CG12744/cbx	CB05091	5762817	2R	+	hv	RA	1	1	2	RB	0.5	2	3									3
CG12797	CB03502	10742504	2R	+	hv	RA	0.5	1	1													3
CG13295	CB04422	6068307	3L	+	hv	RA	1	1	3													4
CG13895	CB04175	708235	3L	−	hv	RA	0.5	2	3													3
CG14215	CB02255	19546304	X	+	hv	RA	0.5	1	7													3
CG14430	CB02236	6879851	X	+	hv	RA	0.5	1	1													3
CG14478	CB02133	13345190	2R	+	hv	RA	1	2	2													4

(*continued*)

TABLE 5

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
CG14542	CB02373	21878549	3R	−	Lethal	RA	0.5	1	6													3
CG14709	CB04397	7394886	3R	+	hv	RA	0.5	1	8													3
CG1621	CB02042	3380602	2R	−	hv	RA	0.5	1	2													3
CG1667	CB05477	5731280	2R	+	Semilethal	RA	2	1	5													4
CG16708	CB04388	1193533	3R	−	hv	RA	1.5	2	4													2
CG16817	CB02934	5503678	3R	+	hv	RA	1	1	4													4
CG16971	CB03898	224106	3L	+	hv	RB	1	2	3	RC	1	2	3	RD	1	2	3					4
CG17002/vimar	CB05094	2873307	2R	+	hv	RB	1	1	5													4
CG17090	CB03116	544248	3L	+	hv	RB	1.5	2	10													2
CG17090	CB02270	544285	3L	−	hv	RA	1.5	1	2													5
CG17323	CB02962	18823565	2L	+	hv	RA	0.5	2	5													3
CG17836	CB05263	14749549	3R	+	hv	RB	3.5	3	5	RA	3.5	3	5	RC	2.5	3	4	RD	0.5	2	3	1
CG1910/l(3)s1921	CB03702	27573214	3R	−	Lethal	RB	1.5	2	4	RA	0.5	1	4									2
CG2051	CB03625	1614858	3R	+	hv	RB	1	1	3	RA	0.5	1	3	RC	0.5	1	3					4
CG2186	CB05020	10751743	X	−	hv	RA	1	1	7													4
CG2446	CB03703	11608741	X	−	hv	RA	0.5	3	4	RE	1	3	4	RB	1	3	4	RD	1	2	3	3
CG2698	CB03026	3827319	3R	+	hv	RA	1	2	8													4
CG2865	CB03023	2187549	X	−	hv	RA	0.5	1	2													3
CG2926	CB02232	1414090	3R	+	hv	RA	0.5	1	1													3
CG2974	CB04047	9980614	X	−	hv	RA	0.5	1	2													3
CG30055	CB02739	8477121	2R	+	hv	RA	1	1	1													4
CG30497	CB02106	3667177	2R	−	hv	RA	2	2	3													4
CG31241	CB02140	14081632	3R	+	Lethal	RA	2	2	2													4
CG31475	CB04600	15006356	3R	+	hv	RA	0.5	2	4													3
CG31522	CB02693	279018	3R	−	hv	RB	1	2	10	RC	1	2	2									4
CG3164	CB02987	129446	2L	+	hv	RA	1.5	1	8	RC	1.5	1	8	RB	0.5	1	7	RD	0.5	1	7	1
CG31650	CB05467	5043131	2L	−	hv	RA	1.5	2	4	RB	1	2	4	RC	1	2	4					2
CG31688	CB03570	20429045	2L	+	hv	RA	1	2	6													4
CG31689	CB03239	2735300	2L	+	hv	RC	1.5	2	11	RD	1.5	2	11	RA	1.5	2	11	RB	1.5	2	11	2
CG31782	CB03345	16716141	2L	−	hv	RA	2.5	1	6	RB	2.5	1	7									1
CG32043	CB03247	9455668	3L	+	hv	RB	2	2	5	RA	2	2	5									4
CG3209	CB05445	19956511	2R	+	hv	RA	1	1	7	RB	1	1	6									4
CG32345	CB03988	628140	3L	−	hv	RA	0.5	1	1													5
CG32436	CB02614	21340125	3L	−	hv	RA	1.5	1	5													1
CG32436	CB04148	21340268	3L	+	hv	RA	1.5	1	5													1
CG32486	CB03414	3070840	3L	+	Lethal	RD	0.5	1	8													3
CG3321	CB05150	10153408	3R	+	hv	RA	2	2	2	RB	2	2	2									3
CG33214	CB04173	21500637	3L	−	hv	RA	1	1	6													4
CG33232	CB03740	2466875	3L	+	hv	RA	0.5	3	9													3
CG33558	CB05689	2859128	2R	−	hv	RA	0.5	2	16													3
CG33967	CB04401	10549498	3R	−	Lethal	RA	1	1	9													4

(*continued*)

