The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes

Hugo J. Bellen[*], Robert W. Levis[†], Guochun Liao[‡,1], Yuchun He[*], Joseph W. Carlson[§], Garson Tsang[‡], Martha Evans-Holm[‡], P. Robin Hiesinger[*], Karen L. Schulze[*], Gerald M Rubin[‡], Roger A. Hoskins[§] and Allan C. Spradling[†]

[*]Howard Hughes Medical Institute

Department of Molecular and Human Genetics

Program in Developmental Biology

Baylor College of Medicine

Houston, TX 77030

[†]Howard Hughes Medical Institute Research Laboratories

Department of Embryology

Carnegie Institution of Washington

115 West University Parkway

Baltimore, MD 21210

[‡]Howard Hughes Medical Institute

Depart. of Molecular and Cellular Biology

Life Sciences Annex Bldg.

University of California, Berkeley

Berkeley, CA 94720-3200

[§]Lawrence Berkeley National Laboratory

Berkeley, CA 94720-3200

[1]Current address: Roche Palo Alto

3431 Hillview Avenue

Palo Alto, CA 94304

Running Head: Drosophila gene disruption project

Keywords: *P*-element, *piggyBac*, insertion, mutation

Corresponding author:

Allan Spradling

Department of Embryology

Carnegie Institute

Baltimore, MD 21210

Email: spradling@ciwemb.edu

ABSTRACT

The Berkeley Drosophila Genome Project (BDGP) strives to disrupt each Drosophila gene by the insertion of a single transposable element. As part of this effort, transposons in more than 30,000 fly strains were localized and analyzed relative to predicted Drosophila gene structures. Approximately 6,300 lines that maximize genomic coverage were selected to be sent to the Bloomington Stock Center for public distribution, bringing the size of the BDGP gene disruption collection to 7,140 lines. It now includes individual lines predicted to disrupt 5,362 of the 13,666 currently annotated Drosophila genes (39%). Other lines contain an insertion at least 2 kb from others in the collection and likely mutate additional incompletely annotated or uncharacterized genes and chromosomal regulatory elements. The remaining strains contain insertions likely to disrupt alternative gene promoters or to allow gene mis-expression. The expanded BDGP gene disruption collection provides a public resource that will facilitate the application of Drosophila genetics to diverse biological problems. Finally, the project reveals new insight into how transposons interact with a eukaryotic genome and helps define optimal strategies for using insertional mutagenesis as a genomic tool.

INTRODUCTION

Mutations represent an essential tool for analyzing gene function. In recent years, organized efforts to generate genome-wide mutant collections have progressed substantially in model organisms such as *S. cerevisiae* (GIAEVER *et al.* 2002; BIDLINGMAIER and SNYDER 2002; reviewed in VIDAN AND SNYDER, 2001), *C. elegans* (JANSEN *et al.* 1997), *A. thaliana* (ALONSO *et al.* 2003), *D. rerio* (GOLLING *et al.* 2002; reviewed in AMSTERDAM, 2003), *M. musculus* (MITCHELL *et al.* 2001; MIKKERS *et al.* 2002; reviewed in STANFORD *et al.* 2001), and many other organisms (ROOS *et al.* 1997; AKERLEY *et al.* 2002; FIRON *et al.* 2003; UHL *et al.* 2003). Transposable elements are now widely used in such efforts (GUEIROS-FILHO and BEVERLEY

1997; FADOOL *et al.* 1998; KLINAKIS *et al.* 2000; ZAGORAIOU *et al.* 2001; BESSEREAU *et al.* 2001).

Insertional mutagenesis using engineered transposable elements has proved to be one of the most productive and versatile approaches to disrupting and manipulating Drosophila genes on a genome-wide scale (COOLEY *et al.* 1988; BIER *et al.* 1989; BELLEN *et al.* 1989; GROSSNICKLAUS *et al.* 1989, BERG and SPRADLING 1991; KARPEN and SPRADLING 1992, GAUL *et al.* 1992, TÖROK *et al.* 1993, CHANG *et al.* 1993; ERDELYI *et al.* 1995; RØRTH 1996; DEAK *et al.* 1997; SALZBERG et al. 1997; RØRTH *et al.* 1998; SEKELSKY *et al.* 1999; MATA *et al.* 2000; BOURBON *et al.* 2002; OH *et al.* 2003; HÄCKER *et al.* 2003). Collections of insertion mutations have been created with independently scorable genetic markers such as eye color, body color, drug resistance, or dominant visible characters, allowing multiple insertions to be easily manipulated. Moreover, specialized transposons have been utilized that trap enhancers (O'KANE and GEHRING 1987; WILSON *et al.* 1989, BIER et al 1989), drive GAL4 production (BRAND and PERRIMON 1993; MANSEAU et al 1997, LUKACSOVICH *et al.* 2001; HORN *et al.* 2003), mis-express adjacent genes (Rorth, 1996; Crisp and Merriam, 1997, TOBA *et al.* 1998; MATA *et al.* 2000; AIGAKI *et al.* 2001; BRENNECKE *et al.* 2003), fuse endogenous proteins to GFP (MORIN *et al.* 2002) or a combination of these properties.

The *P* transposable element has been the vehicle most widely used to disrupt Drosophila genes because it transposes at high rates, depends completely on exogenous transposase, inserts in heterochromatic as well as euchromatic regions (ZHANG and SPRADLING, 1994; ROSEMAN *et al.* 1995; WALLRATH and ELGIN, 1995; YAN *et al.* 2002; KONEV et al. 2003), preferentially transposes near promoters (SPRADLING *et al.* 1995), excises imprecisely, generates local deletions from single elements or between element pairs (PRESTON *et al.* 1996; COOLEY *et al.*

1990; HUET *et al.* 2003; reviewed in GRAY, 2000), transposes locally (TOWER *et al.* 1997; TIMAKOV et al. 2002), induces male recombination (PRESTON and ENGELS, 1996), preferentially replaces existing elements (HESLIP and HODGES, 1994; SEPP and AULD, 1999), and induces unequal recombination in tandem repeats (THOMPSON-STEWART *et al.* 1994). However, these advantages must be balanced against the inefficiency resulting from transposon hotspots (SPRADLING *et al.* 1999), and the possibility that not all genes are *P* element targets. Recently, the TTAA-specific *piggyBac* element (CARY *et al.* 1989) has been shown to function as an alternative insertion vector with many attractive features (HORN *et al.* 2003), including compatibility with *P*-containing strains (HÄCKER *et al.* 2003).

Beginning in 1993, the Berkeley Drosophila Genome Project established a gene disruption library encompassing 1,045 genes with mostly vital function (SPRADLING *et al.* 1995; 1999). These lines were selected from seven *P* element insertional mutagenesis screens, and following insert localization by polytene chromosome in situ hybridization, were verified and associated with genes by complementation tests. While this collection proved extremely useful, its coverage was limited by the requirement for mutations with a scorable phenotype and by the amount of time required for extensive complementation testing.

The sequencing of the eukaryotic portion of the Drosophila genome (ADAMS *et al.* 2000; CELNIKER *et al.* 2002) and the partial sequencing of heterochromatic portion (CELNIKER *et al.* 2002), as well as the detailed annotation of these sequences (MISRA *et al.* 2002; HOSKINS *et al.* 2002) using EST and full-length cDNA sequences (STAPLETON *et al.* 2002) provided an opportunity to greatly expand the collection's coverage. Transposon insertions in newly generated lines could now be precisely localized by sequencing genomic DNA flanking the insertions and computationally associated with known or predicted genes. Using this approach

to rapidly select a subset of lines bearing insertions in genes that had not previously been disrupted was proposed as a way to further grow the BDGP collection (SPRADLING *et al.* 1999).

There are significant challenges to applying a sequence-based strategy successfully on a large scale, however. The *P* element target sequences are broad but nonrandom (LIAO *et al.* 2000). Why certain genes act as hotspots, while others are rarely targeted, remains unknown. How design parameters such as the structure and location of the starting mutator transposon affect the spectrum of hotspots and the diversity of genes that are targeted remains poorly characterized. The *piggyBac* transposon has been suggested to be superior to the *P* element as an insertional mutagen with a broad specificity (HÄCKER *et al.* 2003), but variables affecting *piggyBac* screens are also little known. Nor can such information be easily determined. In a production-oriented project, the number of different transposition schemes that can be evaluated is limited. Preparing and testing new screen designs requires months of lead-time and risks productivity should the screen prove to be inefficient in practice.

Another challenge in a sequence-based strategy is selecting which insertion lines to save. Because of the limited capacity of public Drosophila stock centers it is crucial to preserve lines whose insertions are most likely to disrupt independent genes. Without phenotypes and complementation tests to serve as guides, choosing lines that disrupt distinct genes depends on having a highly accurate annotation of the genome sequence. Drosophila genes undergo complex splicing patterns, reside close to their neighbors and often overlap. Line selection based on inaccurate or incomplete annotation would substantially reduce the project's output by mistakenly causing genetically redundant strains to be retained and novel strains to be discarded.

Here we report expanding the BDGP gene disruption collection from 1,045 to 7,140 strains using a sequence-based strategy. Lines in the collection are predicted to bring at least

5,362 of the 13,666 annotated genes under experimental control. In the process we have begun to answer some of the questions concerning the efficient design of insertional mutagenesis screens.

MATERIALS AND METHODS

The EP collection: The original EP screen (RØRTH 1996; RØRTH *et al.* 1998) was carried out in collaboration with BDGP. The 2,266 lines generated in this project served as the test bed for developing high throughput methods for sequencing transposon flanks (LIAO *et al.* 2000). This screen utilized the original EP element (RØRTH 1996) whose heat shock promoter-derived mis-expression cassette cannot be activated in the female germ line (MATA *et al.* 2000; see Table 1). Because of this limitation, lines from other sources were favored and only 374 EP lines remain in the primary collection (Tables 2 and 3).

The BG (Baylor Genetrap) collection: The BG screen used the "gene trap" *P{GT1}* element developed by LUKACSOVICH *et al.* (2001). *P{GT1}* is designed to express the *white*$^{+}$ gene only when inserted within a gene and to fuse a Gal4-containing exon with this target gene (Table 1). The BG screen was carried out as shown below in one of six isogenized backgrounds. The *w; Iso2A/Iso2A; Iso3A/Iso3A* isogenized stocks (*Iso A* to *Iso F*) were obtained from Cahir O'Kane at the University of Cambridge (personal communication). They were tested in the following behavioral assays and most were judged similar to wild type *Canton S* flies: 1) benzaldehyde jump responses at different drug concentrations; 2) locomotor activity; 3) circadian rhythm; and 4) heat avoidance in an associative learning paradigm. Six pairs of isogenized male and female starting stocks (see first cross below) were constructed thereby avoiding mixing of genetic backgrounds. The starting site of the mutator element (which we termed BG00000) was sequenced (GenBank accession: CL004094), but found to reside entirely

within a blastopia transposon and hence its exact genomic position on *CyO* was not be localized. The crossing scheme used was as follows:

$$\frac{w}{Y};\frac{Sp}{CyO,\ P\{w^{+mGT}=GT1\}};\frac{Iso3A}{Iso3A}\ \mathbf{X}\ \frac{w}{w};\frac{Iso2A}{Iso2A};\frac{TMS,\ P\{ry^{+t7.2},\ Delta2\text{-}3\}99B,\ Sb^{1}}{TM3,\ Ser^{1}}$$

$$\downarrow$$

$$\frac{w}{Y};\frac{Iso2A}{CyO,\ P\{w^{+mGT}=GT1\}};\frac{Iso3A}{TMS,\ P\{ry^{+t7.2},\ Delta2\text{-}3\}99B,\ Sb^{1}}\ \mathbf{X}\ \frac{w}{w};\frac{Iso2A}{Iso2A};\frac{Iso3A}{Iso3A}$$

A single w^+; Cy^+; Sb^+ fly was selected per vial to avoid clusters. The jumping rate was 1 or more in 15% of the vials. These flies have the following genotype:

$$\frac{w}{w};\frac{Iso2A}{Iso2A};\frac{Iso3A}{Iso3A}$$

and have an insert of *P{GT1}* that is w^+. They were crossed to $y^1\ w^{67c23};\ L^2/CyO;\ D^1/TM3,\ Sb^1$ and w^+, Cy^- and Sb^- progeny were kept. Upon determination of the insertion site, the appropriate chromosome was balanced by backcrossing to the *y w; L/CyO; D/TM3, Sb* flies. We generated 2,869 BG strains, 482 of which were selected for the primary collection (Table 2 and 3). Approximately 1,500 of the original BG stocks are available from Trudi MacKay at North Carolina State University. Because of their uniform genetic background, the BG collection has proved useful in studies of quantitative traits including bristle number (NORGA *et al.* 2003) and starvation resistance (HARBISON *et al.* 2004).

The KG (Karpen Genome) collection: The KG screen made use of the *P{SUPor-P}* element (ROSEMAN et al. 1995; FlyBase ID FBtp0001587; Table 1), which was designed to

facilitate insertion recovery by reducing position effects on the white gene due to the presence of two suppressor of Hairy-wing [su(Hw)] binding regions that can act as chromatin insulators. Another potential benefit was the possibility that this *P* element may enhance the rate of mutagenesis as reported previously (ROSEMAN et al. 1995; BELLEN, 1999). Indeed, when inserted between an enhancer and its cognate promoter, a situation likely to be common due to the *P* element's strong promoter target preference (SPRADLING et al. 1995), the insulators may alter gene expression.

The laboratory of Gary Karpen generated 1,236 of the 10,587 KG strains (strains KG00001 – KG00560 and KG01121 – KG01798) as a byproduct of their screen for heterochromatic *P*-element insertions (YAN *et al.* 2002, KONEV *et al.* 2003). They used nine mating schemes with three different *P{SUPor-P}* starting sites and saved exceptional progeny in which there was variegated expression of the *yellow* transgene. They sent us progeny with new insertions in which the *yellow* transgene was expressed normally. We created lines by crossing them to *y; ry*^{506} flies of the opposite sex. The insertion-bearing chromosome of lines selected for the primary collection was balanced, using *FM4/Df(1)260-1, y* for *X*-chromosome insertions; *y*^{1}*; SM6a/ In(2LR)Gla, wg*$^{Gla-1}$ for chromosome *2* insertions; *y*^{1}*; TM3, Sb*^{1} */D*^{1} for chromosome *3* insertions and *y*^{1} *; ci*^{D}*/ey*^{D} for chromosome *4* insertions.

We generated the other KG strains using stocks provided by Gary Karpen's lab. These stocks employed an isogenized *y; ry*^{506} background that had been found in a previous large screen to be uniform and free of *hobo* elements and other sources of "background" mutations (KARPEN AND SPRADLING 1992; SPRADLING *et al.* 1999). We mapped the starting site of the *P{SUPor-P}* mutator element (which we term KG00000) to chromosome *2L* position 2748009

(CELNIKER *et al.* 2002) (equivalent to scaffold segment AE003582.3 position 220758; GenBank accession: CL004095).

The crossing scheme was:

$$\frac{y}{y};\frac{Sp}{CyO,\ P\{y^{+}mDint2\ w^{BR.E.BR}=SUPor\text{-}P\}};\frac{ry^{506}}{ry^{506}}\ \mathbf{X}\ \frac{X^{\wedge}Y}{Y};\frac{+}{+};\frac{TMS,\ P\{ry^{+t7.2},\ Delta2\text{-}3\}99B,\ Sb^{1}}{ry^{506}}$$

$$\downarrow$$

$$\frac{y}{Y};\frac{+}{CyO,\ P\{y^{+}mDint2\ w^{BR.E.BR}=SUPor\text{-}P\}};\frac{ry^{506}}{TMS,\ P\{ry^{+t7.2},\ Delta2\text{-}3\}99B,\ Sb^{1}}\ \mathbf{X}\ \frac{y}{y};\frac{+}{+};\frac{ry^{506}}{ry^{506}}$$

and y^{+}, ry^{-} flies were selected. The jumping rate was 1 or more in 35% of vials. These flies have the following genotype:

$$\frac{y}{y};\frac{+}{+};\frac{ry^{506}}{ry^{506}}$$

and carry a *P{SUPor-P}* element on one of the chromosomes. They were backcrossed to *y/y; +/+; ry^{506}/ry^{506}*. After having been chosen for the primary collection some KG strains were selected for homozygosity. All *X*-chromosome insertions were kept in this genetic background and balanced with *FM7*. Many but not all second, third and fourth chromosome insertions were rebalanced with *y^{1} w^{67c23}; L^{2}/CyO* or *y^{1} w^{67c23}; D^{1}/TM3, Sb^{1}* or *y^{1} w^{67c23}; ci^{D}/ey^{D}*, respectively.