TABLE 5

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
CG33982	CB04965	15527849	3L	+	hv	RA	1	1	1													3
CG3409	CB04412	2602327	2R	−	hv	RA	1.5	3	7													2
CG34110	CB04263	21010670	3R	+	Lethal	RC	2.5	1	9													1
CG3428	CB02131	9480857	3L	−	hv	RA	1	1	3													4
CG3654/Uch-L3	CB04163	9470143	3L	−	hv	RA	0.5	1	4													3
CG4091	CB02284	19589843	2R	−	Lethal	RA	1	2	3	RC	1	3	4									4
CG4300	CB04249	12270328	3L	−	hv	RA	1	1	4	RB	1	1	4									4
CG4570	CB02124	6682617	3R	+	hv	RA	1	1	1													4
CG4612	CB05331	20413641	2R	−	hv	RA	0.5	2	3													3
CG5381	CB04616	10321949	2L	+	hv	RA	0.5	2	7													3
CG5543	CB04854	19696764	2R	+	hv	RA	0.5	1	1													3
CG5548	CB05667	14957872	X	−	hv	RA	0.5	2	2													3
CG5677	CB02054	20055723	3R	−	Lethal	RA	1	1	1													4
CG6014	CB04785	21390388	3L	+	hv	RA	1.5	3	7													2
CG6218	CB03213	11168745	3R	+	hv	RA	1	2	5													4
CG6311/Nedd4	CB04145	17522938	3L	+	hv	RC	1	1	5													4
CG6439	CB03836	17844824	3R	−	Lethal	RA	1	1	6													4
CG6499	CB05192	11075733	3R	−	Lethal	RA	2	1	5													4
CG6540	CB03922	18542228	X	+	hv	RA	1	1	2													4
CG6770	CB02632	12046132	2L	−	hv	RA	0.5	1	1													3
CG7110	CB04925	13399559	2L	+	hv	RB	2	2	7													4
CG7228	CB03115	7994181	2L	−	hv	RA	1	2	5	RB	1	2	5									4
CG7331	CB05154	4175563	3R	+	hv	RA	1	1	1													4
CG7637	CB03644	6698443	2R	−	hv	RA	0.5	1	2													3
CG8036	CB04958	4495764	3R	+	Lethal	RB	2.5	3	4	RC	1.5	2	3									1
CG8092	CB05336	11101113	2R	−	Lethal	RA	1	1	6	RB	1	1	3									4
CG8128	CB02087	15591543	X	+	hv	RA	1	2	4													4
CG8206	CB04527	15634596	X	+	hv	RA	0.5	1	1													3
CG8444	CB03837	5086366	3R	+	Lethal	RA	1	1	3													4
CG8583	CB04752	7353717	3L	−	hv	RA	1	1	4													4
CG9062	CB02056	7171458	2R	−	hv	RB	1	1	8													4
CG9171	CB02706	5800187	2L	−	hv	RA	1	4	8	RB	1	3	7									4
CG9328	CB03031	20797707	2L	−	hv	RB	0.5	2	3													3
CG9591	CB05694	9508902	3R	−	Lethal	RA	2	2	9													4
CG9666	CB04503	19257579	3L	−	Lethal	RA	1	3	4													4
CG9699	CB02196	16581750	X	−	hv	RA	2.5	2	5	RF	2.5	2	5	RB	2.5	2	5	RD	2.5	2	5	1
CG9821	CB03335	4646100	3R	−	Lethal	RB	1	2	2	RA	1	2	2									4
CG9921	CB02890	16226793	X	−	hv	RA	0.5	2	3													3
CG9924	CB03517	9852398	3R	−	hv	RB	2.5	3	9													2
Chd64	CB03690	4122396	3L	−	hv	RB	1	1	3													4
chrb	CB05429	11480949	3L	+	hv	RC	1	1	4													4

(*continued*)

TABLE 5

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
Cp190	CB04933	11100985	3R	+	hv	RA	0.5	2	5													2
crol	CB03039	11809411	2L	−	hv	RD	1	4	8	RA	0.5	4	8	RB	0.5	4	7	RC	0.5	4	8	4
Csp	CB02217	22260750	3L	+	Lethal	RA	1	1	5	RB	1	1	7	RC	1	1	6					4
CtBP	CB04073	8837940	3R	−	Lethal	RA	1.5	1	6	RB	1.5	2	8	RC	1.5	2	6					1
CycB3	CB02981	20696520	3R	−	Lethal	RA	1	1	1													4
CycD	CB02618	15803675	X	+	hv	RC	1	1	5	RB	1	2	6	RA	0.5	2	6					4
Dad	CB04353	12882290	3R	+	Semilethal	RA	2.5	3	8	RB	1.5	2	7	RC	1.5	3	8					1
dally	CB03495	8820577	3L	+	hv	RA	0.5	1	9													2
dap	CB05233	5599791	2R	+	Lethal	RA	0.5	1	3	RB	0.5	1	3									2
Dap160	CB04282	21142183	2L	−	Lethal	RA	1.5	2	11	RB	1.5	2	10									2
Dcp-1/pita	CB03160	19439026	2R	+	Lethal	RA	1	1	3													4
Dek	CB02288	12743725	2R	−	hv	RA	0.5	1	11	RD	0.5	1	9	RC	0.5	1	11	RB	0.5	1	10	3
desat1	CB02105	8269738	3R	+	hv	RA	1.5	2	5	RC	1.5	2	5	RE	1.5	2	5	RB	1.5	2	5	2
Dl	CB02040	15151950	3R	−	Lethal	RA	0.5	1	8	RB	0.5	1	6									3
Dp1/imd	CB03754	14299509	2R	−	hv	RA	0.5	3	7	RB	0.5	3	7	RC	0.5	3	7					3
Dref/RpL13	CB02226	9967309	2L	−	hv	RA	1	1	4													4
drk	CB02974	9389656	2R	−	Lethal	RE	1.5	2	6	RC	1.5	2	6	RB	1.5	2	6	RA	1.5	2	6	2
dup	CB04090	11268107	2R	−	Lethal	RA	1	1	3													4
eas	CB02620	16172362	X	+	hv	RB	1	2	5	RA	1	2	5	RE	1	2	6	RD	0.5	3	7	4
edl	CB04040	14561061	2R	−	hv	RA	0.5	2	2													3
eIF	CB02125	1426796	3R	+	Lethal	RA	1	2	8	RG	1	2	8	RC	0.5	3	9	RF	0.5	2	8	4
eIF3-S10	CB05493	259579	3R	−	Lethal	RA	1	1	4	RB	0.5	3	5									4
eIF5B	CB05358	3460609	3L	−	hv	RB	0.5	1	9													3
Eip75B	CB05160	18007641	3L	−	Lethal	RB	1.5	1	6													1
emc	CB02035	749295	3L	+	hv	RA	0.5	1	2													3
endoA	CB05084	14732350	3R	−	Semilethal	RA	0.5	1	2													3
Eno	CB02039	1727775	2L	−	Lethal	RB	3	3	4	RE	3	3	4	RC	3	3	4	RD	3	3	4	4
esg	CB02017	15333865	2L	+	Lethal	RA	1	1	1													4
fas	CB02992	9510510	2R	+	hv	RA	0.5	3	13													3
fj	CB04634	14120268	2R	+	hv	RA	1	1	1													4
flfl/CG9591	CB05447	9510094	3R	+	Lethal	RA	1.5	2	11	RB	1.5	2	11									2
fok/neb	CB05794	20085050	2L	−	Lethal	RA	1	1	2													4
for	CB02956	3632190	2L	−	Lethal	RA	4.5	3	10	RB	2.5	1	8	RH	3.5	2	9	RI	4.5	3	10	1
fs(1)K10/kz	CB02790	2136127	X	−	hv	RA	0.5	1	2													3
ftz-f1	CB05043	18758843	3L	−	Lethal	RB	1.5	1	9													1
Fur1	CB03489	21298814	3R	−	hv	RA	0.5	4	12	RB	0.5	4	12	RC	0.5	4	11	RD	0.5	3	10	3
fw	CB05224	11898010	X	−	hv	RA	0.5	2	15													3
fz2	CB02997	19163698	3L	−	Semilethal	RB	1	3	3													4
glec	CB02364	17681968	3R	−	Semilethal	RA	1	1	2													4
Glycogenin	CB05177	17087979	2R	+	Semilethal	RB	0.5	1	5	RA	0.5	1	3									3
Grip163	CB05139	11988609	3L	−	Lethal	RA	4	1	5													4

(*continued*)

TABLE 5

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
grp	CB04894	16683920	2L	+	hv	RB	1.5	2	7	RC	1.5	3	8									2
GstE1	CB04956	14285871	2R	+	Lethal	RA	0.5	1	1													3
GstS1	CB05261	12984903	2R	−	hv	RB	1	2	5	RC	1	2	5	RA	0.5	2	5					4
gukh	CB04444	14809830	3R	−	hv	RA	1	1	6	RB	1	1	5	RC	1	1	6					4
Gyc76C	CB03117	19783669	3L	−	Lethal	RC	2.5	5	19	RA	1.5	4	18	RB	1.5	3	18					2
hdc	CB05774	26103641	3R	+	hv	RC	0.5	1	6													3
HLHm7	CB02243	21862637	3R	+	hv	RA	0.5	1	1													3
HmgD	CB02648	17600985	2R	+	hv	RA	1.5	2	3	RB	1.5	3	4									2
HP1b	CB02849	8975314	X	+	hv	RA	2	2	2													4
Hr39	CB05039	21250790	2L	+	hv	RC	2.5	3	5	RB	2.5	3	8	RA	2.5	3	8	RD	2.5	3	8	2
Hsc70-4	CB05603	11068412	3R	−	hv	RA	1	2	2	RC	0.5	2	2	RE	0.5	2	2					4
Hsr&ohgr	CB03168	17122190	3R	+	hv	RA	0.5	1	1	RB	0.5	1	1	RC	0.5	1	1					3
IP3K1	CB02227	9782577	2L	+	hv	RA	1	1	4													4
Irp	CB02050	6238617	3R	+	hv	RA	1	1	10													4
kuz	CB02000	13550247	2L	+	Lethal	RA	1	2	12	RB	1	2	12	RC	1	2	12					4
l(2)gl	CB02331	18536	2L	−	hv	RB	2.5	2	9	RA	1	3	10	RC	2.5	2	9	RD	1	3	9	1
lama	CB05457	5348459	3L	−	hv	RC	1.5	3	5	RB	0.5	3	5									4
LanA	CB04172	6211152	3L	−	hv	RA	0.5	1	15													3
LBR	CB03173	17607992	2R	−	hv	RB	0.5	3	4	RA	0.5	2	3	RC	0.5	1	2					3
lea	CB02898	1420531	2L	−	Lethal	RA	0.5	1	14													3
Lk6	CB02120	7590184	3R	−	hv	RA	0.5	1	6													3
lola	CB02888	6429215	2R	−	hv	RC	0.5	2	6	RG	0.5	2	6	RH	0.5	2	6	RJ	0.5	2	6	3
Map60	CB03167	5478371	2R	+	hv	RA	1	1	2													4
mbc	CB04603	19608306	3R	−	Lethal	RA	1.5	1	14													1
Mbs	CB02150	16045108	3L	−	Lethal	RB	1	2	18	RA	1	2	18	RC	1	2	19					4
MESR3	CB02595	18617241	2L	+	hv	RA	0.5	3	4													3
MESR4	CB04813	13435844	2R	−	Lethal	RA	0.5	2	3													3
mirr	CB02689	12686547	3L	+	Lethal	RA	0.5	1	5	RB	0.5	1	5									3
Mnf	CB03632	11113376	3L	−	hv	RA	1.5	2	9	RE	1.5	2	10	RD	1.5	2	9					2
Mocs1	CB04106	11063638	3L	+	hv	RA	0.5	1	4	RC	0.5	1	4	RB	0.5	4	4					3
mod	CB02172	27878341	3R	+	hv	RA	3	3	6													4
mRpS17	CB03663	20061952	2R	−	hv	RA	1	1	2													4
Myo31DF	CB04377	10506773	2L	+	hv	RA	1	2	10	RB	1	1	9									4
neb/fok	CB04551	20085119	2L	+	Lethal	RA	1.5	1	2													1
nes	CB05687	19253774	3L	+	hv	RA	1	2	4	RB	1	2	4	RC	1	2	4					4
Neu3	CB02076	10523038	3R	−	Lethal	RB	1	1	8													4
NK7.1	CB02349	10198828	3R	−	hv	RA	0.5	2	4													3
nmo	CB02015	7972207	3L	+		RA	1	3	9	RB	1	2	8	RE	1	2	8	RC	1	4	7	4
nmo	CB04635	7972439	3L	−	Lethal	RA	1.5	3	9	RB	1.5	2	8	RE	1.5	2	8	RC	1.5	4	7	3
Nrg	CB04883	8411401	X	+	hv	RC	0.5	2	8	RB	0.5	2	8	RA	0.5	2	8					3
orb2	CB04897	8946768	3L	−	hv	RD	2.5	3	6	RB	1.5	2	5	RC	0.5	2	5					2

(*continued*)