The EY (EP yellow) collection

In an effort to broaden its target specificity, we modified the second generation EP element of MATA *et al.* (2002) that supports germ cell expression. An intronless *yellow+* gene marker was inserted adjacent to the original mini-*white* gene in this *P{EPg}* element (see below). The resulting element, *P{EPgy2},* was termed the EY element. Mis-expression is still driven from an

outwardly directed promoter at the 3' end (Table 1; rightward-pointing arrow). We localized the starting site for the EY screen (EY00000) at nucleotide position 21451923 on the minus strand of the 2L chromosome arm (equivalent to nucleotide 57866 of scaffold segment AE003781.4; GenBank accession: CL004093).

The following crossing scheme was used to generate the EY lines:

$$\frac{X^{\wedge}Y}{Y};\frac{+}{+};\frac{TMS,\ P\{ry^{+t7.2},\ Delta2\text{-}3\}99B,\ ry^{2}\ Sb^{1}}{ry^{506}}\ \mathbf{X}\ \frac{y\ w}{y\ w};\frac{L}{CyO,\ P\{w^{+mC}\ y^{+mDint2}=EPgy2\}};\frac{+}{+}$$

$$\downarrow$$

$$\frac{y\ w}{Y};\frac{+}{CyO,\ P\{w^{+mC}\ y^{+mDint2}=EPgy2\}};\frac{+}{TMS,\ P\{ry^{+t7.2},\ Delta2\text{-}3\}99B,\ ry^{2}\ Sb^{1}}\ \mathbf{X}\ \frac{y\ w}{y\ w};\frac{+}{+};\frac{+}{+}$$

We selected y+ and w+ flies and crossed them to $y^{1}\ w^{67c23}/y^{1}\ w^{67c23};\ +/+;\ +/+$. The jumping rate was 1 or more in 65% of vials. Upon sequencing the genomic DNA adjacent to the *P* element the insertion chromosomes were balanced with *FM7* or $y^{1}\ w^{67c23};\ L^{2}/CyO$ or $y^{1}\ w^{67c23};\ D^{1}/TM3,\ Sb^{1}$ or $y^{1}\ ;\ ci^{D}/ey^{D}$.

Donated collections: Several collections of strains containing a variety of transposon mutators were donated to the Gene Disruption Project (Table 1). With the exception of the PA and PC collections, the insertion site flanking sequences of all donated strains described in this paper were amplified, sequenced, and mapped using the same procedures, described below, that were used for lines generated within the project. In most cases, we extracted genomic DNA from samples of frozen flies collected from unbalanced stock that were provided by the lab donating the strains.

The PA and PC strains were donated by Brian Ring and Daniel Garza. Each strain contained a single autosomal insertion of a *piggyBac* element. The mutator transposon in the PA

strains was *PBac{5HPw^{+}}* (FlyBase ID FBtp0016567), marked with mini-*white*, while the PC strains carried *PBac{3HPy^{+}}* (FlyBase ID FBtp0016566), marked with *yellow*. DNA segments flanking the insertion sites were amplified and sequenced by Exelixis Corp. The Gene Disruption Project received data on the insertion sites of 1,055 PA and PC lines with insertions that could be mapped to unique euchromatic sites. We initially selected 471 of these lines as candidates for the permanent collection, but some of these lines were lost before balanced stocks could be established. Brian Ring and Kathy Matthews constructed balanced stocks of the 342 surviving lines. Kathy Matthews prepared samples of frozen flies from the balanced stocks, which we used to recheck the insertion site flanking sequences (see below).

The KV strains were generated in the laboratory of Gary Karpen using the *P* element mutator *P{SUPor-P}*. They employed a variety of starting sites and genetic crossing schemes, as described by YAN *et al.* (2002) and KONEV *et al.* (2003), to maximize the recovery of heterochromatic insertions. Many sequences flanking KV insertion sites mapped to WGS3 heterochromatic scaffolds whose chromosomal origin is currently unknown. Gary Karpen provided unpublished FISH mapping data for some of these lines; we mapped others to a chromosome by genetic segregation of the transgene markers. We balanced the insertion-bearing chromosomes using *FM4/Df(1)260-1, y* for *X*-chromosome insertions; $y^{1}\ w^{67c23}$*; SM6a/ In(2LR)Gla,* wg^{Gla-1} for chromosome *2* insertions; $y^{1}\ w^{67c23}$*; TM3,* Sb^{1} */*D^{1} for chromosome *3* insertions and y^{1} *;* ci^{D}*/*ey^{D} for chromosome *4* insertions.

The DG strains were made in the laboratory of William Gelbart using the *P{wHy}* element (HUET *et al.* 2002; FlyBase ID FBtp0016141) as a mutator. This is a compound element with *P*-element ends flanking a non-autonomous *hobo* element. Frozen fly samples of 1,384 lines were provided.

The PL strains have insertions of the *piggyBac pBac{GAL4D, EYFP}* mutator element (HORN *et al.* 2003; FlyBase ID FBtp0017476) that is marked with *EYFP* and can act as an enhancer trap to express the GAL4Δ variant. HÄCKER *et al.* (2003) have described a screen in which 798 lines were created that had an insertion of this mutator on chromosome *3*. The third chromosome used as a target had an *P{FRT}* insertion at the base of both chromosome arms that can be used to generate germ-line clones. Udo Häcker provided samples of 634 lines with homozygous-viable insertions from this collection. Because the samples used for this determination came from stocks in which the insertion-bearing third chromosome had already been made homozygous, the lines that we selected for the primary collection were sent to the Bloomington Stock Center without rechecking the insertion flanks.

The LA strains were made in the laboratories of John Merriam, Judith Lengyel and Stephen Poole using the *P*-element mutator *P{Mae-UAS.6.11}* (Merriam and Poole, unpublished; FlyBase ID FBtp0001327). This vector is similar to *P{EP}* in having a GAL4-inducible promoter for misexpression of flanking genes, but differs in that it is marked with *yellow* rather than mini-*white*. The mutator was mobilized from an *X* –chromosome site in males and transpositions to the autosomes were recovered as exceptional y^+ males (LENGYEL AND MERRIAM, 1997). We determined this *X*-chromosome starting site to be 11734628 on the plus strand (CELNIKER *et al.* , 2002) (equivalent to scaffold AE003487.2 position 295085). Insertions were subsequently screened for phenotypes when combined with a GAL4 driver, usually the *P{*w^{+mC} *=Act5C-GAL4}25FO1* driver expressing GAL4 under control of the *Actin 5C* promoter (AKIEDA AND MERRIAM, 2001). Samples from 1,045 strains displaying a phenotype were sent for sequencing.

Construction of *P{EPgy2}:* The *P{EPgy2}* element used in the EY screen was a derivative of *P{EPg}* (MATA *et al.* 2002) (FlyBase ID FBtp0012862; Table 1). The major differences were that *P{EPgy2}* contained an intronless *yellow* gene module and lacked the plasmid rescue module of *P{EPg}*. The plasmid pP{EPgy2} was constructed from two plasmid precursors, p1462 and *yellow*-BSX. p1462 was an intermediate used in the construction of pP{EPg} that lacks the plasmid rescue module. It was obtained from Pernille Rørth (EMBL, Heidelberg). The *yellow*-BSX plasmid was used as the source of the intronless *yellow* gene for pP{EPgy2} and was obtained from Tim Parnell in the laboratory of Pamela Geyer (University of Iowa). It had a *Sal* I fragment containing the intronless *yellow* gene cassette (PATTON *et al.* 1992) inserted into the *Sal* I site of a modified pBluescript vector, pBS-X. This vector had the *Kpn* I site of the polylinker converted into an *Xba* I site. The *yellow Sal* I fragment was the same as the segment designated by FlyBase as $y^{+mDint2}5.2(S,S)$ (FlyBase ID FBms0003824). DNA of the *yellow*-BSX plasmid was digested with a combination of *Not* I and *PspOM* I, which generate the same 5' overhang. *Not* I cut *yellow*-BSX at a single site in the polylinker sequences closest to the 3' end of the *yellow* gene and *PspOM* I cut *yellow*-BSX at a single site in the polylinker sequences closest to the 5' end of the *yellow* gene. The 5.8 kb *Not* I – *PspOM* I fragment containing the intronless *yellow* gene was gel-purified and ligated with DNA from p1462 that had been cut by *Not* I and dephosphorylated with shrimp alkaline phosphatase. p1462 had a unique *Not* I site located between mini-*white* and the GAGA/GAL4-UAS modules. Transformants of the *E. coli* strain DH5α were recovered in which the intronless *yellow* gene fragment had inserted into the *Not* I site of p1462 in each of the two relative orientations and these were named pP{EPgy1} and pP{EPgy2} . The mini-*white* and intronless *yellow* genes of pP{EPgy2} are transcribed in the same direction, which is opposite to that of the *P*-element

promoter (Table 1). Portions of both plasmids were sequenced, including the junctions between the two fragments. A compiled sequence for *P{EPgy2}* (*P*-element portion only) is available in FlyBase (FlyBase ID FBrf0157089).

Initial transgenic *Drosophila* lines containing *P{EPgy2}* were made by Alexei Tulin, using the transformation method described by TULIN *et al.* (2002). Lines with an insertion of *P{EPgy2}* on the *CyO* second chromosome balancer were generated by mobilizing the element from the *X*-chromosome of one of the initial transgenic lines using the *TMS, P{ry*$^{+t7.2}$*, Delta2-3}99B, Sb*^{1} chromosome as a source of transposase.

Determination of Flanking Sequences: Genomic sequences flanking *P* element or *piggyBac* insertions were determined by sequencing inverse PCR products (LIAO *et al.* 2000). A detailed protocol is available on the *P*-Screen webpage at http://flypush.imgen.bcm.tmc.edu/pscreen/.

Genomic DNA was prepared from about 15 insertion-bearing adults. Flies were collected and frozen at -80° in microfuge tubes. Samples were thawed on ice, and three autoclaved stainless steel ball bearings (BALL-1B, Wheels Manufacturing, Broomfield, CO) and 400 µl of Buffer A (100mM Tris-HCl, pH7.5, 100mM EDTA, 100mM NaCl, 0.5% SDS) were added. Samples were disrupted by vigorous vortexing and incubated at 65° for 30 minutes. Debris was precipitated by addition of 800 µl of a 4.3M LiCl / 1.4M KOAc solution, incubation on ice for 10 minutes, and centrifugation at room temperature in a microcentrifuge at 14,000 rpm for 15 minutes. The supernatant was collected, and DNA was precipitated by addition of 800 µl of isopropanol and centrifugation at 14,000 rpm for 10 min. The precipitate was washed with 70% ethanol, air-dried, and resuspended in 75 µl of TE (10 mM Tris pH7.5, 1 mM EDTA), Subsequent steps were performed in 96-well format.

Genomic DNA samples (10 µl) were digested with an appropriate restriction enzyme (5 to 20 units) and RNase A (8 µg/ml) in a 25 µl reaction at 37° for 2.5 hours. The restriction enzyme was inactivated at 65° for 20 minutes. Digested samples (10 µl) were self-ligated with 2 units of T4 DNA ligase at 4° for 12 hours in a dilute reaction (400 µl) to favor the generation of circular products. Ligated samples were precipitated with the addition of 40 µl 3M NaOAc and 1 ml ethanol, and precipitates were washed in 70% ethanol and resuspended in 150 µl TE, pH7.5. Ligation products (10 µl) were used as templates in inverse PCR reactions (50 µl) with 100 µM dNTPs, 0.2 µM oligonucleotide primers, and 2 units of Taq DNA polymerase (Amersham). Reactions were denatured at 95° for 5 minutes, subjected to 35 cycles of denaturation at 95° for 30 seconds, annealing at the appropriate temperature for 1 minute, and extension at 68° for 2 minutes, and a final extension at 72° for 10 minutes.

Flanking sequences were determined by direct sequencing of the inverse PCR products. To remove excess PCR primers and dNTPs, exonuclease I (5 units) and shrimp alkaline phosphatase (2 units) were added directly to an aliquot of PCR reaction (10 µl), the mixture was incubated at 37° for 30 minutes, and the enzymes were inactivated by incubation at 70° for 15 minutes. Sequencing reactions were performed using BigDye terminator chemistry (Perkin-Elmer) at one-quarter of the manufacturer's recommended scale, and sequence data were collected using an ABI 3700 capillary device. With the exception of the LA screen, amplification and sequencing was attempted on both the 5' and 3' flanks of each insertion.

For the BG collection (*P{GT1}* insertions), genomic DNA was digested with *HinP1*; 3' flanks were amplified with the oligonucleotide primers Pry1 (CCTTAGCATGTCCGTGGGGTTTGAAT) and Pry4 (CAATCATATCGCTGTCTCACTCA) at an annealing temperature of 55° and sequenced with Spep1 (GACACTCAGAATACTATTC);

5’ flanks were amplified with pGT1.5a (CCGCACGTAAGGGTTAATG) and pGT1.5d (GAAGTTAAGCGTCTCCAGG) at an annealing temperature of 55° and sequenced with Sp1 (ACACAACCTTTCCTCTCAACAA).

For the KG and KV collections (*P{SUPorP}* insertions), genomic DNA was digested with Hpa II; 3’ flanks were amplified with Pry4 (CAATCATATCGCTGTCTCACTCA) and 3.rev.hpa2 (TTGCCACTTGCTCATACGTC) at an annealing temperature of 55° and sequenced with 3.SUP.seq1 (TATCGCTGTCTCACTCAG); 5’ flanks were amplified with Plac1 (CACCCAAGGCTCTGCTCCCACATT) and Pwht1 (GTAACGCTAATCACTCCGAACAGGTCACA) at an annealing temperature of 60° and sequenced with 5.SUP.seq1 (TCCAGTCACAGCTTTGCAGC).

For the EY collection (*P{EPgy2}* insertions), genomic DNA was digested with *Hpa II*; 3’ flanks were amplified with Pry1 and Pry4 as described above and sequenced with 3.SUP.seq1; and 5’ flanks were amplified with Plac1 and Pwht1 as described above and sequenced with 5.SUP.seq1.

For the LA collection (*P{Mae-UAS.6.11}* insertions), genomic DNA was digested with *Rsa I*; 5’ flanks were amplified with LA(f).1 (GGGAATTGGGAATTCGTTAA) and LA(r).1 (TAGCGACGTGTTCACTTTGC) at an annealing temperature of 55° and sequenced with LA(f)seq1 (CTCTCAACAAGCAAACGTGC).

For the PL collection (*Pbac{GAL4D, EYFP}* insertions), genomic DNA was digested with *Hae III*; 3’ flanks were amplified with PRF (CCTCGATATACAGACCGATAAAACACATGC) and PRR (AGTCAGTCAGAAACAACTTTGGCACATATC) at an annealing temperature of 65° and sequenced with PRF; 5’ flanks were amplified with PLF

(CTTGACCTTGCCACAGAGGACTATTAGAGG) and PLR (CAGTGACACTTACCGCATTGACAAGCACGC) at an annealing temperature of 65° and sequenced with PLF.

The initial determination of the flanking sequences of the PA and PC strains was done by the Exelixis Corporation in collaboration with Brian Ring and Daniel Garza, prior to the donation of these strains to our project. We re-checked the flanking sequences of balanced or homozygous stocks of strains selected for the primary collection. Genomic DNA was digested with *HinP*1 I (3' flank) or *Sau*3A (5' flank); 3' flanks were amplified with 3F1 (CCTCGATATACAGACCGATAAAAC) and 3R1 (TGCATTTGCCTTTCGCCTTAT) at an annealing temperature of 55° and sequenced with pB-3SEQ (CGATAAAACACATGCGTCAATT); 5' flanks were amplified with 5F1 (GACGCATGATTATCTTTTACGTGAC) and 5R1 (TGACACTTACCGCATTGACA) at an annealing temperature of 55° and sequenced with pB-5SEQ (CGCGCTATTTAGAAAGAGAGAG).

Analysis and Alignment of Flanking Sequences: Sequence traces were processed using phred (EWING et al. , 1998; EWING AND GREEN, 1998) to generate base calls with associated quality scores (error probabilities). Proximal vector-genome junction sequences were identified by text searches for several short sequences from the *P*-element or *piggyBac* ends, allowing as many as three nucleotide mismatches per short sequence match. This approach was taken because sequence quality near the beginnings of the traces was variable, so that exact matches to the vector end sequence were not identified in all cases. It achieved almost the same recognition rate as human curators. Distal genome-vector junction sequences were identified by text searching for the appropriate restriction site. Using this approach, the restriction site could

be missed due to low sequence quality. To avoid extending flanking sequences into the vector sequence in such cases, each sequence was compared to the *P*-element or *piggyBac* sequence using BLASTN (ALTSCHUL et al., 1997), and likely vector sequences were removed.

The beginning and end of the high-quality portion of each sequence were defined by identifying low-quality regions based on phred quality scores. A region of low sequence quality was defined as five or more consecutive nucleotides each with a quality score of less than q20 (error probability greater than 1%). If a high-quality region of less than 25 bases of flanking genomic sequence was obtained, then the quality threshold was lowered to q15 (error probability greater than 3.2%). This ensured that at least 25 bases of high-quality flanking sequence were obtained in most cases.

Sequences were trimmed to remove vector- and low-quality sequences. If the proximal vector-genome junction could not be identified, then the sequence was trimmed to begin at the first base of the high-quality region. The distal sequence was trimmed at the restriction site or the last base in the high-quality region, whichever resulted in the shorter flanking sequence. Excluding EP, PA and PC lines, one or both flanking sequences at least 25 bases in length were obtained for 24,157 insertions in 27,642 lines (87%).

Flanking sequences at least 25 bases in length were aligned to the Release 2 or Release 3 genomic sequence using BLASTN. The 5' and 3' flanking sequences of each insertion were aligned independently. Sequence matches with greater than 90% identity over more than 90% of the flanking sequence were saved as alignments. BLASTN results for flanking sequences that did not yield alignments by these criteria were examined by human curators, and curated alignments were used in some cases. If a sequence aligned to multiple locations, indicating a repetitive sequence, or to no location, usually due to a short sequence, then the results were examined by a

human curator and assigned an insertion coordinate if possible. If both the 5' and 3' flanking sequences of an insertion were available but aligned to different genomic sites separated by more than 10 bp and if neither flanking sequence showed evidence of cross-contamination from samples in nearby wells, then the two alignments were assumed to correspond to separate insertions in the same fly stock.

The orientation of each mapped insertion relative to the genomic sequence was defined relative to each vector as shown in Table 1. The position of a mapped insertion in the genomic sequence was defined as the first base at the 5' end of the 8-bp target site duplication of *P*-element insertions or the 4-bp target site duplication (always TTAA) of *piggyBac* insertions. In cases in which the vector-insert junction was not recovered in the flanking sequence, usually due to low sequence quality, the insertion site was defined as the first base of the alignment to the genome sequence. In some cases, a flanking sequence aligned to the genomic sequence along only a portion of its length, indicating a sequence dimorphism between the strain used in the genetic screen and the strain used to produce the reference genome sequenced (*y; cn bw sp*; ADAMS *et al.* 2000). In most such cases, the dimorphic sequence corresponded to a known transposable element (KAMINKER *et al.*, 2002). When an insertion mapped within a dimorphic sequence, the genomic insertion site was defined as the position of the most 5' base in the flanking sequence that aligned to the reference genomic sequence.

Excluding EP, PA and PC lines, a total of 21,928 insertions (91% of those from which flanking sequences were recovered) were mapped to unique sites in the genome during this phase of the BDGP Gene Disruption Project. Including previously described results (RØRTH et al 1998; SPRADLING *et al.* 1999), new lines, and re-check sequencing, more than 50,000 insertion ends have been successfully mapped to the Release 3 genomic sequence in this on-going project.

Brian Ring and Daniel Garza provided sequence data produced at Exelixis Corp. on the insertion sites of 1,055 PA and PC lines that they donated to the project. The insertion site data were in the form of 1 kb segments of Release 2 genomic sequence centered near the insertion site. The target site for *piggyBac* transposons is TTAA and we were told that the insertion site in each mutant strain corresponds to the TTAA closest to the center of the 1 kb segment. We were able to align the 1 kb genomic segments of R2 genomic sequence within unique segments of the R3 genomic sequence for 1046 of these strains. Upon rechecking the flanking sequences for 242 PA and PC lines selected for the primary collection, we confirmed many of these sites, while others differed from the originally reported site by an average of less than 100 bp. When a difference was found the sequence determined by the project was taken as correct.

Line selection: During the initial phases of the project lines were selected if their insertion was located within or < 2 kb upstream of an annotated transcription unit not previously mutated in the BDGP collection. Lines were also retained if the insertion was between genes or within an intron and > 2 kb from any insertion already in the collection. The Release 2 sequence annotations displayed on the GeneSeen browser (N. Harris and S.E. Lewis, unpublished) were used for these determinations. After completion of the Release 3 genome sequence, all remaining new lines and all previously selected lines were re-analyzed as follows. First, a Perl script was used to record for each insertion those transcripts in which it was located (defined as from 500 bp upstream of the annotated transcript to the 3' end). A FileMaker Pro script was then used to search each annotated euchromatic gene against the transcript list and record all lines meeting this criterion.

Using this information as a starting point, final decisions for retention were based on a manual examination by a curator of the insertion position relative to nearby gene models, cDNAs

and other data using the Apollo genome browser (LEWIS *et al.* , 2002). To display insertions in Apollo, XML files describing the Release 3.1 sequence annotation were modified by addition of new data "tiers" including the insertion sites and associated descriptors. To be selected, a strain had to be judged likely to mutate or mis-express a novel gene not currently in the BDGP collection (see Results). In addition, lines whose elements were inserted more than 2 kb from the nearest neighboring P element in the collection were generally also retained. These criteria were designed to minimize unnecessary stock maintenance without severely compromising the long term utility of the collection. The functionalities of the different transposons vary substantially (Table 1) and there is no general consensus as to which characteristics (e.g. enhancer trapping, gene mis-expression, deletion generation) deserve the highest priority. When multiple lines existed that disrupted a gene, the decision on which line to keep was based on a variety of factors, including its verification status, associated mutant phenotype, and element type. Because gene models are less certain in the current annotation of heterochromatin, only manual annotation was used in these regions. Overall, manual curation increased the total number of genes with associated insertions from 5,045 (automated curation) to 5,362.

For the studies of insertion site distribution that are presented here, automated curation was used exclusively to ensure that uniform criteria were applied to all data.

Verification: Only lines selected for the permanent collection were balanced or made homozygous. The flanking sequences of stocks destined for the primary collection were determined and analyzed again after balancing. In most cases (>90%), the initial and re-check coordinates were consistent. When no readable sequence or a different location was obtained, the line was either re-cycled for another round of sequencing or discarded. When the initial sequence indicated the presence of a second insertion on the same chromosome, we looked for

the presence of both sites on the re-check. Rarely, previously undetected second insertions were discovered in the re-check phase. Lines donated to the project as balanced or homozygous stocks were usually not re-checked. More than 96% of selected BG and KG lines were verified. Re-check verification of the other screens has not yet been completed at the time of this publication (see Table 2).

P-Screen Webpage: All strains generated by the project (BG, KG and EY) were made publicly available via an online database at the time they were selected for the primary collection (**http://flypush.imgen.bcm.tmc.edu/pscreen/**) as well as selected other lines. This site contains the following information: P-element constructs used, strain name (BG, KG or EY number), genomic insertion site in release 3 coordinates, inferred cytological location, associated gene (hit or nearby), and availability status after incorporation into the Bloomington Stock Center collection. Lines selected from the donated collections (see below) were not listed until they could be distributed by the Stock Center because they were not available to the project for early distribution.

Data Submission: Data describing all insertions and stocks selected for the primary collection were submitted to public repositories after the fly strains were sent to the Bloomington Stock Center (Fig. 1). Flanking sequences of the selected insertions were deposited at GenBank (http://www.ncbi.nlm.nih.gov/). Stock descriptions, including phenotype and balancer information, were submitted along with the insertion stocks to the Bloomington Stock Center (http://flystocks.bio.indiana.edu/). Detailed descriptions of the selected lines, including insertion coordinates and associated genes, were deposited at FlyBase (**http://flybase.bio.indiana.edu/**). Data submission to Flybase is on-going and not yet completed at the time of this publication.

RESULTS

Strategy: We initiated a new strategy to expand the coverage of the BDGP gene disruption collection shortly after the Drosophila genome sequence was first released (ADAMS *et al.* 2000). Lines with single transposon insertions would either be newly generated by the project or received from other laboratories. Lines with new insertions would be recognized solely by a change in the genetic linkage of the transposon marker gene, rather than by any phenotype associated with the insertion. DNA would be prepared from adult flies of each unbalanced insertion line and inverse PCR products containing the genomic region flanking the insertion would be sequenced. Lines whose insertion points could be uniquely localized by sequence comparison to the reference genome sequence would then be added to the primary collection or discarded depending on whether they appeared, based on insert location, to mutate a gene that had not already been disrupted. Information on each newly selected line would be posted on the project website (http://flypush.imgen.bcm.tmc.edu/pscreen/) and the unbalanced strains distributed to the community until stable stocks could be generated. After balancing, the flanking sequence would be re-checked to verify that the desired insertion was still present. If so, the line would be sent to the Bloomington Stock Center for public distribution, and associated information forwarded to the Bloomington Stock Center, Flybase and GenBank. An outline of the overall strategy is given in Fig. 1, and a detailed description of each step may be found in Materials and Methods.

Designing screens to generate new lines with the broadest possible gene coverage presented the first major challenge. There was little theoretical or empirical information on how factors such as element structure or starting site affect coverage, yet the ability of the project to test these variables was limited due to the time required. It has been difficult in the past to

compare the intrinsic efficiency of different screens because large numbers of molecularly analyzed lines are necessary to obtain statistically significant information regarding anything but a few highly mutable genes (hotspots) (BERG and SPRADLING, 1991). High levels of transposition have been associated with the generation of secondary mutations (SPRADLING *et al.* 1999), so the products of such screens were excluded from the project. To obtain baseline data on the feasibility and efficiency of our experimental plan, we first molecularly analyzed the EP collection (RØRTH *et al.* 1998). Subsequently, we applied the strategy of Fig. 1 to the products of three other screens carried out by us, as well as to donated lines from six additional screens, including three that utilized the *piggyBac* transposon (see Table 1).

Associating lines with genes: Before the gene coverage of individual screens can be compared, it is necessary to address inherent ambiguities in the association of transposon insertions and genes based on insert location. It would be conceptually simple to score as a hit only insertions lying within the annotated 5' and 3' limits of a given gene. However, particularly in the case of *P* elements, such an approach would be a highly inaccurate measure of gene disruption. One reason is that *P* elements lying a short distance upstream from the 5' end have been shown in many cases to generate a gene mutation (SPRADLING *et al.* 1995). We used 500 bp as a rough guide for the maximum distance a *P* element can be located 5' to a transcription start site and still be likely to disrupt its function. Secondly, the Release 1 and Release 2 versions of the Drosophila genome annotation that were available during the first three years of the project utilized computationally predicted gene models that usually lacked 5' untranslated exons. Since *P* elements systematically insert near the true gene 5' ends of genes in a highly preferential manner (SPRADLING *et al.* 1995), and promoters are located on average 1.4 kb upstream from the start codon (OHLER *et al.* 2002), many insertions at true gene promoters

would appear to lie more than 500 bp upstream from the nearest gene using the available annotation. Anticipating this problem, during the first three years we saved lines at novel "intergenic" positions and re-analyzed all our data after the Release 3 (R3) annotation became available (MISRA *et al.* 2002). All data reported in this paper are based on the most recent sequence and annotation (Release 3.1) which includes many more 5' and 3' UTRs than previous releases. This strategy significantly increased the completeness and accuracy of insertion-gene associations (Fig. 2A).

We used the more complete Release 3.1 gene models to obtain further information on the *P* element 5' preference using these large data sets. The locations of the insertions in 5,630 primary collection lines relative to their associated transcript 5' ends are plotted in Fig. 2B. It can be seen that *P* elements tend strongly to insert within 100 base pairs symmetrically about the transcription start site. This sharp peak in the distribution could not arise by chance, because annotated R3 start sites are separated on average by 5.6 kb in the genome. Moreover, no such preference for start sites is seen when *piggyBac* insertions are analyzed in an identical manner (Fig. 2B). It can also be seen that a large fraction of all *P* element insertions associated with genes fall within 500 bp of the transcript start site.

Several other factors were considered in associating insertions and genes. Many Drosophila genes lie near or within neighboring genes (Fig. 2C) often within large introns (Fig. 2D). Over 1,000 of the R3 euchromatic genes (7.5%) are nested in the introns of other genes and over 2,000 genes (15%) have annotated transcripts overlapping those of other genes (MISRA *et al.* 2002). There are also many divergently transcribed pairs of genes whose 5' ends lie < 500 bp apart. Overall, about 20% of insertions were judged likely to disrupt two rather than just one gene based on our criteria. Other insertions were located within known or predicted RNA genes

(Lai et al. 2003; Fig. 2E). As little knowledge of their cis-regulatory regions is available, lines with insertions located up to 500 bp 5' or 3' of such genes were saved.

Three additional classes of lines were saved even though they were not associated with genes by the criteria described above. First, a significant number of insertions lie outside and more than 500 bp 5' of any known transcript (Fig. 2F). Such insertions might disrupt unannotated genes and/or regulatory elements. Consequently, we saved a skeleton set of insertions in such regions such that no insertion was closer than 2 kb to its nearest neighbor. Second, unannotated genes may lie within introns of known transcripts, so we applied the 2 kb spacing criteria to inserts in large gene introns as well. However, neither of these types of lines were counted as gene disruptions as reported here. Third, about 24% of Release 3 gene models lack an annotated 5' UTR (MISRA *et al.* 2002) and are prone to the same problems we experienced with Release 2 models. For example, in Fig 2G the BG01357 insertion lies 1866 bp upstream from the R3 annotation for *CG32767*, but this annotation only begins at the putative methionine start codon. Sequence data from cDNA RE54443 (which may not be full length) indicates that the true 5' end(s) lies further upstream and closer to the *P* element. To deal with such problems we manually annotated each insertion. Lines with insertions located 500-2000 bp upstream from the annotated 5' end of a novel gene were sometimes retained in the collection if the available cDNA, EST and modeling data indicated to a human curator that it would likely provide a primary reagent to researchers wishing to genetically manipulate the gene in question. This process resulted in about 300 additional gene associations not recognized by automated annotation. For these reasons, all 7,140 permanent lines are currently useful as reagents and will likely prove to disrupt substantially more genes than the current estimate of 5,362.

Insertional mutagenesis screens vary widely in genomic coverage: First, we used the methods described above to determine how many genes are associated with lines in the EP collection. Our results indicated that a sequence-based strategy of gene disruption could be highly efficient. An average of 686 +/- 10 genes are associated with 1,000 EP insertions (Fig. 3A). Moreover, the rate of double insertions is only about 3% of total jumps using this transposon (RØRTH *et al.* 1998). Despite these attractive parameters, we did not elect to continue generating new EP lines because mis-expression from this element is ineffective in some tissues, such as female germ line (MATA *et al.* 2000). Nevertheless, these figures set a standard by which other screens utilizing elements with other desirable properties could be judged, and allowed the primary collection to be expanded by 374 strains.

The initial screen carried out by BDGP ("the BG screen") utilized a "gene trap" mutator element designed to stimulate GAL4 production under the control of a gene near the insertion site (LUKACSOVICH *et al.* 2001; Table 1). The BG screen utilized a genetic background that had been extensively isogenized, generating lines that minimize between-line genetic diversity (NORGA *et al.* 2003). However, after generating 2,869 lines we realized that this approach was not optimal for the purposes of genomic coverage (Table 2). First, the rate of BG element jumping was only 1 jump per 7 vials, less than half the rate observed with the EP screen. Secondly, the rate of genes hit per 1,000 insertions was also much lower, only 339 +/- 40. This provided the first evidence that the intrinsic genomic coverage of insertional mutagenesis screens is highly dependent on the structure and/or location of the mutator element that is mobilized. Finally, we were troubled by the frequent recovery of lines in which GAL4 was oriented in the opposite direction to the targeted gene. While this might indicate the existence of many more unannotated genes than anticipated in the Drosophila genome, the goals of our project required a

more efficient, predictable mutator. Nonetheless, we added 482 new BG strains to the primary collection.