TABLE 5

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
Orc2	CB03773	9790734	3R	−	Lethal	RA	2	2	4													4
Pak3	CB04117	12279337	3R	−	hv	RA	1	1	2	RB	1	1	2									4
Pi3K21B	CB05132	297734	2L	+	hv	RA	0.5	1	3													3
Pka-C1	CB03724	9699251	2L	−	Lethal	RB	1	2	2	RA	1	2	2	RC	1	2	2					4
PNUTS-RB	CB04203	870488	2L	−	Lethal	RB	1	2	5	RC	1	3	4	RD	1	1	5					4
ppa	CB03551	18576512	2R	−	hv	RA	0.5	1	1													3
Psc	CB05083	8868241	2R	−	hv	RA	1	2	5													4
ptc	CB02030	4537352	2R	+	Lethal	RA	1	1	6													4
pxb	CB05625	11491866	3R	−	hv	RB	0.5	2	5	RA	0.5	2	5									3
PyK	CB04247	18194169	3R	+	Lethal	RB	1.5	2	4													2
Rab1	CB03450	17094755	3R	−	Lethal	RA	1	1	5	RB	1	1	4									4
Rab23	CB03781	1509120	3R	−	hv	RA	1.5	2	3													2
rdgB	CB03856	13656772	X	+	hv	RA	0.5	2	12	RC	0.5	3	13	RD	0.5	2	12					3
Rga	CB02139	1438302	3R	−	Lethal	RA	1	2	8	RB	1	2	8									4
rgr	CB02242	4487949	2R	+	hv	RA	1	1	5													4
rho	CB05698	1463699	3L	+	hv	RA	0.5	2	4													3
RhoGDI	CB05226	19922267	3L	+	Lethal	RA	1.5	2	5													2
Rpd3	CB04302	4626763	3L	+	Lethal	RA	1	1	4													4
sas	CB03307	2988505	3R	+	Semilethal	RB	0.5	2	10													3
Sc2	CB04638	3903512	3L	+	hv	RA	1	1	4													4
Sep2	CB05018	16414054	3R	+	hv	RA	1	1	2													4
SF1	CB05052	13366171	3R	+	hv	RA	1	1	6													4
SF2	CB03028	12167740	3R	−	hv	RA	0.5	1	4													3
sgl	CB05101	6957887	3L	−	hv	RA	1	1	3													4
shg	CB02169	16944287	2R	−	Lethal	RA	1	1	2													4
Sir2/DnaJ-H	CB02821	13165875	2L	+	hv	RA	1	1	2													4
siz	CB03288	21027758	3L	−	Lethal	RA	1	1	6													4
skd	CB02029	21018904	3L	−	Lethal	RD	1.5	2	13	RC	1.5	2	13									2
Sod	CB03598	11106792	3L	−	hv	RA	1	1	2													4
Ssdp	CB02353	14022949	3R	−	Lethal	RA	1.5	2	2	RB	1.5	2	2	RC	1.5	2	2					2
stg	CB03726	25081026	3R	−	hv	RA	1	1	2													4
stumps	CB05360	10417302	3R	+	Lethal	RA	0.5	1	5													3
stwl	CB05223	14402982	3L	+	hv	RA	1	1	2													4
tankyrase	CB03249	21486905	3R	−	hv	RA	1	1	7													4
tara	CB02767	12063444	3R	+	hv	RA	1.5	1	2													1
tara	CB03523	12075785	3R	−	Lethal	RA	1.5	2	5	RB	1.5	2	5									1
Tbh	CB04045	7889488	X	+	hv	RB	0.5	2	8													3
Tctp	CB03738	7036007	3R	−	hv	RA	1	1	1													4
Tom	CB02307	14962360	3L	+	hv	RA	0.5	1	1													3
trbl	CB02744	20394721	3L	−	Semilethal	RA	0.5	1	2													3
trn	CB05114	13107142	3L	+	hv	RA	0.5	2	2													3

(*continued*)

TABLE 5

(Continued)

Gene	Line	Site[a]	Chr	Strand	Phenotype	T1	Insert	Met	Stop	T2	Insert	Met	Stop	T3	Insert	Met	Stop	T4	Insert	Met	Stop	Type
trx	CB05156	10108478	3R	−	hv	RA	1.5	3	9	RB	1.5	3	8	RC	1.5	2	7	RD	1.5	2	8	2
ttk	CB02274	27550755	3R	+	hv	RE	1.5	2	5	RF	1.5	2	5	RB	1	2	5	RC	1	2	5	2
Ugt37c1	CB02900	12733228	2R	−	Lethal	RA	0.5	1	1													3
Ugt86Da	CB03314	6982675	3R	+	hv	RA	0.5	1	3													3
Vha100-2	CB03404	14224630	3R	−	hv	RB	1.5	2	6													2
VhaPPA1-1	CB02209	10729723	3R	−	Lethal	RA	1	2	2													4
wun2	CB02267	5301760	2R	+	hv	RA	0.5	1	6													3
Xbp1	CB02061	17031115	2R	+	Lethal	RB	1	1	3	RA	1	1	2									4
Xe7	CB05610	1495043	3R	+	Lethal	RA	1	1	8													4
Z4/CG12974	CB02305	21279245	3L	+	Semilethal	RA	1	1	2													4

[a] Annotation release 5.0.

with CG8947, a Drosophila cathepsin K homolog (Figure 2E), and the membrane location of a trap in Picot, a phosphate symporter (Figure 2F). Proteins that are apically localized in polarized epithelia such as ovarian follicle cells were easily visualized, as observed for Bazooka (Par3) (Figure 2G). Subcompartmentalized nuclear proteins were also readily apparent. For example, a trap of the Stonewall (Stwl) HMG-related protein involved in germ cell chromatin organization (Clark and McKearin 1996) labeled nurse cell nuclei and the oocyte nucleus (inset) differently (Figure 2H). Fs(2)Ket, a protein involved in nuclear import, was localized to the nuclear periphery (Figure 2I). In some cases, cell-specific cellular compartments were labeled, such as in CA07051, which traps Lsd-2 and exhibited EGFP localization to lipid droplets that arise in late-stage nurse cells (Figure 2J).

These studies provide a high-resolution view of known protein locations in living cells and also identify many proteins that were not previously known to reside within these compartments. In addition, the value of protein trapping as a discovery tool was illustrated by the fact that we observed new patterns of localization as well. For example, the HDAC4 protein, fused by the CA07134 trap, labeled a small body often found in only one nurse cell within an egg chamber (Figure 2K). A spindle-like structure in young nurse cells was labeled with a protein trap in the *polo* gene encoding a mitotic kinase (Figure 2L). Antibodies specific for the trapped protein can be used to isolate and further investigate the proteins present in these novel structures.