In search of a better mutator, we switched to generating lines using a previously tested element known as P{SUPor-P} (ROSEMAN *et al.* 1995). We refer to this as the KG element (see Methods). The KG mutator contains two chromatin insulator elements designed to minimize chromosomal position effects and enhance mutability via enhancer blocking. In addition, it houses an intronless yellow gene, which has proven to be much less sensitive to position effects than the mini-*white* gene used in many previous *P* element mutators (ROSEMAN et al 1995). We thought that this might substantially increase screen efficiency because many transpositions, even within euchromatin, may not be detected using the mini-*white* gene due to position effects. We generated and analyzed 10,587 new KG transpositions with generally favorable results, adding 2,129 lines to the final collection (Table 2). However, the efficiency of KG gene disruption remained significantly below the EP benchmark, i.e. 541 +/- 22 vs. 686 +/- 10 genes per 1,000 lines (Fig. 3A), and the KG element does not support gene mis-expression.

Consequently, we switched to generating new lines using a modified version of the EP element (Table 1, Methods). An intronless yellow gene was inserted into the EPg version of EP that allows female germ line expression (MATA *et al.* 2002). We called this element EY and used it to generate 10,310 new lines. As hoped, EY transpositions were linked to genes at the same rate as EP jumps, 691 +/- 25 vs 686 +/- 10 genes/1,000 lines (Fig. 3A). This is significantly more efficient than the KG screen, and allowed the project to add 2,338 new lines to the final collection.

The final 17% of lines analyzed by the project were donated from five external laboratories. The Karpen lab provided additional KG insertion lines, which we termed KV lines,

in which the expression of the yellow gene is variegated. As expected, such lines frequently result from insertion within heterochromatin (YAN *et al.* 2003; KONEV et al. 2003). The Gelbart lab contributed 1,384 lines, which we refer to as DG lines, containing the hybrid *P-hobo* element *P{wHy},* that facilitates the generation of local deletions at the site of insertion (MOHR and GELBART, 2002; HUET *et al.* 2002). The Garza lab contributed 1,055 lines generated using two *piggyBac* mutators we refer to as PA and PC (Table 1), while Udo Häcker donated 634 *piggyBac* lines produced from a screen with a different vector we termed PL (Häcker *et al.* 2003). To gain an initial comparison of mutagenesis using *piggyBac* vs *P* element, we calculated the gene disruption efficiency of the PA/PC *piggyBac* lines. Somewhat, surprisingly, our standard efficiency measure indicated that they hit 679 genes, about the same number as 1,000 EP or EY lines. Thus, by this initial test, the efficiency of *piggyBac* mutagenesis equaled, but did not exceed that of the best *P* elements.

Synergy between element types: Further insight into screen strategy came from examining the cumulative number of genes disrupted for different elements over time in a large screen (Fig. 3B). In a large screen, the incremental yield of new gene disruptions continually decreases during the course of the screen as more and more of the preferred target genes have already been hit. Thus, in designing a screening strategy, consideration must be given not only to the initial gene targeting efficiency, but also to how rapidly the yield decreases as new insertions are added. As expected from the initial efficiency measures, more genes were disrupted by EY jumps compared to KG jumps at each point in the screens. BG transpositions were far less efficient than either.

Next, we investigated whether there was any advantage to using a combination of elements rather than a single element (Fig. 3C). We calculated the incremental gene yield

resulting from 1,000 new KG, EY or PC/PA lines, in a project that had already incorporated 7,000 KG or EY lines. If all mutator elements target the same universe of genes, then their relative efficiency would always be proportional to their initial efficiency. However, if elements target sets of genes that only partially overlap, then the element used initially will become less efficient with time (due to the saturation of its targets) in comparison to a new element. This is in fact what was observed. After 7,000 KG insertions, switching to PA/PC lines was even more favorable than expected from the initial rate measurements. 1,000 *piggyBac* lines at this point added 421 more gene associations compared to 248 for 1,000 added EY lines or just 162 for 1,000 more KG lines. The high synergy between *P* and *piggyBac* elements was also seen with EY elements. After 7,000 EY lines, 1,000 PA/PC lines added 358 new genes vs 188 for an equal number of new EY lines. These results begin to quantitate the broader spectrum of gene targeting exhibited by *piggyBac* compared to *P* elements (HÄCKER *et al.* 2003).

In contrast, comparison between the KG and EY mutators revealed only limited synergy. 1,000 KG lines became somewhat more efficient relative to 1,000 EY lines after 7,000 previous EY insertions, now associating with only 25 fewer (163 vs. 188) rather than 100 fewer genes. Curiously, almost no synergy was seen in the reverse direction. After 7,000 KG lines, 1,000 additional EY lines targeted about 100 more genes than 1,000 added KG lines. This is about the same as the number of additional genes hit by EY vs KG lines initially. The very limited synergy indicates that different *P* element screens target substantially the same subsets of total genes (at least in the case of these two elements).

Screen-specific hotspots affect screen efficiency: As documented above, we observed large differences between screens in the total number of insert-associated genes. Thus, a sample of 1,000 unselected KG lines hit an average of 541 +/- 22 genes compared to 686 +/- 25 genes for

an equal number of EY lines. To determine if more insertions in a less efficient screen land between genes, we calculated the fraction of insertions in different screens that actually hit a gene (Fig. 3B). For comparison, we note that Release 3.1 identifies 52,560 kb of the Drosophila genome as intergenic (42%). Correcting for the 500 bp upstream of 13,666 genes that we also scored as potential gene hits, indicates that 63% of random insertions would be associated with a gene. The KG and EY screens hit genes at much higher frequencies (80 +/- 1.3 % and 81 +/- 1.1 % respectively). These values, which reflect P element gene targeting, are very similar and cannot explain the differences in efficiency. However, the rate of gene targeting appears to differ somewhat in other screens. BG elements hit genes only 72% of the time. This result is paradoxical as the *white* transgene within this element, which has a splice donor but no 3' polyadenylation site, was designed to be expressed only when its transcript is spliced to an endogenous 3' exon(s). Apparently, this system actually reduced rather than increased the frequency with which annotated genes are targeted. About 75% of *piggyBac* (PA/PC) jumps hit genes. Thus, *piggyBac* mutators are unlikely to target TTAA sequences randomly within the genome, but insert preferentially in genes, although to a lesser extent than *P* elements and with a reduced 5' bias.

The major source of efficiency differences between *P* screens proved to be transposon hotspots. We analyzed the frequency with which genes are hit in all the screens analyzed during the current phase of the project. Some of these results are shown in Table 4, where it can be seen that the number of times a gene is hit varies widely. The most frequently hit genes were considered "hotspots" (Table 5) and the fraction of all insertions in this class varied significantly between screens (Table 4).

Our results suggest that there are two previously unrecognized subclasses of hotspots.

All *P* element screens (and frequently also *piggyBac* screens) hit certain hotspot genes at elevated frequencies (Table 5, "common hotpots"). These loci must possess some intrinsic affinity for transposon binding and/or integration, perhaps due to the local chromatin state or the presence of particular proteins. Strikingly, however, a second class of hotspots was highly preferential for a particular screen or screens (Table 5). Most dramatically, the KG screen displayed a class of "super-hotspots." For example, *CG9894* alone accounts for a staggering 10% of all KG lines (Fig. 4A) and *Hr39* for another 2.5%. Five other sites are hit more frequently in the KG screen (>0.56%) than any of the common hotspots. These screen-preferential hotspots most likely explain the relative inefficiency in the KG screen. All the super-hotspots, and almost all of a larger number of less dramatic KG-associated hotspots are located on chromosome 2, and clustered in three small regions: 22F-23A, 38B-44A and 49F. A few screen-preferential hotspots were also detected in certain other screens, although their specificity appeared to be lower than for the KG hotspots (Table 5).

Some screen-enriched hotspots resemble local transpositions: *P* elements and many other transposons preferentially jump locally on their starting chromosome (TOWER et al 1993). We considered whether a relationship exists between screen-enriched hotspots and the site of the starting transposon. In the case of the KG screen, the *CG9894* "super-hotspot" corresponds exactly to the position of the starting insertion on the *CyO* balancer chromosome, which contains multiple chromosome inversions to block the recovery of recombinants. As in the case of local jumping, elevated frequencies of integration are not confined to a single nucleotide site, but extend along the chromosome in both directions (Fig. 4A). The broad distribution of insertions seen in Fig. 4A continues along the chromosome. Indeed, the elevated number of KG insertions recovered on chromosome 2 compared to chromosome 3 (Table 3) is due primarily to the

recovery of a higher density of disrupted genes in the vicinity of the super-hotspots, rather than uniformly across the entire chromosome. Thus, in their frequency, site dependence, and regional specificity, the KG screen-enriched hotspots resemble local transpositions, but on the homologous chromosome (TOWER and KURAPATI 1994).

However, these results could not be explained by a simple "homolog hopping" model (TOWER and KURAPATI 1994). Similar hotspots were not generally observed near the starting site in the case of other screens. For example, the starting site for the EY screen was localized in region 39C, yet this is not among the hotspots in this screen (Table 5). Moreover, most of the KG hotspots do not lie directly opposite the starting site, but are located at several distant sites, including a few on other chromosomes. One possibility is that some aspect of the local chromatin structure near the starting site is the critical variable. When plotted on a diagram showing the pairing pattern expected for *CyO* we noticed that the KG (but not the EY) starting site and all three super hotspot-containing regions were located near *CyO* breakpoints that may associate in vivo (Fig. 4C). These chromatin surrounding these sites may have been altered in a manner that enhances local jumping, allowing nearby sites on the homolog, and even on other chromosomes to be targeted. We suggest that screen-associated hotspots may generally arise via local jumping to sites that happen to reside close to the starting transposon in the chromatin of germ cell nuclei (Fig. 4D).

***P* element gene class preferences:** We examined the spacing of insertions in the primary collection throughout the genome by calculating the inter-element distances (average = 16.7 kb). Sometimes, as expected, the insertion density seemed to correlate with the density of genes/promoters, which varies significantly over relatively short regions (ASHBURNER *et al.* 1999). The likely existence of other influences was indicated by the nature of the two largest

gaps, both measuring about 290 kb in length. These correspond to the Antennapedia and Bithorax complexes, neither of which was hit by *P* elements in this phase of the project. (A single *P* insertion, *fs(3)05649* at *AbdB* is in the collection from the earlier phase). The fact that homeotic clusters are insertion cold spots provides further evidence that even in germ cells the genome presents a non-uniform target for transposition.

Previously, we noted that the frequency of insertion seems to vary for different classes of genes (SPRADLING *et al.* 1999). To investigate further, we calculated the number of disrupted genes in various functional classes (Fig, 5A). Common signaling pathways, stress response genes and other genes likely to be active in early germ cells (but not ribosomal protein genes) generally had an above average probability of being disrupted. In contrast, genes encoding cell type-specific proteins expressed late in development such as cuticle proteins, glue proteins or chorion proteins were rarely if ever hit. Although insertions in ribosomal proteins might be haplo-insufficient, there should have been no selection against insertions in structural proteins, and chemically-induced mutations in some of these genes have been recovered. The arrangement of these infrequently hit structural protein genes in chromosomal clusters suggests that some distinctive aspect of their chromatin structure or their promoter elements (OHLER et al. 2002) reduces their susceptibility to *P* element insertion. Unlike the homeotic clusters, the dearth of inserts in clustered cell-specific genes is unlikely to simply reflect low promoter density (Fig. 5B).

It has frequently been suggested that gene activity in germ cells might influence transposon accessibility. To study this variable we examined genes whose embryonic expression has been characterized using whole mount in situ hybridization by BDGP (http://www.fruitfly.org/cgi-bin/ex/insitu.pl). Of 104 genes expressed in pole cells or embryonic

germ cells, 70% contained an associated insertion in the primary collection, far more than the overall average of 40%. However, 61% of 123 genes that are expressed in the embryonic salivary gland. but not germ cells, also have an associated insertion in our project, so the importance of germ cell expression remains uncertain. Taking another tack, we examined if hotspot genes share any diagnostic features of their expression programs. No commonalities were observed. While some hotspots such as CG9894 are highly expressed maternally, and/or in embryos, RNA from others (*Hr39*, *cpo*) was weak or not detected. Consequently, a simple explanation for the gene selectivity observed in the project remains elusive.

The new primary collection: By analyzing lines from all ten screens, 7,140 lines have been designated for the primary collection (Table 3). Most of these lines have already been verified and forwarded to the Bloomington Stock Center for distribution. Insertions in the collection are distributed rather uniformly across the entire genome, including heterochromatic contigs and the 4th chromosome. EY, EP and LA insertions within the collection are positioned to mis-express 1,400 different genes. The BDGP primary collection will provide important reagents for a wide range of biological research.

DISCUSSION

Status of the project: During the last 3.5 years the BDGP gene disruption project primary collection has expanded from 1,000 to more than 7,100 strains and now contains insertions associated with at least 5,362 genes. It has proved possible to generate and molecularly analyze large numbers of lines and to produce a collection of high quality strains with diverse capabilities. High throughput methods were developed for generating, tracking and mapping large numbers of insertions, as well as bioinformatic methods for recording and

manipulating the data (see MATERIALS AND METHODS). We find that a large number of protein-coding genes can be targeted near their promoters and that transposon insertions in RNA gene and heterochromatic genes can be obtained as well. The generation and distribution of these new mutants throughout the course of the project are greatly assisting Drosophila researchers to investigate diverse biological questions. Finally, the project's large, well characterized datasets from multiple large screens utilizing different mutator elements and starting sites, allowed us to gain a better idea of how to optimally design transposon mutagenesis projects.

Two classes of insertion hotspots: Our work suggests the existence of two classes of genes that act as transposon hotspots. The first class comprises genes that evidently possess favorable chromatin accessibility, DNA target sequences, or bound proteins that mediate high efficiency association with freely diffusing transposition complexes. These sites may be highly specific at the nucleotide level (Fig. 4B) and may be responsible for the non-random primary DNA sequence context of *P* element integration sites (BELLEN *et al.* 1992; Liao *et al.* 2000). Many common hotspots appear to be hit frequently in multiple mutagenesis screens utilizing structurally different mutators. Our experiments better documented many members of this class, most of which were already well known as *P* element hotspots (Table 5).

The disruption rates of genes in the second class are highly screen-dependent (Table 5). These screen-associated hotspots may arise in a variety of ways. One mechanism is likely to be physical proximity to the starting transposon, as suggested by the location of the KG hotspots with respect to the re-arranged *CyO* chromosome. This class may primarily depend on the specific transposon starting site. Besides the KG super-hotspots, we found several other potential examples of this type of hotspot in the other screens. It is known that the specific sequences within a mutator may influence target sites (KASSIS 2002). In our experiments, the

EY and EP elements are similar in structure but were launched from different starting sites. We found that the hotspots in these two screens (and often also BG) were very similar, as expected if element structure was also important.

Screen-associated hotspots may provide insight into nuclear organization: Insertional mutagens typically do not integrate with equal efficiency across genomes (SANDMEYER *et al.* 1990; SPRADLING *et al.* 1995; ALONSO *et al.* 2003). Now that the products of large mutagenesis screens can be thoroughly analyzed without prior selection, it may be possible to use insertional preferences as tools for probing chromosome organization and function. Generating new insertions from a starting site located close to a chromosome rearrangement might generate super-hotspots within predictable regions of the chromosome. Such a procedure might increase the rate of mutagenesis in the targeted region by more than tenfold, as we observed in the 23B region, allowing genes in the vicinity to be mutated to saturation and chromatin structure to be probed.

The importance of genome annotation: Molecularly based insertional mutagenesis projects for some species have the luxury that all such lines can be safely stored for later retrieval and use. However, in many other species, including Drosophila, it is necessary to analyze newly generated lines and preserve only those with special value as experimental reagents. Our results illustrate how this latter type of screen depends crucially on accurate genome sequence annotation. The difficulty of making accurate gene-insertion associations is further exacerbated in organisms such as Drosophila that contain small dense genomes rich in overlapping and differentially spliced transcription units. The use of a transposon that inserts preferentially near promoters compounds the difficulty, as promoter prediction programs are accurate only about 50% of the time, even when large, accurate training sets are available (OHLER *et al.* 2002).

During the first three years of the project we worked with gene models based largely on computational predictions that frequently provided incomplete information on gene structure and location. Approximately 100 genes were lost from the project when lines located less than 2 kb

from existing lines were discarded and only later found to disrupt a separate, previously unannotated gene. In retrospect, it would probably have been worthwhile to maintain a higher density of insertions in intergenic regions and large introns. Our project suggests that a high priority should be placed on transcript mapping in combination with insertional mutagenesis projects.

Making stocks publicly available: As insertional mutagenesis of the Drosophila genome progresses, the issue of how to maintain all the valuable lines becomes increasingly acute. Frequently, multiple alleles of a gene are obtained that might each provide unique and valuable information regarding gene function. In the case of genes with multiple promoters, often encoding distinct protein splice variants in different tissues, insertions near the start site of each distinct transcript would allow their individual roles to be investigated. Complex patterns of gene expression during development might be efficiently studied using other alleles that sensitively report patterns of gene expression and, in some cases, reveal the sub-cellular location of the protein product(s) by fusing transcripts or protein domains to reporters. Much valuable information on gene function can likewise be derived from insertion alleles bearing regulatory elements that allow a gene to be mis-expressed under experimental control. Thus, an average of four alleles per gene, rather than one would likely be necessary to take full advantage of the experimental potential of Drosophila gene disruption collections. Unfortunately, at present, the world capacity for public storage and distribution of Drosophila stock is much more limited than this. Unless a solution to this problem is found, it is likely that many valuable tools will have to be discarded and the full value of publicly supported projects will be diminished.

The future of Drosophila gene disruption: Despite the progress toward genetic saturation reported here, many genes remain to be disrupted and still lack readily available tools

for understanding their biological roles throughout the life cycle. How should our project continue to address these remaining needs in an efficient manner? First, it is clear that a simple continuation of the current strategy using EY elements would be well worthwhile. The last set of 1,000 EY insertions scored still yielded 188 new genes, along with another 50-70 lines hitting previously missed intergenic regions or allowing gene mis-expression. Consequently, the "yield" of worthwhile lines remains above 20%, so that another 30,000 lines might be expected to yield 26,000 single insertions and up to perhaps 4,000 additional genes (15%). Switching to a *piggyBac* vector for the next 30,000 lines would yield insertions associated with a significantly large number of genes. This conclusion is strongly reinforced by the successful construction of several large collections of *piggyBac* insertions. (HÄCKER *et al.* 2003).

At what point does working to attain further genomic coverage using transposon mutagenesis becomes unattractive? Experimental data suggests that ultimately even *P* element mutagenesis can disrupt the great majority of Drosophila genes. Recently, OH *et al.* (2002) reported that they had increased the coverage of second chromosome vital genes from 25% to 80%. Likewise, TIMAKOV *et al.* (2002) recently demonstrated that a high fraction of genes are susceptible to *P* element insertion when rates are elevated by local hopping. However, our data suggest that some gene subclasses such as the cuticle protein genes may be refractory to this approach. Consequently, to disrupt every Drosophila gene will likely require a directed finishing strategy. Fortunately, there are several methods available in Drosophila that should be adequate for this task (RONG *et al.* 2002; MCCALLUM *et al.* 2000). Indeed, we can now look forward to a period when attention can shift from obtaining mutations to analyzing and understanding the biological processes they disrupt.

ACKNOWLEDGEMENTS

We thank Meiying Ji, Lara Chetkovich, Ruidong Ma, Ping Dang, Yaojuan Lu, Hongyuan Zhang, Jin Yue, Xingjie Shen and Mengfei Huang for creating, balancing and maintaining the fly stocks. Nicole Mozden, Dianne Thompson and Tiffany Jackson, assisted in the line maintenance and balancing at Carnegie. We are grateful to Alexei Tulin for transforming *P{Epgy2}* into flies, and to Christine Norman for help with the figures. We thank Soo Park and Kenneth Wan at LBNL for sequencing the inverse PCR products. Pernille Rorth, Tim Parnell and Pamela Geyer provided plasmids and unpublished data used in the construction of the pP{EPgy2}. We are grateful to Exelixis Corp. for providing a protocol for inverse PCR and sequencing of the flanking sequences of the PA and PC insertions. We are indebted to Gary Karpen, Daniel Garza, Udo Häcker, John Merriam, Stephen Poole, Judith Lengyel, William Gelbart and the members of their laboratories for donating their collections of insertion mutants to this project. We thank Kathy Matthews for balancing some PA/PC lines and for sending frozen samples for sequencing. This work was supported by an NIH Drosophila genome center grant (GMR, PI), and by a special supplement to this grant (ACS, PI) from NIGMS. Additional funds were provided through the support of the Spradling, Bellen and Rubin labs from the Howard Hughes Medical Institute.

REFERENCES

ADAMS, M. D., S. E. CELNIKER, R. A. HOLT, C. A. EVANS, J. D. GOCAYNE *et al.* , 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185-2195.

AIGAKI, T., T. OHSAKO, G. TOBA, K. SEONG and T. MATSUO, 2001 The gene search system: its application to functional genomics in *Drosophila melanogaster*. J Neurogenet **15:** 169-178.

ALTSCHUL, S.F., T.L. MADDEN, A.A., SCHAFFER, J. ZHANG, Z. ZHANG *et al.* 1997**,** Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, **25**: 3389-3402

AKERLEY, B. J., E. J. RUBIN, V. L. NOVICK, K. AMAYA, N. JUDSON *et al.* , 2002 A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. Proc Natl Acad Sci U S A **99:** 966-971.

AKIEDA, Y. and J. MERRIAM, 2001 Genes with ectopic expression phenotypes are common, not rare. D.I.S. **84:** 130-132.

ALONSO, J. M., A. N. STEPANOVA, T. J. LEISSE, C. J. KIM, H. CHEN *et al.* , 2003 Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science **301:** 653-657.

AMSTERDAM, A., 2003 Insertional mutagenesis in zebrafish. Dev Dyn **228:** 523-534.

ASHBURNER, M., S. MISRA, J. ROOTE, S. E. LEWIS, R. BLAZEJ *et al.* , 1999 An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*: the Adh region. Genetics **153:** 179-219.

BELLEN, H. J., C. J. O'KANE, C. WILSON, U. GROSSNIKLAUS, R. K. PEARSON *et al.* , 1989 P-element-mediated enhancer detection: a versatile method to study development in Drosophila. Genes Dev **3:** 1288-1300.

BELLEN, H. J., S. KOOYER, D. D'EVELYN AND J. PEARLMAN, 1992 The Drosophila couch potato protein is expressed in nuclei of peripheral neuronal precursors and shows homology to RNA-binding proteins. Genes Dev **6:** 2125-2136.

BELLEN, H. J., 1999 Ten years of enhancer detection: lessons from the fly. Plant Cell **11:** 2271-2281.

BERG, C. A. and A. C. SPRADLING, 1991 Studies on the rate and site-specificity of P element transposition. Genetics **127:** 515-524.

BESSEREAU, J. L., A. WRIGHT, D. C. WILLIAMS, K. SCHUSKE, M. W. DAVIS *et al.* , 2001 Mobilization of a Drosophila transposon in the *Caenorhabditis elegans* germ line. Nature **413:** 70-74.

BIDLINGMAIER S. and M. SNYDER, 2002, Large-scale identification of genes important for apical growth in *Saccharomyces cerevisiae* by directed allele replacement technology (DART) screening. Funct Integr Genomics 6:345-56.

BIER, E., H. VAESSIN, S. SHEPHERD, K. LEE, K. MCCALL *et al.* , 1989 Searching for pattern and mutation in the Drosophila genome with a P-lacZ vector. Genes Dev **3:** 1273-1287.

BOURBON, H. M., G. GONZY-TREBOUL, F. PERONNET, M. F. ALIN, C. ARDOUREL *et al.* , 2002 A P-insertion screen identifying novel X-linked essential genes in Drosophila. Mech Dev **110:** 71-83.

BRAND, A. H., AND N. PERRIMON, 1993 Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development **118:** 401-415.

BRENNECKE, J., D. R. HIPFNER, A. STARK, R. B. RUSSELL and S. M. COHEN, 2003 bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell **113:** 25-36.

CARY, L. C., M. GOEBEL, B. G. CORSARO, H. G. WANG, E. ROSEN *et al.* , 1989 Transposon mutagenesis of baculoviruses: analysis of *Trichoplusia ni* transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology **172:** 156-169.

CELNIKER, S. E., D. A. WHEELER, B. KRONMILLER, J. W. CARLSON, A. HALPERN *et al.* , 2002 Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. Genome Biol **3:** RESEARCH0079.

CHANG, Z., B. D. PRICE, S. BOCKHEIM, M. J. BOEDIGHEIMER, R. SMITH *et al.* , 1993 Molecular and genetic characterization of the Drosophila tartan gene. Dev Biol **160:** 315-332.

COOLEY, L., R. KELLEY and A. SPRADLING, 1988 Insertional mutagenesis of the Drosophila genome with single P elements. Science **239:** 1121-1128.

COOLEY, L., D. THOMPSON and A. C. SPRADLING, 1990 Constructing deletions with defined endpoints in Drosophila. Proc Natl Acad Sci U S A **87:** 3170-3173.

DEAK, P., M. M. OMAR, R. D. SAUNDERS, M. PAL, O. KOMONYI *et al.* , 1997 P-element insertion alleles of essential genes on the third chromosome of *Drosophila melanogaster*: correlation of physical and cytogenetic maps in chromosomal region 86E-87F. Genetics **147:** 1697-1722.

ERDELYI, M., A. M. MICHON, A. GUICHET, J. B. GLOTZER and A. EPHRUSSI, 1995 Requirement for Drosophila cytoplasmic tropomyosin in oskar mRNA localization. Nature **377:** 524-527.

EWING, B., *et al.* , 1998, Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research **8:** 175-185.

EWING, B. and GREEN, P., 1998, Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research **8:** 186-194.

FADOOL, J. M., D. L. HARTL AND J. E. DOWLING, 1998 Transposition of the mariner element from *Drosophila mauritiana* in zebrafish. Proc Natl Acad Sci U S A **95:** 5182-5186.

FIRON, A., F. VILLALBA, R. BEFFA AND C. D'ENFERT, 2003 Identification of essential genes in the human fungal pathogen *Aspergillus fumigatus* by transposon mutagenesis. Eukaryot Cell **2:** 247-255.

GAUL, U., MARDON, G. and G.M. RUBIN, 1992. A putative Ras GTPase activating protein acts as a negative regulator of signaling by the Sevenless receptor tyrosine kinase. Cell **68:** 1007-1019**.**

GIAEVER, G., A. M. CHU, L. NI, C. CONNELLY, L. RILES *et al.* , 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. Nature **418:** 387-391.

GOLLING, G., A. AMSTERDAM, Z. SUN, M. ANTONELLI, E. MALDONADO *et al.,* 2002 Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. Nat Genet **31:** 135-140.

GRAY, Y. H., 2000 It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. Trends Genet **16:** 461-468.

GROSSNIKLAUS, U., H. J. BELLEN, C. WILSON and W. J. GEHRING, 1989 *P*-element-mediated enhancer detection applied to the study of oogenesis in Drosophila. Development **107:** 189-200.

GUEIROS-FILHO, F. J., and S. M. BEVERLEY, 1997 Trans-kingdom transposition of the Drosophila element mariner within the protozoan Leishmania. Science **276:** 1716-1719.

HÄCKER, U., S. NYSTEDT, M. P. BARMCHI, C. HORN and E. A. WIMMER, 2003 *piggyBac*-based insertional mutagenesis in the presence of stably integrated *P* elements in Drosophila. Proc Natl Acad Sci U S A **100:** 7720-7725.

HARBISON, S. T., A. H. YAMAMOTO, J. J. FANARA, K. K. NORGA AND T. F. C. MACKAY, 2004 Quantitative trait loci affecting starvation resistance in *Drosophila melanogaster*. **Submitted**.

HESLIP, T. R., and R. B. HODGETTS, 1994 Targeted transposition at the vestigial locus of Drosophila melanogaster. Genetics **138:** 1127-1135.

HORN, C., N. OFFEN, S. NYSTEDT, U. HACKER AND E. A. WIMMER, 2003 piggyBac-based insertional mutagenesis and enhancer detection as a tool for functional insect genomics. Genetics **163:** 647-661.

HOSKINS, R. A., C. D. SMITH, J. W. CARLSON, A. B. CARVALHO, A. HALPERN *et al.* , 2002 Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. Genome Biol **3:** RESEARCH0085.

HUET, F., J. T. LU, K. V. MYRICK, L. R. BAUGH, M. A. CROSBY *et al.* , 2002 A deletion-generator compound element allows deletion saturation analysis for genome wide phenotypic annotation. Proc Natl Acad Sci U S A **99:** 9948-9953.

JANSEN G, E. HAZENDONK, K.L. THIJSSEN K.L. and R.H. PLASTERK, 1997 Reverse genetics by chemical mutagenesis in *Caenorhabditis elegans*. Nat Genet. **17:** 119-121.

KAMINKER, J.S., C.M. BERGMAN, B. KRONMILLER, J. CARLSON, R. SVIRSKAS, *et al*. 2002, The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. **3:** RESEARCH0084.

KARPEN, G. H., and A. C. SPRADLING, 1992 Analysis of subtelomeric heterochromatin in the Drosophila minichromosome Dp1187 by single P element insertional mutagenesis. Genetics **132:** 737-753.

KASSIS JA. (2002). Pairing-sensitive silencing, polycomb group response elements, and transposon homing in Drosophila. Adv Genet. 46: 421-38.

KLINAKIS, A.G., L. ZAGORAIOU, D.K. VASSILATIS, and C. SAVAKIS, 2000 Genome-wide insertional mutagensis in human cells by the Drosophila mobile element Minos. Embo Report 5: 416-421.

KONEV, A.Y, , C.M. YAN, D. ACEVEDO, C. KENNEDY, E. WARD *et al*., 2003 Genetics of P-Element Transposition Into Drosophila melanogaster Centric Heterochromatin. Genetics. **165:** 2039-2053.

LAI, E.C., P. TOMANCAK, R.W. WILLIAMS and G.M. RUBIN, 2003 Computational identification of Drosophila microRNA genes. Genome Biol. 4: RESEARCH42.

LENGYEL, J.A. and J. MERRIAM, 2001 Translation of a new genetic screening method (ectopic expression) from the research lab to a teaching lab. D.I.S. **84:** 218-224.

LEWIS, S. E., S. M. SEARLE, N. HARRIS, M. GIBSON, V. LYER *et al.* , 2002 Apollo: a sequence annotation editor. Genome Biol **3:** RESERACH 0082.

LIAO, G. C., E. J. REHM and G. M. RUBIN, 2000 Insertion site preferences of the *P* transposable element in *Drosophila melanogaster*. Proc Natl Acad Sci U S A **97:** 3347-3351.

LUKACSOVICH, T., Z. ASZTALOS, W. AWANO, K. BABA, S. KONDO *et al.* , 2001 Dual-tagging gene trap of novel genes in *Drosophila melanogaster*. Genetics **157:** 727-742.

MANSEAU, L., A. BARADARAN, D. BROWER, A. BUDHU, F. ELEFANT *et al.* , 1997 GAL4 enhancer traps expressed in the embryo, larval brain, imaginal discs, and ovary of Drosophila. Dev Dyn **209:** 310-322.

MATA, J., S. CURADO, A. EPHRUSSI AND P. RØRTH, 2000 Tribbles coordinates mitosis and morphogenesis in Drosophila by regulating string/CDC25 proteolysis. Cell **101:** 511-522.

MCCALLUM, C. M., L. COMAI, E. A. GREENE AND S. HENIKOFF, 2000 Targeted screening for induced mutations. Nat Biotechnol **18:** 455-457.

MIKKERS, H., J. ALLEN, P. KNIPSCHEER, L. ROMEIJN, A. HART *et al.,* 2002 High-throughput retroviral tagging to identify components of specific signaling pathways in cancer. Nat Genet **32:** 153-159.

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.