Analysis of developmental regulation—ovarian cells: All the lines in the core collection were characterized on the basis of their patterns of expression in germ cells and follicle cells during oogenesis. These experiments identified lines expressing in the major classes of somatic cells, including terminal filament cells (Figure 3B), cap cells (Figure 3C), escort cells (Figure 3D), profollicle cells (Figure 3E), and border and posterior cells (Figure 3F). Other lines expressed in germ cells of various ages, including some that were highly enriched in the oocyte nucleus (germinal vesicle) (Figure 3G). As in the case of subcellular compartments, these studies documented patterns of developmental expression for many genes that were not previously known. These genes become attractive candidates for study of their function in the corresponding processes.

Strikingly, the collection also revealed the likely existence of new cell types and novel biological processes previously unrecognized despite many years of study of ovarian biology. Line CC01646 traps the CG12920 protein and is expressed in a small subset of ovarian sheath cells that likely represent a novel cell type (Figure 3H). In line CC00523 we observed accumulation of Msn-EGFP preferentially at the center of mid-stage growing follicles (Figure 3I). It was not previously known that this region was the site of unique protein

accumulation. Msn encodes a protein involved in Jun kinase signaling, suggesting that a special intercellular junction may assemble in this region to structurally organize the nurse cells. We observed a similar expression program (not shown) for the line CB03040 that fuses the *Pli* gene, encoding a protein associated with the NF-κB signaling response.

Developmentally specific subcellular structures, including the fusome (Figure 3J, arrow) and both somatic and germline ring canals (Figure 3K), were also labeled by rare lines. A general and extremely useful application of the collection is to identify new proteins that are associated with such structures and analyze the effect of mutations. For example, the preferential accumulation of Sec61 in the fusome observed here has been validated in recent studies (Snapp *et al.* 2004). Proof of principle experiments of this type that focus on the fusome will be described elsewhere.

Another valuable capacity of protein trap lines is the ability to follow important developmental processes at high resolution and in living cells. During oogenesis, at least four major clusters of chorion genes undergo specific gene amplification in stage 10B follicles, a process that can be visualized as small "amplification dots" of BrdU incorporation (Calvi *et al.* 1998). The amplifying genes specifically contain substantial amounts of replication initiation proteins such as Orc2 at this time, whereas normally Orc2 is found throughout the cell nuclei (Royzman *et al.* 1999). A protein trap line in Orc2 allows the amplifying loci to be directly visualized (Figure 3L). Inspection shows that the dots are not present in preamplification stage follicle cells but strongly label amplifying gene loci at stage 10B (Figure 3L, arrow).

The Drosophila oocyte represents an important model system for studying RNA and protein localization. Several biochemical and genetic studies have identified proteins enriched at either the anterior or the posterior pole of the oocyte (Lasko 2003; Wilhelm and Smibert 2005). Protein traps in genes identified in these studies, including EIF-4E (Figure 3M) and Tral (Figure 3N), faithfully recapitulate the localization patterns of their endogenous counterparts to the posterior pole of the oocyte (Wilhelm *et al.* 2003, 2005). Several other proteins tagged in the collection display posterior localization patterns including a trap in CG32423 (Figure 3O), a largely uncharacterized RNA-binding protein. In addition, a trap in CG6151 appears to be enriched at the anterior of the oocyte (Figure 3P). While future work will clarify the role of these proteins in oocyte patterning, these examples show that protein trapping can complement other approaches and be used to identify new components of localized RNP complexes within the cells.

Protein traps provide unique opportunities to analyze gene regulation during development in populations of cells that are difficult to isolate and in those that are sensitive to loss of cellular context. We illustrate the potential of this approach using the regulation of germ cell development within and just downstream from the germline stem cells (GSCs). Many protein trap lines, including CC01961 in Actn, showed uniform expression in GSCs, CBs, and developing germline cysts (Figure 3Q). However, it was possible to find other examples where expression levels between GSCs and early germ cells differed from those in other germ cells within the germarium. One of the most striking examples was line CC06238 that traps the putative Drosophila succinate CoA ligase gene. Expression was stronger in stem cells (and sometimes early cystoblasts) than in later germ cells as illustrated in Figure 3R. Several other genes were downregulated shortly after GSC division, including *effete* (Figure 3S), encoding the UbcD1 ubiquitin-conjugating enzyme that has been shown to affect germline cyst formation (Lilly *et al.* 2000). Another line whose EGFP expression was downregulated slightly later, at the completion of cyst formation, trapped the Drosophila eIF-4E gene (Figure 3T). Downregulation of a related gene CG8023 was previously observed at a slightly earlier time, during cyst divisions (Kai *et al.* 2005).

Studies on the limitations of current protein trap methods: We also tested the sensitivity of the approach used here to detect Drosophila genes by looking at the expression of the lines in the core collection in germline stem cells. First, although lines were selected on the basis of expression in embryos, we found ovarian expression above background in >90% of lines in the core collection. However, this does not address whether many other genes exist that were fused but expressed EGFP at levels too low to detect in either tissue. Analysis of germline stem cell RNA by hybridization to Affymetrix arrays detected transcripts from ~6500 Drosophila genes over an ~1000-fold dynamic range (Kai *et al.* 2005). Although translational regulation and differential protein stability, not to mention differences in staining sensitivity between different preparations, would be expected to introduce potential variation, we were curious whether protein trap lines could detect stem cell gene transcripts across the full range of expression levels.