MITCHELL, K. J., K. I. PINSON, O. G. KELLY, J. BRENNAN, J. ZUPICICH *et al.* , 2001 Functional analysis of secreted and transmembrane proteins critical to mouse development. Nat Genet **28:** 241-249.

MOHR, S. E., and W. M. GELBART, 2002 Using the P[wHy] hybrid transposable element to disrupt genes in region 54D-55B in *Drosophila melanogaster*. Genetics **162:** 165-176.

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 U S A **98:** 15050-15055.

NORGA, K. K., M. C. GURGANUS, C. L. DILDA, A. YAMAMOTO, R. F. LYMAN *et al.* , 2003 Quantitative analysis of bristle number in Drosophila mutants identifies genes involved in neural development. Curr Biol **13:** 1388-1396.

OH, S. W., T. KINGSLEY, H. H. SHIN, Z. ZHENG, H. W. CHEN *et al.* , 2003 A P-element insertion screen identified mutations in 455 novel essential genes in Drosophila. Genetics **163:** 195-201.

OHLER, U., G. C. LIAO, H. NIEMANN and G. M. RUBIN, 2002 Computational analysis of core promoters in the Drosophila genome. Genome Biol **3:** 1-87.

O'Kane, C..J., and W.J. Gehring, 1987 Detection in situ of genomic regulatory elements in Drosophila. Proc Natl Acad Sci U S A. **84:** 9123-9127.

PATTON, J.S., X.V. GOMES and P.K. GEYER, 1992 Position-independent germline transformation in Drosophila using a cuticle pigmentation gene as a selectable marker. Nucleic Acids Res. **20:** 5859-60.

PRESTON, C.R. and W.R. ENGELS, 1996 P-element-induced male recombination and gene conversion in Drosophila. Genetics. **144:** 1611-22.

PRESTON, C. R., J. A. SVED and W. R. ENGELS, 1996 Flanking duplications and deletions associated with *P*-induced male recombination in Drosophila. Genetics **144:** 1623-1638.

RONG, Y. S., S. W. TITEN, H. B. XIE, M. M. GOLIC, M. BASTIANI *et al.,* 2002 Targeted mutagenesis by homologous recombination in *D. melanogaster*. Genes Dev **16:** 1568-1581.

ROOS, D. S., W. J. SULLIVAN, B. STRIEPEN, W. BOHNE and R. G. DONALD, 1997 Tagging genes and trapping promoters in *Toxoplasma gondii* by insertional mutagenesis. Methods **13:** 112-122.

RØRTH, P., 1996 A modular misexpression screen in Drosophila detecting tissue-specific phenotypes. Proc Natl Acad Sci U S A **93:** 12418-12422.

RØRTH, P., K. SZABO, A. BAILEY, T. LAVERTY, J. REHM *et al.* , 1998 Systematic gain-of-function genetics in Drosophila. Development **125:** 1049-1057.

ROSEMAN, R. R., E. A. JOHNSON, C. K. RODESCH, M. BJERKE, R. N. NAGOSHI *et al.* , 1995 A P element containing suppressor of hairy-wing binding regions has novel properties for mutagenesis in *Drosophila melanogaster*. Genetics **141:** 1061-1074.

SALZBERG, A., S. N. PROKOPENKO, Y. HE, P. TSAI, M. PAL *et al.* , 1997 P-element insertion alleles of essential genes on the third chromosome of *Drosophila melanogaster*: mutations affecting embryonic PNS development. Genetics **147:** 1723-1741.

SANDMEYER, S. B., L. J. HANSEN and D. L. CHALKER, 1990 Integration specificity of retrotransposons and retroviruses. Annu Rev Genet **24:** 491-518.

SEKELSKY, J. J., K. S. MCKIM, L. MESSINA, R. L. FRENCH, W. D. HURLEY *et al.* , 1999 Identification of novel Drosophila meiotic genes recovered in a P-element screen. Genetics **152:** 529-542.

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.

SPRADLING, A. C., D. M. STERN, I. KISS, J. ROOTE, T. LAVERTY *et al.* 1995 Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. Proc Natl Acad Sci U S A **92:** 10824-10830.

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.

STANFORD, W. L., J. B. COHN and S. P. CORDES, 2001 Gene-trap mutagenesis: past, present and beyond. Nat Rev Genet **2:** 756-768.

STAPLETON, M., J. CARLSON, P. BROKSTEIN, C. YU, M. CHAMPE *et al.* , 2002 A Drosophila full-length cDNA resource. Genome Biol **3:** RESEARCH0080.

THOMPSON-STEWART, D., G. H. KARPEN and A. C. SPRADLING, 1994 A transposable element can drive the concerted evolution of tandemly repetitious DNA. Proc Natl Acad Sci USA **91:** 9042-9046.

TIMAKOV, B., X. LIU, I. TURGUT AND P. ZHANG, 2002 Timing and targeting of P-element local transposition in the male germline cells of *Drosophila melanogaster*. Genetics **160:** 1011-1022.

TOBA, G., T. OHSAKO, N. MIYATA, T. OHTSUKA, K. H. SEONG *et al.* , 1999 The gene search system. A method for efficient detection and rapid molecular identification of genes in *Drosophila melanogaster*. Genetics **151:** 725-737.

TÖROK, T., G. TICK, M. ALVARADO and I. KISS, 1993 P-lacW insertional mutagenesis on the second chromosome of *Drosophila melanogaster*: isolation of lethals with different overgrowth phenotypes. Genetics **135:** 71-80.

TOWER, J., G. H. KARPEN, N. CRAIG and A. C. SPRADLING, 1993 Preferential transposition of Drosophila *P* elements to nearby chromosomal sites. Genetics **133:** 347-359.

TOWER, J., AND R. KURAPATI, 1994 Preferential transposition of a Drosophila *P* element to the corresponding region of the homologous chromosome. Mol Gen Genet **244:** 484-490.

TULIN A, D. STEWART and A.C. SPRADLING, 2002. The Drosophila heterochromatic gene encoding poly(ADP-ribose) polymerase (PARP) is required to modulate chromatin structure during development. Genes Dev. **16:** 2108-19.

UHL, M. A., M. BIERY, N. CRAIG and A. D. JOHNSON, 2003 Haploinsufficiency-based large-scale forward genetic analysis of filamentous growth in the diploid human fungal pathogen C.albicans. Embo J **22:** 2668-2678.

VIDAN, S. and M. SNYDER, 2001 Large-scale mutagenesis: yeast genetics in the genome era. Curr Opin Biotechnol **12:** 28-34.

WALLRATH, L. L. and S. C. ELGIN, 1995 Position effect variegation in Drosophila is associated with an altered chromatin structure. Genes Dev **9:** 1263-1277.

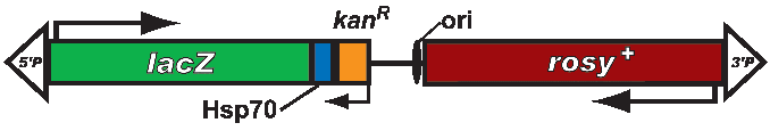
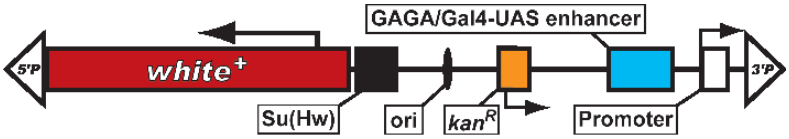
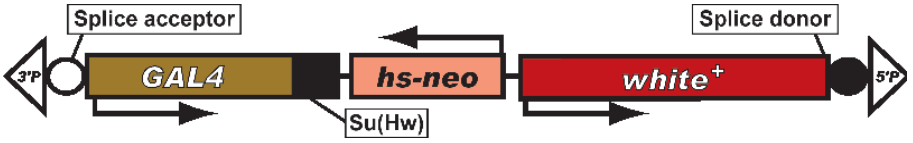
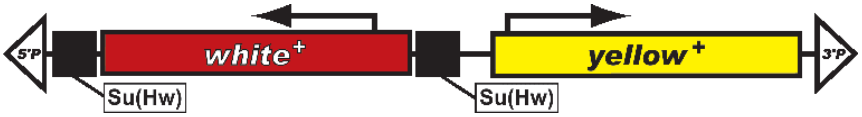
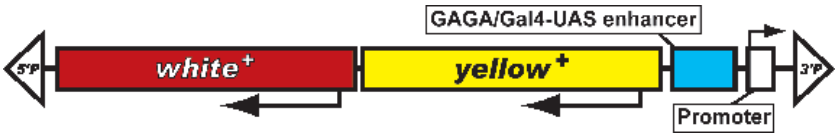
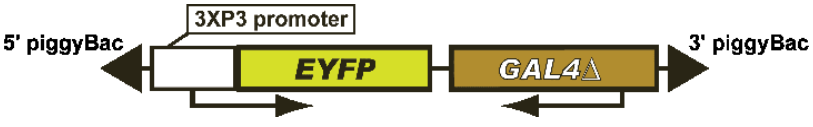
WILSON, C., R. K. PEARSON, H. J. BELLEN, C. J. O'KANE, U. GROSSNIKLAUS *et al.* , 1989 P-element-mediated enhancer detection: an efficient method for isolating and characterizing developmentally regulated genes in Drosophila. Genes Dev **3:** 1301-1313.

YAN, C. M., K. W. DOBIE, H. D. LE, A. Y. KONEV and G. H. KARPEN, 2002 Efficient recovery of centric heterochromatin P-element insertions in *Drosophila melanogaster*. Genetics **161:** 217-229.

ZAGORAIOU, L., D. DRABEK, S. ALEXAKI, J. A. GUY, A. G. KLINAKIS *et al.,* 2001 In vivo transposition of Minos, a Drosophila mobile element, in mammalian tissues. Proc Natl Acad Sci U S A **98:** 11474-11478.

ZHANG, P., and A. C. SPRADLING, 1994 Insertional mutagenesis of Drosophila heterochromatin with single P elements. Proc Natl Acad Sci U S A. **91:** 3539-43.

Table 1. Mutator transposons

Symbol	Marker	Transposon	Ref	Map
PZ	*rosy*	*P{PZ}*	1	
PlacW	*white*	*P{lacW}*	2	
EP	*white*	*P{EP}*	3	
BG	*white*	*P{GT1}*	4	
KG, KV	*white, yellow*	*P{SUPor-P}*	5	
EY	*white, yellow*	*P{EPgy2}*	6	
DG	*white, yellow*	*P{wHy}*	7	
PA	*white*	*PBac{5HPw[+]}*	8	
PC	*yellow*	*PBac{3HPy[+]}*	8	
PL	*EYFP*	*PBac{GAL4D,EYFP}*	9	

Symbol	Marker	Transposon	Ref	Map
LA	*white*	*P{Mae-UAS.6.11}*	10	

The schematic diagrams are not drawn to scale and are meant only to indicate the components present in each transposon. Thin lines separating some components have been added to prevent labels from overlapping and are not intended to indicate spacers between components. Please refer to the original publications and curated FlyBase reports for details.

[1] Mlodzik and Hiromi, 1992
[2] Bier et al., 1989
[3] Rørth, 1996
[4] Lukacsovich et al. 2001
[5] Roseman et al. 1995
[6] this work
[7] Huet et al. 2002
[8] B. Ring and D. Garza, unpublished
[9] Horn et al. 2003
[10] J. Merriam and S. Poole, unpublished

Table 2. Line Summary

Symbol	Number Received	Sequence Recovered	Unique hit	Repetitive[4]	Double hit[4]	Selected	Confirmed[5]	Genes/ 1,000	% lethal (semilethal)[6]
EP	2,266	2,241	2,012	79	24	374	N/A	686 ± 10	
BG	2,869	2,333	2,086	165	78	482	461	339 ± 40	8.0 (1.5)
KG	10,587	9,501	8,838	430	357	2,129	2,073	541 ± 22	16.3 (3.4)
EY	10,310	8,941	8,309	337	411	2,338	1,696	691 ± 25	12.0 (1.5)
KV	813	658	379	245	6	108	0		
DG	1,384	1,194	1,030	96	48	154	0	648	
PA/PC [1]	1.055	1,055	1046	N/A	N/A	342	284	689	
PL	634	617	533	29	38	266	N/A		
LA [1]	1045	913	753	34	N/A	101	0		
Subtotals	**30,963**	**27,453**	**24,986**	**1,415**	**962**	**6,294**	**4,514**		
placZ [2]						459	459		
placW [2,3]						387	387		
Totals	**30,963**	**27,453**	**24,986**	**1,415**	**962**	**7,140**	**5,360**		

[1] only previously sequenced lines with unique hits were received
[2] Spradling *et al.* (1999)
[3] includes 9 neo lines
[4] "Repetitive" means the flanking sequence matched 2 or more separate genomic sites; "Double hit" means the 5' and 3' flanking sequences matched two distinct genomic sites indicating the likely presence of 2 insertions
[5] data complete only for BG and KG lines; others still in progress
[6] lethality data based only on lines selected for balancing and distribution; % semilethal shown in parentheses

Table 3. Primary Collection Summary

Arm	R3 genes	placZ	placW	EP	BG	KG	EY	KV	DG	PA/PC	PL	LA	Lines	tagged R3 genes
X	2,232	0	0	91	148	384	376	5	3	5	1	0	1,013	725
2L	2,428	94	120	48	87	562	351	8	34	63	3	18	1,388	1,070
2R	2,665	111	148	97	69	499	498	8	26	54	2	25	1,537	1,204
3L	2,607	101	49	63	102	325	503	6	43	102	127	20	1,441	1,065
3R	3,377	143	69	74	74	317	598	7	45	110	129	37	1,603	1,243
4	82	0	0	0	2	25	6	0	3	8	0	1	45	25
U[2]	275	10	1	1	0	17	6	74	0	0	4	0	113	30
	13,666	**459**	**387** [1]	**374**	**482**	**2,129**	**2,338**	**108**	**154**	**342**	**266**	**101**	**7,140**	**5,362**

[1]includes 9 neo lines

[2]WGS3 heterochromatic sequence (see Hoskins et al. 2002).

Table 4. Frequency distribution of targeted genes between screens

Number	BG	KG	EY	PA/PC[1]
0	12961	11186	10596	12690
1	402	1174	1424	545
2	122	397	526	102
3	51	168	284	23
4	34	127	148	5
5	18	73	96	1
6	15	60	63	1
7	9	37	40	0
8	5	25	33	1
9	4	19	21	1
10	3	17	27	0
11	4	20	12	0
12	0	6	9	0
13	2	8	16	0
14	1	5	13	0
15	0	9	10	0
16	3	1	6	0
17	1	5	5	0
18	1	0	7	0
19	0	4	4	0
20-29	0	13	22	0
30-39	2	6	5	0
40-49	1	0	2	0
50-59	0	0	0	0
60-69	0	2	0	0
70-79	0	4	0	0
80-89	0	0	0	0
90-99	0	0	0	0
100-199	0	2	0	0
200-299	0	0	0	0
300-399	0	0	0	0
400-499	0	0	0	0
500-599	0	0	0	0
600-699	0	0	0	0
700-799	0	1	0	0
Total genes	**678**	**2183**	**2773**	**679**

[1]Duplicates due to pre-meiotic clusters were not excluded.

Table 5. Gene targeting rates

Gene	FlyBase	Site	Arm	BG rate	KG rate	EY rate	EP rate	PA/PC rate
KG hotspots (22F-23A)								
CG9894		23A3	2L	20	996	19	45	0
CG16987	Alp23B	23B1	2L	0	89	10	5	0
CG9884	oaf	22F3	2L	0	95	15	0	0
CG3539	Slh	22F3	2L	0	24	0	0	0
CG3104		23B5	2L	0	21	1	0	0
CG31690	CG31690	23A2	2L	0	16	0	0	0
KG hotspots (38B-44A)								
CG8676	Hr39	39B4	2L	35	246	15	10	20
CG11546	l(2)02045	44B7	2R	15	136	4	5	0
CG8709		44B5	2R	0	96	8	5	0
CG31611/CG31613		39E1	2L	0	80	11	0	0
CG8678		39B3	2L	0	45	3	0	0
CG8677	BEST:LD14959	39B3	2L	0	39	1	5	0
CG15845	Adf1	42C3	2R	0	34	8	10	0
CG31626		39B4	2L	5	29	1	0	10
CG2163	Pabp2	44B4	2R	0	28	4	5	0
CG12110	Pld	42A15	2R	0	26	5	0	0
CG9243/CG9244	Acon	39A7	2L	0	21	1	0	0
CG10718	neb	38B4	2L	5	19	5	5	0
CG10746	fok	38B4	2L	5	19	5	0	0
KG hotspots (49B-F)								
CG4654	Dp	49F10	2R	0	38	5	5	0
CG4670		49F11	2R	0	24	0	0	0
CG4663		49F11	2R	0	13	0	0	10
EY/EP/BG hotspots (85C-91F)								
CG9429	Crc	850E1	3R	55	5	31	25	0
CG10120	Men	87C6	3R	15	9	46	30	0
CG5555/CG31475		91F11	3R	0	8	29	60	0
CG11033		85C3	3R	15	4	30	35	0
CG3937	cher	89E13	3R	40	5	20	5	0
CG9366	RhoL	85D18	3R	0	5	16	15	0
EY/EP/BG hotspots (misc)								
CG5723	Ten-m	79E1	3L	10	10	33	10	0
CG5320	Gdh	95C13	3R	5	6	29	15	0
CG3979	Indy	75E1	3L	0	5	28	10	0
CG14450/CG11367		80A1	3L	5	6	21	0	0
CG31522		82B4	3R	5	1	18	5	0
PA/PC hotspots								
CG9216		14A9	X	0	5	4	0	50
CG14307		91A8	3R	0	3	4	0	40
Common hotspots								
CG31243	cpo	90D1	3R	90	14	56	60	10
CG8276	bin3	42A14	2R	60	87	30	0	40
CG17161	grp	36A10	2L	45	25	20	45	50
CG12052	lola	47A11	2R	55	18	13	75	0
CG6889	tara	89B13	3R	25	13	34	50	30
CG8938	GstS1	53F9	2R	25	11	29	55	30
CG32529/amn	amn	18F4	X	10	48	41	50	0
CG9755	pum	85C4	3R	35	19	20	40	30
CG3758	esg	35D1	2L	75	10	15	35	0
CG14709	BEST:CK0122	86E14	3R	30	19	55	30	0

Gene	FlyBase	Site	Arm	BG rate	KG rate	EY rate	EP rate	PA/PC rate
CG8846	Thor	23F6	2L	10	44	24	40	0
CG3696	kis	21B6	2L	15	19	21	45	10
CG12284		72D1	3L	20	9	13	25	40
CG3903	Gli	35D4	2L	10	29	29	25	10
CG1856	ttk	100D1	3R	30	26	18	15	10
CG7437	mub	79A2	3L	35	6	23	35	0
CG31000	heph	100D4	3R	20	14	28	20	10
CG10645	lama	64D1	3L	20	8	29	25	10
CG9432	l(2)01289	42C7	2R	5	30	41	10	0
CG8651	trx	88B1	3R	15	18	25	15	10
CG5393	apt	59F1	2R	0	30	35	10	0
CG17950	HmgD	57F10	2R	20	20	26	5	0
CG17716	fas	50B6	2R	15	36	13	5	0
CG6072	sra	89B12	3R	0	16	26	5	20
CG7481/CG7582	RhoGAP18B	18A3	X	0	24	24	15	0
CG3036		25B1	2L	0	25	16	20	0
CG5461	bun	33E5	2L	10	21	13	15	0
CG8804	wun	45D4	2R	5	21	9	20	0
CG10033	for	24A2	2L	0	26	23	5	0
CG2922	eIF-5C	83B1-2	3R	5	3	15	15	10
CG8815	Sin3A	49B6	2R	0	13	4	20	10

Mutation rates are in hits per 10,000 unique, localized insertions: i.e. % x 100. The number of chromosomes used in the calculations were as follows: BG (2,000), KG (8,000), EY (8,000), PA/PC (1,000). Complete hotspot data is available on request.

FIGURE LEGENDS

Figure 1. Schematic diagram of project workflow. The arrows show how new Drosophila strains from single P element mutagenesis screens are processed by the project. Lines are sequenced, sequences aligned to unique genomic sites, and insertions likely to disrupt genes not already mutated in the collection are selected (central boxes). Selected lines are balanced, re-checked to verify quality and sent to the Bloomington stock center for public distribution. Lines failing to meet these criteria are re-cycled or discarded. The percentages indicate the fraction of lines falling into the indicated categories along each path.

Figure 2. Computationally associating insertions with genes. A. The upper panel shows a sample insertion, KG10308, as it appears in the Release 2-based GeneSeen display. The position of the insertion (triangle and vertical line) is shown relative to the local DNA sequence (horizontal line) and gene models (blue boxes) following the convention that genes above the line are oriented left to right and below the line they are oriented oppositely (arrowheads). KG10308 fails to meet project criteria for a gene association under R2 because it maps about 2 kb upstream from the *CG8249* annotation and 3.5 kb 5' to the *CG8253* annotation. Below, the same region is displayed based on Release 3 sequence annotations (Misra et al. 2002) using the Apollo browser (Lewis *et al.* 2002). The inclusion of more information on 5' UTRs in Release 3 reveals that KG10308 actually lies at the 5' end of *CG8253* and likely mutates this gene. B. A histogram showing the distance between the *P* element insertions in 5,630 gene-associated primary collection lines and their associated transcript 5' ends (blue). For comparison, a similar plot of the 267 lines with transcript-associated *piggyBac* insertions is shown (red). The last point on the right shows all remaining lines more than 500 bp from +1. C. KG00786 is located at -10 relative to *CG8315* and at -49 relative to *CG8320*. Nearby, the gene *ATPCL* (*CG8322*) is seen

overlapping with *CG8320*. Both close gene spacing and overlapping transcription units are quite common in the Drosophila genome and account for the fact that 20% of single insertions in the primary collection likely affect two genes. D. KG05287 is shown near the 5' end of *CG31849*. This gene lies within a large intron of *CG5287*, which is transcribed from the opposite strand. The occurrence of genes within the large introns of other genes is common in Drosophila, and motivated us to retain insertions in the large introns of already mutated transcription units if they were separated from other insertions by at least 2 kb. E. KG02679 is one of 60 insertions predicted to lie within or close to an RNA gene. F. Insertions are shown upstream from *CG12462* that lie more than 10 kb from any annotated gene. Many insertions in this category using Release 2 annotation were later shown to be located near the 5' ends of genes. Because annotation remains highly imperfect, insertions were saved if they lay more than 2 kb from any existing insertion in the primary collection. G. An example of a manually determined insertion-gene association. Many release 3 gene models (such as *CG32767* shown) are still computationally derived only from their protein-coding sequences and lack sequences 5' to the predicted methionine start codon. Automated annotation fails in these cases because *P* elements preferentially insert near 5' ends, which commonly lie more than 500 bp from the start codon. In the example shown, BG01357 lies 1.8 kb 5' to the R3 annotation of *CG32767* but was manually associated by considering cDNAs such as the 5' EST RE54443.5 displayed in Apollo.

Figure 3. Individual screens differ widely in mutagenic efficiency. A. The average number of genes disrupted by 1,000 lines from the indicated screens is shown. Each bar represents an average of from 2-4 sets of 1,000 lines except for PA/PC, where only 1,046 lines were available (668 PA + 332 PC were used). B. The cumulative number of genes disrupted as successive sets of 1,000 lines are added for the indicated screens. C. Screen synergy. The relative number of

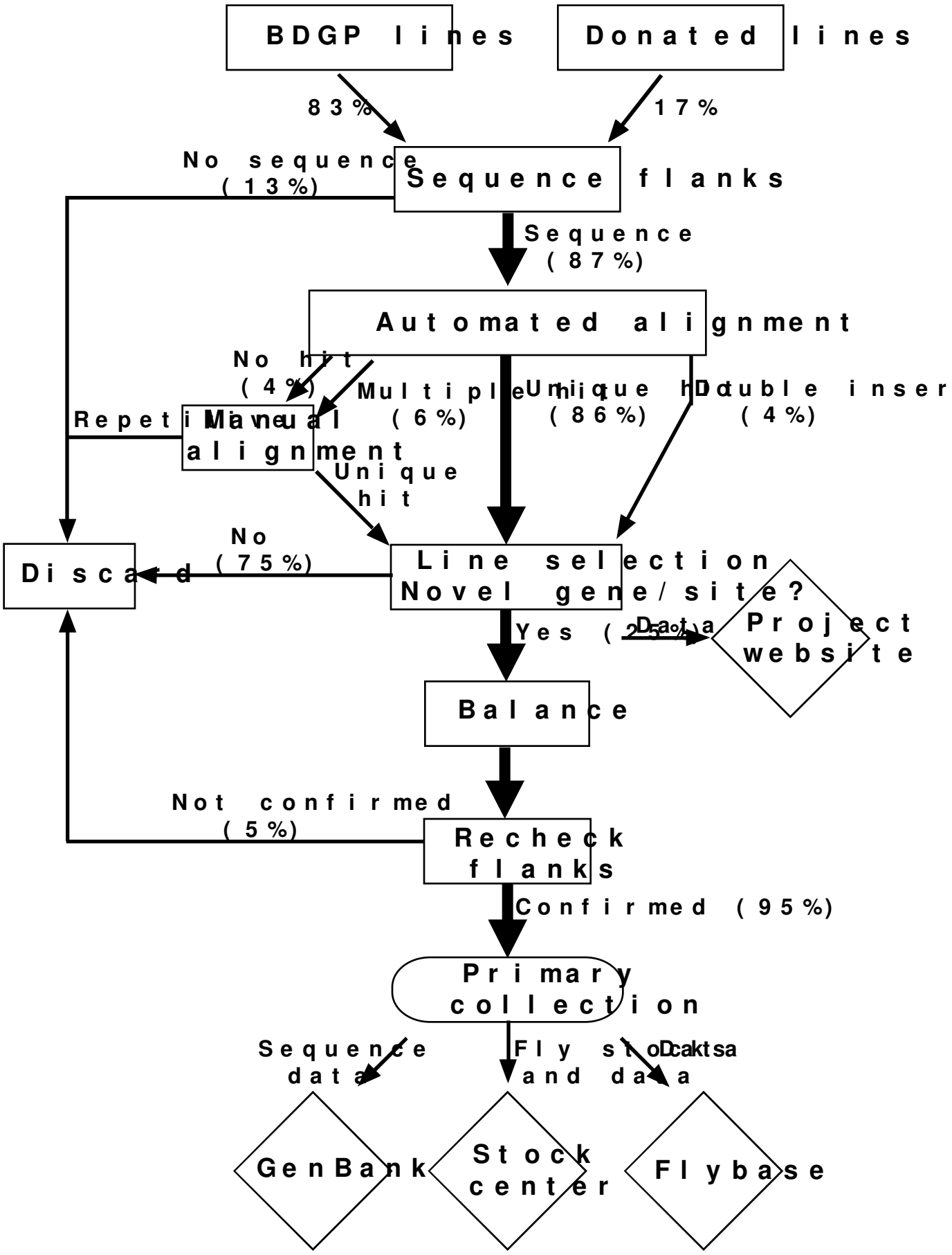
total genes disrupted when a set of 1,000 additional lines from the indicated screens are added to a collection of either 7,000 KG lines (lower set: KG-) or 7,000 EY lines upper set (EY-). D. The mean percentage of 1,000 lines that hit genes is shown for 5 screens. Standard deviations are given in the text. All of the lines used for these analyses were localized to a unique site in the euchromatic genome.

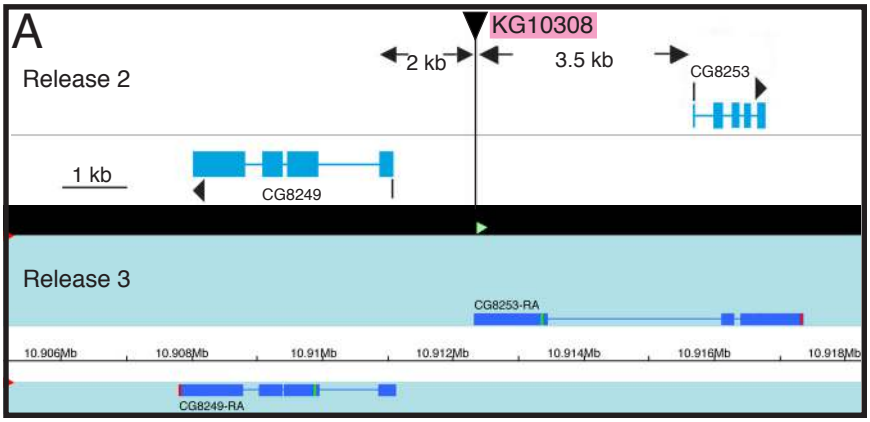
Figure 4. Screen-preferential hotspots. A. An Apollo display of the region surrounding the major KG "super-hotspot" at gene *CG9894*, which contains the screen starting site. Note that insertions (triangles) are distributed on both strands and at multiple sites. Not all the insertions could be represented as separate triangles. B. An Apollo display of a major EY hotspot in gene *CG3979* (Indy). 1360 1044 is a repetitive element. C. Pairing diagram of the *CyO* balancer (black line) with its wild type homolog (green line) in the germ cells in which new jumps occur. The position of the starting transposon (red bar) and four major groups of hotspots (orange bars) are shown. All reside close to the central region where normal pairing is disrupted due to the multiple inversions on *CyO*. D. Model for the generation of screen-associated hotspots by local jumping from the starting site to other chromosome regions that happen to lie nearby in germ cell nuclei.

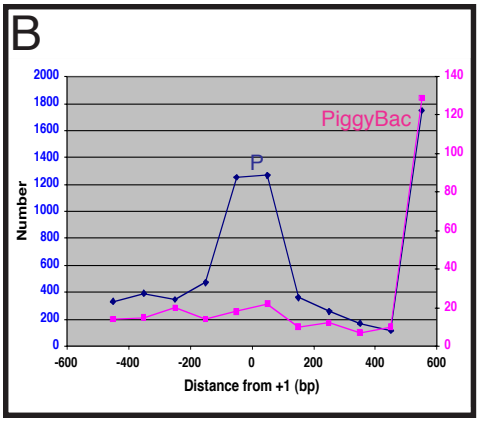
Figure 5. Target selectivity

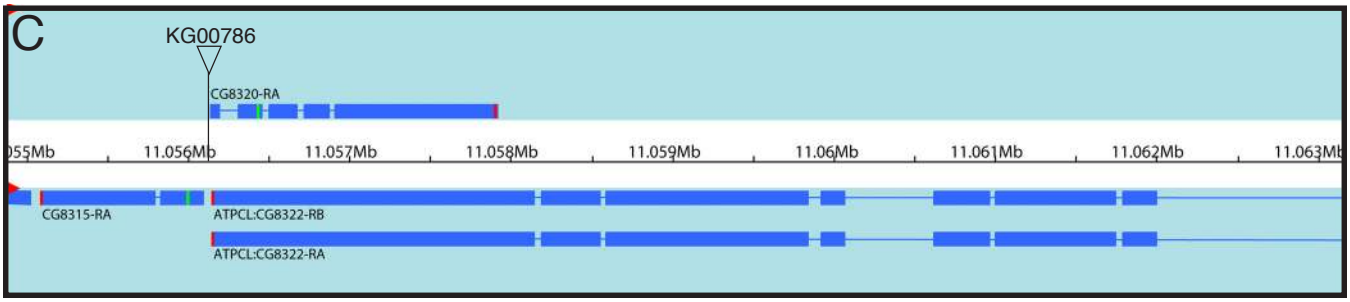
A. Different pathways and gene classes are differentially susceptible to *P* element insertion. The fraction of genes in various classes hit in the primary collection is shown. B. Apollo display from the 65A larval cuticle protein gene cluster spanning approximately 45 kb. No insertions in

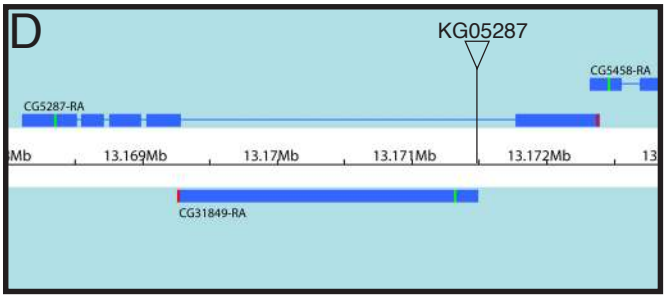
this region were recovered (black regions above and below maps, or in regions housing several other similar clusters of genes expressed in terminally differentiated cells.