We observed a strong correspondence between these two measures of stem cell gene expression (Figure 4, A–E). Lines with very strong EGFP expression tended to have RNA levels at least 10-fold higher than lines with above background but relatively weak expression. These lines in turn had signals higher than most lines scored as below the level of detection on arrays. The correlation was not perfect; for example, some lines showed more EGFP expression in stem cells than might be expected from the microarray study (Figure 4F). The existence of such lines was not unexpected, because some lines likely still carry second insertions, and the microarray used was based on release 1 gene models. Overall, we could detect EGFP above background in GSCs from nearly all

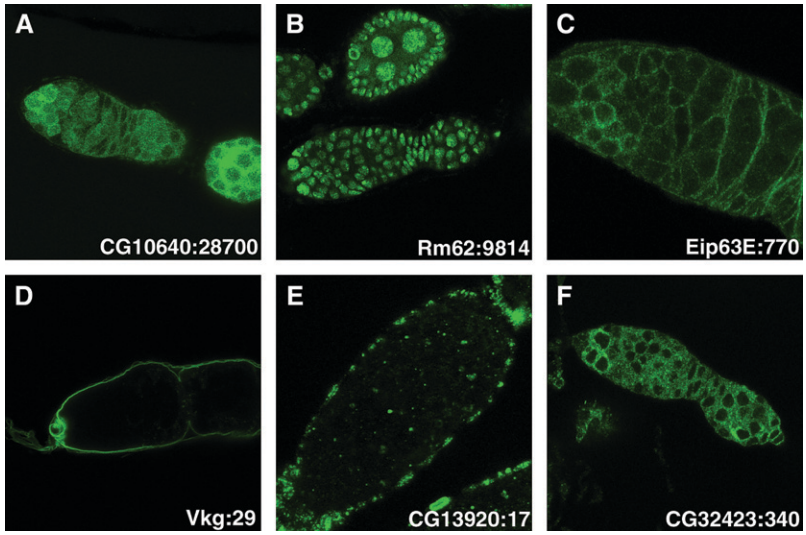

Figure 4.—Studies of protein trap expression We compared the apparent intensity of GFP-protein staining in the GSCs with the RNA level of the corresponding gene as determined by Affymetrix arrays (Kai *et al.* 2005). (A–F) The pattern of protein trap expression of the indicated gene (see Table 4 for strain names). The expression level from Affymetrix software (mean of three measurements) is given.

lines whose levels of mRNA are called as "present" on Affymetrix arrays (Kai *et al.* 2005). This suggests that protein trap lines are not limited to a relatively small number of highly expressed genes, but can be used to follow a large fraction of gene activity.

DISCUSSION

The Carnegie protein trap collection—a versatile research tool: These experiments significantly expand the number of protein trap lines available for studies of gene expression within a complex multicellular animal (also see accompanying article by Quiñones-Coello *et al.* 2007, this issue). Our initial characterization of these lines extends previous documentation that the behavior of the EGFP-tagged protein often corresponds to the behavior of the protein to which it is fused. Moreover, we demonstrate that collections such as ours are exceptionally useful as tools of gene discovery. Candidate genes can be selected on the basis of the developmental expression, subcellular localization, or dynamic behavior of particular protein isoforms. The same line can subsequently be used to purify the protein and its associated complexes and to create deletions for further genetic analysis. The Carnegie collection is available for research use from the Carnegie Institution. Information is available at http://www.ciwemb.edu/resources/proteintrapcollection.html and at http://flytrap.med.yale.edu/.

Subcellular distribution of protein location: Previously, gene and protein trapping in yeast has been used to estimate the fraction of proteins that are localized to various cellular compartments (Ross-Macdonald *et al.* 1999; Kumar *et al.* 2002; Huh *et al.* 2003). More than half of all proteins showed a simple localization to the cytoplasm or nucleus. Other subcellular structures labeled by tagged proteins in yeast included the plasma membrane, the ER, mitochondria, the lysosome, and the perixosome. We obtained similar results. The distribution patterns of EGFP-tagged proteins within Drosophila ovarian cells generally matched those of yeast and most known structures within egg chambers have now been labeled with at least one protein trap line (this study; Morin *et al.* 2001; Clyne *et al.* 2003). Moreover, the large size of ovarian cells often allowed us to distinguish the fine structure of several subcellular compartments labeled by EGFP fusion proteins generated in this screen.

Developmental regulation of protein expression: Despite the fact that only one tissue was examined closely, a large number of proteins in the core collection were expressed and many were developmentally regulated. A diverse array of cell types within the germarium including the terminal filament, cap cells, escort cells, germline stem cells, and prefollicle cells were labeled in various lines. However, expression frequently varies from stage to stage, not only in cell type but also in subcellular location, complicating the problem of accurate annotation. Currently, protein trap images within the ovary are being curated in the FlyTrap database. Because of the relative cellular simplicity of the germarium and developing ovarian follicles, it may be possible to develop tools for displaying expression patterns at single-cell resolution in this tissue. It will be particularly valuable to add data from many other tissues and developmental stages for these same lines, to facilitate comparisons.

Identification of insertions not predicted by genome annotation: One of the surprising results of our studies was the relatively high frequency of EGFP-positive lines that were located at sites not predicted to fuse to any annotated Drosophila transcript. However, similar results were observed in previous studies of gene trap transposons. At least 44% of insertions analyzed by Morin *et al.* (2001) were not within annotated genes; moreover, the reading frame of insertions in genes was

not determined. In yeast, Ross-Macdonald *et al.* (1999) observed that while 1346 EGFP-positive insertions were in the correct frame, another 480 were not. Since most lacked an alternative start site, they postulated that a higher than expected frequency of translational frameshifting may occur. In a recent study in the mouse, 24% of genes were trapped in more than one reading frame (De-Zolt *et al.* 2006). We also observed this phenomenon; however, our studies emphasized the difficulty of drawing final conclusions until the location of every insertion and the actual pattern of splicing within the mutant strains have been characterized.

Sensitivity of gene traps: A potential limitation of protein trapping *in vivo* is that many gene products may be expressed at levels so low that EGFP expressed at the same level could not be detected above background. Only 20% of mouse secretory traps that are G418 resistant express detectable CD2, even though the neophosphotransferase gene is fused to CD2 (De-Zolt *et al.* 2006). Only 33% of β-geo lines resistant to G418 express detectable lacZ. This probably indicates that many genes exist that generate enough neophosphotransferase to confer G418 resistance, but not enough β-galactosidase to be scored as lacZ positive (De-Zolt *et al.* 2006). Consistent with this, Ross-Macdonald *et al.* (1999) found that 415 of 1340 in-frame fusions (31%) could be detected above background by immunofluorescence. In contrast, Huh *et al.* (2003), who tagged complete proteins, detected signals above background for 4156 of 6029 (69%) genes. The system we employed is also designed to tag full-length proteins, and this may have enhanced its sensitivity.