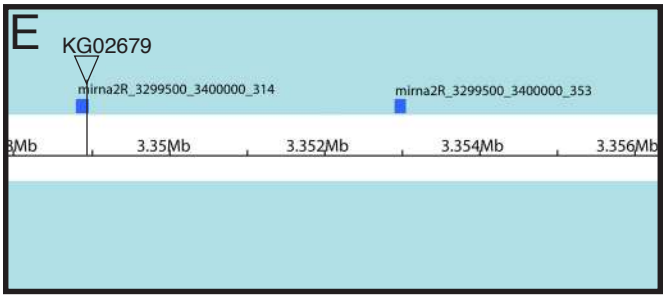

BDGP lines
Donated lines
83%
17%
No sequence (13%)
Sequence flanks
Sequence (87%)
Automated alignment
No hit (4%)
Multiple hit (6%)
Unique hit (86%)
Double insert (4%)
Repeat
Manual alignment
Unique hit
No (75%)
Discard
Line selection
Novel gene/site?
Yes (25%)
Data
Project website
Balance
Not confirmed (5%)
Recheck flanks
Confirmed (95%)
Primary collection
Sequence data
Fly stocks and data
Data
GenBank
Stock center
Flybase

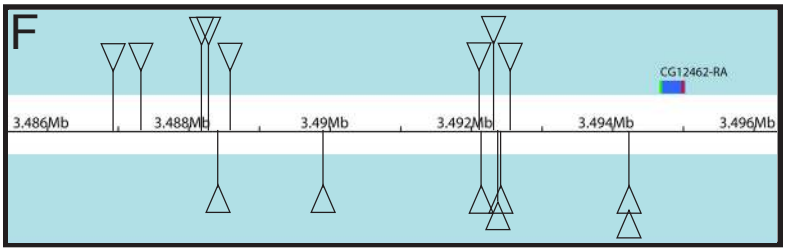

A
Release 2
KG10308
2 kb
3.5 kb
CG8253
1 kb
CG8249
Release 3
CG8253-RA
10.906Mb
10.908Mb
10.91Mb
10.912Mb
10.914Mb
10.916Mb
10.918Mb
CG8249-RA

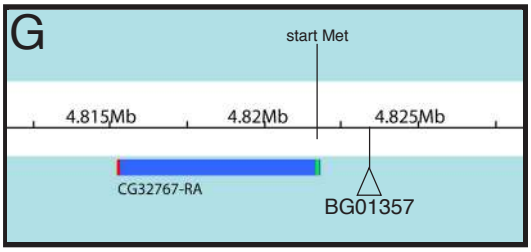

B
PiggyBac
P
Number
Distance from +1 (bp)

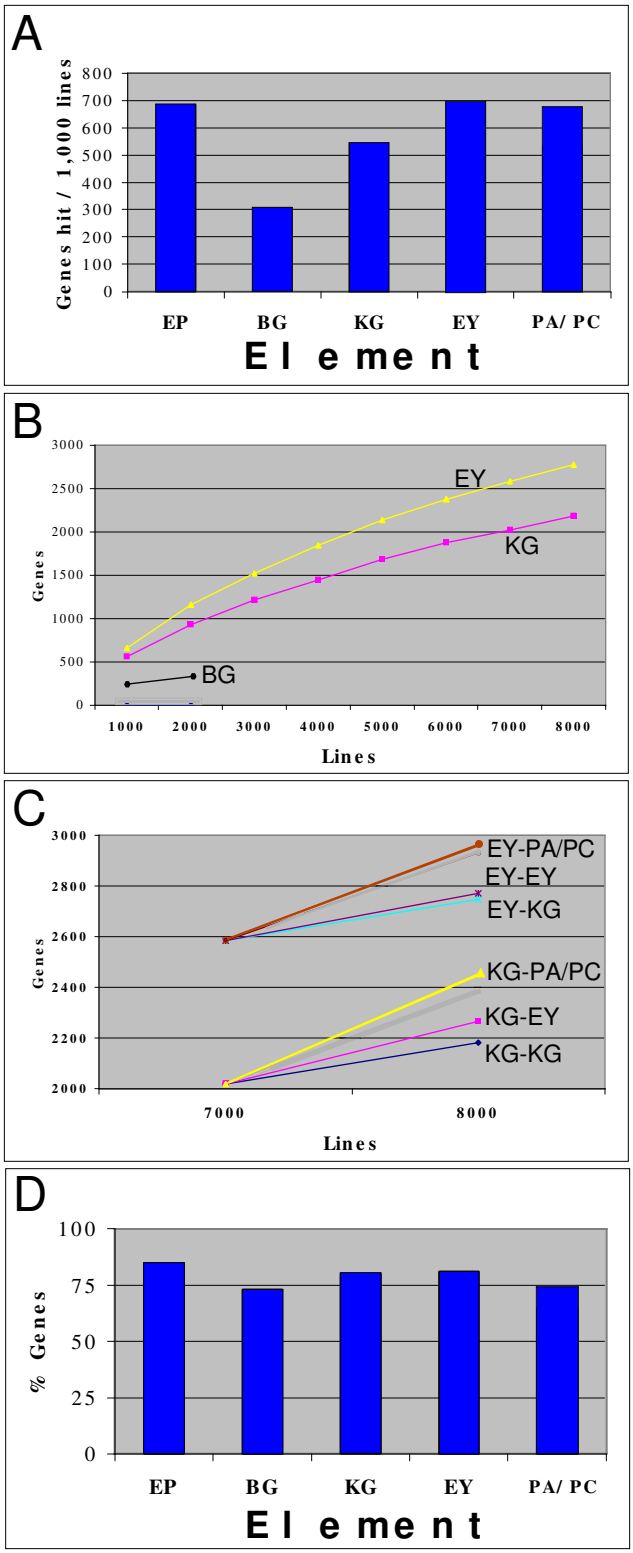

C
KG00786
CG8320-RA
11.056Mb
11.057Mb
11.058Mb
11.059Mb
11.06Mb
11.061Mb
11.062Mb
CG8315-RA
ATPCL:CG8322-RB
ATPCL:CG8322-RA

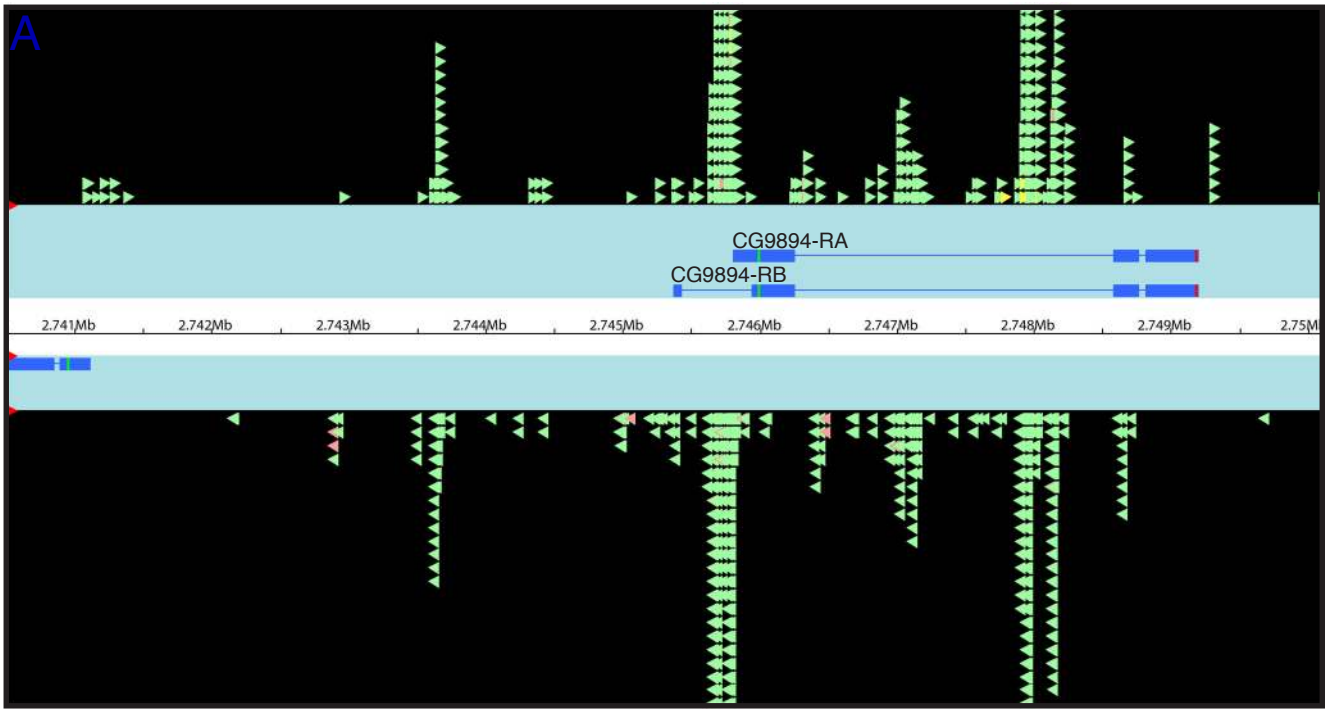

D
KG05287
CG5458-RA
CG5287-RA
13.169Mb
13.17Mb
13.171Mb
13.172Mb
CG31849-RA

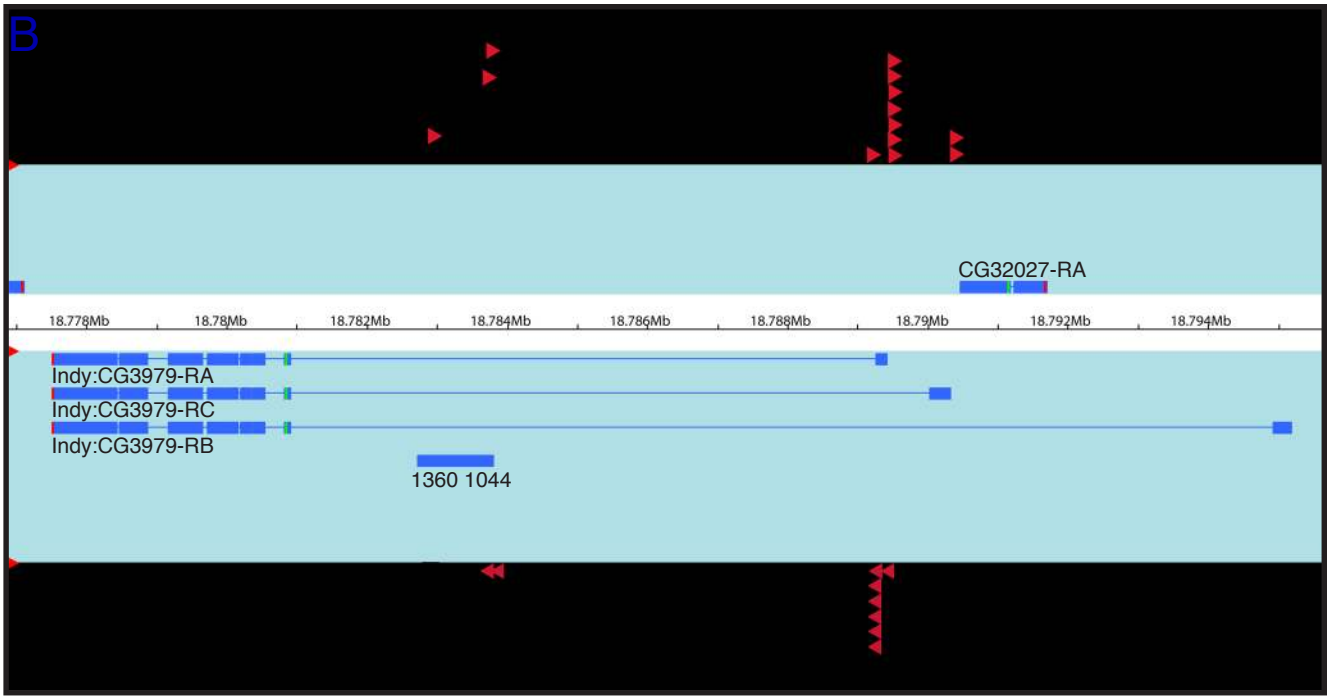

E
KG02679
mirna2R_3299500_3400000_314
mirna2R_3299500_3400000_353
3.35Mb
3.352Mb
3.354Mb
3.356Mb

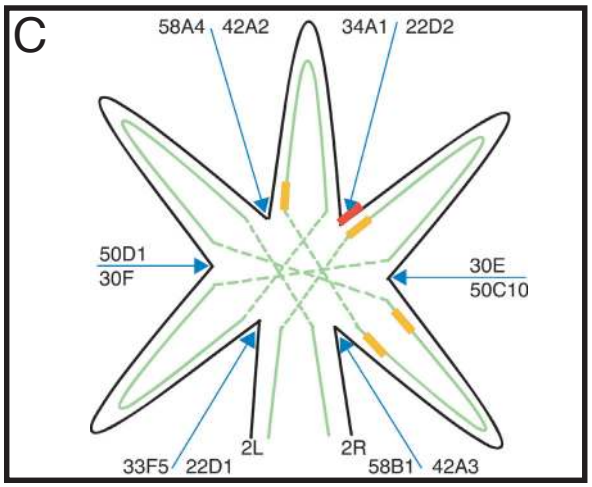

F
CG12462-RA
3.486Mb
3.488Mb
3.49Mb
3.492Mb
3.494Mb
3.496Mb

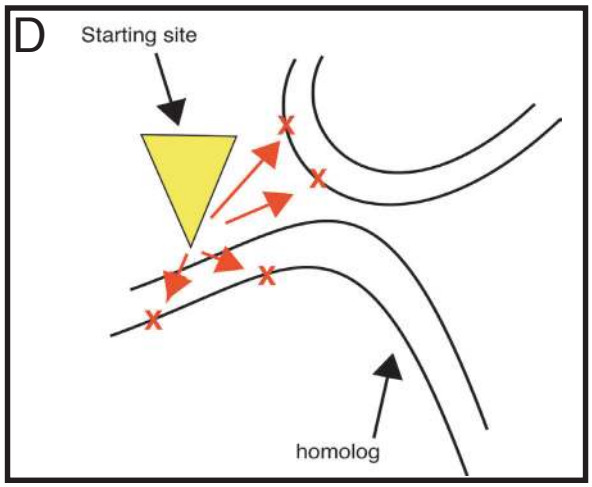

G
start Met
4.815Mb
4.82Mb
4.825Mb
CG32767-RA
BG01357


A
Genes hit / 1,000 lines
800
700
600
500
400
300
200
100
0
EP
BG
KG
EY
PA/ PC
Element
B
Genes
3000
2500
2000
1500
1000
500
0
EY
KG
BG
1000
2000
3000
4000
5000
6000
7000
8000
Lines
C
Genes
3000
2800
2600
2400
2200
2000
EY-PA/PC
EY-EY
EY-KG
KG-PA/PC
KG-EY
KG-KG
7000
8000
Lines
D
% Genes
100
75
50
25
0
EP
BG
KG
EY
PA/ PC
Element


A
CG9894-RA
CG9894-RB
2.741Mb
2.742Mb
2.743Mb
2.744Mb
2.745Mb
2.746Mb
2.747Mb
2.748Mb
2.749Mb


B
CG32027-RA
18.778Mb
18.78Mb
18.782Mb
18.784Mb
18.786Mb
18.788Mb
18.79Mb
18.792Mb
18.794Mb
Indy:CG3979-RA
Indy:CG3979-RC
Indy:CG3979-RB
1360 1044


C
58A4
42A2
34A1
22D2
50D1
30F
30E
50C10
33F5
22D1
2L
2R
58B1
42A3


D
Starting site
homolog

A

Pathway	Genes	Hit
Hedgehog signaling	13	13 (100%)
Wingless signaling	16	12 (75%)
Dpp signaling	10	8 (80%)
Posterior group	14	12 (88%)
Fusome	4	4 (100%)
Heat shock genes	25	18 (75%)
cell cycle	14	23 (61%)
chorion genes	10	0 (0%)
cuticle genes	49	0 (0%)
glue genes	6	0 (0%)
86D Ugt cluster	9	2 (22%)
ribosomal genes	116	11 (9.4%)

B