The requirement that each line generate EGFP fluorescence in embryos might provide a limitation on the number of genes that could be tagged. However, our experiments argue that this poses relatively little selection on which genes can be fused. We found that genes expressed in germline stem cells at a wide variety of levels on the basis of microarray studies had been fused in our collection of protein trap lines. There was a rough correlation between the levels observed using antibody staining in these cells and the microarray results. This would indicate that the protein trap methodology can potentially be used to analyze thousands of diverse Drosophila genes.

Increasing proteome coverage: Our analysis revealed two major limitations of the current strategy for generating protein traps using *P* elements. Despite the advantages of embryo sorting, the inherent 5′ bias of *P*-element insertion (Bellen *et al.* 2004) greatly limits the rate at which new genes can be trapped. Many of the insertions were recovered when an insertion occurred at an internal promoter that lies within a coding intron of another gene isoform. Many genes lack such alternative promoters, so it will be necessary to use different methods to efficiently recover a more diverse collection of protein trap strains.

At least two alternative approaches are worthy of consideration. First, it should be possible to take advantage of the extensive collection of *P*-element insertions in Drosophila genes that have been generated by the BDGP gene disruption project and other members of the Drosophila community (reviewed in Matthews *et al.* 2005). We calculate that ~2000 genes already have an existing *P*-element insertion within a coding intron. Moreover, *P* elements can recombine into the sites of existing *P* elements in the presence of transposase (Sepp and Auld 1999). Consequently, a protein trap allele of each of these genes could, in principle, be generated by combining a protein trap insertion of the appropriate reading frame with the "target" gene insertion in a single strain, ideally using inserts bearing scorable markers and then screening for replacement.

Alternatively, transposons with a broader insertional specificity than the *P* element would be worthwhile. *piggyBac* elements are suitable for widespread mutagenesis of Drosophila genes (Thibault *et al.* 2004) and as gene traps (Bonin and Mann 2004). Minimal sequences for *piggyBac* transposition were defined recently (Li *et al.* 2005). We obtained several hundred *piggyBac* protein trap insertions that express EGFP, but the *piggyBac* gene trap vector appeared to be less efficient, on the basis of EGFP intensity and Western blotting, than *P*-element gene traps in nearby locations (also see accompanying article by Quiñones-Coello *et al.* 2007, this issue). New vectors containing different splice acceptor and donor sites should be tested within the context of a *piggyBac* element. Several new transposons with diverse insertional specificities, including *Minos* (Arensburger *et al.* 2005; Metaxakis *et al.* 2005), are also worthy of consideration. Continued efforts to provide greater coverage within the Drosophila proteome are warranted because of the exceptional utility of protein traps in analyzing the development and physiology of multicellular organisms.

We thank Lynn Cooley for helpful discussions. We thank X. Morin, W. Chia, A. Hudson, R. Lehmann, and A. Handler for reagents. We are grateful to the following people for their assistance with diverse aspects of the project: Alison Pinder, Joseph Carlson, Martha Evans-Holm, Crista Sewald, Becca Sheng, Melanie Issigonis, Megan Kutzer, Emily Seay, and Dianne Williams. We thank the Howard Hughes Medical Institute, Carnegie Institution, and the National Institutes of Health for support. M.B. was a fellow of the American Cancer Society. T.G.N. is a Howard Hughes Fellow of Life Sciences Research Foundation.

LITERATURE CITED

Arensburger, P., Y. J. Kim, J. Orsetti, C. Aluvihare, D. A. O'Brochta *et al.*, 2005 An active transposable element, Herves, from the African malaria mosquito *Anopheles gambiae*. Genetics **169:** 697–708.

Aspenstrom, P., 1997 A Cdc42 target protein with homology to the non-kinase domain of FER has a potential role in regulating the actin cytoskeleton. Curr. Biol. **7:** 479–487.

Bellen, H. J., R. W. Levis, G. Liao, Y. He, J. W. Carlson *et al.*, 2004 The BDGP gene disruption project: single transposon

insertions associated with 40% of Drosophila genes. Genetics **167:** 761–781.

Bonin, C. P., and R. S. Mann, 2004 A piggyBac transposon gene trap for the analysis of gene expression and function in Drosophila. Genetics **167:** 1801–1811.

Buszczak, M., and A. C. Spradling, 2006 The Drosophila P68 RNA helicase regulates transcriptional deactivation by promoting RNA release from chromatin. Genes Dev. **20:** 977–989.

Calvi, B. R., M. A. Lilly and A. C. Spradling, 1998 Cell cycle control of chorion gene amplification. Genes Dev. **12:** 734–744.

Casanueva, M. O., and E. L. Ferguson, 2004 Germline stem cell number in the Drosophila ovary is regulated by redundant mechanisms that control Dpp signaling. Development **131:** 1881–1890.

Clark, K. A., and D. M. McKearin, 1996 The Drosophila stonewall gene encodes a putative transcription factor essential for germ cell development. Development **122:** 937–950.

Clyne, P. J., J. S. Brotman, S. T. Sweeney and G. Davis, 2003 Green fluorescent protein tagging Drosophila proteins at their native genomic loci with small *P* elements. Genetics **165:** 1433–1441.

De-Zolt, S., F. Schnutgen, C. Seisenberger, J. Hansen, M. Hollatz *et al.*, 2006 High-throughput trapping of secretory pathway genes in mouse embryonic stem cells. Nucleic Acids Res. **34:** e25.

Dirks, R. W., and H. J. Tanke, 2006 Advances in fluorescent tracking of nucleic acids in living cells. Biotechniques **40:** 489–496.

Espina, V., J. Milia, G. Wu, S. Cowherd and L. A. Liotta, 2006 Laser capture microdissection. Methods Mol. Biol. **319:** 213–229.

Forbes, A. J., A. C. Spradling, P. W. Ingham and H. Lin, 1996 The role of segment polarity genes during early oogenesis in Drosophila. Development **122:** 3283–3294.

Friedrich, G., and P. Soriano, 1991 Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. Genes Dev. **5:** 1513–1523.

Gossler, A., A. L. Joyner, J. Rossant and W. C. Skarnes, 1989 Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes. Science **244:** 463–465.

Handler, A. M., and R. A. Harrell, II, 1999 Germline transformation of *Drosophila melanogaster* with the piggyBac transposon vector. Insect Mol. Biol. **8:** 449–457.

Herschman, H. R., 2003 Molecular imaging: looking at problems, seeing solutions. Science **302:** 605–608.

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

Huh, W. K., J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson *et al.*, 2003 Global analysis of protein localization in budding yeast. Nature **425:** 686–691.

Kai, T., and A. Spradling, 2003 An empty Drosophila stem cell niche reactivates the proliferation of ectopic cells. Proc. Natl. Acad. Sci. USA **100:** 4633–4638.

Kai, T., D. Williams and A. C. Spradling, 2005 The expression profile of purified Drosophila germline stem cells. Dev. Biol. **283:** 486–502.

Kelso, R. J., M. Buszczak, A. T. Quinones, C. Castiblanco, S. Mazzalupo *et al.*, 2004 Flytrap, a database documenting a GFP protein-trap insertion screen in *Drosophila melanogaster*. Nucleic Acids Res. **32:** D418–D420.

Kumar, A., S. Agarwal, J. A. Heyman, S. Matson, M. Heidtman *et al.*, 2002 Subcellular localization of the yeast proteome. Genes Dev. **16:** 707–719.

Lasko, P., 2003 Cup-ling oskar RNA localization and translational control. J. Cell Biol. **163:** 1189–1191.

Li, X., R. A. Harrell, A. M. Handler, T. Beam, K. Hennessy *et al.*, 2005 piggyBac internal sequences are necessary for efficient transformation of target genomes. Insect Mol. Biol. **14:** 17–30.

Lilly, M. A., M. de Cuevas and A. C. Spradling, 2000 Cyclin A associates with the fusome during germline cyst formation in the Drosophila ovary. Dev. Biol. **218:** 53–63.

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

Metaxakis, A., S. Oehler, A. Klinakis and C. Savakis, 2005 Minos as a genetic and genomic tool in *Drosophila melanogaster*. Genetics **171:** 571–581.

Misra, S., M. A. Crosby, C. J. Mungall, B. B. Matthews, K. S. Campbell *et al.*, 2002 Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. **3:** RESEARCH0083.

Morin, X., R. Daneman, M. Zavortink and W. Chia, 2001 A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in Drosophila. Proc. Natl. Acad. Sci. USA **98:** 15050–15055.

Quiñones-Coello, A. T., L. N. Petrella, K. Ayers, A. Melillo, S. Mazzalupo *et al.*, 2007 Exploring strategies for protein trapping in Drosophila. Genetics **175:** 1089–1104.

Ross-Macdonald, P., A. Sheehan, C. Friddle, G. S. Roeder and M. Snyder, 1999 Transposon mutagenesis for the analysis of protein production, function, and localization. Methods Enzymol. **303:** 512–532.

Royzman, I., R. J. Austin, G. Bosco, S. P. Bell and T. L. Orr-Weaver, 1999 ORC localization in Drosophila follicle cells and the effects of mutations in dE2F and dDP. Genes Dev. **13:** 827–840.

Sepp, K. J., and V. J. Auld, 1999 Conversion of lacZ enhancer trap lines to GAL4 lines using targeted transposition in *Drosophila melanogaster*. Genetics **151:** 1093–1101.

Singer, T., and E. Burke, 2003 High-throughput TAIL-PCR as a tool to identify DNA flanking insertions. Methods Mol. Biol. **236:** 241–272.

Skarnes, W. C., J. E. Moss, S. M. Hurtley and R. S. Beddington, 1995 Capturing genes encoding membrane and secreted proteins important for mouse development. Proc. Natl. Acad. Sci. USA **92:** 6592–6596.

Skarnes, W. C., H. von Melchner, W. Wurst, G. Hicks, A. S. Nord *et al.*, 2004 A public gene trap resource for mouse functional genomics. Nat. Genet. **36:** 543–544.

Snapp, E. L., T. Iida, D. Frescas, J. Lippincott-Schwartz and M. A. Lilly, 2004 The fusome mediates intercellular endoplasmic reticulum connectivity in Drosophila ovarian cysts. Mol. Biol. Cell **15:** 4512–4521.

Spradling, A. C., D. Stern, A. Beaton, E. J. Rhem, T. Laverty *et al.*, 1999 The Berkeley Drosophila Genome Project gene disruption project: single *P*-element insertions mutating 25% of vital Drosophila genes. Genetics **153:** 135–177.

Stathopoulos, A., and M. Levine, 2005 Genomic regulatory networks and animal development. Dev. Cell **9:** 449–462.

Thibault, S. T., M. A. Singer, W. Y. Miyazaki, B. Milash, N. A. Dompe *et al.*, 2004 A complementary transposon tool kit for *Drosophila melanogaster* using P and piggyBac. Nat. Genet. **36:** 283–287.

Tomancak, P., A. Beaton, R. Weiszmann, E. Kwan, S. Shu *et al.*, 2002 Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol. **3:** RESEARCH0088.

Vasudevan, S., and S. W. Peltz, 2003 Nuclear mRNA surveillance. Curr. Opin. Cell Biol. **15:** 332–337.

Ward, R. E., IV, R. S. Lamb and R. G. Fehon, 1998 A conserved functional domain of Drosophila coracle is required for localization at the septate junction and has membrane-organizing activity. J. Cell Biol. **140:** 1463–1473.

Wilhelm, J. E., and C. A. Smibert, 2005 Mechanisms of translational regulation in Drosophila. Biol. Cell **97:** 235–252.

Wilhelm, J. E., M. Hilton, Q. Amos and W. J. Henzel, 2003 Cup is an eIF4E binding protein required for both the translational repression of oskar and the recruitment of Barentsz. J. Cell Biol. **163:** 1197–1204.

Wilhelm, J. E., M. Buszczak and S. Sayles, 2005 Efficient protein trafficking requires trailer hitch, a component of a ribonucleoprotein complex localized to the ER in Drosophila. Dev. Cell **9:** 675–685.

Communicating editor: T. Schüpbach